Class Dependent Factor Analysis and Its Application to Face Recognition

Birkan Tunç[a], Volkan Dağlı[a], Muhittin Gökmen[b]

[a]*Istanbul Technical University, Informatics Institute, 34469, Turkey*
[b]*Istanbul Technical University, Computer Engineering Department, 34469, Turkey*

Abstract

We propose a class dependent factor analysis model (CDFA) which can be used in the general face recognition task under certain variations. The model utilizes the class information in a supervised manner to define a separate manifold for each class. Inside each manifold, a mixture of Gaussians is designated to handle the variation. The proposed model learns the system parameters in a probabilistic framework, allowing a Bayesian decision model. A manifold embedding technique is incorporated to handle the nonlinearity introduced by the variation; hence, a novel connection between manifold learning and probabilistic generative models is proposed. CDFA has better recognition accuracy and scalability over a classical factor analysis model. Experimental evaluations on the face recognition under changing illumination conditions and facial expressions indicate the ability of the proposed model to handle different types of variation. The achieved recognition rates are comparable to the state-of-art results, while it is also shown that the recognition rate does not decrease critically as the number of gallery identities increases.

Keywords:
Factor analysis, face recognition, manifold learning, generative models, probabilistic inference

1. Introduction

The face recognition domain still needs robust and scalable algorithms to handle real life variations like pose, illumination, and facial expressions. Although various approaches have been studied by many authors, there is not a generic method which can deal with different variations with a promising scalability.

In this paper, we propose a novel object recognition framework called class dependent factor analysis model (CDFA) to increase the scalability of a recognition system and to overcome certain variations that may be present during the data acquisition. The primary goal is to develope a generic method for the face recognition problem under a selected type of variation. Combining different types of variation is another challenge that is left to be investigated as a future work. To this extend, the grammatical structure of the classical factor analysis model is redesigned to enhance its semantic setup.

The method is initially developed to handle illumination and expression changes; however, the viewpoint change can also be analyzed with the same approach after preprocessing or feature point selection steps as introduced in [1, 2, 3]. Such requirements for the pose variation are inevitable due to the high nonlinearity of image formation with changing view points.

1.1. The Classical Approaches And Their Limitations

Many popular face recognition algorithms use holistic approaches in conjunction with appearence-based models [4]. Appearance-based models utilize the actual pixel intensities, and this fact alone is enough to damage the effective signal-noise ratio since individual pixels tend to change dramatically under certain variations like illumination and facial expression. A common approach to handle such variations is to define a lower dimensional subspace in which the useful statistics are more definite compared to the noise. As an example, Principal Component Analysis (PCA) [5] is used to define a subspace where the variance on principal axes is stimulated.

When the utilized appearance-based method depends on a dimensionality reduction technique, factor analysis happens to be the main actor. Factor analysis is a well known and commonly used approach in the data analysis community. Although its early development traces to the beginning of the century, it is still one of the most popular multivariate statistical analysis tools in applied science domain [6]. Its main formulation is a linear generative model

$$\mathbf{x} = \mathbf{W}\mathbf{c} + \boldsymbol{\epsilon}, \tag{1}$$

where the weighted average of lower dimensional factors, $\mathbf{c}$, is taken to generate a higher dimensional signal, $\mathbf{x}$. In this view, factor analysis can be seen as a dimensionality reduction technique when the inverse mapping of $\mathbf{W}$ is considered.

Factor analysis is a powerful tool, especially when it is used for the dimensionality reduction. The classification is achieved in the lower dimensional subspace instead of the noisy higher dimensional pixel space. The very same idea is exploited in PCA and Linear Discriminant Analysis (LDA). They both have the same underlying generative model but different ways to get the mapping $\mathbf{W}$.

The first limitation of such approaches arises with the *common subspace* constraint: The mapping, $\mathbf{W}$, is common for all classes. The discrimination among classes is achieved by the deployment of the class centroids on the coordinate system. Such a modeling is insufficient when the effect of the varia-

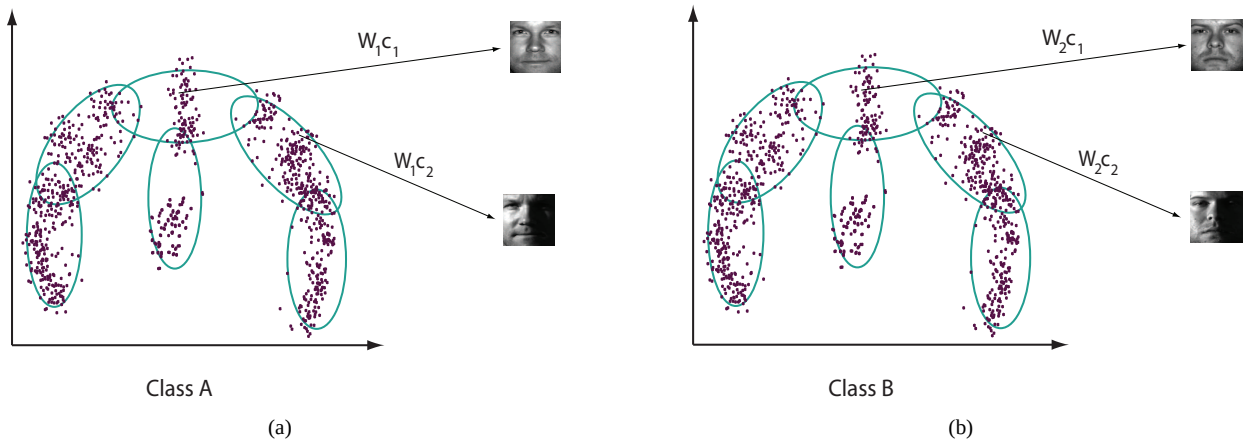


Figure 1: Illustration of individual manifolds of different identities. Any point on the manifold corresponds to a variation type. The intrinsic geometry is common among different manifolds. This behavior results in the same variation type for same coordinate values.

tion is more dominant than the class characteristics. In such a case, the coordinates of the points are mostly determined by the variation type. A well known example is the fact that the images of different people under same illumination lie closer in such subspaces compared to the images of a single person under different illumination.

Embeddings like PCA can solve problems caused by statistically well behaving noise terms. However, under a problematic variation, individual or interclass statistics may be altered dramatically preventing a useful discrimination. An elegant idea is to distinguish the real signal (identity of the image) and the noise caused by variation (differences imposed by illumination). In LDA [7], this idea is exploited by controlling the inter-class and intra-class variances. However, it keeps the same generative model and still tries to assign a unique coordinate vector, $\mathbf{c}$, for different samples of the same class.

In the real life, recognition and classification tasks usually deal with variations that result in nonlinear data geometries. Therefore, one requires more sophisticated mathematical tools to investigate [8]. While the most of the methods use a linearity constraint over the data geometry, relatively new techniques called "Manifold Learning" have been developed to eliminate the linearity constraints [9, 10, 11]. The main idea behind manifold learning is to utilize local geodesic distances instead of global Euclidean distances.

1.2. Overview of the CDFA Framework

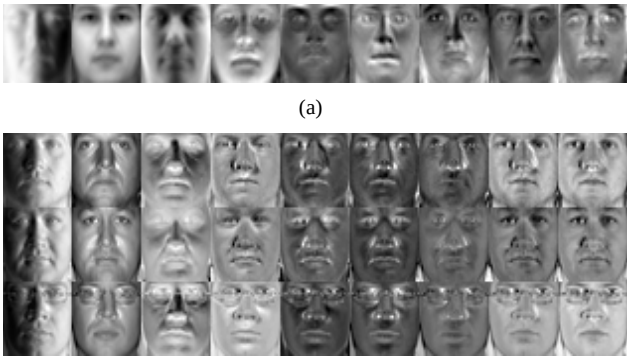
The design of the framework starts with the reformulation of the factor analysis model under a variation such as illumination. An observation $\mathbf{x}_{ik}$, which belongs to the class i and has a variation k, is generated by the model

$$\mathbf{x}_{ik} = \mathbf{W}_i \mathbf{c}_k + \boldsymbol{\epsilon}_k \,. \tag{2}$$

With this formulation, we introduce individual factor loadings, $\mathbf{W}_i$, for each class i, instead of a common loading matrix for all classes. However, the factors, $\mathbf{c}_k$ (coordinates on the lower dimensional subspace), are common for all classes and related to the variation type. The geometric interpretation yields different manifolds for different classes while all manifolds have exactly same intrinsic geometries. Inside two manifolds, points having same local coordinates correspond to the same variation type. This interpretation is illustrated in Figure 1. Several important aspects of this formulation should be mentioned:

- Each class has its own subspace/manifold. Therefore, discrimination between classes is performed by the distance to the manifold instead of the distance within the manifold. Inside each individual manifold, a mixture of Gaussians may be defined to model the variation.
- Coordinate vectors, $\mathbf{c}_k$, represent the variation type instead of class identities. Thus, the determination of the variation value is explicitly provided.
- Class identities are stored as factor loadings in matrix $\mathbf{W}_i$. The variation does not condition the structure of the matrix since it is already modeled by the factors.
- The intrinsic dimensionality of manifolds is fixed once determined during the bootstrap. Nevertheless, the actual dimensionality in which the recognition is performed is n since the manifolds are embedded in $\Re^n$, where n is the number of pixels in images.
- A manifold learning step is employed to derive the reduced dimensional coordinates, $\mathbf{c}_k$. Thus, a connection between manifold learning and probabilistic generative models is proposed. This can be seen as an initial step towards nonlinear probabilistic models.

