Probabilistic Identity Characterization for Face Recognition*

Shaohua Kevin Zhou and Rama Chellappa
Center for Automation Research (CfAR) and
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20742
{shaohua, rama}@cfar.umd.edu

Abstract

We present a general framework for characterizing the object identity in a single image or a group of images with each image containing a transformed version of the object, with applications to face recognition. In terms of the transformation, the group is made of either many still images or frames of a video sequence. The object identity is either discrete- or continuous-valued. This probabilistic framework integrates all the evidence of the set and handles the localization problem, illumination and pose variations through subspace identity encoding. Issues and challenges arising in this framework are addressed and efficient computational schemes are presented. Good face recognition results using the PIE database are reported.

1 Introduction

Visual face recognition is an important task. Even though a lot of research has been carried out, current state-of-the-art recognizers still yield unsatisfactory results especially when confronted with pose and illumination variations. In addition, the recognizers are further complicated by the registration requirement as the images that the recognizers process contain transformed appearances of the object. Below, we simply use the term 'transformation' to model the variations involved, be it registration, pose and/or illumination variations.

While most recognizers process a single image, there is a growing interest in using a group of images [11, 20, 14, 12, 10, 17, 6]. In terms of the transformations embedded in the group or the temporal continuity between the transformations, the group can be either independent or not. Examples of the independent group (I-group) are face databases that store multiple appearances for one object. Examples of the dependent group are video sequence with temporal continuity. If the temporal information is stripped, video sequences reduce to I-groups. In this paper, whenever we mention video sequences, we mean dependent groups of images.

Approaches that use I-groups can be rough divided into two categories. The first category is based on manifold matching. In [11], hypothetical identity surfaces are constructed by computing the linear coefficients of view space. Illumination variations are not accounted for. Discriminant features are then extracted to overcome other variations. In [6], manifolds are formed for every I-group. Recognition is performed by computing the shortest distance between two manifolds. The manifold takes a certain parameterized form and the parameters are directly learned from the visual appearances. Robustness to pose and illumination variations are not reported. The second category is based on statistical learning. In [14], a multi-variate Gaussian density is fitted for every I-group. Recognition is achieved by computing the Kullback-Leibler distance [4] between two Gaussian densities. However, the Gaussian assumption is easily violated if pose and illumination variations exist. In [17], principal subspaces are learned for each I-group and principal angel between two principal subspace are used for recognition. The computation of principal angle is also carried on the feature space embedded by kernel functions. One common disadvantage of the above approaches is that they also assume that the face regions have already been cropped beforehand, using either a detector or a tracker.

Approaches using video sequences utilize temporal information for recognition as well. In [20], simultaneous tracking and recognition is implemented in a probabilistic framework. The joint posterior probability of the tracking parameter and the identity variable is approximated using the sequential important sampling (SIS) algorithm and the marginal posterior probability of the identity variable is used for recognition. However, only affine localization parameter is used for tracking and handling pose and illumination variations are not reported. In addition, exemplars are learned from the gallery videos to cover pose and illumination variations. In [12], hidden Markov models are used to learn the dynamics before successive appearances. In [10], pose variations are handled by learning the view-discretized appearance manifolds from the training ensemble. Transi-

*Partially supported by the DARPA/ONR Grant N00014-03-1-0520.

tion probabilities from one view to another view are used to regularize the search space. However, in [12, 10], the cropped images are used for testing.

In this paper, we propose a generic framework which possesses the following features: (i) It processes either a single image or a group of images (including I-group and video sequence). (ii) It handles the localization problem, illumination and pose variations. (iii) The identity description could be either discrete or continuous. The continuous identity encoding typically arises from a subspace modeling. (iv) It is probabilistic and integrates all the available evidence.

In Section 2 we introduce the generic framework which provides a probabilistic characterization of the object identity. In Section 3 we address issues and challenges arising in this framework. In Section 4 we focus on how to achieve an identity encoding which is invariant to localization, illumination and pose variations. In Section 5, we present some efficient computational methods. In Section 6, we present experimental results. In Section 7, we conclude our paper and briefly summarize potential future works.

2 Probabilistic Identity Characterization

Suppose α is the identity signature, which represents the identity in an abstract manner. It can be either be discrete- or continuous- valued. If we have a C-class problem, α is discrete taking value in $\{1, 2, ..., \mathsf{C}\}$. If we associate the identity with image intensity or feature vectors derived from say subspace projections, α is continuous-valued. Given a group of images $\mathsf{y}_{1:N} \doteq \{\mathsf{y}_1, ..., \mathsf{y}_N\}$ containing the appearances of the same but unknown identity, *probabilistic identity characterization is equivalent to finding the posterior probability* $p(\alpha|y_{1:N})$. Probabilistic modeling is commonly used in computer vision and applications. See [9].

As the image only contains a transformed version of the object, we also need to associate it a transformation parameter θ, which lies in a transformation space Θ. The transformation space Θ is usually application dependent. Affine transformation is often used to compensate for the localization problem. To handle illumination variation, the lighting direction is used. If pose variation is involved, 3D transformation is needed or a discrete set is used if we quantize the continuous view space.

We assume that the prior probability of α is $\pi(\alpha)$, which is assumed to be, in practice, a *non-informative* prior. A non-informative prior is uniform in the discrete case and treated as a constant, say 1, in the continuous case.

