Probabilistic models for inference about identity

Peng Li, *Member, IEEE,* Yun Fu, *Member, IEEE,* Umar Mohammed, *Member, IEEE,*
James Elder, *Member, IEEE*and Simon J.D. Prince, *Member, IEEE,*

Abstract—Many face recognition algorithms use "distance-based" methods: feature vectors are extracted from each face and distances in feature space are compared to determine matches. In this paper we argue for a fundamentally different approach. We consider each image as having been generated from several underlying causes, some of which are due to identity (latent identity variables, or LIVs) and some of which are not. In recognition we evaluate the probability that two faces have the same underlying identity cause. We make these ideas concrete by developing a series of novel generative models which incorporate both within-individual and between-individual variation. We consider both the linear case where signal and noise are represented by a subspace, and 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 of faces across quite different viewing conditions. We demonstrate that our model produces results that are comparable or better than the state of the art for both frontal face recognition and face recognition under varying pose.

Index Terms—Computing Methodologies, Pattern Recognition, Applications, Face and Gesture Recognition

1 Introduction

Generative probabilistic formulations have proved successful in many areas of computer vision including object recognition [12], image segmentation [37] and tracking [3]. These algorithms assume that the measurements are indirectly created from a set of underlying variables in noisy conditions. Vision problems are formulated as inverting this process: we attempt to estimate the variables from the measurements. This is no different to non-probabilistic approaches, but Bayesian generative formulations have three desirable characteristics. First, they encourage careful modeling of the measurement noise. Second, they allow us to ignore variables that we are not interested in by marginalizing over them. Third, they provide a coherent way to compare models of different sizes using Bayesian model comparison. In this paper we describe a probabilistic approach to face recognition that exploits all of these characteristics.

1.1 Distance-Based Face Recognition

Most contemporary face recognition methods are based on the *distance-based* approach (see Figure 1). A low-dimensional feature vector is extracted from each face image. The distances between these vectors are used to identify faces of the same person. For example, in closed set identification we choose the gallery feature vector with minimum distance from the probe feature vector. In face verification, two face images are ascribed to the same individual if the distance in feature space between them is less than some threshold. The logic of this approach is that for a judicious choice of feature space, the signal-to-noise ratio will be improved relative to that in the original space.

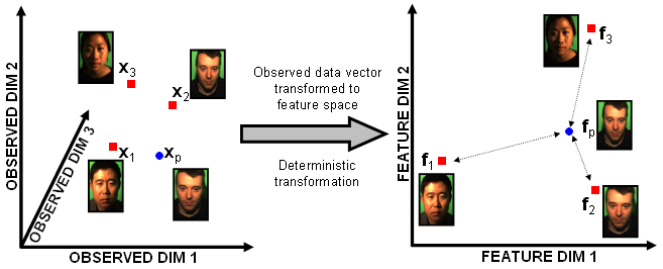


Fig. 1. Conventional distance-based approach. Observed probe $\mathbf{x}_p$ and gallery $\mathbf{x}_{1...3}$ images are deterministically transformed from pixel space (left) to a lower dimensional *feature space* (right). Distance in feature space between the probe image $\mathbf{f}_p$ and each of the gallery images $\mathbf{f}_{1...3}$ is calculated (arrows). The probe vector $\mathbf{f}_p$ is associated with the nearest neighbor gallery vector (here $\mathbf{f}_2$).

Within the class of distance-based methods, the dominant paradigm is the "appearance based" approach which uses weighted sums of pixel values as features for the recognition decision. Turk and Pentland [39] transformed the image data to a feature space based on the principal components of the pixel covariance. Other work has variously investigated using different linear weighted pixel sums, [1], [2], [17], non-linear techniques [45], and different distance measures [32].

A notable sub-category of these methods consists of approaches based on linear discriminant analysis (LDA). The Fisherfaces algorithm [2] 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

- *P. Li, S.J.D. Prince , Y.Fu and U. Mohammed are with the Department of Computer Science, University College London, United Kingdom. E-mail: {ucacpli, yun.fu, u.mohammed,s.prince}@cs.ucl.ac.uk*
- *J. Elder is with the Center for Vision Research, York University, Canada. E-mail: jelder@yorku.ca*

Manuscript received ???; revised ???.

problem). The null-space LDA approach [8] exploited the signal in the remaining subspace. The Dual-Space LDA approach [40] combined these two approaches.

These models perform excellently for frontal face recognition (e.g. see [40]). However, they perform poorly when the probe face is viewed with a very different pose, expression, or illumination to the corresponding gallery image [46]. Distance-based methods fail because the extracted feature vector changes with these variables, making the measured distances between vectors unreliable. Indeed the variation attributable to these factors may dwarf the variation due to differences in identity, rendering the nearest-neighbor decision meaningless. Even LDA-based approaches such as [2], [40] fail as the signal lies in part of the subspace where the noise is also great and is hence down-weighted or discarded.

Many other approaches have been proposed to cope with variable pose and illumination. Important categories include algorithms which (i) require more than one input image of each face [14] (ii) create a 3D model from the 2D image and estimate pose and lighting explicitly [5], [4] and (iii) learn a statistical relation between faces viewed under different conditions [16], [27], [34].

1.2 Probabilistic Face Recognition

The aforementioned distance-based models provide a hard matching decision - however, it would be better to assign a posterior probability to each explanation of the data. In a practical system (e.g. access control), we could defer the final decision and collect more data if the uncertainty is too great. Moreover, a probabilistic solution means that we can easily combine information from different measurement modalities and apply priors over the possible matching configurations.

Generative probabilistic approaches have yielded considerable progress in the closely-related problem of object recognition (e.g. [12]). Nonetheless, there have been few attempts to construct probabilistic algorithms for face recognition. One of the reasons for the paucity of probabilistic approaches is the diversity of tasks in face recognition. These include:

(i) Closed Set Recognition: choose one of N gallery faces that matches a probe face.
(ii) Open Set Recognition: choose one of N gallery faces that matches a probe *or* identify that there is no match.
(iii) Verification: given two face images indicate whether they belong to the same person or not.
(iv) Clustering: given N faces, find how many different people are present and which person is in which image.

Until recently, recognition algorithms could not provide posterior probabilities over different hypotheses for all of these tasks. Liu and Wechsler [26] described a probabilistic method in which they model the data for each individual (after projection to a subspace) as a Gaussian with identical and diagonal variance. However, this method is only suitable for closed set recognition and is not a full probabilistic model as it only describes the data after projection. The scheme of Moghaddam [28] considered pixel-wise difference between probe and gallery images. They modeled distributions of "within-individual" and "between-individual" differences. For two new images, they find the posterior probability that the difference belongs to each. This is well suited to face verification, but does not provide a posterior over possible matches for other tasks. Moreover, performance in uncontrolled conditions is poor [27]. There is no obvious way to remedy these problems.

Recently probabilistic recognition algorithms have been proposed which can address all the above tasks [19], [35], [34]. The key idea is to construct a model describing how the data was generated from identity. Prince and Elder [34] developed a probabilistic "tied factor analysis" model for face recognition. This was specialized to the case of large pose changes and is a special case of one of the models presented in this paper. Ioffe et al. [19] presented a probabilistic LDA model for face recognition that is also closely related to one of the models presented in this paper.

In this paper we develop a model in which identity is represented as a hidden variable in a generative description of the image data. Remaining variation that is not attributed to identity is described as noise. The model is learnt with the EM algorithm [9] and face recognition is framed as a model comparison task. An earlier version of this work was published in [35]. Code is available via http://pvl.cs.ucl.ac.uk.

In Section 2 we introduce a probabilistic framework in which to solve face recognition problems. In Section 3 we introduce a probabilistic version of Fisherfaces [2], which we term probabilistic LDA (or PLDA). We show that this approach sidesteps the small sample problem and produces good results for frontal faces. In Section 4 we introduce a non-linear generalization of this approach. In Section 5 we introduce "Tied PLDA" which allows us to compare faces captured at very different poses. In Section 7 we discuss the relationship between these models and other work.

2 Generative Models for Face Data

Our approach is founded on the following four premises.

(i) Faces images depend on several interacting factors: these include the person's identity (signal) and the pose, illumination etc. (nuisance variables).
(ii) Image generation is noisy: even in matching conditions, images of the same person differ. This remaining variation comprises un-modeled factors and sensor noise.
(iii) Identity cannot be known exactly: since generation is noisy, there will always be uncertainty on any estimate of identity, regardless of how we form this estimate.
(iv) Recognition tasks do not require identity estimates: in face recognition, we can ask whether two faces have the *same* identity, regardless of what this identity is.

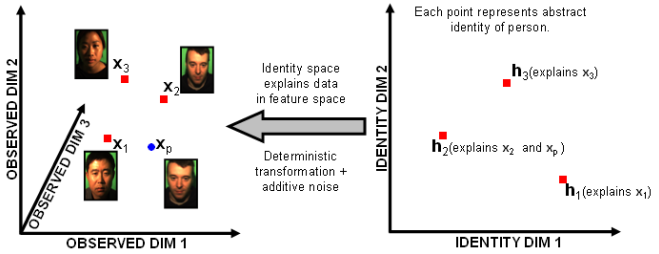


Fig. 2. Latent Identity Variable (LIV) approach. Observed face data vectors $\mathbf{x}$ (left) are generated from the underlying identity space $\mathbf{h}$ (right). The model, $\mathbf{x} = f(\mathbf{h}) + \epsilon$, explains the face data $\mathbf{x}$ as generated by a deterministic transformation $f()$ of the identity variable $\mathbf{h}$ followed by the addition of a stochastic noise term ϵ. In this case the faces $\mathbf{x}_2$ and $\mathbf{x}_p$ are deemed to match as they were generated from the same underlying identity variable $\mathbf{h}_2$.

2.1 Latent Identity Variables

At the core of our approach is the notion that there exists a multidimensional variable $\mathbf{h}$ that represents the identity of the individual. We term this a *latent identity variable* (LIV) and the space that it resides in *identity space*. Latent identity variables have this key property: if two variables have identical values, they describe the same individual. If two variables differ, they describe different people. Crucially, the identity variable is constant for an individual regardless of pose, illumination or any other factors that effect the image.

