10.6

Face Recognition from Still Images and Videos

Shaohua Kevin Zhou
Siemens Corporate Research, Princeton

Rama Chellappa
University of Maryland

1 Introduction

In most situations, identifying humans using faces is an effortless task for humans. Is this true for computers? This very question defines the field of automatic face recognition [7, 31, 62], one of the most active research areas in computer vision, pattern recognition, and image understanding.

Over the past decade, the problem of face recognition has attracted substantial attention from various disciplines and has witnessed a skyrocketing growth of the literature. In this chapter, we mainly emphasize some key perspectives of the face recognition problem.

1.1 Biometric Perspective

Face is a biometric. As a consequence, face recognition finds wide applications in authentication, security, and so on. One recent application is the US-VISIT system by the Department of Homeland Security (DHS), collecting foreign passengers' fingerprints and face images.

Biometric signatures of a person characterize the physiologic or behavioral characteristics. Physiologic biometrics are innate or naturally occuring, while behavioral biometrics arise from mannerisms or traits that are learned or acquired. Table 1 lists commonly used biometrics. Biometric technologies provide the foundation for an extensive array of highly secure identification and personal verification solutions. Compared with conventional identification and verification methods based on personal identification numbers (PINs) or passwords, biometric technologies offer many advantages. First, biometrics are individualized traits while passwords may be used or stolen by someone other than the authorized user. Also, a biometric signature is very convenient since there is nothing to carry or remember. In addition, biometric technologies are becoming more accurate and less expensive.

Among all biometrics listed in Table 1, the face is a very unique one because it is the only biometric belonging to both physiologic and behavioral categories. While the physiologic part of the face has been widely exploited for face recognition, the behavioral part has not yet been fully investigated. In addition, as reported in [19, 43], face enjoys many advantages over other biometrics because it is a natural, nonintrusive, and easy-to-use biometric. For example [19], among six biometrics of face, finger, hand, voice, eye, and signature, face biometric ranks the first in the compatibility evaluation of a machine-readable travel document (MRTD) system in terms of six criteria: enrollment, renewal, machine-assisted identity verification requirements, redundancy, public

TABLE 1 Physiologic and behavioral biometrics

Type	Examples
Physiologic biometrics	DNA, face, fingerprint, hand geometry, iris, pulse, retinal, and body odor
Behavioral biometrics	Face, gait, handwriting, signature, and voice

perception, and storage requirements and performance. Probably the most important feature of acquiring the face biometric signature is that less cooperation is required during data acquisition.

In addition to applications related to identification and verification such as access control, law enforcement, identification (ID), licensing, surveillance, and so forth, face recognition is also useful in human–computer interaction, virtual reality, database retrieval, multimedia, computer entertainment, and so forth. See [31, 62] for recent summaries on face recognition applications.

1.2 Experimental Perspective

Face recognition mainly involves the following three tasks [46]:

- Verification. The recognition system determines if the query face image and the claimed identity match.
- Identification. The recognition system determines the identity of the query face image.
- Watch list. The recognition system first determines if the identity of the query face image is in the watch list and, if yes, then identifies the individual.

Figure 1 illustrates the above three tasks and corresponding metrics used for evaluation. Among three tasks, the watch list task is the most difficult one.

This chapter focuses only on the identification task. We introduce a face recognition test protocol FERET [45] widely followed in the face recognition literature. FERET stands for "facial recognition technology". FERET assumes the availability of the following three sets: a training set, a gallery set, and a probe set. The training set is provided for the recognition algorithm to learn features that are capable of characterizing the whole human face space. The gallery and probe sets are used in the testing stage. The gallery set contains images with known identities and the probe set with unknown identities. The algorithm associates descriptive features with images in the gallery and probe sets and determines the identities of the probe images by comparing their associated features with features associated with gallery images.

1.3 Theoretical Perspective

Face recognition is by nature an interdisciplinary research area, involving researchers from pattern recognition, computer vision and graphics, image processing/understanding, statistical computing and machine learning. In addition, automatic face recognition algorithms/systems are often guided by the psychophysics and neural studies on how