The key to our probabilistic identity characterization is as follows:

$$\begin{aligned} \mathsf{p}(\alpha|\mathsf{y}_{1:N}) &\propto \pi(\alpha)\mathsf{p}(\mathsf{y}_{1:N}|\alpha) \\ &= \pi(\alpha)\int_{\theta_{1:N}} \mathsf{p}(\mathsf{y}_{1:N}|\theta_{1:N},\alpha)\mathsf{p}(\theta_{1:N})\mathsf{d}\theta_{1:N} \\ &= \pi(\alpha)\int_{\theta_{1:N}} \prod_{t=1}^{N} \mathsf{p}(\mathsf{y}_t|\theta_t,\alpha)\mathsf{p}(\theta_t|\theta_{1:t-1})\mathsf{d}\theta_{1:N}, \quad (1) \end{aligned}$$

where the following rules, namely (a) *observational conditional independence* and (b) *chain rule*, are applied:

$$(a)\ \mathsf{p}(\mathsf{y}_{1:N}|\theta_{1:N},\alpha) = \prod_{t=1}^{N} \mathsf{p}(\mathsf{y}_t|\theta_t,\alpha); \quad (2)$$

$$(b)\ \mathsf{p}(\theta_{1:N}) = \prod_{t=1}^{N} \mathsf{p}(\theta_t|\theta_{1:t-1});\ \mathsf{p}(\theta_1|\theta_0) \doteq \mathsf{p}(\theta_1). \quad (3)$$

Equation (1) involves two key quantities: the *observation likelihood* $\mathsf{p}(\mathsf{y}_t|\theta_t,\alpha)$ and the *state transition probability* $\mathsf{p}(\theta_t|\theta_{1:t-1})$. The former is essential to a recognition task, the ideal case being that it possesses a discriminative power in the sense that it always favors the correct identity and disfavors the others; the latter is also very helpful especially when processing video sequences, which constrains the search space.

We now study two special cases of $\mathsf{p}(\theta_t|\theta_{1:t-1})$.

2.1 Independent group (I-group)

In this case, the transformations $\{\theta_t; t = 1, \ldots, N\}$ are independent of each other, i.e.

$$\mathsf{p}(\theta_t|\theta_{1:t-1}) = \mathsf{p}(\theta_t). \quad (4)$$

Eq. (1) becomes

$$\mathsf{p}(\alpha|\mathsf{y}_{1:N}) \propto \pi(\alpha)\prod_{t=1}^{N}\int_{\theta_t} \mathsf{p}(\mathsf{y}_t|\theta_t,\alpha)\mathsf{p}(\theta_t)\mathsf{d}\theta_t. \quad (5)$$

In this context, the probability $\mathsf{p}(\theta_t)$ can be regarded as a prior for θ_t, which is often assumed to be Gaussian with mean $\hat{\theta}$ or non-informative.

The most widely studied case in the literature is $N = 1$, i.e. there is only a single image in the group. Due to its importance, sometime we will distinguish it from the I-group (with $N > 1$) depending on the context. We will present in Section 3 the shortcomings of many contemporary approaches.

It all boils down to how to compute the integral in (5) in real application. In the sequel, we will show how to approximate it efficiently.

2.2 Video sequence

In the case of video sequence, temporal continuity between successive video frames implies that the transformations

$\{\theta_t; t = 1, \ldots, N\}$ follow a Markov chain. Without loss of generality, we assume a first-order Markov chain, i.e.

$$\mathsf{p}(\theta_t|\theta_{1:t-1}) = \mathsf{p}(\theta_t|\theta_{t-1}). \quad (6)$$

Eq. (1) becomes

$$\mathsf{p}(\alpha|\mathsf{y}_{1:N}) \propto \pi(\alpha) \int_{\theta_{1:N}} \prod_{t=1}^{N} \mathsf{p}(\mathsf{y}_t|\theta_t, \alpha)\mathsf{p}(\theta_t|\theta_{t-1})\mathsf{d}\theta_{1:N}. \quad (7)$$

The difference between (5) and (7) is whether the product lies inside or outside the integral. In (5), the product lies outside the integral, which divides the quantity of interest into 'small' integrals that can be computed efficiently; while (7) does not have such a decomposition, causing computational difficulty.

2.3 Difference from Bayesian estimation

Our framework is very different from the traditional Bayesian parameter estimation setting, where a certain parameter β should be estimated from the i.i.d. observations $\{\mathsf{x}_1, \mathsf{x}_2, ..., \mathsf{x}_N\}$ generated from a parametric density $\mathsf{p}(\mathsf{x}|\beta)$. If we assume that β has a prior probability $\pi(\beta)$, then the posterior probability $\mathsf{p}(\beta|\mathsf{x}_{1:N})$ is computed as

$$\mathsf{p}(\beta|\mathsf{x}_{1:N}) \propto \pi(\beta)\mathsf{p}(\mathsf{x}_{1:N}|\beta) = \pi(\beta) \prod_{t=1}^{N} \mathsf{p}(\mathsf{x}_t|\beta) \quad (8)$$

and used to derive the parameter estimate $\hat{\beta}$. One should not confuse our transformation parameter θ with the parameter β. Notice that β is fixed in $\mathsf{p}(\mathsf{x}_t|\beta)$ for different t's. However, each y_t is associates with a θ_t. Also, α is different from β in the sense that α describes the identity and β helps to describe the parametric density.

To make our framework more general, we can also incorporate the β parameter by letting the observation likelihood be $\mathsf{p}(\mathsf{y}|\theta, \alpha, \beta)$. Equation (1) then becomes

$$\begin{aligned} & \mathsf{p}(\alpha|\mathsf{y}_{1:N}) \propto \pi(\alpha)\mathsf{p}(\mathsf{y}_{1:N}|\alpha) \quad (9) \\ = \; & \pi(\alpha) \int_{\beta, \theta_{1:N}} \mathsf{p}(\mathsf{y}_{1:N}|\theta_{1:N}, \alpha, \beta)\mathsf{p}(\theta_{1:N})\pi(\beta)\mathsf{d}\theta_{1:N}\mathsf{d}\beta \\ = \; & \pi(\alpha) \int \prod_{t=1}^{N} \mathsf{p}(\mathsf{y}_t|\theta_t, \alpha, \beta)\mathsf{p}(\theta_t|\theta_{1:t-1})\pi(\beta)\mathsf{d}\theta_{1:N}\mathsf{d}\beta, \end{aligned}$$

where $\theta_{1:N}$ and β are assumed to be statistically independent. In this paper, we will focus only on (1) as if we already know the true parameter β in (9). This greatly simplifies our computation.

3 Recognition Setting and Issues

Equation (1) lays a theoretical foundation, which is universal for all recognition settings: (i) recognition is based on a single image (an I-group with $N = 1$), an I-group with $N \geq 2$, or a video sequence; (ii) the identity signature is either discrete- or continuous-valued; and (iii) the transformation space takes into account all available variations, such as localization and variations in illumination and pose.