We never observe identity variables directly but we consider the observed faces to have been generated from the latent identity variable by a noisy process (see Figure 2). Our goal is not necessarily to describe the true generative process, but to obtain a model that describes the image data, within which we can obtain accurate predictions and valid uncertainty estimates. In this paper, we consider models of the form:

$$\mathbf{x}_{ij} = f(\mathbf{h}_i, \mathbf{w}_{ij}, \Theta) + \epsilon_{ij} \quad (1)$$

where $\mathbf{x}_{ij}$ is the vectorized data from the jth image of the ith person. The term $\mathbf{h}_i$ is the LIV and is constant for every image of that person (i.e. it is not indexed by j). The term $\mathbf{w}_{ij}$ is another latent variable representing the viewing conditions (pose, illumination, expression etc.) for the jth image of the ith person. The term ϵ_{ij} is an axis oriented Gaussian noise term and is used to explain any remaining variation. The term Θ is a vector of model parameters, which are learnt during a training phase and remain constant during recognition.

Assuming that we know the model parameters, Θ, how can we then identify if a gallery and probe face match? In the next two sections we consider two alternative strategies, based on (i) evaluating the joint probability of probe and gallery images and (ii) forming class-conditional predictive distributions.

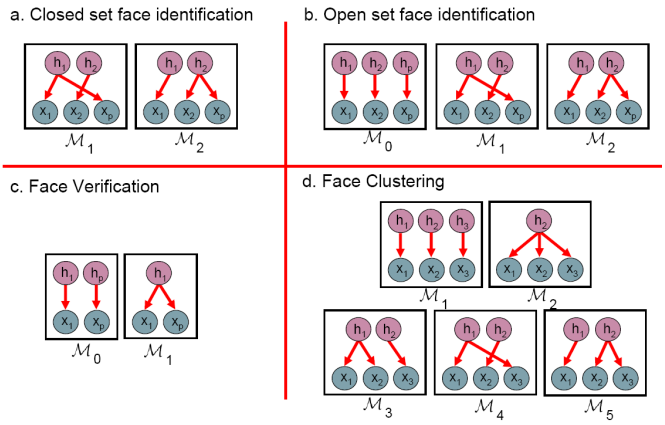


Fig. 3. Inference by comparing data likelihood under different models. Each model represents a different relationship between the LIVs $\mathbf{h}$ and observations $\mathbf{x}$. (a) Closed set identification with gallery of two faces. In Model $\mathcal{M}_1$ the probe $\mathbf{x}_p$ matches gallery face $\mathbf{x}_1$. In model $\mathcal{M}_2$ the probe $\mathbf{x}_p$ matches $\mathbf{x}_2$. (b) Open set identification. We add the possibility $\mathcal{M}_0$ that the probe face $\mathbf{x}_p$ matches neither gallery face $\mathbf{x}_1$ nor $\mathbf{x}_2$. Verification (c) and face clustering (d) can also be expressed as model comparison. Note that there is also a noise variable $\mathbf{w}$ associated with each datum $\mathbf{x}$ which is not shown.

2.2 Recognition: joint perspective

Our framework infers whether two observed images $\mathbf{x}_1$ and $\mathbf{x}_2$ were generated from the same identity variable $\mathbf{h}$ and hence belong to the same individual. Unfortunately, this presents a problem: the data $\mathbf{x}$ was generated in noisy conditions, so we can never be certain of the underlying value of $\mathbf{h}$ or $\mathbf{w}$. To resolve this, we consider all possible values of $\mathbf{h}$ and $\mathbf{w}$.

More formally, the recognition process compares the likelihood of the data under different models $\mathcal{M}$. Each model assigns identity variables $\mathbf{h}$ to explain the observed faces $\mathbf{x}$ in a different way. If the current model ascribes two face images to belong to the same person then they will have the same identity variable. If not then they will each have their own identity variables.

Figure 3a shows the model construction for closed set identification. We are given a probe face $\mathbf{x}_p$ and the N gallery faces (here $N = 2$) each representing a different person, $\mathbf{x}_{1\ldots N}$. In model $\mathcal{M}_n$ the nth gallery face is forced to share its latent identity variable $\mathbf{h}_n$ with the probe indicating that these faces belong to the same person. Figure 3c shows the models for face verification. Here Model $\mathcal{M}_0$ represents the case where the two faces do not match (each image has a separate identity). Model $\mathcal{M}_1$ represents the case where they do match (they share an identity). In fact, all four recognition tasks from Section 1.2 can be expressed in terms of model comparison: for open set identification (Figure 3b) , we start with the closed set case and add a model $\mathcal{M}_0$ representing the situation where the probe has its own unique identity. In face clustering (Figure 3d) we are given N faces, and

there may be N different people (N identity variables), just one person (1 identity variable) or anything between. In this paper we concentrate on closed set identification, verification and clustering.

We combine the likelihoods of these models with suitable priors $Pr(\mathcal{M})$ (always uniform in this paper) and find a posterior probability for the match using Bayes' rule. However, the question remains as to how to calculate the model likelihoods. Noise in the generation process means that we can never exactly know either the identity variables $\mathbf{h}$ or the noise variables $\mathbf{w}$ in these models. Hence we marginalize (integrate out) these variables:

$$Pr(\mathbf{x}_{1...N,p}|\mathcal{M}_0) = \prod_{n=1}^{N} Pr(\mathbf{x}_n)Pr(\mathbf{x}_p) \quad (2)$$

$$Pr(\mathbf{x}_{1...N,p}|\mathcal{M}_m) = \prod_{n=1,n\neq m}^{N} Pr(\mathbf{x}_n)Pr(\mathbf{x}_p,\mathbf{x}_m) \quad (3)$$

where

$$Pr(\mathbf{x}_n) = \iint Pr(\mathbf{x}_n,\mathbf{h}_n,\mathbf{w}_n)d\mathbf{h}_n d\mathbf{w}_n \quad (4)$$

$$Pr(\mathbf{x}_p) = \iint Pr(\mathbf{x}_p,\mathbf{h}_p,\mathbf{w}_p)d\mathbf{h}_p d\mathbf{w}_p \quad (5)$$

$$Pr(\mathbf{x}_p,\mathbf{x}_m) = \iiint Pr(\mathbf{x}_p,\mathbf{x}_m,\mathbf{h}_m,\mathbf{w}_p,\mathbf{w}_m)d\mathbf{h}_m d\mathbf{w}_p d\mathbf{w}_m \quad (6)$$

An important aspect of this formulation is that the final likelihood expression does not explicitly depend on the latent identity variables $\mathbf{h}$. This makes it valid to compare models with different numbers of latent identity variables. For example in face verification we compare a model with two underlying latent variables (no match) to one (match). This is an example of Bayesian model selection in which we compare the *evidence* for the different explanations of the data.

2.3 Recognition: class conditional perspective

In the previous treatment, we evaluate the joint likelihood of the probe and gallery images under different models (Equations 2 and 3). An alternative perspective is to consider the predictive distribution for the probe image $\mathbf{x}_p$ induced by the matching gallery data $\mathbf{x}_g$ in each of the models. To make face recognition decisions, we evaluate the likelihood of the probe image under each of these class conditional density functions and combine with priors. Now we write:

$$\begin{aligned} Pr(\mathbf{x}_{1...N,p}|\mathcal{M}_0) &= Pr(\mathbf{x}_p|\mathbf{x}_{1...N},\mathcal{M}_0)Pr(\mathbf{x}_{1...N}) \\ &= Pr(\mathbf{x}_p)\prod_{n=1}^{N} Pr(\mathbf{x}_n) \quad (7) \\ Pr(\mathbf{x}_{1...N,p}|\mathcal{M}_m) &= Pr(\mathbf{x}_p|\mathbf{x}_{1...N},\mathcal{M}_m)Pr(\mathbf{x}_{1...N}) \\ &= Pr(\mathbf{x}_p|\mathbf{x}_m)\prod_{n=1}^{N} Pr(\mathbf{x}_n) \quad (8) \end{aligned}$$

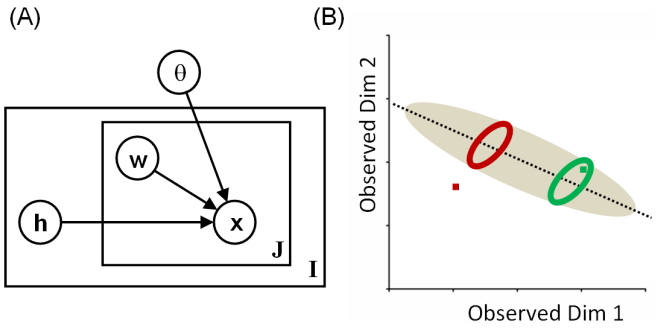


Fig. 4. PLDA Model (A) Graphical model relating data $\mathbf{x}$ to identities $\mathbf{h}$, noise variables $\mathbf{w}$ and parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$. (B) Predictive distribution for subspace model with one identity factor $\mathbf{F}$ (dotted line) and one noise factor $\mathbf{G}$ (not shown). Gray region represents associated Gaussian face manifold. New gallery images (red and green dots) induce Gaussian predictive distributions (red and green ellipses).

where $Pr(\mathbf{x}_p|\mathbf{x}_n)$ is found by taking the conditional of Equation 6. This approach is closely related to object recognition: generative models such as [12] create a separate probability density for each class. However, in object recognition there are usually numerous training examples of each class (e.g. cars). For face recognition, we often only have a single example of each class (individual). Hence face recognition models necessarily deal with the situation of "one shot" learning [23].

2.4 Tractability of Integrals

We are assuming that the integrals in Equations 4-6 can be computed. This is true for all models in this paper. When they cannot be computed one approach is to approximate the distributions over the hidden variables $\mathbf{h}$ and/or $\mathbf{w}$ by point estimates $\hat{\mathbf{h}}$ and $\hat{\mathbf{w}}$. The choice of the joint or class-conditional methods now becomes important: in the joint method, the point estimate of the identity will be based on both the gallery and probe images, whereas in the class-conditional method, the identity will be based on the gallery alone.