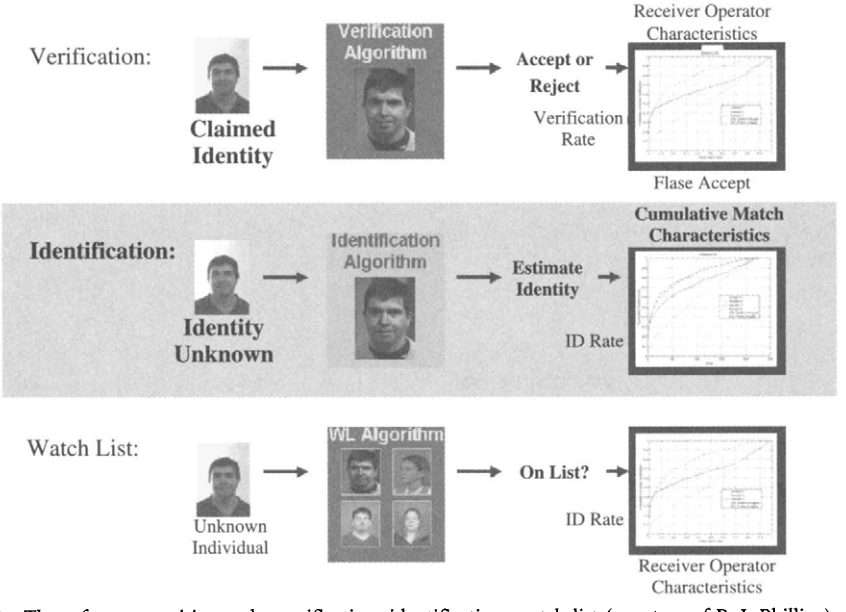

FIGURE 1 Three face recognition tasks: verification, identification, watch list (courtesy of P. J. Phillips).

humans perceive faces. A good summary of research on face perception is presented in [40]. We now focus on the theoretical implication of pattern recognition for the task of face recognition.

We present a hierarchic study of face pattern. There are three levels forming the hierarchy: pattern, visual pattern, and face pattern, each associated with a corresponding theory of recognition. Accordingly, face recognition approaches can be grouped into three categories.

1. *Pattern and pattern recognition*: Because face is first a pattern, any pattern recognition theory [11] can be directly applied to the face recognition problem. In general, a vector representation is used in pattern recognition. A common way of deriving such a vector representation from a two-dimensional (2D) face image, say of size $M \times N$, is through a "vectorization" operator that stacks all pixels in a particular order, say a raster-scanning order, to an $MN \times 1$ vector. Obviously, given an arbitrary $MN \times 1$ vector, it can be decoded into an $M \times N$ image by an inverse-vectorization operator. Such a vector representation corresponds to a holistic perception viewpoint in psychophysics literature [6].

Subspace methods are pattern recognition techniques widely invoked in various face recognition approaches. Two well-known appearance-based recognition schemes use principal component analysis (PCA) and linear discriminant analysis (LDA). PCA performs an eigen-decomposition of the covariance matrix and consequently minimizes the reconstruction error in the mean square sense. LDA minimizes the within-class scatter while maximizing the between-class scatter [23]. The PCA approach used in face recognition is also known as the "Eigenface" approach [54]. The LDA approach [12] used in face recognition is referred to as the "Fisherface" approach [3] since LDA is also known as Fisher discriminant analysis. Further, PCA and LDA have been combined (LDA after PCA) as in [60] to obtain improved recognition. Other subspace methods such as independent component analysis (ICA) [1], local feature analysis (LFA) [41], probabilistic subspace [38, 39], multiexemplar discriminant analysis [69] have been used. A comparison of these subspace methods is reported in [39]. Other than subspace methods, classic pattern recognition tools such as neural networks [33], learning methods [44], and evolutionary pursuit/genetic algorithms [35] have also been applied.

One concern in a regular pattern recognition problem is the "curse of dimensionality" since usually M and N themselves are quite large numbers. In face recognition, because of limitations in image acquisition, practical face recognition systems store only a small number of samples per subject. This further aggravates the curse of dimensionality problem.

2. *Visual pattern and visual recognition:* In the middle of the hierarchy sits visual pattern. Face is a visual pattern in the sense that it is a 2D appearance of a three-dimensional (3D) object captured by an imaging system. Certainly, visual appearance is affected by the configuration of the imaging system. An illustration of the imaging system is presented in Fig. 2.