3.1 Discrete identity signature

In a typical pattern recognition scenario, say a C-class problem, the identity signature for $\mathsf{y}_{1:N}$, $\hat{\alpha}$, is determined by the Bayesian decision rule:

$$\hat{\alpha} = \arg \max_{\{1,2,\ldots,\mathsf{C}\}} \mathsf{p}(\alpha|\mathsf{y}_{1:N}). \quad (10)$$

Usually $\mathsf{p}(\mathsf{y}|\theta, \alpha)$ is a class-dependent density, either prespecified or learned. This is a well studied problem and we will not focus on this.

3.2 Continuous identity signature

If the identity signature is continuous-valued, two recognition schemes are possible. The first is to derive a point estimate $\hat{\alpha}$ (e.g. conditional mean, mode) from $\mathsf{p}(\alpha|\mathsf{y}_{1:N})$ to represent the identity of image group $\mathsf{y}_{1:N}$. Recognition is performed by matching $\hat{\alpha}$'s belonging to different groups of images using a metric $\mathsf{k}(.,.)$. Say, $\hat{\alpha}_1$ is for group 1 and $\hat{\alpha}_2$ for group 2, the point distance

$$\hat{\mathsf{k}}_{1,2} \doteq \mathsf{k}(\hat{\alpha}_1, \hat{\alpha}_2)$$

is computed to characterize the difference between groups 1 and 2.

Instead of comparing the point estimates, the second scheme directly compares different distributions that characterize the identities for different groups of images. Therefore, for two groups 1 and 2 with the corresponding posterior probabilities $\mathsf{p}(\alpha_1)$ and $\mathsf{p}(\alpha_2)$, we use the following expected distance [18]

$$\bar{\mathsf{k}}_{1,2} \doteq \int_{\alpha_1} \int_{\alpha_2} \mathsf{k}(\alpha_1, \alpha_2)\mathsf{p}(\alpha_1)\mathsf{p}(\alpha_2)\mathsf{d}\alpha_1\mathsf{d}\alpha_2.$$

Ideally, we wish to compare the two probability distributions using quantities such as the Kullback-Leibler distance [4]. However, computing such quantities is numerically prohibitive when α is of high dimensionality.

The second scheme is preferred as it utilizes the complete statistical information, while in the first one, point estimates use partial information. For examples, if only the conditional mean is used, the covariance structure or higher-order statistics is thrown away. However, there are circumstances when the first scheme makes sense: the posterior distribution $\mathsf{p}(\alpha|Y_{1:N})$ is highly peaked or even degenerate at $\hat{\alpha}$. This might occur when (i) the variance parameters are taken to be very small; or (ii) we let N go to ∞, i.e. keep observing the same object for a long time.

3.3 The effects of the transformation

Even though recognition based on single images has been conducted for a long time, most efforts assume only one alignment parameter $\hat{\theta}$ and compute the probability $\mathsf{p}(\mathsf{y}|\hat{\theta}, \alpha)$. Any recognition algorithm computing some distance measures can be thought of as using a properly defined Gibbs distribution. The underlying assumption is that

$$\mathsf{p}(\theta) = \delta(\theta - \hat{\theta}), \tag{11}$$

where $\delta(.)$ is an impulse function. Using (11), (5) becomes

$$\mathsf{p}(\alpha|\mathsf{y}) \propto \pi(\alpha) \int_{\theta} \mathsf{p}(\mathsf{y}|\theta, \alpha)\delta(\theta - \hat{\theta})\mathsf{d}\theta = \pi(\alpha)\mathsf{p}(\mathsf{y}|\hat{\theta}, \alpha). \tag{12}$$

Incidentally, if the Laplace's method is used to approximate the integral (refer to the Appendix I for details) and the maximizer $\hat{\theta}_\alpha = \arg\max_\theta \mathsf{p}(\mathsf{y}|\theta, \alpha)\mathsf{p}(\theta)$ does not depend on α, say $\hat{\theta}_\alpha = \hat{\theta}$, then

$$\begin{aligned} \mathsf{p}(\alpha|\mathsf{y}) &\propto \pi(\alpha) \int_{\theta} \mathsf{p}(\mathsf{y}|\theta, \alpha)\mathsf{p}(\theta)\mathsf{d}\theta \\ &\simeq \pi(\alpha)\mathsf{p}(\mathsf{y}|\hat{\theta}, \alpha)\mathsf{p}(\hat{\theta})\sqrt{(2\pi)^r/|\mathtt{I}(\hat{\theta})|}. \end{aligned} \tag{13}$$

This gives rise to the same decision rule as implied by (12) and also partly explains why the simple assumption (11) can work in practice.

The alignment parameter is therefore very crucial for a good recognition performance. Even a slightly erroneous $\hat{\theta}$ may affect the recognition system significantly. It is very beneficial to have a continuous density $\mathsf{p}(\theta)$ such as a Gaussian or even a non-informative since marginalization of $\mathsf{p}(\theta, \alpha|\mathsf{y})$ over θ yields a robust estimate of $\mathsf{p}(\alpha|\mathsf{y})$.

In addition, our Bayesian framework also provides a way to estimate the best alignment parameter through the posterior probability:

$$\mathsf{p}(\theta|\mathsf{y}) \propto \int_{\alpha} \mathsf{p}(\mathsf{y}|\theta, \alpha)\pi(\alpha)\mathsf{d}\alpha. \tag{14}$$

3.4 Asymptotic behaviors

When we have an I-group or a video sequence, we are often interested in discovering the asymptotic (or large-sample) behaviors of the posterior distribution $\mathsf{p}(\alpha|\mathsf{y}_{1:N})$ when N is large. In [20], the discrete case of α in a video sequence is studied. However it is very challenging to extend this study to a continuous case. Experimentally (refer to Section 6), we find that $\mathsf{p}(\alpha|\mathsf{y}_{1:N})$ becomes more and more peaked as N increase, which seems to suggest a degenerancy in the true value α_{true}.