3 Model 1: Probabilistic LDA (PLDA)

To make these ideas concrete, we investigate a probabilistic model that is closely related to linear discriminant analysis (LDA). LDA is a technique that models 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. We refer to our version of this algorithm as probabilistic linear discriminant analysis (PLDA). The relationship between PLDA and standard LDA is similar 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 jth image of the ith individual by $\mathbf{x}_{ij}$. We model the data generation as:

$$\mathbf{x}_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij} \quad (9)$$

where μ is the mean of the data, $\mathbf{F}$ is a factor matrix with the basis vectors of the between individual subspace in its columns and $\mathbf{h}_i$ is the latent identity variable that is constant for all images $\mathbf{x}_{i1} \ldots \mathbf{x}_{iJ}$ of person i. Just as the matrix $\mathbf{F}$ contains a matrix determining the between-individual subspace, the matrix $\mathbf{G}$ contains a basis for the within-individual subspace. The term $\mathbf{w}_{ij}$ represents the position in this subspace. The term ϵ_{ij} is a stochastic noise term, with diagonal covariance Σ.

The term $\mu + \mathbf{F}\mathbf{h}_i$ is the signal and accounts for between-individual variance. For a given individual, this term is constant. The term $\mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}$ consists of the noise or within-individual variance. It explains why two images of the same individual do not look identical.

More formally, we can describe the model in Equation 9 in terms of conditional probabilities:

$$Pr(\mathbf{x}_{ij}|\mathbf{h}_i, \mathbf{w}_{ij}, \theta) = \mathcal{G}_{\mathbf{x}}\left[\mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij}, \Sigma\right] \quad (10)$$

$$Pr(\mathbf{h}_i) = \mathcal{G}_{\mathbf{h}}\left[0, \mathbf{I}\right] \quad (11)$$

$$Pr(\mathbf{w}_{ij}) = \mathcal{G}_{\mathbf{w}}\left[0, \mathbf{I}\right] \quad (12)$$

where $\mathcal{G}_{\mathbf{a}}[\mathbf{b}, \mathbf{C}]$ denotes a Gaussian in $\mathbf{a}$ with mean $\mathbf{b}$ and covariance $\mathbf{C}$. In Equations 11 and 12 we have defined simple priors on the latent variables $\mathbf{h}_i$ and $\mathbf{w}_{ij}$. The relationship between the variables is indicated in Figure 4a. It is important to note that Equations 10, 11 and 12, implicitly define the joint probability distribution required for Equations 4-6. The Gaussian forms for this model have been chosen because they provide clean closed form solutions to these integrals, rather than because they represent the true generative process.

3.1 Learning

In the *learning* stage we aim to learn the parameters, $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$ given data $\mathbf{x}_{ij}$. It would be easy to estimate these parameters if we knew the hidden identity variables $\mathbf{h}_i$ and hidden noise variables $\mathbf{w}_{ij}$. Likewise, it would be easy to infer the identity variables $\mathbf{h}_i$ and noise variables $\mathbf{w}_{ij}$ if we knew the parameters Θ. This type of "chicken and egg" problem is well suited to the expectation-maximization (EM) algorithm [9]. Details of this process are given in Appendix A.

Figure 5 shows the results of 10 iterations of learning from the first 195 individuals from the XM2VTS database with minimal preprocessing. We show several positions in the between-individual subspace (samples where $\mathbf{h}$ varies but $\mathbf{w}$ is constant) and these look like different people. We also show positions in the within-individual subspace (samples where $\mathbf{h}$ is constant and $\mathbf{w}$ varies). These look like the same person under slightly different illuminations and poses.

Figure 6 shows a visualization of the model for four faces. In each case, we decompose the image into signal $\mathbf{F}\mathbf{h}_i$ and noise components $\mathbf{G}\mathbf{w}_{ij}$ and ϵ_{ij} using the final MAP estimate of the hidden variables from the E-Step in training.

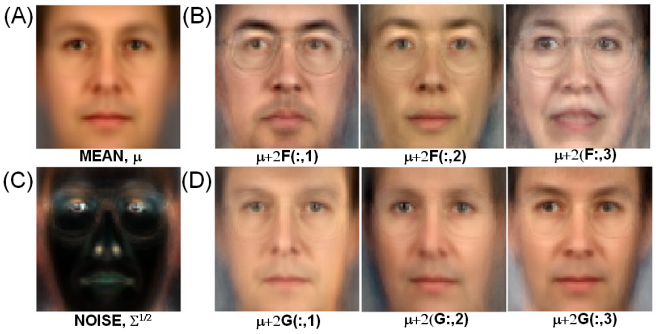


Fig. 5. PLDA Model. (A) Mean face (B) Three directions in between-individual subspace. Each image looks like a different person. (C) Per-pixel noise covariance (D) Three directions in within-individual subspace. Each image looks like the same person under minor pose and lighting changes.

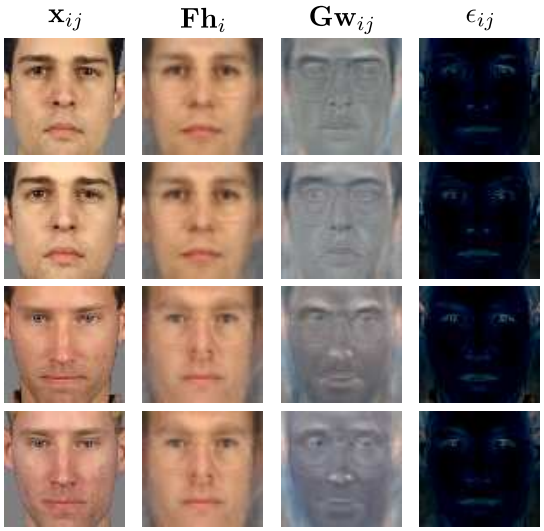


Fig. 6. PLDA. The first column shows the original images $\mathbf{x}_{ij}$. These are broken down into a signal subspace component $\mathbf{F}\mathbf{h}_i$ which is the same for each identity, a noise subspace component $\mathbf{G}\mathbf{w}_{ij}$ and a per pixel noise ϵ_{ij}.

3.2 Recognition with Joint Method

In recogntion we must evaluate the integrals in Equation 4-6. The general problem is to evaluate the likelihood that N images $\mathbf{x}_{1\ldots N}$ share the same identity variable, $\mathbf{h}$, regardless of the noise variables $\mathbf{w}_1 \ldots \mathbf{w}_N$. Our approach is to re-write the equations in the form of a factor analyzer and use a standard result for the integral. To this end we combine the generative equations for all N images:

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} + \begin{bmatrix} \mathbf{F} & \mathbf{G} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{F} & \mathbf{0} & \mathbf{G} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{F} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_N \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \quad (13)$$

or, giving names to these composite matrices:

$$\mathbf{x}' = \mu' + \mathbf{A}\mathbf{y} + \epsilon' \tag{14}$$

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

$$\begin{aligned} Pr(\mathbf{x}'|\mathbf{y}) &= \mathcal{G}_{\mathbf{x}'}\left[\mu + \mathbf{A}\mathbf{y}, \Sigma'\right] & (15)\\ Pr(\mathbf{y}) &= \mathcal{G}_{\mathbf{y}}\left[0, \mathbf{I}\right] & (16)\end{aligned}$$

where

$$\Sigma' = \begin{bmatrix} \Sigma & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \Sigma & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \Sigma \end{bmatrix} \tag{17}$$

This now has the form of a factor analyzer. From Equation 14 it is easy to see that the first two moments of the distribution of the compound vector $\mathbf{x}'$ are given by:

$$\begin{aligned} E\left[\mathbf{x}'\right] &= \mu' \\ E\left[(\mathbf{x}' - \mu')(\mathbf{x}' - \mu')^T\right] &= E\left[(\mathbf{A}\mathbf{y} + \epsilon')(\mathbf{A}\mathbf{y} + \epsilon')^T\right] \\ &= \mathbf{A}\mathbf{A}^T + \Sigma' & (18)\end{aligned}$$

and it can be shown that when we marginalize over the hidden variable $\mathbf{y}$ the form of the resulting distribution is Gaussian with these moments:

$$Pr(\mathbf{x}_{1\ldots N}) = Pr(\mathbf{x}') = \mathcal{G}_{\mathbf{x}'}\left[\mu', \mathbf{A}\mathbf{A}^T + \Sigma'\right] \tag{19}$$

3.3 Recognition with Predictive Distribution

Instead of calculating the expressions in Equations 2-6. We could equivalently have performed this experiment by calculating the predictive distributions $Pr(\mathbf{x}_p|\mathbf{x}_1)\ldots Pr(\mathbf{x}_p|\mathbf{x}_n)$. We then assess the likelihood of the probe image under each of these distributions.

The predictive distributions can be calculated by taking the joint distribution in Equation 19 and finding the conditional distribution of $\mathbf{x}_p$ given all the other variables. If the mean and information matrix of the joint distribution in Equation 19 are partitioned so that

$$\mu' = \begin{bmatrix} \mathbf{m}_p \\ \mathbf{m}_g \end{bmatrix} \qquad (\mathbf{A}\mathbf{A}^T + \Sigma')^{-1} = \begin{bmatrix} \Lambda_{pp} & \Lambda_{gp} \\ \Lambda_{pg} & \Lambda_{gg} \end{bmatrix}, \tag{20}$$

then the conditional distribution is also Gaussian (see [6]) with mean and covariance

$$Pr(\mathbf{x}_p|\mathbf{x}_g) = \mathbf{G}_{\mathbf{x}_p}\left[\mathbf{m}_p - \Lambda_{pp}^{-1}\Lambda_{pg}(\mathbf{x}_g - \mathbf{m}_p), \Lambda_{pp}^{-1}\right]. \tag{21}$$

It is possible to evaluate this Gaussian probability efficiently (appendix B) making the complexity of this algorithm similar to that of the original LDA method.

Example predictive distributions for the case where the subspaces $\mathbf{F}$ and $\mathbf{G}$ are one dimensional are shown in Figure 4b. The learned face manifold (indicated in gray) is given by $Pr(\mathbf{x}) = \mathcal{G}_x[\mu, \mathbf{F}\mathbf{F}^T + \mathbf{G}\mathbf{G}^T + \Sigma]$. A new gallery image induces a Gaussian predictive distribution, with a mean that is projected onto the subspace spanned by the factor in $\mathbf{F}$ (dotted line). The projection direction depends on the noise parameters Σ and the noise subspace $\mathbf{G}$. To perform face recognition we would compare the likelihood of a new probe point under each of the predictive distributions. The likelihood for not matching either gallery image is found by evaluating the probe point under the whole manifold distribution (gray area).