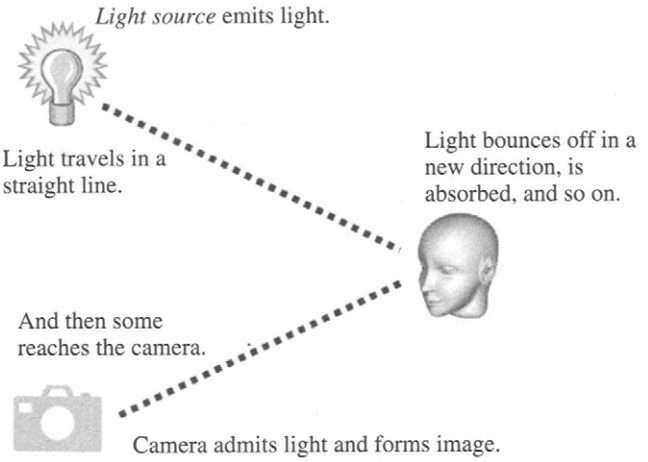

FIGURE 2 An illustration of the imaging system.

There are two distinct characteristics of the imaging system: photometric and geometric.

- The photometric characteristics are related to the lighting source distribution in the scene. Figure 3 shows the face images of a person captured under varying illumination conditions. Numerous models have been proposed to describe the illumination phenomenon (i.e., how the light travels when it hits the object). In addition to its relationship with the light distribution such as light direction and intensity, an illumination model is in general also relevant to surface material properties of the illuminated object.
- The geometric characteristic is about camera properties and relative positioning of the camera and the object. Camera properties include camera intrinsic parameters and camera imaging models. The imaging models widely studied in the computer vision literature are orthographic, scale orthographic, and perspective models. Due to the projective nature of the perspective model, the orthographic or scale-orthograhic models are used in the face recognition community. The relative positioning of the camera and the object results in pose variation, a key factor in determining how the 2D appearances are produced. Figure 3 shows the face images of a person captured at varying poses.

Studying photometric and geometric characteristics is one of the key problems in the object recognition literature and consequently visual recognition under illumination and pose variations is the main challenge for object recognition. A full

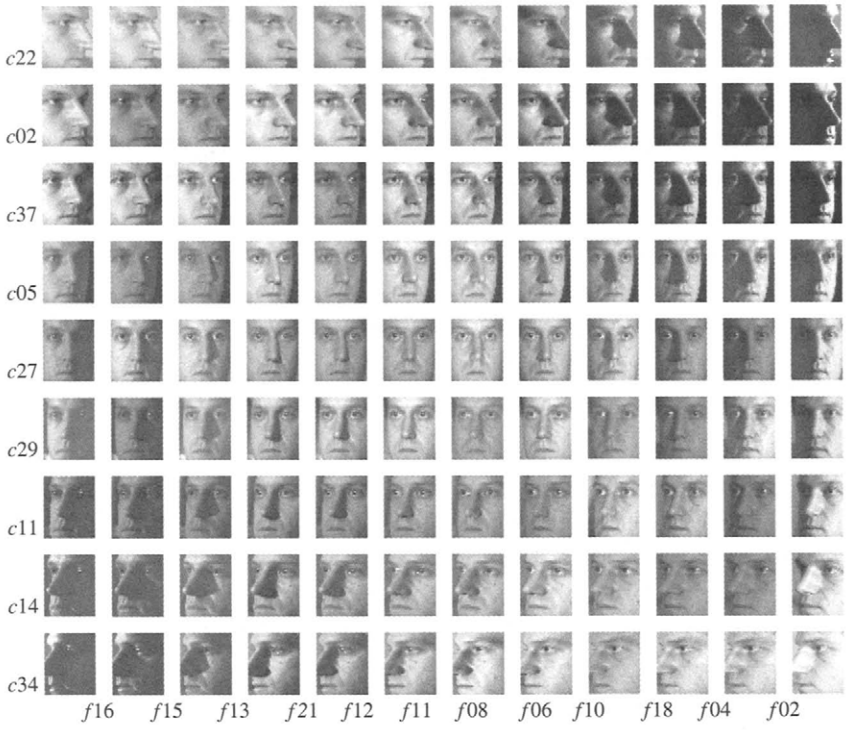

FIGURE 3 Examples of the face images of one PIE [52] object under the different illumination and poses. The images are of size 48 × 40.

review of visual recognition literature is beyond the scope of the chapter. However, face recognition addressing the photometric and geometric challenges is still at a nascent stage.