The difference between individualized and common factor loadings can be observed in Figure 2. The proposed method introduces basis sets which are specific to their corresponding classes. With this setting, one can synthesize different images of a person under different conditions like changing light source positions.



(a)

(b)

Figure 2: Demonstration of the semantic difference between (a) a common basis set generated by a classical approach (SVD was used for this example) and (b) class dependent basis sets generated by the proposed approach. Each basis set includes the class information intrinsically. For this example, images under changing illumination conditions were used.

A critical feature of the method is its generic structure. We do not employ any physical or geometrical attributes of the concerned variation. Hence, any variation lying on a smooth manifold can be modeled by the proposed method.

2. Connections to Previous Works

The proposed method has an analogous formulation with the probabilistic interpretation of PCA [12, 13]. Both approaches tackle with finding lower dimensional representations of observations under some prior assumptions. The main difference is that the proposed method derives class specific coordinates and accounts for the variation explicitly.

Similar frameworks were introduced in [14] and [15]. Both works dealt with individualized subspaces. The actual improvement over [14] is that CDFA has a more generic structure which can be used for the general classification problem whereas only illumination was considered in [14]. The authors of [14] used spherical harmonics to calculate class specific bases. The results are limited to illumination as the spherical harmonics can not be generalized to other types of variation. Having a complete probabilistic framework is our advantage over the work done in [15].

The authors of [16] developed a cone model to solve the face recognition problem with varying illumination. They argued that the set of images of an object in a fixed pose but under all possible illumination define a convex cone. The approach requires a few images of each gallery identity to estimate its surface geometry and albedo map. That model illustrates the real power of the subspace analysis; nevertheless, it is again constrained to be useful only for illumination and may not work with a single observation.

Other techniques such as [17, 18, 19, 20, 21] suffer from being useful only for the specific variation type that they have been developed for. We try to propose a method which can be used for different variations.

A comparable work was performed in [1]. Authors defined a common subspace for class identities yet different transformation matrices (factor loadings) for different poses. Keeping the class information inside the coordinate vectors inherits an important disadvantage of classical subspace methods: as the number of classes increases, the subspace dimensions also need to be increased to sustain the scalability. The same idea was used in [22] again for pose variations.

The probabilistic approaches for the discriminative subspace analysis were proposed in [23] and [24]. Both solutions were based on LDA with different settings. In [23], authors defined a three layer decision process. At the initial layer, identity is drawn from a common Gaussian distribution. Then, at the second layer a perturbation is applied by another Gaussian. Finally, the third layer defines a projection from the latent space to the observation space. In [24], the model introduced in [1] was improved by employing different projections from the latent space to the observation space: one for the between-individual subspace and one for the within-individual subspace. Both models still assume common subspaces for different identities.

Compressive sensing and sparse representation were utilized in [25] and [26]. The subspace analysis was performed on the basis of compressive sensing theory. Both techniques can be used for different types of variation. We use these methods in benchmarks against facial expressions.

3. Mathematical Background

The proposed method can be summarized as a two step probabilistic framework. The first step is a bootstrap phase in which useful statistics are calculated. A manifold embedding technique is employed at this step to define the geometry of the subspace. The second step includes regular training and testing tasks. Framework starts with analyzing the underlying manifold. A bootstrap database, consisting of identities with several observations (people with several images), is collected for this purpose. The identities of the bootstrap database are different than the ones to be recognized; any suitable database can be selected.

To simplify the calculations, the equation (2) may be rewritten in an element-wise form as

$$x_{ik} = \mathbf{w}_i^T \mathbf{c}_k + \epsilon_k \,, \tag{3}$$

where x_{ik} is an element of the observation vector, $\mathbf{x}_{ik}$. Similarly, the vector $\mathbf{w}_i$ is the corresponding row of the matrix $\mathbf{W}_i$. Again, ϵ_k is the corresponding element of the error vector, $\boldsymbol{\epsilon}_k$. Such an element-wise formulation ignores the correlations among pixels while introducing new correlations among columns of $\mathbf{W}_i$. Unlike the classical factor analysis model, the factors are treated as deterministic variables which are calculated during the manifold learning step. Moreover, Gaussian priors are defined on the vector $\mathbf{w}$ and the constant ϵ_k as

$$\begin{aligned} p(\mathbf{w}) &\sim \mathcal{G}(\boldsymbol{\mu}, \boldsymbol{\Omega}^{-1}), \\ p(\epsilon_k) &\sim \mathcal{G}(0, \sigma_k^2). \end{aligned} \tag{4}$$

The proposed method is detailed through the following sections and summarized in Table 1 at the end of Section 3. For all

formulations a single variation such as illumination is considered to be effective.

3.1. Manifold Learning: Bootstrap

The aim of this step is to define a mapping, $\mathbf{M}$, from the high dimensional image space to the lower dimensional variation space as in

$$\mathbf{c}_k = \mathbf{M}^T \mathbf{x}_k \ . \tag{5}$$

The term *variation space* is chosen to emphasize that the coordinates of the subspace are related to the variation. Locality Preserving Projections (LPP) [9] is employed as a manifold embedding technique. This technique tries to preserve the intrinsic geometry and the local structure of the underlying manifold. The error function of LPP can be interpretated as a summation over distances between close data points.

$$\varepsilon = \sum_k \sum_j \left(c_k - c_j\right)^2 S_{kj} \ , \tag{6}$$

where c_k is the one-dimensional representation of the data point, $\mathbf{x_k}$. The relation between $\mathbf{x_k}$ and c_k is defined as $c_k = \mathbf{m}^T \mathbf{x_k}$, where the vector $\mathbf{m}$ is a column of the mapping $\mathbf{M}$. The coefficients, S_{kj}, represent the similarity index. They may be defined as

$$S_{kj} = \begin{cases} \exp(-\|\mathbf{x_k} - \mathbf{x_j}\|^2/t), & \|\mathbf{x_k} - \mathbf{x_j}\|^2 < \epsilon \ , \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where ϵ defines the radius of the local neighborhood. The cost function (6) can be rewritten as

$$\begin{aligned} \varepsilon &= \tfrac{1}{2} \textstyle\sum_k \sum_j \left(c_k - c_j\right)^2 S_{kj} \\ &= \tfrac{1}{2} \textstyle\sum_k \sum_j \left(\mathbf{m}^T \mathbf{x_k} - \mathbf{m}^T \mathbf{x_j}\right)^2 S_{kj} \\ &= \mathbf{m}^T \mathbf{A}(\mathbf{D} - \mathbf{S})\mathbf{A}^T \mathbf{m} \\ &= \mathbf{m}^T \mathbf{A}\mathbf{L}\mathbf{A}^T \mathbf{m}, \end{aligned} \tag{8}$$

where the matrix $\mathbf{A}$ has data points, $\mathbf{x}_i$, as its columns. $\mathbf{D}$ is a diagonal matrix, and its entries are column sums of $\mathbf{S}$. $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the laplacian matrix. By introducing a constraint as $\mathbf{m}^T \mathbf{A}\mathbf{D}\mathbf{A}^T \mathbf{m} = 1$, the minimization of (8) is transformed to the generalized eigenvalue problem,

$$\mathbf{A}\mathbf{L}\mathbf{A}^T \mathbf{m} = \lambda \mathbf{A}\mathbf{D}\mathbf{A}^T \mathbf{m}. \tag{9}$$

Then, the eigenvectors corresponding to minimum eigenvalues are selected to construct a linear mapping, $\mathbf{M}$.