3.4 Experiment 1: Frontal face identification

To explore the properties of the algorithm, we trained the algorithm with all of the data from the first 195 individuals in the XM2VTS database. We tested with the last 100 individuals, using the one image from the first capture session as the gallery set and the one image from the last for the probe set. Hence, model must generalize from the training set to new individuals.

To ensure that the experiments are easy to replicate, we used 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 $70 \times 70 \times 3$. Raw pixel values form the input vector. There was no photometric normalization. For each probe, we compute the likelihood that it matches each face in the gallery using Equation 3. We calculate a posterior for the match, assuming uniform priors. We take the MAP solution as the match.

In Figure 7A we plot $\%$ correct first match results as a function of the subspace dimensions. For each line of the graph, the dimension of the signal Dim(F) is constant, but the dimension of the noise Dim(G) varies. Increasing the signal dimension improves performance. Increasing the dimension of the noise subspace has a more complex effect: performance is always worst when Dim(G) is zero and is best when Dim(G) is roughly the same as Dim(F). Hence we set the signal and noise subspace size to be identical in all remaining experiments.

In Figure 7B we decompose the model into constituent parts. We force the noise Σ to be a multiple of the identity rather than diagonal (dashed vs solid lines). This reduces performance: the full algorithm learns which parts of the image are most variable and downweights them in the decision. We also remove the noise subspace by setting Dim(G) to zero (blue lines vs red lines) which decreases performance even further.

These restrictions can be easily related to other models. When the covariance Σ is uniform and there is no noise subspace the model takes the form of probabilistic PCA and the results are very similar to that for eigenfaces. In fact, the PPCA model has very slightly superior performance due to regularization induced by the prior

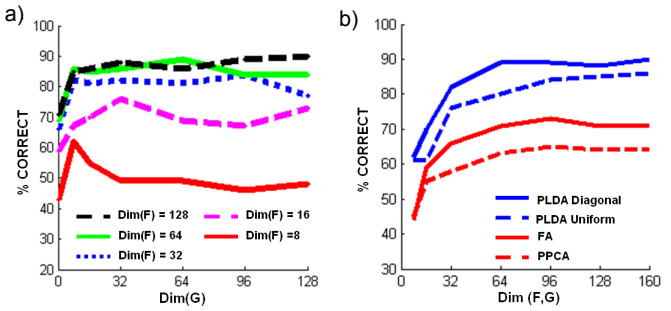


Fig. 7. (A) Identification performance as with minimal preprocessing as a function of signal and noise subspace size. (B) Identification performance for progressive simplifications of the model.

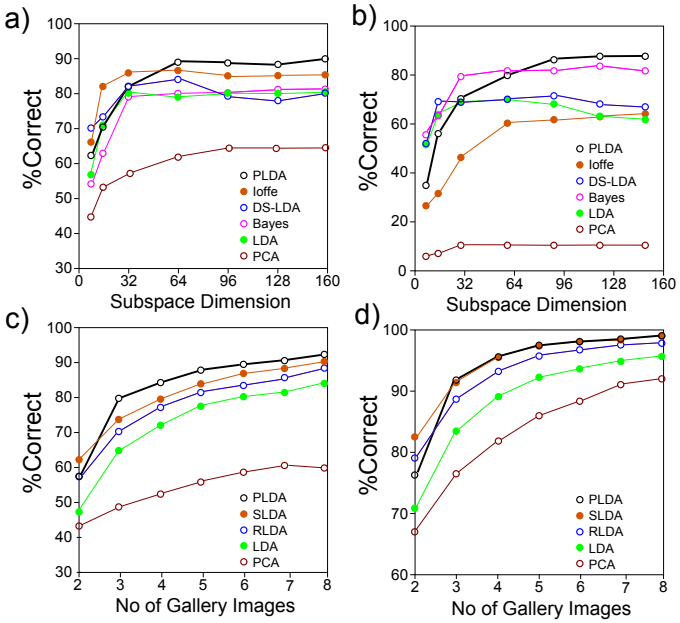


Fig. 8. (A) Comparison of algorithms for XM2VTS database. PLDA outperforms PCA [39], LDA [2], the Bayesian approach [28], Dual-Space (DS) LDA [40] and the probabilistic approach of Ioffe [19] (B) Performance comparison for XM2VTS lighting subset. (C) Comparison for YALE database as a function of gallery images (nearest neighbour approach) to RLDA [7], SLDA [7], LDA [2] and PCA [39]. (D) Comparison for ORL database.

over $\mathbf{h}$. When we allow Σ to have independent diagonal terms, the model takes the form of a factor analyzer. When we provide allow the noise subspace dimension Dim(G) to be non-zero and restrict Σ to be diagonal, our model is similar to to that of Ioffe [19].

3.4.1 Comparison to Other Algorithms

In Figure 8 we compare the performance of PLDA to other algorithms. We emphasize here that the preprocessing of the data is exactly the same in each case, so this is a pure test of recognition ability when the remaining parts of the pipeline are held constant.

Figure 8a shows that PLDA outperforms our implementations of five other algorithms on the XM2VTS database The closest competing methods are dual-space LDA [40] and the probabilistic approach of Ioffe [19].

In figure 8b we investigate performance for the same algorithms on the lighting subset of the XM2VTS database. The training set consisted of 7 images each of the first 195 individuals and contained 2 lighting conditions. For each individual there were 5 images under frontal lighting and 2 under side-lighting. The test set consisted of 100 different individuals, where the gallery images were taken from the first recording session and were under frontal lighting and the probe images were taken from the fourth session and were lit from the side. All other preprocessing was the same as for the original XM2VTS data. Once more, the PLDA algorithm outperforms the five competing algorithms.

In figure 8c we present results from the Yale database which also contains lighting variation. which was preprocessed as in [7]. We compare to published data from [7] and show that performance is superior to the RLDA, SLDA, LDA and PCA algorithms. Finally, in 8d we compare results to the same algorithms on ORL database (also as preprocessed by [7]) which contains both pose and lighting variation. Here the PLDA algorithm provides performance that is comparable to SLDA and superior to RLDA, LDA and PCA.

These experiments make a strong case for PLDA: over four different databases and seven algorithms it produces reliably better performance when all other parts of the face recognition pipeline are held constant.

Our technique outperforms other LDA methods for three reasons. First, the per-pixel noise term Σ means we have a more sophisticated model of within-individual variation (see figure 7b). Second, our method does not suffer from the small sample problem: the signal subspace $\mathbf{F}$ and noise subspace $\mathbf{G}$ may be completely parallel or entirely orthogonal. There is no need for two separate procedures as in the dual-space LDA algorithm [40]. Third, a slight benefit results from the regularizing effect of the prior over the identity and noise variables.

We also investigated identification performance for the PLDA algorithm for the XM2VTS with more elaborate preprocessing. Eight keypoints on each face were identified by hand or automatically using the method described in [34], depending on the condition. The images were registered using a piecewise triangular warp. The final image size was 400×400. We extracted feature vectors consisting of image gradients at 8 orientations and 3 scales at points in a 6×6 grid around each keypoint. A separate recognition model was built for each and these were treated as independent.

Here, the training data consisted of images from the first three capture sessions from all 295 individuals in the database. We use images from (i) capture session 1 or (ii) capture sessions 1-3 to form the gallery. We use images from capture session 4 to form the probe set. This protocol was chosen to facilitate comparison with [25].

With a single gallery image, peak performance was

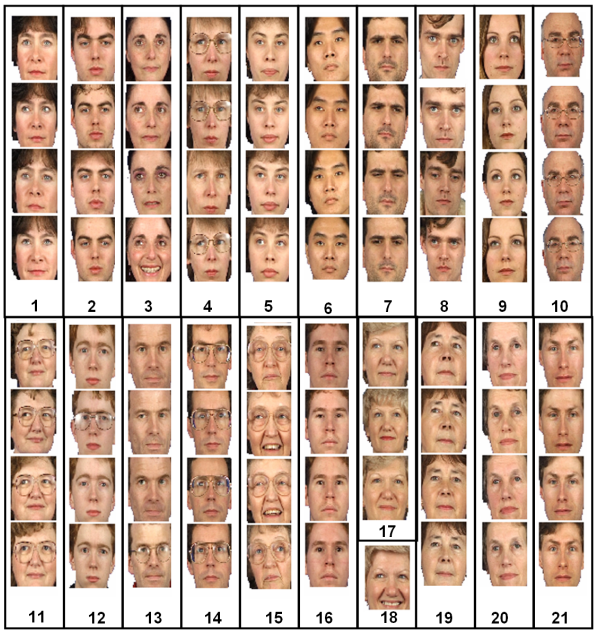


Fig. 9. Experiment 2: clustering results for 80 frontal images consisting of 4 images each of 20 people. Blue lines divide clusters. The algorithm has found 21 clusters - one of the original clusters have erroneously been split.

99.7%: we mis-classified one face (image 169.4.1) where the pose deviated from frontal. In Section 5 we present an algorithm to cope with pose changes. With three gallery images we achieved 100% performance.

In Table 1 we compare PLDA performance to published results from [25]. Here, the experimental protocol was identical, but the whole preprocessing pipeline differs. Our method compares favorably to other algorithms, although it is unwise to draw strong conclusions where the difference in performance is only small. We believe that Figure 8 provides more information about the relative strengths of these algorithms. Nonetheless, these results suggest that PLDA can support strong recognition performance and that the results of figure 8 were not just an artefact of the simple preprocessing.

METHOD	N	Error Rate
PCA [39]	1	33.9
LDA [2]	1	11.9
Bayesian [28]	1	11.5
Unified Subspace (US)[41]	1	6.8
Adaptive Clustering US SVM [25]	1	1.0
LIV Subspace Model	1	12.0
LIV PLDA Model	1	0.3 (0.7)
Bayesian Gabor [42]	3	2.9
LIV PLDA Model	3	0.0 (0.0)

Table 1 - Results for XM2VTS database with N gallery images PCA, LDA, Bayesian and Unified Subspace results from [25]. Bracketed results indicate results from our approach with automated feature finding.