4 Subspace Identity Encoding

The main challenge is to specify the likelihood $\mathsf{p}(\mathsf{y}|\theta, \alpha)$. Practical considerations require that (i) the identity encoding coefficient α is compact so that our target space where α resides is of low dimensional; and (ii) α should be invariant to transformations and tightly clustered so that we can safely focus on a small portion of the spaces.

Inspired by the popularity of subspace analysis, we assume that the observation y can be well explained by a subspace, whose basis vectors are encoded in a matrix denoted by B, i.e. there exists linear coefficients α such that $\mathsf{y} \approx \mathsf{B}\alpha$. Clearly, α naturally encodes the identity. However, the observation under the transformation condition (parameterized by θ) deviates from the canonical condition (parameterized by say $\bar{\theta}$) under which the B matrix is defined. To achieve an identity encoding that is invariant to the transformation, there are two possible ways. One way is to inverse-warp the observation y from the transformation condition θ to the canonical condition $\bar{\theta}$ and the other way is to warp the basis matrix B from the canonical condition $\bar{\theta}$ to the transformation condition θ. In practice, inverse-warping is typically difficult. For example, we cannot easily warp an off-frontal view to a frontal view without explicit 3D depth information that is unavailable. Hence, we follow the second approach, which is also known as *analysis-by-synthesis* approach. We denote the basis matrix under the transformation condition θ by B_θ.

4.1 Invariant to localization, illumination, and pose

Localization parameter, denoted by ε, includes the face location, scale and in-plane rotation. Typically, an affine transformation is used. We absorb the localization parameter ε in the observation using $\mathcal{T}_\varepsilon\{\mathsf{y}\}$, where the $\mathcal{T}_\varepsilon$ is a localization operator, cropping the patch of interest and normalizing it match with the size of the basis.

The illumination parameter, denoted by λ, is a vector specifying the illuminant direction (and intensity if required). The pose parameter, denoted by υ, is a continuous-valued random variable. However, practical systems [3, 8] often discretize this due to the difficulty in handling 3D to 2D projection. Suppose the quantized pose set is $\{1, \ldots, V\}$. To achieve pose invariance, we concatenate all the images [8] $\{\mathsf{y}^1, \ldots, \mathsf{y}^V\}$ under all the views and a fixed illumination λ to form a 'huge' vector $\mathsf{Y}^\lambda = [\mathsf{y}^{1,\lambda}, ..., \mathsf{y}^{V,\lambda}]^\mathsf{T}$. To further achieve invariance to illumination, we invoke the Lambertian reflectance model, ignoring shadow pixels. Now, λ is actually a 3-D vector describing the illuminant. Since all y^υ's are illuminated by the same λ, the Lambertian model gives,

$$\mathsf{Y}^\lambda = \mathsf{W}\lambda. \tag{15}$$

Following [19] we assume that

$$\mathsf{W} = \sum_{i=1}^{m} \alpha_i \mathsf{W}_i, \tag{16}$$

we have

$$\mathsf{Y}^{\lambda} = \sum_{i=1}^{m} \alpha_i \mathsf{W}_i \lambda, \tag{17}$$

where W_i's are illumination-invariant bilinear basis and $\alpha = [\alpha_1, \ldots, \alpha_m]^{\mathsf{T}}$ provides an illuminant-invariant identity signature. Those bilinear basis can be easily learned as shown in [7, 19]. Thus α is also pose-invariant because, for a given view υ, we take the part in Y corresponding to this view and still have

$$\mathsf{y}^{\lambda,\upsilon} = \sum_{i=1}^{m} \alpha_i \mathsf{W}_i^{\upsilon} \lambda. \tag{18}$$

In summary, the basis matrix B_θ for $\theta = (\varepsilon, \lambda, \upsilon)$ with ε absorbed in y is expressed as $\mathsf{B}_{\lambda,\upsilon} = [\mathsf{W}_1^{\upsilon}\lambda, \ldots, \mathsf{W}_m^{\upsilon}\lambda]$.

We focus on the following likelihood:

$$\begin{aligned} \mathsf{p}(\mathsf{y}|\theta) &= \mathsf{p}(\mathsf{y}|\varepsilon, \lambda, \upsilon, \alpha) \\ &= \mathsf{Z}_{\lambda,\upsilon,\alpha}^{-1} \exp\{-\mathsf{D}(\mathcal{T}_\varepsilon\{\mathsf{y}\}, \mathsf{B}_{\lambda,\upsilon}\alpha)\}, \end{aligned} \tag{19}$$

where $\mathsf{D}(\mathsf{y}, \mathsf{B}_\theta\alpha)$ is some distance measure and $\mathsf{Z}_{\lambda,\upsilon,\alpha}$ is the so-called partition function which plays a normalization role. In particular, if we take D as

$$\mathsf{D}(\mathcal{T}_\varepsilon\{\mathsf{y}\}, \mathsf{B}_{\lambda,\upsilon}\alpha) = (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\lambda,\upsilon}\alpha)^{\mathsf{T}} \Sigma^{-1} (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\lambda,\upsilon}\alpha)/2, \tag{20}$$

with a given Σ (say $\Sigma = \sigma^2 \mathtt{I}$ where $\mathtt{I}$ is an identity matrix), then (19) becomes a multivariate Gaussian and the partition function $\mathsf{Z}_{\lambda,\upsilon,\alpha}$ does not depend on the parameters any more. However, even though (19) is a multivariate Gaussian, the posterior distribution $\mathsf{p}(\alpha|\mathsf{y}_{1:N})$ is no longer Gaussian.

5 Computational Issues

5.1 The integral

If the transformation space Θ is discrete, it is easy to evaluate the integral[1] $\int_\theta \mathsf{p}(\mathsf{y}|\theta, \alpha)\mathsf{p}(\theta)\mathsf{d}\theta$, which becomes a sum. If Θ is continuous, in general, computing integral $\int_\theta \mathsf{p}(\mathsf{y}|\theta, \alpha)\mathsf{p}(\theta)\mathsf{d}\theta$ is a difficult task. Many techniques are available in the literature. Here we mainly focus on two techniques: Monte Carlo simulation [13] and Laplace's method [2, 13].