During our experiments, the following settings are used: A bootstrap database, $\{\mathbf{x}_{ik}\}$, is collected for the concerned variation type. Each identity i has several images corresponding to different values of the variation. The distances between images are calculated in a supervised manner. In other words, the similarity indexes in (7) are calculated based on variation labels. Details can be gathered from [27, 9].

Using such a supervised approach draws an upper bound to the dimensionality of the manifold. Since the rank of the generalized eigenvalue problem in (9) is determined by the number of discretized variation labels (different types of illumination), the dimensionality is at most the number of different variation labels in the bootstrap database.

An example embedding of the bootstrap database into two dimensional subspace is illustrated in Figure 3(a). A further averaging step is performed to discard the effect of the identity completely. As shown in Figure 3(b), averages over identities are calculated to represent each variation type.

The averaging is applied as follows: For each observation, $\mathbf{x}_{ik}$, the reduced dimensional coordinates, $\mathbf{c}_{ik}$, are calculated by $\mathbf{c}_{ik} = \mathbf{M}^T \mathbf{x}_{ik}$. Then, for each variation label, k, the average over all identities is taken by

$$\mathbf{c}_k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{c}_{ik} \ , \tag{10}$$

where N is the total number of identities in the bootstrap database.

3.2. Learning Factors and Other Statistics: Bootstrap

In this stage, the parameters of prior distributions defined in (4) are calculated using the bootstrap database, $\mathbf{X} = \{\mathbf{x}_{ik}\}$. Considering the element-wise formulation (3) and priors, the conditional and the marginal distributions over the variable x_k are

$$\begin{aligned} p(x_k|\mathbf{w}, \mathbf{c}_k) &\sim \mathcal{G}(\mathbf{w}^T \mathbf{c}_k, \sigma_k^2), \\ p(x_k) &= \int p(x_k|\mathbf{w}, \mathbf{c}_k) p(\mathbf{w}) d\mathbf{w}. \end{aligned} \tag{11}$$

Both the prior and the conditional distributions are Gaussians in (11), and this makes the resulting marginal distribution, $p(x_k)$, to be another Gaussian. Indeed, we do not need to solve this integral form analytically since the mean value and the variance can be easily evaluated by the following identities which employ the equation (3).

$$\begin{aligned} E[x_k] &= E[\mathbf{w}^T \mathbf{c}_k + \epsilon_k] = \boldsymbol{\mu}^T \mathbf{c}_k, \\ E\left[(x_k - E[x_k])^2\right] &= \mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2. \end{aligned} \tag{12}$$

These two parameters are sufficient to define the marginal as

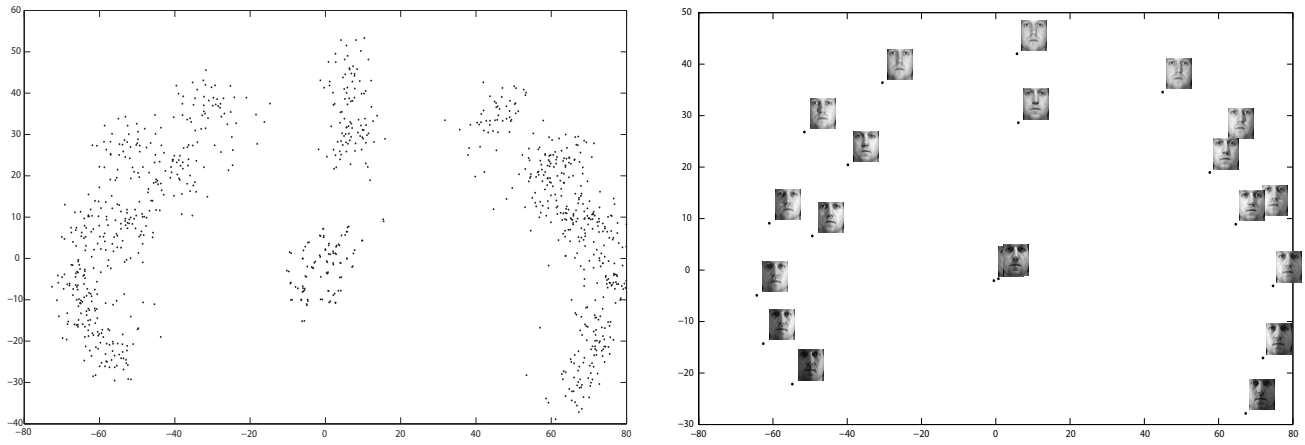
$$p(x_k) \sim \mathcal{G}(\boldsymbol{\mu}^T \mathbf{c}_k, \mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2). \tag{13}$$

The bootstrap database can be used at this point to calculate the unknown parameters, $\boldsymbol{\Omega}$, $\boldsymbol{\mu}$, and σ_k^2 by maximizing the likelihoods. The Likelihood to be maximized is the empirical likelihood of the observed points, x_{ik}. Assuming i.i.d observations, the total log likelihood over observations is

$$\ln p(X|\boldsymbol{\mu}, \boldsymbol{\Omega}, \sigma_k^2) = \sum_i^N \sum_k^K \ln p(x_{ik}), \tag{14}$$

where the upper bounds N and K denote the number of identities and different values of the variation in the bootstrap gallery, respectively. After omitting the constant terms which are not related to the unknown parameters, the cost functional becomes

$$\mathcal{J} = -\sum_i^N \sum_k^K \ln\left(\mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2\right) - \sum_i^N \sum_k^K \frac{\left(x_{ik} - \boldsymbol{\mu}^T \mathbf{c}_k\right)^2}{\mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2} . \tag{15}$$



(a) (b)

Figure 3: Embedding results of LPP: (a) 2D embedding of the bootstrap database with changing illumination. (b) Average coordinates corresponding to different illumination conditions. These coordinates are invariant to the identity.

In order to determine the unknowns which minimize the cost functional, we simply take partial derivatives with respect to those and set them equal to zero. By this way, a system of nonlinear equations is obtained as

$$\sigma_k^2 = \frac{1}{N}\sum_i^N \left(x_{ik} - \boldsymbol{\mu}^T \mathbf{c}_k\right)^2 - \mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k \,, \tag{16}$$

$$N \sum_k^K \frac{\mathbf{c}_k \mathbf{c}_k^T}{\mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2} = \sum_k^K \frac{\mathbf{c}_k \mathbf{c}_k^T}{\left(\mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2\right)^2} \sum_i^N \left(x_{ik} - \boldsymbol{\mu}^T \mathbf{c}_k\right)^2 \,, \tag{17}$$

$$\left(\sum_k^K \frac{\mathbf{c}_k \mathbf{c}_k^T}{\mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2}\right) \boldsymbol{\mu} = \frac{1}{N} \sum_i^N \sum_k^K \frac{x_{ik} \mathbf{c}_k}{\mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2} \,. \tag{18}$$

The solution for (16) is also a solution for (17), thus the system is rank deficient. It has infinite solutions, and we can not assume any optimality. To overcome this problem, one may calculate the emprical covariance matrix $\boldsymbol{\Omega}_e$ (See Appendix A). It is expected that the emprical covariance leads to an optimal solution. Our experiments on changing illumination conditions and facial expressions indicate that this assumption holds for real life scenarios. Finally, two useful equations emerge as

$$\sigma_k^2 = \frac{1}{N}\sum_i^N \left(x_{ik} - \boldsymbol{\mu}^T \mathbf{c}_k\right)^2 - \mathbf{c}_k^T \boldsymbol{\Omega}_e \mathbf{c}_k \,, \tag{19}$$

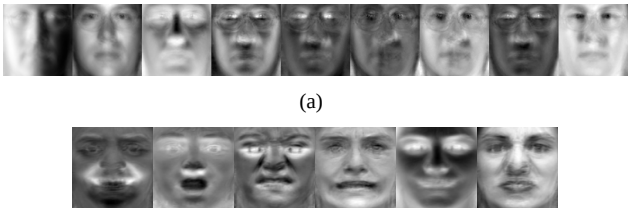
$$\left(\sum_k^K \frac{\mathbf{c}_k \mathbf{c}_k^T}{\mathbf{c}_k^T \boldsymbol{\Omega}_e \mathbf{c}_k + \sigma_k^2}\right) \boldsymbol{\mu} = \frac{1}{N} \sum_i^N \sum_k^K \frac{x_{ik} \mathbf{c}_k}{\mathbf{c}_k^T \boldsymbol{\Omega}_e \mathbf{c}_k + \sigma_k^2} \,. \tag{20}$$

Analytic solution to these nonlinear equations is not trivial. Thus, a fixed point iteration is employed to approximate the solution. Let $\zeta_k = \mathbf{c}_k^T \boldsymbol{\Omega} \mathbf{c}_k + \sigma_k^2$ and $a(t)$ indicates the value of the variable a at t^{th} iteration step, then

$$\left(\sum_k^K \frac{\mathbf{c}_k \mathbf{c}_k^T}{\zeta_k(t)}\right) \boldsymbol{\mu}(t) = \frac{1}{N} \sum_i^N \sum_k^K \frac{x_{ik} \mathbf{c}_k}{\zeta_k(t)} \,, \tag{21}$$

$$\zeta_k(t+1) = \frac{1}{N}\sum_i^N \left(x_{ik} - \boldsymbol{\mu}(t)^T \mathbf{c}_k\right)^2 \,. \tag{22}$$

With an appropriate initial guess, this procedure converges fast. Two example solutions for $\boldsymbol{\mu}$ corresponding to different variation types are illustrated in Figure 4. For all experiments, we have used $\zeta_k(1) = 1$ as the initialization and stopped the iteration when $|\zeta_k(t+1) - \zeta_k(t)| \leq 10^{-6}$. These calculations must be repeated for each pixel location as the element-wise formulation is utilized.