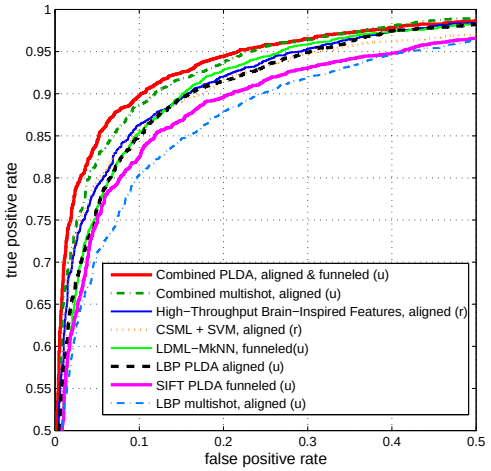


Fig. 10. Experiment 3 results: ROC curve of PLDA and other the state of arts methods for face verification on LFW dataset.

3.5 Experiment 2: Frontal face clustering

In Experiment 2 (Figure 9) we demonstrate clustering using the elaborately preprocessed XM2VTS data. We train the system using only the first 195 individuals from the XM2VTS database and signal and noise subspaces of size 64. The algorithm is presented with 80 images taken from the last 100 individuals. In principle it is possible to calculate the likelihood for each possible clustering of the data using Equation 19: for example we can calculate the likelihood that there are 80 different individuals or that the 80 images are all of the same individual.

Unfortunately, in practice there are far too many possible configurations. Hence we adopt a greedy agglomerative strategy. We start with the hypothesis that there are 80 different individuals. We then consider merging all pairs of individuals and choose the combination that increases the likelihood the most. We continue this process until the likelihood cannot be improved. In order to test the clustering performance we randomly select 4 images each from 20 individuals and apply our algorithm. We can quantify performance by counting the number of splits and merges required to change our estimated clustering to the ground truth. Averaged over 100 datasets, the mean number of split/merges was 1.60.

Typical results are shown in Figure 9 (here the number of split/merges required is 1). The algorithm slightly over-partitions the data but does not erroneously associate images from different individuals. We conclude that our model can cope with complex compound decisions about identity and can select model size without the need for extra parameters.

3.6 Experiment 3: Face verification

In experiment 3 we investigate face verification using the Labeled Faces in the Wild [18] database which contains large variations in pose, expression and lighting. Images

were gray scale and were prepared in two ways: 1) aligned using commercial face alignment software by Taigman et al [38] and 2) funneled which is available on the LFW website [18]. There are a total of 13233 images and 5749 people in the database. The number of images varies from one to 530 images.

The images are divided into ten groups where the subject identities are mutually exclusive. In each group, there are 300 pairs of images from the same identity and 300 pair from different identities. There are two possible training configurations. In the "restricted configuration" only the same/not-same labels are used no information about the actual identities is used. Most previous work has restricted protocol (e.g. [44] and [22]). In the "unrestricted configuration" all available information including the identities can be used for training. The studies of [13], [38] used this configuration.

The aligned images were cropped to 80×150 pixels following Nguyen and Bai [29]. Each image was normalized by passing it through a log function ($log(x+1)$) to suppress the effect of shadows and lighting. In addition, we localize four keypoints following [10], [24] and estimate the facial pose by projecting the keypoint positions to the first principal component following [38]. The images of large right profile faces are swapped to left profile faces so that all the images are left profile or near frontal. We investigated two types of descriptors on the aligned images: local binary patterns (LBP) [31] and three-patch local binary patterns (TPLBP) [44]. We used the same parameters as [29], [38]. In addition, we also investigated the SIFT descriptors computed at the 9 facial keypoints on the funneled images. The SIFT data are available from [13]. The original dimensionality of the features was quite high (7080 for LBP and TPLBP and 3456 for SIFT) so we reduced the dimension to 200 using PCA.

For each pair of images, we compute the likelihood that they match each other using Equation 3 and likelihood that they do not match using Equation 2. There are two views of the LFW database. The images in View 1 are used for model selection (subspace dimension of PLDA) and the images in View 2 are used for training and test. We followed the "unrestricted configuration". For each of the ten–fold cross-validation test, we used identities with at least two images for training. The number of training images is around 8000. A threshold of the log-likelihood ratio is learned using 5400 pairs of images in the 9 folds of the data. The learned model is then tested using the 600 pairs of held–out data. Table 2 and Figure 10 are the comparison of PLDA with the state of the art methods on LFW database evaluated using average verification rate and ROC curves of the 10–fold cross validation test respectively, where "u" and "r" denotes unrestricted and restricted configuration respectively.

The optimal subspace dimension of PLDA is 128, 96 and 48 for the LBP, TPLBP and SIFT descriptors respectively and these settings were used in Table 2 and Figure 10. The best performance of PLDA based on a single descriptor is 87.3% using the LBP descriptor, which is 2.2% better than the result of multishot learning (also using LBP) in [38]. In addition, PLDA outperforms LDML using a single SIFT descriptor [13] (3.0% higher in terms of verification rate). Note that we have used the same SIFT data as [13] and a similar LBP descriptor (but different image size) as [38]. Therefore, the performance of our PLDA model outperforms the current best model based on single descriptor [38] that is reported on the result page of LFW database [18].

The combination of different descriptors is straightforward for PLDA. We treat these descriptors independently and the likelihoods of match and not–match are just the product of those calculated on each descriptor. So it is unnecessary to train another classifier such as SVM to do the final decision as in [44], [13], [38]. The performance of PLDA by combining LBP, TPLBP and SIFT descriptors (Combined PLDA in Table 2 and Figure 10) is 90.1%, which was consistently better than that using each individual descriptor alone, agreeing with [44], [13], [38]. Furthermore, this is better than the state of the art result: multishot [38] (89.5%) and LDML-MkNN [13] (87.5%) in the unrestricted setting and High-Throughput Brain-Inspired (HTBI) Features [33] and CSML + SVM [29] in the restricted setting. Note that the current top two methods in unrestricted setting are based on four types of descriptors and it is possible that PLDA's performance might increase further if we also introduced more discriminative descriptors.

Method	Accuracy
Combined PLDA, aligned & funneled (u)	0.901 ± 0.005
Combined multishot, aligned (u) [38]	0.895 ± 0.005
Combined LDML–MkNN, funneled (u) [13]	0.875 ± 0.004
HTBI Features, aligned (r) [33]	0.881 ± 0.006
CSML + SVM, aligned (r) [29]	0.880 ± 0.004
TPLBP PLDA, aligned (u)	0.837 ± 0.007
LBP PLDA, aligned (u)	0.873 ± 0.006
LBP multishot, aligned (u) [38]	0.851 ± 0.006
SIFT PLDA, funneled (u)	0.862 ± 0.012
SIFT LDML, funneled (u) [13]	0.832 ± 0.004

Table 2 - Results of PLDA and other the state of the art methods for LFW database (mean classification accuracy and standard error of the mean). The top 5 rows are based on multiple descriptors and the bottom 5 rows are based on single descriptor.

We also note that multi-shot learning [38] needs to train two classifiers during testing and the marginalized k nearest neighbors (MkNN) [13] needs to find a set of nearest neighbors. Compared to these methods, the PLDA algorithm is relatively efficient (see Appendix B).

We encourage caution in comparing these results which compare pipeline to pipeline rather than algorithm to algorithm: the remaining differences may be due to preprocessing or the recognition algorithm. We can conclude that the PLDA algorithm can produce verification results that are at least comparable to the state of the art using this challenging real-world database.

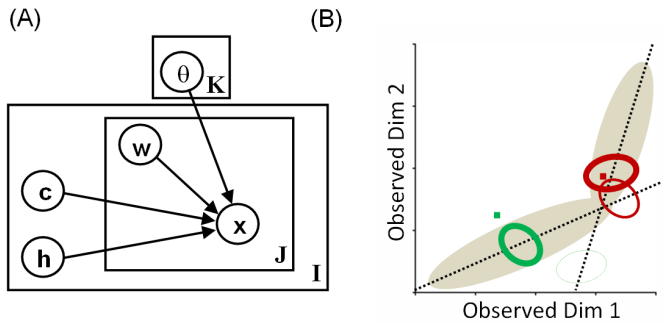


Fig. 11. Mixtures of PLDA Model (A) Graphical model relating data $\mathbf{x}$ to identities $\mathbf{h}$, noise variables $\mathbf{w}$ and parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$. (B) Predictive distribution for subspace model with clusters. Each contains a one-dimensional identity subspace (dotted lines) and one noise factor (not shown). Gray region represents face manifold. New gallery images (red and green dots) induce predictive distributions that are themselves mixtures of Gaussians (red and green ellipses). Weight of ellipses correspond to weights of component.

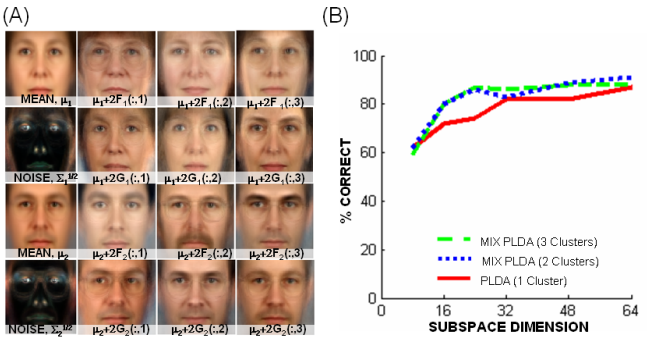


Fig. 12. (A) Mixtures of PLDA model. Top two rows show elements of mixture component 1. Bottom two rows show component 2. Interestingly, the two clusters correspond to the two sexes. The mean of cluster 1 and images representing directions in the signal subspace all look like women (top row). For cluster 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.

4 Model 2: Mixtures of PLDAs

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. We resolve these problems by describing the face manifold as a weighted additive mixture of K PLDA distributions (see figure 11A).

$$\begin{aligned} Pr(\mathbf{x}_{ij}) &= \mathcal{G}_x\left[\mu_{c_i} + \mathbf{F}_{c_i}\mathbf{h}_i + \mathbf{G}_{c_i}\mathbf{w}_{ij}, \Sigma_{c_i}\right] \\ Pr(\mathbf{h}_i) &= \mathcal{G}_h\left[\mathbf{0}, \mathbf{I}\right] \\ Pr(\mathbf{w}_{ij}) &= \mathcal{G}_w\left[\mathbf{0}, \mathbf{I}\right] \\ Pr(c_i = k) &= \pi_k \qquad\qquad k = \{0 \ldots K\}. \end{aligned} \tag{22}$$

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. There are now two latent identity variables associated with an individual: the discrete variable c_i determines which cluster the individual belongs to and the identity vector $\mathbf{h}_i$ determines the position within this cluster. For two faces to belong to the same individual *both* of these variables must match.