Approaches to face recognition under illumination variation are usually treated as extensions of research efforts on illumination models. For example, if a simplified Lambertian reflectance model ignoring the shadow pixels [49, 59] is used, a rank-3 subspace can be constructed to cover appearances arbitrarily illuminated by a distant point source. Similar low-dimensional subspace [2] can be found in the Lambertian model that includes attached shadows. Face recognition is conducted by checking if a query face image lies in the object-specific illumination subspace. To generalize from the object-specific illumination subspace to a class-specific illumination subspace, bilinear models are used in [15, 50, 66]. Most face recognition approaches addressing pose variations use view-based appearance representations [8, 17, 42]. Face recognition under variations in illumination and poses is more difficult compared to recognition when only one variation is present. Proposed approaches in the literature include [5, 16, 18, 55, 67], among which the 3D morphable model [5] yields the best recognition performance. The feature-based approach [27] is reported to be partially robust to illumination and pose variations.

Another important extension of visual pattern recognition is in exploiting video. The ubiquitousness of video sequences calls upon novel recognition algorithms based on videos. Because a video sequence is a collection of still images, face recognition from still images certainly applies to video sequences. However, an important property of a video sequence is its temporal dimension or dynamics. Recent psychophysical and neural studies [25] demonstrate the role of movement in face recognition: Famous faces are easier to recognize when presented in moving sequences than in still photographs, even under a range of different types of degradations. Computational approaches using such temporal information include [26, 30, 63–65]. Clearly, due to the free movement of human face and uncontrolled environments, issues like illumination and pose variations still exist. Besides these issues, localizing faces or face segmentation in a cluttered environment in video sequences is very challenging too.

In surveillance scenarios, further challenges include poor video quality and lower resolution. For example, the face region can be as small as 15 × 15. Most feature-based approaches [5, 27] need face images of size as large as 128 × 128. The attractiveness of the video sequence is that the video provides multiple observations with temporal continuity.

3. *Face pattern and face recognition:* At the top of the hierarchy lies the face pattern. The face pattern specializes the visual pattern by specializing the object to be a human face.

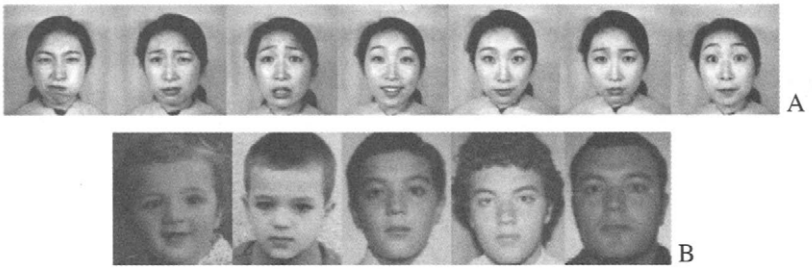

FIGURE 4 **A:** Appearances of a person with different facial expressions (from [37]). **B:** Appearances of a person at different ages (from [28]).

Therefore, face-specific properties or characteristics should be taken into account when performing face recognition.

- *Expression and deformation.* Humans exhibit emotions. The natural way to express the emotions is through facial expressions, yielding patterns under nonrigid deformations. This nonrigidity introduces very high degrees of freedom and perplexes the recognition task. Figure 4A shows the face images of a person exhibiting different expressions. While face expression analysis has attracted a lot of attention [4, 53], recognition under facial expression variation has not been fully explored.
- *Aging.* Face appearances vary significantly with age and such variations are specific to an individual. Theoretical modeling of aging [28] is very difficult due to the individualized variation. Figure 4B shows the face images of a person at different ages.
- *Face surface.* One speciality of the face surface is its bilateral symmetry. The symmetry constraint has been widely exploited in [51, 61, 66]. In addition, surface integrability is an inherent property of any surface, which has also been used in [14, 59, 66].
- *Self-similarity.* There is a strong visual similarity among face images of different individuals. Geometric positioning of facial features such as eyes, noses, and mouths are similar across individuals. Early face recognition approaches in the 1970s [21, 22] used the distances between feature points to describe the face and achieved some success. Also, the properties of face surface materials are similar within the same race. As a consequence of visual similarity, the "shapes" of the face appearance manifolds belonging to different subjects are similar. This is the foundation of approaches [39, 69] that attempt to capture the "shape" characteristics using the so-called intraperson space.
- *Makeup and cosmetics.* These factors are very individualized and unpredictable. Other than the effect of glasses, which has been studied in [3], effects induced by other factors are not widely understood in the recognition literature. However, modeling these factors can be useful for face animation in the computer graphics literature.

1.4 A Unified Approach

A wide array of face recognition approaches has been proposed in the literature. Early face recognizers [1, 3, 12, 23, 38, 39, 41, 54] yielded unsatisfactory results especially when confronted with variations in pose, illumination, and expression. In addition, the recognizers have been further hampered by the registration requirement as images that the recognizers process contain transformed appearances of the object.