(a) (b)

Figure 4: Mean parameter, $\boldsymbol{\mu}$, is illustrated for two different variation types: (a) for illumination and (b) for expression.

3.3. Recovering Class Factors: Training

Having the conditional probability $p(x_{gk}|\mathbf{w}_g, \mathbf{c}_k)$ and the prior probability $p(\mathbf{w}_g)$ defined in the bootstrap, the MAP estimation may be applied to recover the factor loadings of a gallery identity g, given an observation x_{gk} by

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}_g} \quad p(\mathbf{w}_g|x_{gk}, \mathbf{c}_k) \,.$$

Using Bayes' rule we get

$$\mathbf{w}_{MAP} = \arg\max_{\mathbf{w}_g} \quad p(x_{gk}|\mathbf{w}_g, \mathbf{c}_k) p(\mathbf{w}_g) \,,$$

where the constant term $p(x_{gk})$ is omitted. The MAP estimate for $\mathbf{w}_g$ is the solution to the following set of linear equations [14].

$$\mathbf{A}\mathbf{w}_g = \mathbf{b}, \tag{23}$$

where

$$\mathbf{A} = \frac{1}{\sigma_k^2}\mathbf{c}_k\mathbf{c}_k^T + \mathbf{\Omega}^{-1}, \quad \mathbf{b} = \frac{x_{gk}}{\sigma_k^2}\mathbf{c}_k + \mathbf{\Omega}^{-1}\boldsymbol{\mu}. \tag{24}$$

In this formulation, a single observation is enough for each class while having more points will increase the reliability of the recovery. When multiple observations for an identity g exist, the coefficient matrix and the right-hand side vector are determined by summations over observations.

$$\mathbf{A} = \sum_k \frac{1}{\sigma_k^2}\mathbf{c}_k\mathbf{c}_k^T + \mathbf{\Omega}^{-1}, \quad \mathbf{b} = \sum_k \frac{x_{gk}}{\sigma_k^2}\mathbf{c}_k + \mathbf{\Omega}^{-1}\boldsymbol{\mu}. \tag{25}$$

Factors, $\mathbf{c}_k$, are assumed to be calculated by the mapping $\mathbf{M}$ of LPP. First, the identity dependent factors, $\mathbf{c}_{gk}$, are calculated by

$$\mathbf{c}_{gk} = \mathbf{M}^T\mathbf{x}_{gk}. \tag{26}$$

Then, the identity invariant factors are obtained by finding the closest (in terms of Euclidean distance) $\mathbf{c}_k$ that is calculated by (10) during the bootstrap. Instead, one may take the average of k nearest $\mathbf{c}_k$ to increase the ability of handling novel values. During our tests, we took the average of 3 nearest $\mathbf{c}_k$.

3.4. Classification of novel points: Testing

Given a novel observation $\mathbf{x}_{pk}$, the class label can be determined by assigning the class with the maximum likelihood $p(\mathbf{x}_{pk}|\mathbf{W}_g, \mathbf{c}_k)$.

Another approach which is used during our experiments is to minimize the distance between the novel point and its synthesized counterparts (distance to manifold).

$$d_g = ||\mathbf{x}_{pk} - \mathbf{x}_{gk}||, \tag{27}$$

where $\mathbf{x}_{gk} = \mathbf{W}_g\mathbf{W}_g^T\mathbf{x}_{pk}$ is calculated for each gallery identity, g.

As a third choice, posterior probabilities may be used to decide the identity of the novel point. The decision is made by selecting the maximum posterior $p(\mathbf{W}_g|\mathbf{x}_{pk}, \mathbf{c}_k)$. Bayes' rule transforms the posterior into the multiplication of the likelihood and the prior: $p(\mathbf{W}_g|\mathbf{x}_{pk}, \mathbf{c}_k) = p(\mathbf{x}_{pk}|\mathbf{W}_g, \mathbf{c}_k).p(\mathbf{W}_g)$ (the constant denominator $p(\mathbf{x}_{pk})$ is omitted). This approach can be very useful in large scale real life scenarios as it lets us to employ priors over gallery identities.

The second approach was used for all of our experiments. For this approach, the orthonormality is assumed for matrices $\mathbf{W}_g$ whereas no such constraint was considered during the recovery. Therefore, Gram-Schmidt orthonormalization process is employed after solving (23). The detailed algorithm of the CDFA is given in Table 1.

Table 1: Detailed algorithm of the CDFA.

Bootstrap: Given a bootstrap database, $\mathbf{X} = \{\mathbf{x}_{ik}\}$

- Calculate the lower dimensional coordinates, $\mathbf{c}_k$ by (10)
- For each pixel location
  - Calculate the emprical covariance matrix, $\mathbf{\Omega}_e$ by (A.5)
  - Calculate $\boldsymbol{\mu}$ and σ_k^2 using (21) and (22)

Training: Given gallery observations, $\mathbf{G} = \{\mathbf{x}_{gk}\}$, for each identity g

- Calculate the lower dimensional coordinates, $\mathbf{c}_k$ by (26)
- Recover $\mathbf{w}_g$ for each pixel location by (23)
- Construct the matrix $\mathbf{W}_g$ so that it has vectors $\mathbf{w}_g$ as its rows
- Apply Gram-Schmidt orthonormalization to the columns of $\mathbf{W}_g$

Testing: Given a probe observation $\mathbf{x}_{pk}$,

- Calculate d_g for each gallery identity g using (27)
- Select the identity with the minimum distance

4. Interpretation of Governing Distributions

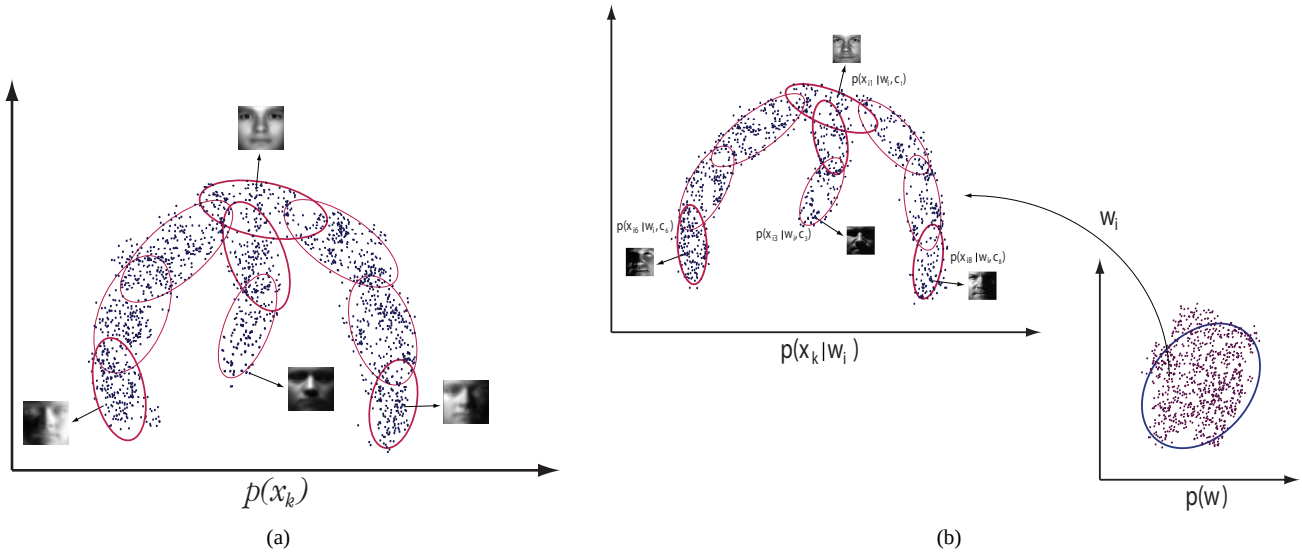
Beside the geometrical interpretation of the generative model described in Section 1.2, another probabilistic interpretation is given here, regarding the formulation of the CDFA framework. The marginal distribution $p(x_k)$ specifies a mixture of Gaussians in which Gaussians are determined by the variation label k. Each Gaussian is characterized by parameters $\boldsymbol{\mu}^T\mathbf{c}_k$ and $\mathbf{c}_k^T\mathbf{\Omega}\mathbf{c}_k + \sigma_k^2$. Hence, the variation defines the shape of each Gaussian.