4.1 Learning and Recognition

To learn the MixPLDA model we apply the standard recipe for learning mixtures of distributions (e.g. see [15] and [6]). We embed the PLDA learning algorithm inside a second instance of the EM algorithm. In the E-Step we find which cluster is responsible for each identity. In the M-Step we learn the PLDA models for each cluster based on all of the associated data. More formally: (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 an individual i belongs to the kth cluster using the likelihood term in Equation 19, where the matrix $\mathbf{A}$ has the same structure as in Equations 13 and 14. (ii) M-Step: for each cluster k, learn the associated PLDA model using data weighted by the posterior probability of belonging to the cluster.

Figure 12A shows the results of learning a model from the first 195 individuals in the XM2VTS database with minimal preprocessing. For this case we used K=2 clusters and noise and identity subspaces of dimension 8. We used 10 iterations of the outer loop of the EM algorithm, and updated the PLDA model at each iteration with 6 iterations. Interestingly, the algorithm has organized the clusters to separate men from women.

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 6 is:

$$Pr(\mathbf{x}_p, \mathbf{x}_m) = \tag{23}$$
$$\sum_{c_1=1}^{K} \int\int\int Pr(\mathbf{x}_p, \mathbf{x}_m, \mathbf{h}_m, c_m, \mathbf{w}_p, \mathbf{w}_m,) d\mathbf{h}_m d\mathbf{w}_p d\mathbf{w}_m$$

Figure 11B shows the data manifold (gray region) and predictive distributions (ellipses) for two data points. The predictive distributions are mixtures of Gaussians: there will be one contribution from each mixture component of the model. However, many data points (e.g. green point) will be almost entirely associated with the nearest mixture component and will effectively have a Gaussian predictive distribution. Notice that this is quite a sophisticated model. The data manifold is non-linear.

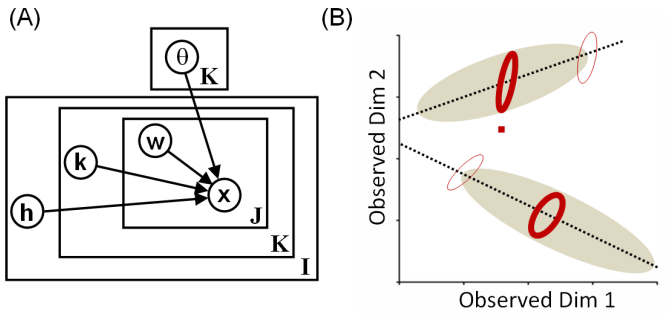


Fig. 13. Tied PLDA Model (A) Graphical model relating data $\mathbf{x}$ to identities $\mathbf{h}$, noise variables $\mathbf{w}$ and parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \Sigma\}$. (B) Predictive distribution for Tied model with two clusters. Each contains a one-dimensional identity subspace (dotted lines) and one noise factor (not shown). Gray region represents face manifold. A new gallery image (red dot) induces a predictive distribution that is a mixture of four Gaussians. Ellipse weight indicates weight of component in predictive distribution.

The shape of the predictive distribution is complex and varies depending on the gallery data.

4.2 Experiment 4

In Experiment 4, we repeat the XM2VTS experiment (figure 8a) for the mixture model. Percent correct performance improves as we move from 1 to 2 clusters (figure 12B), but adding a third 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 principle, it is possible for the two clusters with N/2 dimensions to approximate the same solution as the PLDA model with N dimensions. However, the clusters found in Figure 12A suggest that this did not happen.

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.

It would not have been easy to construct this model with a conventional distance-based approach. The representation of identity consists of one discrete variable c_i and one continuous variable $\mathbf{h}_i$ and hence measurements of distance are no longer straightforward.

5 Model 3: Tied PLDA

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

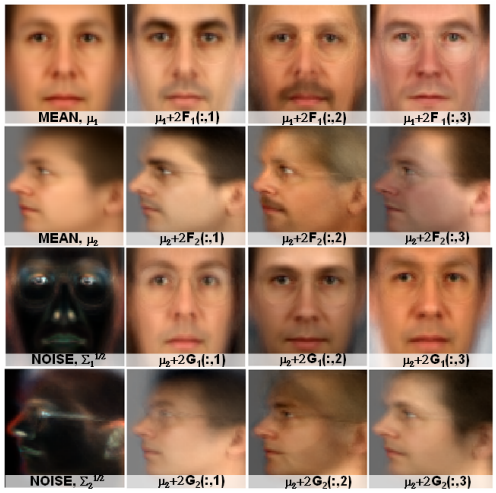


Fig. 14. Tied PLDA model for face recognition across pose. The position $\mathbf{h}_i$ in the identity subspace $\mathbf{F}$ is forced to be constant for both poses: as we move along the dimensions of the signal subspace, the basis functions look like the same person, regardless of pose (top two rows). Position in the noise subspaces $\mathbf{G}$ is not tied, so these basis functions are unrelated (bottom two rows).

are so disparate that a more powerful technique must be applied. In "tied" models [34], 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. Here, we will assume that the pose $\mathbf{k}$ is known for each observed datum $\mathbf{x}_{ijk}$ although this is not necessary. The generative model for this data is:

$$\begin{aligned} Pr(\mathbf{x}_{ijk}|\mathbf{h}_i, \mathbf{w}_{ijk}) &= \mathcal{G}_{\mathbf{x}}\left[\mu_k + \mathbf{F}_k\mathbf{h}_i + \mathbf{G}_k\mathbf{w}_{ijk}, \Sigma_k\right] \\ Pr(\mathbf{h}_i) &= \mathcal{G}_{\mathbf{h}}\left[\mathbf{0}, \mathbf{I}\right] \\ Pr(\mathbf{w}_{ijk}) &= \mathcal{G}_{\mathbf{w}}\left[\mathbf{0}, \mathbf{I}\right] \end{aligned} \tag{24}$$

The graphical model for Tied PLDA is given in Figure 13A. 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".

5.1 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 [34].

We train using 195 individuals from the XM2VTS database, with 4 frontal and 4 profile faces of each individual. Figure 14 shows the results of training the model with two clusters using 4 frontal and 4 profile images each from the first 195 individuals from the XM2VTS database with minimal preprocessing. 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.

Recognition proceeds exactly as in the PLDA model, but now likelihood terms are calculated by marginalizing the joint likelihood implicitly defined by Equations 24. The analogue of Equation 13 is:

$$\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_N\end{bmatrix}=\begin{bmatrix}\mu_{k1}\\ \mu_{k2}\\ \vdots\\ \mu_{kn}\end{bmatrix}+\begin{bmatrix}\mathbf{F}_{k1} & \mathbf{G}_{k1} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{F}_{k2} & \mathbf{0} & \mathbf{G}_{k2} & \dots & \mathbf{0}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{F}_{kn} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{G}_{kn}\end{bmatrix}\begin{bmatrix}\mathbf{h}\\ \mathbf{w}_1\\ \mathbf{w}_2\\ \vdots\\ \mathbf{w}_N\end{bmatrix}+\begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \vdots\\ \epsilon_N\end{bmatrix} \tag{25}$$

where $\mathbf{F}_{kn}$ indicates the factor matrix associated with the known pose of image n.

It is instructive to examine the predictive distribution for a gallery image (Figure 13B). If there are K viewing conditions (clusters), then this will be a mixture of K^2 Gaussians. The observed gallery datum may be associated with each of the clusters (it may be in any pose). For each possible association, it makes a prediction in every cluster (the probe image may also be in any pose). When the pose of the probe and gallery are known, the predictive distribution becomes just a single Gaussian.

5.2 Experiments 5-6: Cross Pose Identification

In Experiment 5, we use full images with the same minimal preprocessing as in experiment 1. We trained using the first 195 individuals from the XM2VTS database and 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 assumed to be known. In Figure 15A we plot % correct first match results as a function of the subspace dimension for both the tied PLDA and PLDA models. The tied PLDA model doubles performance but only from roughly 20% to 40%.

In Experiment 6, we apply the elaborate preprocessing method from Experiment 1. However, as in the previous experiment, we train using the first 195 individuals and test with the last 100. 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 15B. Peak performance for our algorithm is 87% which compares favorably to prior work by [34].

In Table 3 we compare our method to other published results across different databases. Note that that most

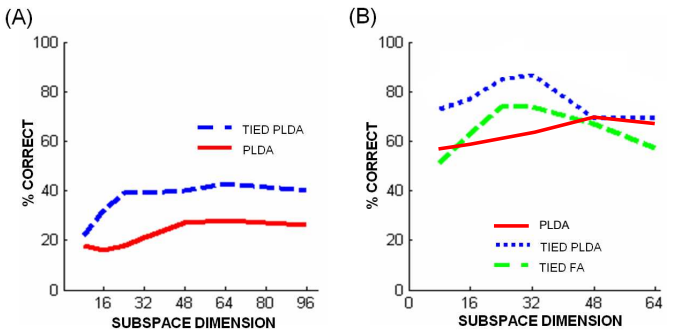


Fig. 15. Results for recognition across a 90^o pose change using Tied PLDA. (A) Experiment 5: Minimal preprocessing (B) Experiment 6: Full preprocessing. Results from PLDA and tied factor analysis model [34] plotted for comparison.

of the closest comparable algorithms (e.g. [16], [4], [34]) also use manual feature positioning.

STUDY	DATABASE	POSE DIFF(^{o})	%
Gross [16]	FERET (100)	30($Ave.$)	75
Blanz [4]	FRVT (87)	45	86
Kim [21]	XM2VTS (125)	30	53
Prince [34]	XM2VTS (100)	90	77
Our Approach	XM2VTS (100)	90	87 (82)

Table 3: Results for % correct face identification across large pose changes. Number of gallery images given in brackets after database name. Bracketed results from our approach with automated feature finding.

5.3 Experiment 7: Cross-pose Face Clustering

In experiment 7 we investigate clustering for the cross pose case using the elaborately preprocessed data. This is a challenging task involving compound recognition decisions across widely varying viewing conditions and a choice of model order. Once more, we train with the first 195 individuals in the database. On each trial we select four images each of 15 different individuals from the last 100 people in the database. Each image may be either frontal or profile and may come from any of the four recording sessions. We cluster the data using the greedy agglomerative method described in Section 3.5, replacing the PLDA model with the Tied PLDA model.