[1] We drop the subscript $[.]_t$ notation as this is a general treatment.

Monte Carlo simulation. The underlying principle is the law of large number (LLN). If $\{\mathsf{x}^1, \mathsf{x}^2, \ldots, \mathsf{x}^K\}$ are K i.i.d. samples of the density $\mathsf{p}(\mathsf{x})$, for any bounded function $\mathsf{h}(\mathsf{x})$,

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathsf{h}(\mathsf{x}^k) = \int_{\mathsf{x}} \mathsf{h}(\mathsf{x})\mathsf{p}(\mathsf{x})\mathsf{d}\mathsf{x} = \mathsf{E}_{\mathsf{p}}[\mathsf{h}]. \tag{21}$$

Alternatively, when drawing i.i.d. samples from $\mathsf{p}(\mathsf{x})$ is difficult, we can use importance sampling [13]. Suppose that the *importance function* $\mathsf{q}(\mathsf{x})$ has i.i.d. realizations $\{\mathsf{x}^1, \mathsf{x}^2, \ldots, \mathsf{x}^K\}$. The pdf $\mathsf{p}(\mathsf{x})$ can be represented by a weighted set samples $\{(\mathsf{x}^k, w_{\mathsf{p}}^k)\}_{k=1}^K$, where the weight for the sample x^k is

$$w_{\mathsf{p}}^k = \mathsf{p}(\mathsf{x}^k)/\mathsf{q}(\mathsf{x}^k), \tag{22}$$

in the sense that for any bounded function $\mathsf{h}(\mathsf{x})$,

$$\lim_{K\to\infty} \sum_{k=1}^{K} w_{\mathsf{p}}^k \mathsf{h}(\mathsf{x}^k) = \sum_{k=1}^{K} \frac{\mathsf{p}(\mathsf{x}^k)}{\mathsf{q}(\mathsf{x}^k)} \mathsf{h}(\mathsf{x}^k) = \mathsf{E}_{\mathsf{p}}[\mathsf{h}]. \tag{23}$$

Laplace's method [2, 13]. The general approach of this method is presented in Appendix-I. This is a good approximation to the integral, only if the integrand is uniquely peaked and reasonably mimics the Gaussian function.

In our context, we use importance sampling (or i.i.d sampling if possible) for ε and the Laplace's method for λ and enumerate υ. We draw i.i.d. samples $\{\varepsilon^1, \varepsilon^2, \ldots, \varepsilon^K\}$ from $\mathsf{q}(\varepsilon)$ and, for each sample ε^k, compute the weight $w_{\varepsilon^k} = \mathsf{p}(\varepsilon^k)/\mathsf{q}(\varepsilon^k)$. If the i.i.d. sampling is used, the weights are always ones. Putting things together, we have (assuming $\pi(\alpha)$ is a non-informative prior)

$$\begin{aligned} \mathsf{p}(\alpha|\mathsf{y}) &\propto \int_{\varepsilon,\lambda,\upsilon} \mathsf{p}(\mathsf{y}|\varepsilon, \lambda, \upsilon, \alpha)\mathsf{p}(\varepsilon)\mathsf{p}(\lambda)\mathsf{p}(\upsilon)\mathsf{d}\varepsilon\mathsf{d}\lambda\mathsf{d}\upsilon \\ &\simeq \frac{1}{K}\sum_{k=1}^{K} w_{\varepsilon^k} \frac{1}{V}\sum_{\upsilon=1}^{V} \mathsf{p}(\mathsf{y}|\varepsilon^k, \hat{\lambda}_{\varepsilon^k,\upsilon,\alpha}, \upsilon, \alpha) \times \\ &\qquad \mathsf{p}(\hat{\lambda}_{\varepsilon^k,\upsilon,\alpha})\sqrt{(2\pi)^r/|\mathtt{I}(\hat{\lambda}_{\varepsilon^k,\upsilon,\alpha})|}, \end{aligned} \tag{24}$$

where $\hat{\lambda}_{\varepsilon^k,\upsilon,\alpha}$ is the maximizer

$$\hat{\lambda}_{\varepsilon^k,\upsilon,\alpha} = \arg\min_{\lambda} \mathsf{p}(\mathsf{y}|\varepsilon^k, \lambda, \upsilon, \alpha)\mathsf{p}(\lambda), \tag{25}$$

r is the dimensionality of λ, and $\mathtt{I}(\hat{\lambda}_{\varepsilon,\upsilon,\alpha})$ is a properly defined matrix. Refer to Appendix-II for computing $\hat{\lambda}_{\varepsilon,\upsilon,\alpha}$ and $\mathtt{I}(\hat{\lambda}_{\varepsilon,\upsilon,\alpha})$ if the likelihood is given as (19) and (20) and a non-informative prior $\mathsf{p}(\lambda)$ is assumed. Similar derivations can be conducted for an I-group of observations $\mathsf{y}_{1:N}$.

5.2 The distances $\bar{\mathsf{k}}$ and $\hat{\mathsf{k}}$

To evaluate the expected distance $\bar{\mathsf{k}}$, we resort to Monte Carlo method. In our context, the target distribution is

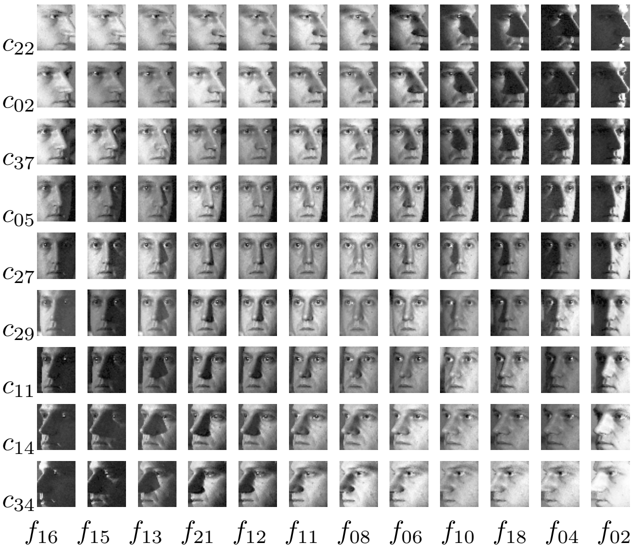

Figure 1: Examples of the face images of one PIE object under the selected illumination and poses actually used in recognition.