Initially, the geometry of the manifold consisting of this mixture does not depend on the identities, but only on the mean identity. Thus, the manifold can be considered as a template that will be customized after selection of an identity. When an identity is drawn from the prior distribution $p(\mathbf{w})$, it redefines the mixture by the conditional distributions $p(x_k|\mathbf{w}, \mathbf{c}_k)$. This procedure also eliminates a considerable amount of uncertainty in each Gaussian as the variance decreases to σ^2 from $\mathbf{c}_k^T\mathbf{\Omega}\mathbf{c}_k + \sigma_k^2$. Whole process is illustrated in Figure 5.

CDFA is defined as a two-layer decision process. At the first layer, class identities are drawn from a prior distribution. The second layer defines a mixture of Gaussians depending on a template manifold characterized by $p(x_k)$, and the conditional distributions $p(x_k|\mathbf{w}, \mathbf{c}_k)$. The assignment of observations to each Gaussian is achieved by the manifold embedding. In this view, the manifold embedding can be considered as a clustering scheme.

5. Experimental Evaluations

Several experiments were conducted to explore two important aspects of the CDFA framework: (1) the recognition performance against extreme variations and (2) scalability in relatively large databases. For the first evaluation, we selected



(a) (b)

Figure 5: Illustration of the governing distributions: (a) A template manifold is defined by the marginal distribution, $p(x_k)$. (b) This template is customized by the identity drawn form the prior distribution, $p(\mathbf{w})$.

databases with extreme variations. Nevertheless, the sizes of such databases are usually small, including at most 30-40 identities. To analyze the real life performance of the method, a second group of experiments was performed on another set of databases with moderate variations but large number of identities.

The main characteristic of the method is its ability to be used for different types of variation. This claim was verified by different experiments under different types of variation. Two types of variation were used during tests: (1) changing illumination and (2) changing facial expressions.

5.1. Tuning the Bootstrap Parameters

Each test begins with the manifold embedding on the selected bootstrap database to decide the geometrical features of the manifold. One parameter that should be determined is the dimension of the underlying manifold. The manifold learning technique LPP relies on the solution of a generalized eigenvalue problem; therefore, the spectrum of eigenvalues may help with determining the dimension. However, using an evaluation dataset is a better choice since the characteristics of the variation may prevent a meaninful spectrum analysis.

As indicated in Section 3.1, the intrinsic dimensionality is bounded by the number of different variation labels present in the bootstrap database. For instance, when using Multi-PIE [28] as the bootstrap database, the dimensionality is bounded by 20 since there are 20 different illumination conditions. However, this does not mean that the recognition is performed in a 20 dimensional subspace. This value represents the number of basis vectors to span the variation subspace of each identity. It is only related to the range of the generative model, *i.e.* how the method deals with novel variations. The recognition is performed by the point-to-manifold measure which is calculated in the original observation space $\Re^n$, where n is the number of pixels of the input images.

Certain properties of the manifolds like dimensionality are totally determined by the bootstrap database. This is a clear and an understandable behavior since the bootstrap database reflects the way that the operative variation is modeled. The best practice is to use a bootstrap database that is the most compatible with the testing requirements.

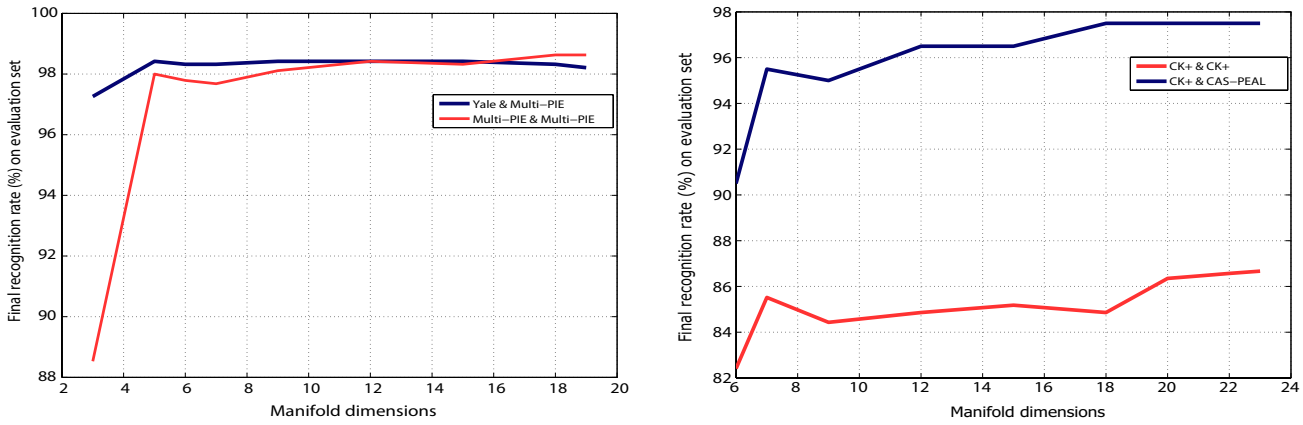
The effect of the manifold dimension is given in Figure 6. For two types of variation (illumination and facial expression), evaluation datasets were collected. Scenarios with different bootstrap and evaluation sets are demonstrated to grasp the characteristics completely. All tests were performed with evaluation sets containing 50 identities. A single image was selected as the gallery and all remaining images were used as probes. Those identities collected for the evaluation sets were not used during the further experiments to reflect a real life behavior.

Experiments indicate that the method behaves similarly in terms of dimensionality even if the bootstrap database is changed. The results are comparable when the dimension is fixed among different evaluation sets. Moreover, slight changes in dimension do not affect the recognition performance, considerably.

5.2. Classification Performance against Illumination

Tests with changing illumination conditions were performed with Yale B Database [16]. This database includes 10 identities with 45 different illumination conditions. The database can be split into 4 subsets according to the illumination direction, which also highlights the difficulty of the recognition.

The Extended Yale Face Database [29] was used as the bootstrap database. This database is an extension of the original Yale B with 28 identities which are not present in the original database. At the bootstrap phase, a subset of 41 illumina-



(a) (b)

Figure 6: Recognition rates on evaluation sets with different manifold dimensions under (a) illumination and (b) facial expression changes. Yale & Multi-PIE means that the bootstrap set is from Yale and the evaluation set is from Multi-PIE.

tion types out of 45 was used due to several corrupted images. Hence, the gallery and probe images had novel variations which were not present in the bootstrap database.

The size of images used in the experiments was 100×90. As a preprocessing step, all images were normalized so that they have zero mean and unit variance. The dimension of the manifold was fixed as 9. We performed 19 tests, and the average was taken as the final performance. For each test, a single image from subset 1 or subset 2 was selected as the gallery image, and all remaining images were used as probes. In other words, 440 recognition attempts were performed for each test, resulting in 8360 recognition attempts in total. The recognition rates with this configuration are given in Table 2.

Table 2: Face recognition rates for Yale B Database. Performances of the other methods were taken from [14].

Methods	Subset 1-2	Subset 3	Subset 4
Correlation	100	76.7	26.4
Eigenfaces	100	74.2	24.3
Linear Subspaces	100	100	85
Cones-attached	100	100	91.4
Cones-cast	100	100	100
9PL	100	100	97.2
Spherical Harmonics	100	99.7	96.9
CDFA	100	99.2	95

Recognition rates are very promising considering the moderate requirements for the bootstrap and the training. CDFA is trained by a single image for each identity unlike methods Cones-attached, Cones-cast, and 9PL which need number of images between 5 and 9. Compared to the spherical harmonics, CDFA is a more generic approach since it is not related to the physical aspects of the variation. The behavior of the CDFA with increasing number of gallery images is demonstrated in Table 3. Random images from subsets 1 and 2 are selected as gallery images for each test. The increase in the recognition performance makes the proposed method more comparable to other methods.