Typical results are shown in Figure 16. The algorithm successfully identified most clusters regardless of whether the faces are all frontal, all profile or a mix of both (10/15 clusters are correct). The number of splits and merges required to move to the correct clustering for this example is 8 which is slightly worse than the experimental average of 7.32 over 100 repetitions.

6 Manual vs. Automatic Keypoints

Throughout this paper, we constructed models using manually localized image keypoints. We take this approach as it makes the results easier to replicate and defines a clear upper bound on performance. However,

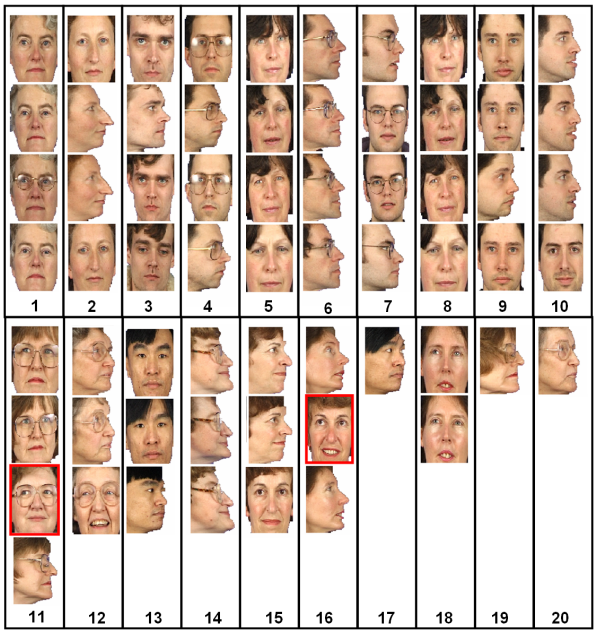


Fig. 16. Experiment 7: Clustering results for 60 frontal and profile images, consisting of 4 images each of 15 people. Blue lines divide clusters, red boxes indicate images erroneously associated with the wrong group. The algorithm found 20 clusters. Several clusters have erroneously been split (bottom right) and two erroneously merged (red boxes). In this case we would require 8 splits and merges to associate the data correctly.

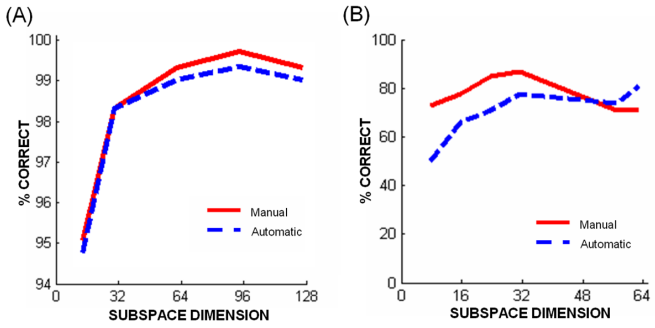


Fig. 17. Experiments 4 and 7 are repeated with automatically localized features (using the method of [34]). Performance declines slightly but remains very high.

it might be that our models are unusually susceptible to inaccurate keypoint localization. In order to check that this was not the case, we repeated the frontal and cross-pose XM2VTS experiments using gallery images that were manually labeled, but probe images that were automatically labeled using the method of [34]. The results can be seen in Figure 17. Performance declines slightly with automatic labeling, but is still very high. Our feature localizer is not particularly sophisticated and it is likely that these results could be improved upon.

7 Discussion

In this paper, we have presented a series of probabilistic models for face recognition. Our key contribution is a probabilistic approach to LDA that sidesteps the small-sample problem and has a more sophisticated noise model. We have demonstrated that there are empirical reasons to favor our approach for both frontal and cross-pose face recognition. Linear discriminant analysis is a very general technique and this method could find application in many other areas of computer vision. Our probabilistic approach also leads to two non-linear extensions that are more expressive: mixtures of PLDA models and tied PLDA models.

7.1 Relation to other methods

A probabilistic framework is beneficial for several reasons. First, the posterior distribution provides a measure of the uncertainty on the decision. Second, we can apply priors to different matching hypotheses. For example, in a surveillance system, identification in one camera could be used to modify the prior probability of the same person appearing in a nearby camera. Third, probabilities make it easy to combine different sources such as image sub-models or data from different biometric domains without the need to learn weights.

There have been other probabilistic algorithms for face recognition, most notably the work of Liu and Wechsler [26] and Moghaddam [28], which we considered in Section 1.2. Zhou and Chellapa [47] presented a system which also acknowledged uncertainty in the identity representation, but inference was quite different and did not produce a posterior probability over matches.

The most closely related work to this paper is by Ioffe [19] who presented a similar algorithm to Model 1, which he also termed probabilistic linear discriminant analysis. The main differences are (i) his model requires projection of the data down onto a few principal components rather than modelling the original data vector and is hence not a fully probabilistic description of the data. If the dimensions of the data have very different scales, it is quite possible to erroneously throw out critical information with this approach. (ii) there is no equivalent of the per-pixel noise vector Σ (and hence it misses out on the performance gain illustrated in Figure 7b): the results of Ioffe's algorithm are consequently very similar to the blue dashed line in this figure. (iii) he proposes a closed form learning algorithm. No such solution is known for the PLDA models proposed here with diagonal covariance - to get the extra performance boost of our method we must resort to iterative optimization techniques such as the EM algorithm.

The tied PLDA model is a generalization of Tied Factor Analysis [34]. The latter model uses a bilinear mechanism to compare images across disparate viewing conditions. Tied PLDA uses the same mechanism, but additionally models a within-individual subspace

in each viewing condition. This is shown to improve matching performance across large pose differences.

7.2 Further Work

Many other methods can also be understood in terms of LIVs. For example, distance-based methods based on independent components analysis [1] have a clear generative interpretation. The models described here have been purely statistical in nature. However, a sensible future research direction would be to combine this form of inference with work which attempts to formally model the physical process of face image creation, such as that of Blanz et al. [5].

Some successful approaches to face recognition in unconstrained conditions (e.g. [11], [30]) do not build a probabilistic description of the data, but seek discriminative patch-based features which classify pairs of faces into "same" or "different". These methods have complementary properties to our system: they are specialized to face verification and cannot easily be adapted for other tasks such as clustering. They may also struggle with extreme variations such as 90^o pose differences as feature extraction is based on cross-correlation between images. An interesting direction for future work would be to construct a LIV model based on similar features which retains the strengths of both approaches.

Appendix A: Learning LIV Models

The goal of this section is to present the EM algorithm updates for learning models 1 and 3. The basic approach is to rewrite both E-Step and M-Step to resemble the simpler factor analysis model by assimilating terms. More details about learning factor analysis models can be found in [6], [36].

E-Step: We simultaneously estimate the joint probability distribution of all latent variables $\mathbf{h}, \mathbf{w}$ that pertain to each given individual i. We combine together the generative equations for all of the data pertaining to individual i. For models 1 and 3 this is given in Equations 13 and 25 resulting in a likelihood and prior terms:

$$Pr(\mathbf{x}_i|\mathbf{y}_i,\theta) = \mathcal{G}_{\mathbf{x}}\left[\mu' + \mathbf{A}\mathbf{y}_i, \Sigma'\right] \quad (26)$$

$$Pr(\mathbf{y}_i) = \mathcal{G}_{\mathbf{h}}\left[0, \mathbf{I}\right] \quad (27)$$

where $\mathbf{A}$, Σ' and $\mathbf{y}_i$ are defined as in Equations 14 and 17. The model defined in Equations 26 and 27 takes the form of a factor analysis model. Applying Bayes' rule to calculate the posterior, we get:

$$Pr(\mathbf{y}_i|\mathbf{x}_i,\theta) \propto Pr(\mathbf{x}_i|\mathbf{y}_i,\theta)Pr(\mathbf{y}_i) \quad (28)$$

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:

$$E[\mathbf{y}_i] = (\mathbf{A}^T\Sigma'^{-1}\mathbf{A} + \mathbf{I})^{-1}\mathbf{A}^T\Sigma'^{-1}(\mathbf{x}_i - \mu') \quad (29)$$

$$E[\mathbf{y}_i\mathbf{y}_i^T] = (\mathbf{A}^T\Sigma'^{-1}\mathbf{A}^T + \mathbf{I})^{-1} + E[\mathbf{y}_i]E[\mathbf{y}_i]^T \quad (30)$$

M-Step: In the M-Step, we aim to update the values of the parameters $\theta = \{\mu, \mathbf{F}, \mathbf{G}, \mathbf{\Sigma}\}$. To do this, we write a single equation for each observed data point. For Models 1 and 3, these are respectively:

$$\mathbf{x}_{ij} = \mu + \begin{bmatrix}\mathbf{F} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\mathbf{h}_i \\ \mathbf{w}_{ij}\end{bmatrix} + \epsilon_{ij} \quad (31)$$

$$\mathbf{x}_{ijk} = \mu + \begin{bmatrix}\mathbf{F}_k & \mathbf{G}_k\end{bmatrix}\begin{bmatrix}\mathbf{h}_i \\ \mathbf{w}_{ijk}\end{bmatrix} + \epsilon_{ijk} \quad (32)$$

Each has the form:

$$\mathbf{x} = \mu + \mathbf{B}\mathbf{z}_{ij} + \epsilon_{ij} \quad (33)$$

For Model 1 we optimize:

$$Q(\theta_t, \theta_{t-1}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\int Pr(\mathbf{z}_i|\mathbf{x}_{i1\ldots iJ}, \theta_{t-1}) \log[Pr(\mathbf{x}_{ij}|\mathbf{z}_i)Pr(\mathbf{z}_i)]d\mathbf{z}_i \quad (34)$$

where t is the iteration index. The first log probability term in Equation 34 can be written as:

$$\log[Pr(\mathbf{x}_{ij}|\mathbf{z}_i\theta_t)] = K - 0.5\left(\log|\Sigma^{-1}| + (\mathbf{x}_{ij} - \mu - \mathbf{B}\mathbf{z}_i)^T\Sigma^{-1}(\mathbf{x}_{ij} - \mu - \mathbf{B}\mathbf{z}_i)\right) \quad (35)$$

where K is an unimportant constant. We substitute this term into Equation 34 and take derivatives with respect to $\mathbf{B}$ and Σ. The second log term in Equation 34 has no dependence on these parameters. We equate these derivatives to zero and re-arrange to provide the following update rules:

$$\mu = \frac{1}{IJ}\sum_{i,j}\mathbf{x}_{ij} \quad (36)$$

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

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

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 29 and 30 using the equivalence between Equations 13 and 14. The updated values of $\mathbf{F}$ and $\mathbf{G}$ are retrieved from the new value of $\mathbf{B}$. The M-Step for Model 3 proceeds in exactly the same way, but we separately estimate the parameters for each cluster.