$\mathsf{p}(\alpha|\mathsf{y}_{1:N})$. Based on the above derivations, we know how to evaluate the target distribution, but not to draw sample from it. Therefore, we use the importance sampling. Other sampling techniques such as Monte Carlo Markov chain [13] can be applied too.

Suppose that, say for group 1, the importance function is $\mathsf{q}_1(\alpha_1)$, and weighted sample set is $\{\alpha_1^i, w_1^i\}_{i=1}^I$, the expected distance [18] is approximated as

$$\bar{\mathsf{k}}_{1,2} \simeq \frac{\sum_{i=1}^I \sum_{j=1}^J w_1^i w_2^j \mathsf{k}(\alpha_1^i, \alpha_2^j)}{\sum_{i=1}^I w_1^i \sum_{j=1}^J w_2^j}. \qquad (26)$$

The point distance is approximated as

$$\hat{\mathsf{k}}_{1,2} \simeq \mathsf{k}\left(\frac{\sum_{i=1}^I w_1^i \alpha_1^i}{\sum_{i=1}^I w_1^i}, \frac{\sum_{j=1}^J w_2^j \alpha_2^j}{\sum_{j=1}^J w_2^j}\right). \qquad (27)$$

6 Experimental Results

We use the 'illum' subset of the PIE database [15] in our experiments. This subset has 68 subjects under 21 illumination and 13 poses. Out of the 21 illumination, we select 12 of them denoted by F,

$$F = \{f_{16}, f_{15}, f_{13}, f_{21}, f_{12}, f_{11}, f_{08}, f_{06}, f_{10}, f_{18}, f_{04}, f_{02}\},$$

which typically span the set of variations. Out of the 13 poses, we select 9 of them denoted by C,

$$C = \{c_{22}, c_{02}, c_{37}, c_{05}, c_{27}, c_{29}, c_{11}, c_{14}, c_{34}\},$$

which cover from the left profile to the frontal to the right profile. In total, we have $68 * 12 * 9 = 7344$ images. Fig 1 displays one PIE object under the illumination and pose variations.

We randomly divide the 68 subjects into two parts. The first 34 subjects are used in the training set and the remaining 34 subjects are used in the gallery and probe sets. It is guaranteed that there is no identity overlap between the training set and the gallery and probe sets.

During training, the images are pre-preprocessed by aligning the eyes and mouth to desired positions. No flow computation is carried on for further alignment. After the pre-processing step, the used face image is of size 48 by 40, i.e. $d = 48 * 40 = 1920$. Also, we only study gray images by taking the average of the red, green, and blue channels of their color versions.

The training set is used to learn the basis matrix B_θ or the bilinear basis W_i's. As mentioned before, θ includes the illumination direction λ and the view pose υ, where s is a continuous-valued random vector and υ is a discrete random variable taking values in $\{1, \ldots, V\}$ with $p = 9$ (corresponding to C).

The images belonging to the remaining 34 subjects are used in the gallery and probe sets. The construction of the gallery and probe sets conforms the following: To form a gallery set of the 34 subjects, for each subject, we use an I-group of 12 images under all the illumination under one pose υ_p; to form a probe set, we use I-groups under the other pose υ_g. We mainly concentrate on the case with $\upsilon_p \neq \upsilon_g$. Thus, we have $9 * 8 = 72$ tests, with each test giving rise to a recognition score. The 1-NN (nearest neighbor) rule is applied to find the identity for a probe I-group.

During testing, we no longer use the pre-processed images and therefore the unknown transformation parameter includes the affine localization parameter, the light direction, and the discrete view pose. The prior distribution $\mathsf{p}(\varepsilon_t)$ is assumed to be a Gaussian, whose mean is found by a background subtraction algorithm and whose covariance matrix is manually specified. We use i.i.d. sampling from $\mathsf{p}(\varepsilon_t)$ since it is Gaussian. The metric $\mathsf{k}(.,.)$ actually used in our experiments is the correlation coefficient:

$$\mathsf{k}(\mathsf{x}, \mathsf{y}) = \{(\mathsf{x}^\mathsf{T}\mathsf{y})^2\}/\{(\mathsf{x}^\mathsf{T}\mathsf{x})(\mathsf{y}^\mathsf{T}\mathsf{y})\}.$$

Fig. 2 shows the marginal posterior distribution of the first element α^1 of the identity variable α, i.e., $\mathsf{p}(\alpha^1|\mathsf{y}_{1:N})$, with different N's. From Fig. 2, we notice that (i) the posterior probability $\mathsf{p}(\alpha^1|\mathsf{y}_{1:N})$ has two modes, which might fail those algorithms using the point estimate, and (ii) it becomes more peaked and tightly-supported as N increases, which empirically supports the asymptotic behavior mentioned in Section 3.

Fig. 3 shows the recognition rates for all the 72 tests. In general, when the poses of the gallery and probe sets are

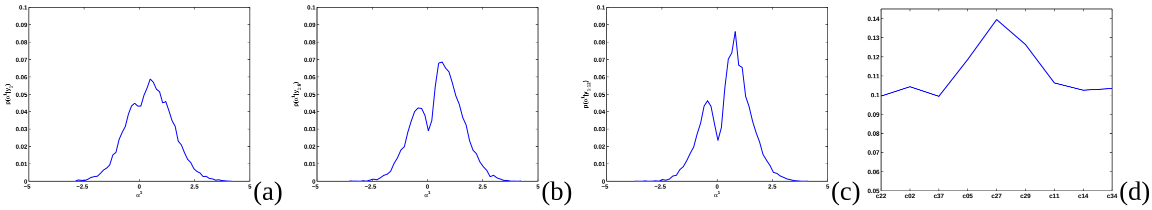

Figure 2: The posterior distributions $\mathsf{p}(\alpha^1|\mathsf{y}_{1:N})$ with different N's: (a) $\mathsf{p}(\alpha^1|\mathsf{y}_1)$; (b) $\mathsf{p}(\alpha^1|\mathsf{y}_{1:6})$; and (c) $\mathsf{p}(\alpha^1|\mathsf{y}_{1:12})$, and (d) the posterior distribution $\mathsf{p}(\upsilon|\mathsf{y}_{1:12})$. Notice that $\mathsf{p}(\alpha^1|\mathsf{y}_{1:N})$ has two modes and becomes more peaked as N increases.