Table 3: Recognition error rates for Yale B Database with multiple gallery images.

# images	Subset 1-2	Subset 3	Subset4
1	0.0	0.8	5.0
2	0.0	0.2	1.4
3	0.0	0.1	0.6
4	0.0	0.0	0.3
5	0.0	0.0	0.1
6	0.0	0.0	0.0

5.3. Classification Performance against Facial Expressions

As a second set of experiments, the performance of the CDFA with facial expressions was analyzed. For this purpose, three databases were selected: Cohn-Kanade AU-Coded facial expression database (CK+) [30], Japanese female facial expression database (JAFFE) [31], and CMU AMP face expression database [32].

CK+ is a collection of video sequences starting with a neutral pose and ending with a peak expression. This database is used as a common bootstrap gallery. Inside each sequence, 4 images were sampled. Including one additional neutral image, at most 25 different images were collected for each identity (24 images corresponding to 6 expression and 1 neutral image). The manifold dimension was determined to be 20.

Two groups of tests were performed using databases JAFFE and CMU AMP. JAFFE includes 213 images of 10 Japanese women with number of facial expressions varying between 20 and 23. These expressions can be very different from the expressions which exist in the bootstrap database. Therefore, we also showed the ability of the method with handling novel variations. CMU AMP have 13 identities with 75 different expressions. Expressions present in this database are extremely severe as they also cause slight pose changes along with changes in face geometries.

CDFA is compared against two state-of-art techniques CS [25] and SRC [26]. To make fair comparisons, we followed the same scenarios with the compared methods, and the gallery selection procedure and the structure of random tests were kept same. Image size was set to be 32×32 since the compared methods had selected to use such a small image size. For each identity, several gallery images were selected randomly, and the remaining images were used as probes. Images were used after *zero mean-unit variance* normalization. Results of two classical subspace techniques, PCA and LPP, are also analyzed to understand the marginal improvements. The transformation matrices for PCA and LPP are obtained using the CK+ bootstrap database. LPP is trained in kNN mode with distances being calculated by the heat kernel. Table 4 and Table 5 show test results for JAFFE and CMU AMP. Results for CS and SRC were taken from [25]. To give an impression of the significance of the presented results, the second columns list the number of actual recognition attempts for each experiment. These values are simply calculated as (the number of test images $\times$ the number of random trials).

Table 4: Average face recognition rates on JAFFE database. 40 trials with randomly chosen gallery images were performed for each row.

# Gallery Images	Recognition Attempts	CDFA	CS	SRC	PCA	LPP
2	7720	93.04	89.94	90.1	85.84	83.84
3	7320	94.50	93.22	92.1	89.1	89.32
4	6920	96.17	95.12	95.13	91.62	91.33
5	6520	96.33	96.12	96.01	93.54	93.87

Table 5: Average face recognition rates on CMU AMP database. 10 trials with randomly chosen gallery images were performed for each row.

# Gallery Images	Recognition Attempts	CDFA	CS	SRC	PCA	LPP
4	9230	99.92	98.95	98.9	99.6	99.91
5	9100	100	99.91	99.8	99.66	99.71
6	8970	99.99	99.97	99.75	99.68	99.84
7	8840	100	100	99.74	99.71	99.75
8	8710	100	100	99.87	99.89	99.87
9	8580	100	100	100	99.94	99.97
10	8450	100	100	99.49	99.85	99.95

CDFA steadily outperforms others for both databases. However, the main intention here is to highlight that the same framework can be utilized for different types of variation without any modification in the base configuration. Indeed, these databases happend to be trivial although they include severe variations. Even a classical approach like PCA can achieve high recognition rates on them.

5.4. Scalability

Further experiments were performed to examine the scalability of the proposed method. Two relatively large databases were selected for the testing: CMU Multi-PIE Database [28] and CAS-PEAL Database [33]. Both databases consist of images of more than 200 people. CAS-PEAL was used for the evaluation against facial expressions and Multi-PIE for the illumination. Multi-PIE includes 20 different illumination conditions, and CAS-PEAL serves 5 facial expressions for each identity.

The behavior of a classical subspace method against the increasing number of gallery identities is demonstrated in Figure 7 (a). LDA against illumination was used for the demonstration. All tests were performed on Multi-PIE with 2 random images of each identity being selected as the gallery and the remaining 18 images as probes.

LDA can perform steadily in terms of recognition rate with its usual configuration. The subspace is re-constructed with each new identity, and the subspace dimension becomes $(ni-1)$ where *ni* is the number identities. However, as new identities are introduced, LDA needs to be re-trained to get a promising recognition rate. This behavior is illustrated in Figure 7 (a) with "No bootstrap" label. One may eliminate such a training requirement by using a bootstrap database. In this new setting, the subspace is constructed only once by using the bootstrap database, yet the recognition rate decreases as the number of gallery identities is increased. Moreover, different bootstrap databases may result in significantly different recognition rates.

CDFA framework can improve the scalability as shown in Figure 7 (b). The method was tested with several scenarios both for illumination and facial expressions. When Yale[1] or CK+ was used as the bootstrap database, all settings like manifold dimensionality were kept same as the ones in Section 5.2 and Section 5.3. We observe that the final recognition rates are not affected significantly as bootstrap databases are switched. The largest performance difference caused by changing the bootstrap database was between 1% – 2%.

The results in Figure 7 (b) also suggest that it is possible to fix the template manifold for a certain type of variation since same bootstrap database can be used in different tests: CK+ was employed successfully in tests with CAS-PEAL, JAFFE, and CMU AMP while the Yale database is compatible both for Multi-PIE and Yale itself.

Figure 8 gives recognition rates of several methods with increasing number of identities in the gallery. CDFA was compared with PCA [5], LDA [7], and Tied Factor Analysis (TFA) [1] since they share very common aspects with CDFA, in terms of subspace analysis. The method in [1] was initially developed to handle the pose variation; however, the authors proposed the algorithm as a generic factor analysis framework just like CDFA. Multi-PIE and CAS-PEAL were used for testing against illumination and facial expression, respectively. To provide a fair comparison, a common bootstrap database with 50 identities was collected to learn the subspace parameters for all methods. For tests with Multi-PIE, the bootstrap includes 1000 images while this value is 250 for the tests with CAS-

[1]There are two different Yale databases used during tests: Yale B Database [16] and Extended Yale Face Database [29]. However, when a common name 'Yale' is mentioned, it means that an augmented database which is established by concatenating two is used.

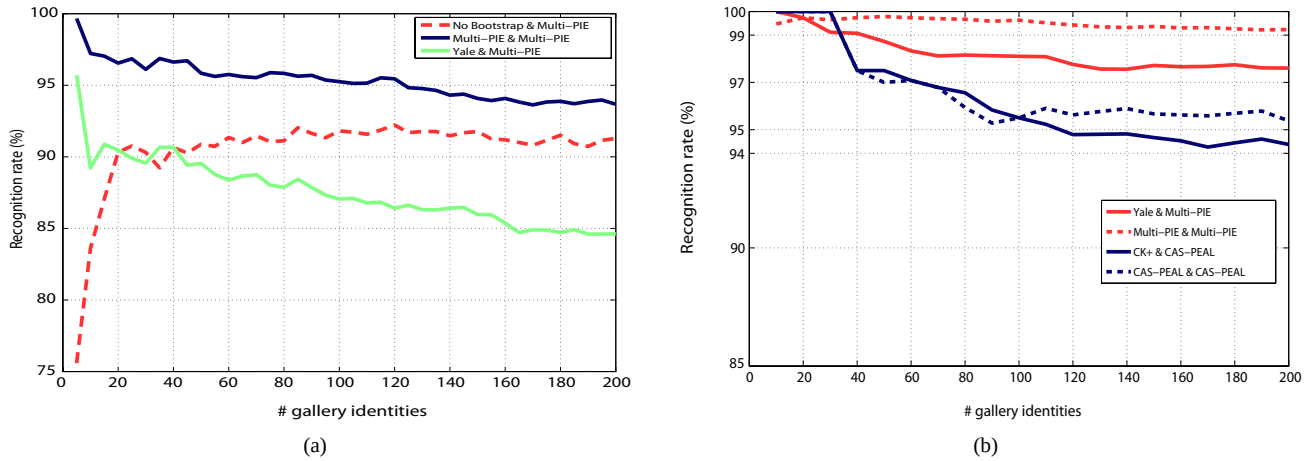


Figure 7: (a) Behavior of LDA against the illumination with increasing number of identities. Three scenarios were tried: with no bootstrap, with a bootstrap drawn from Multi-PIE, and with a bootstrap drawn from Yale. (b) Behavior of CDFA against illumination and facial expressions. Yale & Multi-PIE means that the bootstrap set is from Yale and the evaluation set is from Multi-PIE.