Appendix B: Efficient Evaluation of Likelihood

In Section 3.3 we showed that the likelihood of the probe matching a gallery datum takes a Gaussian form. To calculate this likelihood we must evaluate the quadratic term in the exponent of the Gaussian. Since the measurements are of high dimension this appears to be costly. Here, we show that this is not the case - in fact the quadratic term may be calculated by (i) projecting the data onto a subspace and (ii) calculating a dot product.

To demonstrate this, we use the Woodbury matrix identity (matrix inversion lemma) to re-write the information matrix of the Gaussian in question:

$$\begin{aligned}\Lambda &= (\mathbf{A}\mathbf{A}^T + \Sigma')^{-1} \\ &= \Sigma'^{-1} - \Sigma'^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}^T\Sigma'^{-1}\mathbf{A})^{-1}\mathbf{A}^T\Sigma'^{-1} \\ &= \Sigma'^{-1} - \mathbf{D}\mathbf{D}^T \end{aligned} \quad (37)$$

where we define $\mathbf{D} = \Sigma^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}^T\Sigma'^{-1}\mathbf{A})^{-1/2}$ As discussed in Section 3.3 we need to calculate a quadratic term $\mathbf{x}_p\Lambda_{pp}\mathbf{x}_p$ where Λ_{pp} is the top left part of the information matrix Λ. In other words, we need to calculate part of $\mathbf{x}^T\Sigma^{-1}\mathbf{x} - \mathbf{x}^T\mathbf{D}_p\mathbf{D}_p^T\mathbf{x}$. The first term is efficiently calculated as Σ is diagonal. The second part is also efficiently calculated as the term $\mathbf{D}_p$ has a number of columns that depends on the noise and signal subspace. Hence we can first calculate $\mathbf{D}_p^T\mathbf{x}$ and then calculate the magnitude of this vector. The final computation is of the same order as traditional LDA methods which project into a subspace and then measure a distance. Calculations under the joint perspective can also be made efficient using similar techniques.

REFERENCES

[1] M.S. Bartlett, H.M. Lades and T.J. Sejnowski, "Independent component representations for face recognition," *Proc. SPIE*, Vol. 3299, pp. 528-539, 1998.

[2] 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.

[3] M. Isard and A. Blake, "CONDENSATION – conditional density propagation for visual tracking,", *IJCV* Vol. 29, pp. 5-28, 1998.

[4] 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.

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

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

[7] D. Cai, X.F. He, Y.X. Hu, J.W. Han,"Learning a Spatially Smooth Subspace for Face Recognition," *CVPR*, pp. 1-7, 2007.

[8] 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.

[9] 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.

[10] M. Everingham, J. Sivic and A. Zisserman, ""Hello! My name is... Buffy" C automatic naming of characters in TV video", *BMVC*, 2006.

[11] A. Ferencz, E. Learned-Miller and J. Malik. "Learning to locate informative features for visual identification," *IJCV*, Vol.77, pp. 3-24, 2008.

[12] R. Fergus, P. Perona and A. Zisserman, "Object class recognition by unsupervised scale invariant learning," *CVPR*, pp.264-271, 2003.

[13] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid, "Is that you? Metric Learning Approaches for Face Identification," *ICCV*, 2009.

[14] A.S. Georghiades, P.N. Belhumeur and D.J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose. *PAMI,* Vol 23, pp. 643-660, 2001

[15] 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.

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

[17] X. He , S. Yan, Y. Hu, P. Nihogi and H. Zhang, "Face recognition using Laplacianfaces,", *PAMI,* Vol. 27, pp. 328-340, 2005.

[18] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," *University of Massachusetts, Amherst, Technical Report 07-49*, October, 2007.

[19] S. Ioffe, "Probabilistic linear discriminant analysis," *ECCV,* pp. 531-542, 2006.

[20] E. Jones and S. Soatto, "Layered active appearance models," *ICCV*, pp. 17-21, 2005.

[21] 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.

[22] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar, "Attribute and Simile Classifiers for Face Verification," *ICCV*, 2009.

[23] L. Fei-Fei, R. Fergus and P. Perona, "A Bayesian approach to unsupervised one-shot learning of object categories," *CVPR*, pp. 1134-1141, 2003.

[24] P. Li, J. Warrell, J. Aghajanian and S.J.D. Prince, "Context based additive logistic model for facial keypoint localization", *BMVC*, 2010.

[25] Z. Li and X. Tang, "Bayesian face recognition using support vector machine and face clustering," *CVPR*, pp. 259-265, 2004.

[26] C. Liu and H. Wechsler, "Probabilistic reasoning models for face recognition," *CVPR*, pp. 827-832, 1998.

[27] S. Lucey and T. Chen, "Learning Patch Dependencies for Improved Pose Mismatched Face Verification," *CVPR*, Vol. 1, pp. 17-22, 2006.

[28] B. Moghaddam, T. Jebara and A. Pentland, "Bayesian face recognition," *Pattern Recognition,* Vol. 33, pp. 1771-1782, 2000.

[29] Hieu V. Nguyen and Li Bai, Cosine Similarity Metric Learning for Face Verification, *Asian Conference on Computer Vision (ACCV)*, 2010.

[30] E. Nowak and F. Jurie,"Learning visual similarity measures for comparing never seen objects",*CVPR* pp. 1–8, 2007.

[31] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, pp. 971-987, 2002.

[32] V. Perlibakas, "Distance Measures for PCA-Based Face Recognition," *Patt. Rec. Letters*, Vol. 25, pp. 711-724, 2004.

[33] Nicolas Pinto and David Cox, Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition, *FG*, 2011.

[34] S.J.D. Prince, J. Warrell, J.H. Elder and F.M. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE PAMI*, Vol. 30, pp. 970-984, 2008.

[35] S.J.D. Prince and J.H. Elder, Probabilistic linear discriminant analysis for inferences about identity, *ICCV*, 2007.

[36] R. Rubin and D. Thayer, "EM Algorithms for ML Factor Analysis," *Psychometrica,* Vol 47, pp. 69-76, 1982.

[37] C. Rother, V. Kolmogorov and A. Blake, "GrabCut: interactive foreground extraction using iterated graph cuts," in Proc. *SIGGRAPH*, pp. 309-314, 2004.

[38] Yaniv Taigman, Lior Wolf, and Tal Hassner, "Multiple One-Shots for Utilizing Class Label Information," *BMVC*, 2009.

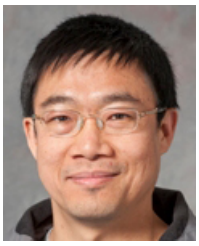
[39] M. Turk and A.P. Pentland, "Face recognition Using eigenfaces," *CVPR*, pp.586-591, 1991.

[40] X.Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," *CVPR,* Vol. 2, pp.564-569, 2004.

[41] X. Wang and X. Tang, "Unified Framework for Subspace Face Recognition," *PAMI*, Vol. 26, pp. 1222-1228, 2004.

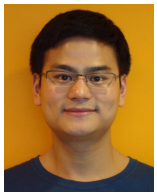
[42] X. Wang and X. Tang, "Bayesian Face Recognition Using Gabor Features," *WBMA,* pp. 70-73, 2003.

[43] X. Wang and X. Tang, "Random sampling for subspace face recognition ," *IJCV*, Vol. 70, pp. 91-104, 2006.
[44] Lior Wolf, Tal Hassner, and Yaniv Taigman, "Descriptor Based Methods in the Wild," *ECCV*, 2008.
[45] M.H. Yang "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods" *ICFAG*, pp. 215-220, 2002.
[46] W. Zhao, R.Chellappa, A.Rosenfeld and J. Phillips, "Face Recognition: A literature Survey," *ACM Computing Surveys*, Vol. 12, pp. 399-458, 2003.
[47] S.K. Zhou and R. Chellappa, "Probabilistic Identity characterization for face recognition," *CVPR*, Vol. 2, pp. 805-812, 2004.
[48] http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/



Peng Li Li received the B. Eng. and M. Eng. degrees in automation from North China Electric Power University in 1993 and 1998, and the PhD degree in Electrical and Electronic Engineering from the Nanyang Technological University in 2006. He was a lecturer in North China Electric Power University from 1998 to 2002. From 2005 to 2007, he was a computer vision engineer in Stratech Systems Ltd. Singapore. He was a research associate in University of Bristol in 2007. From 2008 to 2010, he was a research fellow in the Department of Computer Science, University College London. He is now a research associate in University of Bristol. His research interests include computer vision, machine learning and bioinformatics, particularly in face recognition and Bayesian methods, etc. He is a member of IEEE.

Umar Mohammed Mohammed is a PhD student at University College London in the dept of computer science. He has a BSc in Mathematics from City University and an MSc in Vision, Imaging and Virtual Environments from University College London. His research interests include face recognition and image based rendering.



Yun Fu Fu is currently pursuing a PhD degree in dept of Computer Science at University College London. He received a BSc in Computer Science from Xi'an University of Science and Technology and an MSc in Vision, Image and Virtual Environments from University College London. His research interest is in face recognition.

James H. Elder Elder received the B.A.Sc. degree in Electrical Engineering from the University of British Columbia in 1987 and the Ph.D. degree in Electrical Engineering from McGill University in 1995. From 1995 to 1996 he was with the NEC Research Institute in Princeton, NJ. He joined the faculty of York University in 1996, where he is presently Associate Professor. His research interests are in computer and human vision. Recent work has focused on natural scene statistics, perceptual organization, contour processing, attentive vision systems, and face detection.

Simon J.D. Prince Prince received his Ph.D. in 1999 from the University of Oxford for work concerning human stereo vision. He has a diverse background in biological and computing sciences and has published papers across the fields of biometrics, psychology, physiology, medical imaging, computer vision, computer graphics and HCI. He worked as a research scientist in Oxford, Singapore and Toronto. Prince is currently a Reader in Computer Science at University College London.