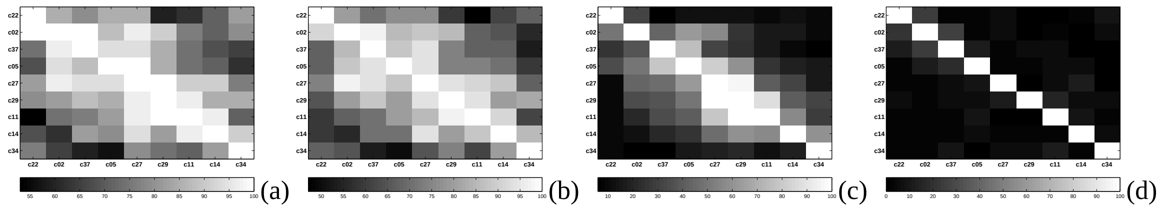

Figure 3: The recognition rates of all tests. (a) Our method based on $\bar{\mathsf{k}}$. (b) Our method based on $\hat{\mathsf{k}}$. (c) The PCA approach [16]. (d) The KL approach. Notice the different ranges of values for different methods and the diagonal entries should be ignored.

far apart, the recognition rates decrease. The best gallery sets for recognition are those in frontal poses and the worst gallery sets are those in profile views. For a comparison, Table 1 shows the average recognition rates for four different methods: our two probabilistic approaches using $\bar{\mathsf{k}}$ and $\hat{\mathsf{k}}$, respectively, the PCA approach [16], and the statistical approach [14] using the KL distance. When implementing the PCA approach, we learned a generic face subspace from all the training images, stripping their illumination and pose conditions; while implementing the KL approach, we fit a Gaussian density on every I-group and the learning set is not used. Our approaches outperform the other two approaches significantly due to the transformation-invariant subspace modeling. The KL approach [14] performs even worse than the PCA approach simply because no illumination and pose learning is used in the KL approach while the PCA approach has a learning algorithm based on image ensembles taken under different illumination and poses (though this specific information is stripped). The state-of-the-art face recognition algorithm using the PIE database is [1]. However, they used the 'lights' portion of the PIE database with an ambient light always present and the color images, which make the recognition task relatively easy.

Method	$\bar{\mathsf{k}}$	$\hat{\mathsf{k}}$	PCA	KL [14]
Rec. Rate (top 1)	82%	76%	36%	6%
Rec. Rate (top 3)	94%	91%	56%	15%

Table 1: Recognition rates of different methods.

As earlier mentioned in Section 3.3, we can infer the transformation parameters using the posterior probability $\mathsf{p}(\theta|\mathsf{y}_{1:N})$. Fig. 2 also shows the obtained $\mathsf{p}(\upsilon|\mathsf{y}_{1:12})$ for one probe I-group. In this case, the actual pose is $\upsilon = 5$ (i.e. camera c_{27}), which has the maximum probability in Fig. 2(d). Similarly, we can find an estimation for ε, which is quite accurate as the back ground subtraction algorithm already provides a clean position.

7 Conclusions and Discussions

We presented a generic framework of modeling human identity for a group of images. This framework provides a complete statistic description of the identity. We also proposed a subspace identity encoding that is invariant to location, illumination and pose variations. Our experimental results confirm the effectiveness of this approach.

We are now investigating the following: (i) Robustness to occlusions. We currently take input images without occlusions. The existence of occlusions will degrade our performances. Robust statistics can be used to handle occlusions. (ii) Without too much difficulty, we can extend our approach to perform recognition from video sequences with localization, illumination, and pose variations. We plan to use Sequential Monte Carlo methods [5] to exploit temporal continuity.

References

[1] V. Blanz and T. Vetter. Face recognition base on fitting a 3D morphable model. *IEEE Trans. PAMI*, 25:1063–1074, 2003.

[2] R. Bolle and D. Cooper. On optimally combining pieces of information with application to estimating 3-d complex-object position from range data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8:619–638, 1986.

[3] T. Cootes, K. Walker, and C. Taylor. View-base active appearance models. *Proc. Intl. Conf. on Automatic Face and Gesture Recognition*, 2000.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[5] A. Doucet, N. d. Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[6] A. Fitzgibbon and A. Zisserman. Joint manifold distance: a new approach to appearance based clustering. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[7] W. T. Freeman and J. B. Tenenbaum. Learning bilinear models for two-factor problems in vision. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[8] R. Gross, I. Matthews, and S. Baker. Eigen light-fields and face recognition across pose. *Proc. of Intl. Conf. on Automatic Face and Gesture Recognition*, 2002.

[9] J. Hornegger, D. Paulus, and N. H. Probabilistic modeling in computer vision. *Handbook of Computer Vision and Applications (Eds. Jaehne and Haussecker)*, 1999.

[10] K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[11] Y. Li, S. Gong, and H. Liddell. Constructing facial identity surfaces in a nonlinear discriminating space. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[12] X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[13] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

[14] G. Shakhnarovich, J. Fisher, and T. Darrell. Face recognition from long-term observations. *Proc. European Conference on Computer Vision*, 2002.

[15] T. Sim, S. Baker, and M. Bsat. The CMU pose, illuminatin, and expression (PIE) database. *Proc. Intl. Conf. on Face and Gesture Recognition*, 2002.

[16] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:72–86, 1991.