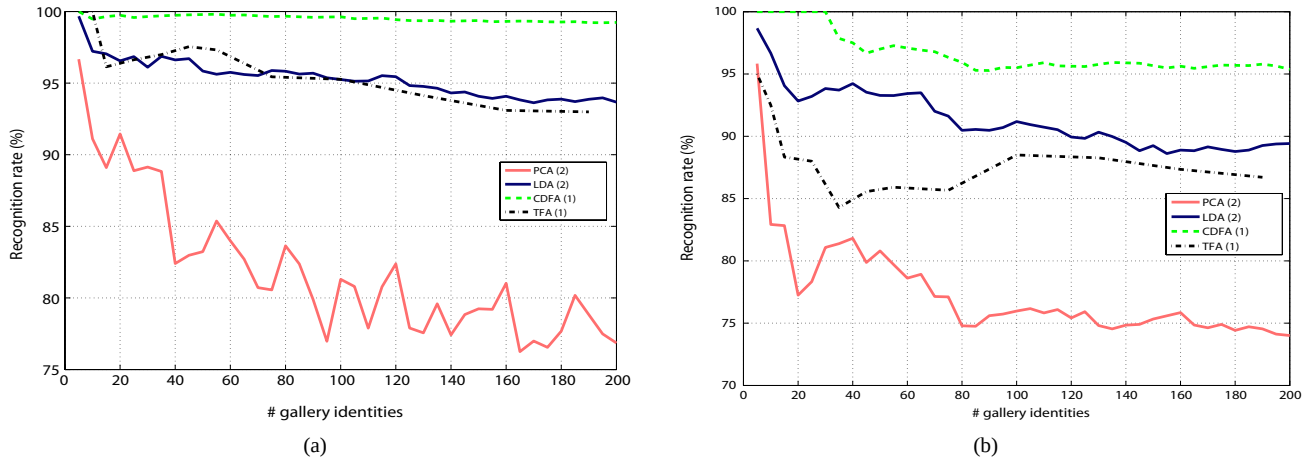


Figure 8: Recognition performance of different methods on (a) Multi-PIE illumination database and (b) CAS-PEAL expression database. Values in parentheses shows the number of gallery images.

PEAL. Subspace dimensions were optimized individually for each method.

For both tests, bootstrap and the training/testing images were drawn from the same databases. Therefore, the manifold dimension was 4 for tests with CAS-PEAL since there are 5 different expressions in database, and the upper bound is limited by the number of expressions. In both sets of experiments, the image size was 100×90. Images were normalized with *zero mean-unit variance* normalization.

For CDFA and TFA, a single gallery image was selected and all remaining images were used as probes. Then, for a test having N gallery identities, $19 \times N$ recognition attempts were performed for Multi-PIE and $4 \times N$ recognition attempts were performed for CAS-PEAL. These attempts were repeated for each random gallery image selection, and the averages were noted.

The recognition rates tend to decrease with other methods whereas CDFA performs steadily as the number of identities increases. This fact is depicted in Figure 8.

5.5. *Discussions on Experimental Results*

We performed several experiments to analyze the performance of the proposed method against different variation types and with relatively large databases. In both cases, the results are very promising.

Several advantages of the method can be summarized as follows: (1) different types of variation that lie on smooth manifolds can be handled by the method, (2) the scalability of the classical factor analysis is improved by a class dependent scheme, (3) the decision process is fully probabilistic, and posterior probabilities can be utilized for large scale and domain specific real life applications by incorporating priors on the identities, (4) bootstrap has less time complexity compared to 3D rendering approaches, and finally (5) a single observation for each identity is sufficient to perform reliable recognition while a way to favor more images is also introduced.

The main drawback of the proposed framework is its space complexity. For each gallery identity, the whole subspace is

defined. Compared to classical methods, which store a low dimensional vector for each identity, storing a high dimensional matrix requires more space. Moreover, the testing has relatively higher time complexity since at least two matrix-vector products are required ($\mathbf{M}^T\mathbf{x}_p$ and $\mathbf{W}_g\mathbf{c}_k$) to make decision while the classical factor analysis only performs a norm calculation. When speaking in terms of *wall clock time*, the training and the testing per image take approximately 0.3 seconds and 40 milliseconds, respectively on a regular PC (Intel Core 2 Duo 2.2 GHz and 3 GB RAM). These values are valid on a development environment. The real life performance is better with approximately 20 milliseconds for testing on the same PC.

6. Conclusions

A linear generative model was developed to improve the general factor analysis framework. The main novelty is the complete probabilistic structure that individualizes manifold charts resulting in a class dependent design. Modeling nonlinear variations like illumination and facial expression is achieved by incorporating a manifold embedding technique to obtain a linear representation of the effective variation. This is not a surprising result considering the fact that such variations can be modeled linearly on some geometries. For instance, illumination can be modeled as a linear combination of spherical harmonics on a unit sphere.

We propose a probabilistic framework that can be employed in general classification problems when a problematic variation is exhibited on class samples. The only assumption which is used implicitly is that the variation can be modeled on a smooth manifold. If the nonlinear embedding fails, the resulting lower dimensional coordinates may disturb the final performance.

The initial results are very promising indicating the potential of the proposed framework as a replacement to regular subspace analysis methods. The proposed approach defines a novel connection between the manifold embedding and the probabilistic models.

Combining different variations is left as a future work. The first step towards this goal may be using factor tensors instead of factor matrices.

Appendix A. Approximating the covariance matrix of W distribution

To calculate the covariance matrix of the distribution defined on the factor loadings, $\mathbf{W}$, a way similar to the one proposed in [15] is followed. The factor loadings, $\mathbf{W}$, are considered as a basis set of the variation subspace. Therefore, factors, $\mathbf{c}_k$, are assumed to be coordinates *i.e.* linear combination coefficients.

Let's assume that we have K images of an identity i in the bootstrap database. Then the total reconstruction error for the identity i is

$$\begin{aligned}\mathcal{E} &= \sum_{k=1}^{K} \|\mathbf{x}_{ik} - \mathbf{W}_i\mathbf{c}_k\| \\ &= \sum_{k=1}^{K} \|\mathbf{x}_{ik} - \mathbf{w}_{i1}c_{k1} - \mathbf{w}_{i2}c_{k2} - \ldots - \mathbf{w}_{in}c_{kn}\|, \end{aligned} \tag{A.1}$$

where $\mathbf{w}_{ij}$ indicates j^{th} column of the matrix $\mathbf{W}_i$, and c_{kj} is j^{th} element of vector $\mathbf{c}_k$.

Normalization constraints $\|\mathbf{w}_{ij}\| = 1$ are not introduced, since the scaling factors, c_{kj}, are already known and fixed. Thus, relaxations on the norms of the vectors are required to assure a global minimum. Similarly, orthogonality is not considered.

The optimization problem can be restated as the following trace minimization to simplify calculations.

$$\mathcal{J} = Tr[\mathbf{X}^T - \mathbf{c}_1\mathbf{w}_1^T - \mathbf{c}_2\mathbf{w}_2^T - \ldots - \mathbf{c}_n\mathbf{w}_n^T]. \tag{A.2}$$

Here the notation is changed slightly. The matrix $\mathbf{X}$ has the vector $\mathbf{x}_{ik}$ as it k^{th} column. The vector $\mathbf{c}_j$ is the collection of constants c_{kj}. The index i of vectors $\mathbf{w}_{ij}$ is dropped for the clarity. By rewriting the equation we get

$$\begin{aligned}\mathcal{J} &= Tr(\mathbf{X}\mathbf{X}^T) - 2\mathbf{c}_1^T\mathbf{X}^T\mathbf{w}_1 - \ldots - 2\mathbf{c}_n^T\mathbf{X}^T\mathbf{w}_n \\ &+ \mathbf{c}_1^T\mathbf{c}_1\mathbf{w}_1^T\mathbf{w}_1 + 2\mathbf{c}_1^T\mathbf{c}_2\mathbf{w}_1^T\mathbf{w}_2 + \ldots + 2\mathbf{c}_1^T\mathbf{c}_n\mathbf{w}_1^T\mathbf{w}_n \\ &+ 2\mathbf{c}_2^T\mathbf{c}_1\mathbf{w}_2^T\mathbf{w}_1 + \mathbf{c}_2^T\mathbf{c}_2\mathbf{w}_2^T\mathbf{w}_2 + \ldots + 2\mathbf{c}_2^T\mathbf{c}_n\mathbf{w}_2^T\mathbf{w}_n \\ &+ \ldots \\ &+ 2\mathbf{c}_n^T\mathbf{c}_1\mathbf{w}_n^T\mathbf{w}_1 + \ldots + \mathbf{c}_n^T\mathbf{c}_n\mathbf{w}_n^T\mathbf{w}_n. \end{aligned} \tag{A.3}$$

Derivatives with respect to each basis vector yield the following set of linear equations.

$$\begin{aligned} -\mathbf{X}\mathbf{c}_1 + \mathbf{c}_1^T\mathbf{c}_1\mathbf{w}_1 + \ldots + \mathbf{c}_1^T\mathbf{c}_n\mathbf{w}_n &= 0, \\ &\vdots \\ -\mathbf{X}\mathbf{c}_n + \mathbf{c}_n^T\mathbf{c}_1\mathbf{w}_1 + \ldots + \mathbf{c}_n^T\mathbf{c}_n\mathbf{w}_n &= 0. \end{aligned} \tag{A.4}$$

In the matrix form, we have

$$\begin{bmatrix} \mathbf{c}_1^T\mathbf{c}_1 & \mathbf{c}_1^T\mathbf{c}_2 & \ldots & \mathbf{c}_1^T\mathbf{c}_n \\ \mathbf{c}_2^T\mathbf{c}_1 & \mathbf{c}_2^T\mathbf{c}_2 & \ldots & \mathbf{c}_2^T\mathbf{c}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{c}_n^T\mathbf{c}_1 & \mathbf{c}_n^T\mathbf{c}_2 & \ldots & \mathbf{c}_n^T\mathbf{c}_n \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}\mathbf{c}_1 \\ \mathbf{X}\mathbf{c}_2 \\ \vdots \\ \mathbf{X}\mathbf{c}_n \end{bmatrix}.$$

The size of the system is relatively small depending on the dimension of the subspace. The rank of the coefficient matrix is usually n provided that a linearly independent set of vectors, $\mathbf{c}_i$, exists. Hence, there is a unique solution for the problem. As the complete basis set, $\mathbf{W}_i$, of each identity i is calculated, the covariance matrix for the distribution $p(\mathbf{w})$ can be estimated by the empirical formula

$$\mathbf{\Omega}_e = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_i - \overline{\mathbf{w}})(\mathbf{w}_i - \overline{\mathbf{w}})^T, \tag{A.5}$$

where $\overline{\mathbf{w}}$ is the mean value. One should be careful with this notation. Here, the form defined in (3) is used. Therefore, the vector $\mathbf{w}_i$ is a row (not a column) of the matrix $\mathbf{W}_i$. After calculating the matrices $\mathbf{W}_i$ for all identities of the bootstrap gallery, the covariance matrices corresponding to different rows are calculated independently.

Acknowledgements

This work was partially funded by The Scientific and Technological Research Council of Turkey with the grant number 109E268. The first author was partially supported with the grant ITU-BAP 34385.

References

[1] S. Prince, J. Warrell, J. Elder, F. Felisberti, Tied factor analysis for face recognition across large pose differences, Pattern Analysis and Machine Intelligence 30 (6) (2008) 970–984.

[2] T. Kanade, A. Yamada, Multi-subregion based probabilistic approach toward pose-invariant face recognition, in: International Symposium on Computational Intelligence in Robotics and Automation, 2003, pp. 954–959.

[3] M. Saquib Sarfraz, O. Hellwich, Probabilistic learning for fully automatic face recognition across pose, Image and Vision Computing 28 (2010) 744–753.

[4] W. Zhao, R. Chellappa, A. Rosenfeld, P. J. Phillips, Face recognition: A literature survey, ACM Computing Surveys (2003) 399–458.

[5] M. Turk, A. Pentland, Face recognition using eigenfaces, in: Procs. of Computer Vision and Pattern Recognition, 1991, pp. 586–591.

[6] R. Gnanadesikan, J. R. Kettenring, A pragmatic review of multivariate methods in applications, in: H. A. David, H. T. David (Eds.), In Statistics: An Appraisal, Iowa State Univ. Press, 1984, pp. 309–337.

[7] P. N. Belhumeur, J. Hespanha, D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, in: Procs. of European Conference on Computer Vision, 1996, pp. 45–58.

[8] G. Lebanon, Riemannian geometry and statistical machine learning, PhD Thesis, School of Computer Science - Carnegie Mellon University (2005).

[9] X. He, P. Niyogi, Locality preserving projections, in: Procs. of Advances in Neural Information Processing Systems, 2003, 2003.

[10] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[11] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction., Science 290 (2000) 2319–2323.

[12] M. E. Tipping, C. M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society, Series B 61 (1999) 611–622.

[13] S. Roweis, Em algorithms for PCA and SPCA, in: M. I. Jordan, M. J. Kearns, S. A. Solla (Eds.), Advances in Neural Information Processing Systems, Vol. 10, The MIT Press, 1998.

[14] L. Zhang, D. Samaras, Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics, Pattern Analysis and Machine Intelligence 28 (3) (2006) 351–363.

[15] B. Tunç, M. Gökmen, Manifold learning for face recognition under changing illumination, Telecommunication Systems 47 (3-4) (2011) 185–195.

[16] A. Georghiades, P. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, Pattern Analysis and Machine Intelligence 23 (6) (2001) 643–660.

[17] S. K. Zhou, R. Chellappa, Illuminating light field: image-based face recognition across illuminations and poses, in: Procs. of Automatic Face and Gesture Recognition, 2004, pp. 229–234.

[18] K.-C. Lee, J. Ho, D. Kriegman, Nine points of light: Acquiring subspaces for face recognition under variable lighting, in: Procs. of Computer Vision and Pattern Recognition, Vol. 1, 2001, p. 519.

[19] R. Gross, I. Matthews, S. Baker, Eigen light-fields and face recognition across pose, in: Procs. of Automatic Face and Gesture Recognition, 2002, pp. 1–7.

[20] A. Shashua, T. Riklin-Raviv, The quotient image: class-based rerendering and recognition with varying illuminations, Pattern Analysis and Machine Intelligence 23 (2) (2001) 129–139.

[21] S. K. Zhou, R. Chellappa, Rank constrained recognition under unknown illuminations, in: Procs. of International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 11–18.

[22] T.-K. Kim, J. Kittler, Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image, Pattern Analysis and Machine Intelligence 27 (3) (2005) 318–327.

[23] S. Ioffe, Probabilistic linear discriminant analysis, in: Procs. of European Conference on Computer Vision, 2006, pp. 531–542.

[24] S. Prince, P. Li, Y. Fu, U. Mohammed, J. Elder, Probabilistic models for inference about identity, Pattern Analysis and Machine Intelligence (PrePrints).

[25] P. Nagesh, B. Li, A compressive sensing approach for expression-invariant face recognition, in: Procs. of Computer Vision and Pattern Recognition, 2009, pp. 1518–1525.

[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, Y. Ma, Robust face recognition via sparse representation, Pattern Analysis and Machine Intelligence 31 (2009) 210–227.

[27] X. He, S. Yan, Y. Hu, P. Niyogi, H.-J. Zhang, Face recognition using laplacianfaces, Pattern Analysis and Machine Intelligence 27 (3) (2005) 328–340.

[28] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, S. Baker, Multi-PIE, in: Procs. of International Conference on Automatic Face and Gesture Recognition, 2008.

[29] K. Lee, J. Ho, D. Kriegman, Acquiring linear subspaces for face recognition under variable lighting, Pattern Analysis and Machine Intelligence 27 (5) (2005) 684–698.

[30] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, in: Procs. of Computer Vision and Pattern Recognition Workshops, 2010.

[31] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: Procs. of International Conference on Face & Gesture Recognition, 1998.

[32] X. Liu, T. Chen, B. V. K. V. Kumar, Face authentication for multiple subjects using eigenflow, Pattern Recognition, Special issue on Biometric 36 (2003) 313–328.

[33] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, D. Zhao, The CAS-PEAL large-scale chinese face database and baseline evaluations, IEEE Transactions on Systems, Man, and Cybernetics 38 (1) (2008) 149–161.