[17] L. Wolf and A. Shashua. Kernel principal angles for classification machines with applications to image sequence interpretation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[18] C. Yang, R. Duraiswami, A. Elgammal, and D. L. Real-time kernel-based tracking in joint feature-spatial spaces. *Tech Report CS-TR-4567, Univ. of Maryland*, 2004.

[19] S. Zhou and R. Chellappa. Rank constrained recognition under unknown illuminations. *IEEE Intl. Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[20] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91:214–245, 2003.

Appendix I – Laplace's method [2, 13]

We are interested in computing the following quantity, for $\theta = [\theta_1, \theta_2, \ldots, \theta_r]^{\mathsf{T}} \in \mathcal{R}^r$, $\mathtt{J} = \int \mathsf{p}(\theta)\mathsf{d}\theta$. Suppose that $\hat{\theta}$ is the maximizer of $\mathsf{p}(\theta)$ or equivalently $\log \mathsf{p}(\theta)$ which satisfies

$$\frac{\partial \mathsf{p}(\theta)}{\partial \theta}|_{\hat{\theta}} = 0 \; or \; \frac{\partial \log \mathsf{p}(\theta)}{\partial \theta}|_{\hat{\theta}} = 0. \tag{28}$$

We expand $\log \mathsf{p}(\theta)$ around $\hat{\theta}$ using a Taylor series:

$$\log \mathsf{p}(\theta) \simeq \log \mathsf{p}(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^{\mathsf{T}} \mathtt{I}(\hat{\theta})(\theta - \hat{\theta}), \tag{29}$$

where $\mathtt{I}(\theta)$ is an $r \times r$ matrix whose ij^{th} element is

$$\mathtt{I}_{ij}(\theta) = -\frac{\partial^2 \log \mathsf{p}(\theta)}{\partial \theta_i \partial \theta_j}. \tag{30}$$

Note that the first-order term in (29) is cancelled using (28). If $\mathsf{p}(\theta)$ is a pdf function with parameter θ, then $\mathtt{I}(\theta)$ is the famous Fisher information matrix [13]. Substituting (29) into $\mathtt{J}$ gives

$$\begin{aligned} \mathtt{J} &\simeq \mathsf{p}(\hat{\theta}) \int \exp\{-\frac{1}{2}(\theta - \hat{\theta})^{\mathsf{T}} \mathtt{I}(\hat{\theta})(\theta - \hat{\theta})\}\mathsf{d}\theta \\ &= \mathsf{p}(\hat{\theta})\sqrt{(2\pi)^r / |\mathtt{I}(\hat{\theta})|}. \end{aligned} \tag{31}$$

Appendix II – About $\hat{\lambda}_{\varepsilon,\upsilon,\alpha}$

If a non-information prior $\mathsf{p}(\lambda)$ is assumed[2], the maximizer $\hat{\lambda}_{\varepsilon,\upsilon,\alpha}$ satisfies

$$\begin{aligned} \hat{\lambda}_{\varepsilon,\upsilon,\alpha} &= \arg\max_{\lambda} \mathsf{p}(\mathsf{y}|\varepsilon, \lambda, \upsilon, \alpha) \qquad (32) \\ &= \arg\min_{\lambda} (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\lambda,\upsilon}\alpha)^{\mathsf{T}} (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\lambda,\upsilon}\alpha) \\ &= \arg\min_{\lambda} \mathsf{L}(\varepsilon, \upsilon, \lambda, \alpha) \end{aligned}$$

where $\mathsf{L}(\varepsilon, \upsilon, \lambda, \alpha) \doteq (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\lambda,\upsilon}\alpha)^{\mathsf{T}} (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\lambda,\upsilon}\alpha)$.

Using the fact that

$$\mathsf{B}_{\lambda,\upsilon}\alpha = [\mathsf{W}_1^\upsilon \lambda, \ldots, \mathsf{W}_m^\upsilon \lambda]\alpha = \mathsf{B}_{\alpha,\upsilon}\lambda; \; \mathsf{B}_{\alpha,\upsilon} \doteq \sum_{i=1}^m \alpha_i \mathsf{W}_i^\upsilon, \tag{33}$$

The term $\mathsf{L}(\varepsilon, \upsilon, \lambda, \alpha)$ becomes

$$\mathsf{L}(\varepsilon, \upsilon, \lambda, \alpha) = (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\alpha,\upsilon}\lambda)^{\mathsf{T}} (\mathcal{T}_\varepsilon\{\mathsf{y}\} - \mathsf{B}_{\alpha,\upsilon}\lambda), \tag{34}$$

which is quadratic in λ. The optimum $\hat{\lambda}_{\varepsilon,\upsilon,\alpha}$ is unique and its value is

$$\hat{\lambda}_{\varepsilon,\upsilon,\alpha} = (\mathsf{B}_{\alpha,\upsilon}{}^{\mathsf{T}} \mathsf{B}_{\alpha,\upsilon})^{-1} \mathsf{B}_{\alpha,\upsilon}{}^{\mathsf{T}} \mathsf{y} = \mathsf{B}_{\alpha,\upsilon}{}^{\dagger} \mathcal{T}_\varepsilon\{\mathsf{y}\}. \tag{35}$$

where $[.]^\dagger$ is the pseudo-inverse. Substituting (35) into $\mathsf{L}(\varepsilon, \upsilon, \lambda, \alpha)$ yields

$$\mathsf{L}(\varepsilon, \upsilon, \hat{\lambda}_{\varepsilon,\upsilon,\alpha}, \alpha) = \mathcal{T}_\varepsilon\{\mathsf{y}\}^{\mathsf{T}} (\mathtt{I}_d - \mathsf{B}_{\alpha,\upsilon}\mathsf{B}_{\alpha,\upsilon}{}^{\dagger}) \mathcal{T}_\varepsilon\{\mathsf{y}\}. \tag{36}$$

It is easy to show that $\mathtt{I}(\lambda)$ is no longer a function of λ and equals to

$$\mathtt{I} = \sigma^{-2} \mathsf{B}_{\alpha,\upsilon}{}^{\mathsf{T}} \mathsf{B}_{\alpha,\upsilon}. \tag{37}$$

[2] If a Gaussian prior is assumed, a similar derivation can be carried.