Recent advances in face recognition have focused on face recognition under illumination and pose variations [2, 5, 8, 17, 50, 56, 66, 67]. Face recognition under variations in expression and aging have been less investigated.

While most recognizers process a single image, there is a growing interest in using a group of images [13, 29, 30, 32, 36, 48, 57, 64]. 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 sequences. If temporal information is stripped, the video sequences reduce to I-groups. Whenever we refer to video sequences, we imply dependent groups of images.

We attempt to propose a unified approach [68] that possesses the following features:

- It processes either a single image or a group of images (including the I-group and video sequences) in a universal manner.
- It handles the localization problem, pose, illumination, and expression variations.
- The identity description could be either discrete or continuous. The continuous identity encoding typically arises from subspace modeling.
- It is probabilistic and integrates all the available evidences.

We elaborate on the proposed framework and point out its properties and connections with various approaches in Section 2. In Section 3, we substantiate the framework with two instances: face recognition from a group of still images and from video sequences.

2 Framework of Probabilistic Identity Characterization

Suppose α represents the identity in an abstract manner. It can be either discrete- or continuous-valued. If we have an N-class problem, α is discrete taking value in $\{1, 2, \ldots, N\}$. If we associate the identity with image intensity or feature vectors derived from subspace projections, α is continuous-valued. Given a group of images $\mathbf{y}_{1:T} \doteq \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ containing the appearances of the same but unknown identity, *probabilistic identity characterization is equivalent to finding the posterior probability* $\mathrm{p}(\alpha|\mathbf{y}_{1:T})$.

As the image only contains a transformed version of the object, we also need to associate with it a transformation parameter θ, which lies in a transformation space Θ. Here the term "transformation" is a loose word to model the variations involved, be it warping, pose, illumination, or expressions. The transformation space Θ is usually application-dependent. Affine transformation is often used to compensate for the localization problem. To handle illumination variation, the estimates of lighting direction are used. If pose variations are involved, a 3D transformation is needed or a discrete set is used if we quantize the continuous view space. Suppose that the dimension of the transformation space Θ is r.

We assume that the prior probability of α is $\pi(\alpha)$, which is assumed to be, in practice, a *noninformative* prior. A noninformative 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:

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

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

$$(a)\ \mathrm{p}(\mathbf{y}_{1:T}|\theta_{1:T}, \alpha) = \prod_{t=1}^{T} \mathrm{p}(\mathbf{y}_t|\theta_t, \alpha); \qquad (2)$$

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

Equation (1) involves two key quantities: the *observation likelihood* $\mathrm{p}(\mathbf{y}_t|\theta_t, \alpha)$ and the *state transition probability* $\mathrm{p}(\theta_t|\theta_{1:t-1})$. The former is essential to the 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 $\mathrm{p}(\theta_t|\theta_{1:t-1})$.

1. *Independent group (I-group):* In this case, the transformations $\{\theta_t; t = 1, 2, \ldots, T\}$ are independent of each other, for example

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

Equation (1) then becomes

$$\mathrm{p}(\alpha|\mathbf{y}_{1:T}) \propto \pi(\alpha)\prod_{t=1}^{T}\int_{\theta_t} \mathrm{p}(\mathbf{y}_t|\theta_t, \alpha)\mathrm{p}(\theta_t)\mathrm{d}\theta_t. \qquad (5)$$

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

The most widely studied case in the literature is when $T = 1$ (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 $T > 1$) depending on the context. We will present in Section 2.1 the shortcomings of many contemporary approaches.

It all boils down to how to compute the integral in (5). However, in real applications it is difficult to directly compute it and numeric techniques are often used.

2. *Video sequence:* In the case of a video sequence, temporal continuity between successive video frames implies that the transformations $\{\theta_t; t = 1, 2, \ldots, T\}$ follow a Markov chain. Without loss of generality, we assume a first-order Markov chain, for example

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

Equation (1) becomes

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

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; whereas (7) does not have such a decomposition, causing computational difficulty.

3. *Difference from Bayesian estimation:* Our framework is very different from the traditional Bayesian parameter estimation

setting, where a certain parameter β is estimated from the identical and independantly distributed (i.i.d). observations $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ generated from a parametric density $\mathrm{p}(\mathbf{x}|\beta)$. If we assume that β has a prior probability $\pi(\beta)$, then the posterior probability $\mathrm{p}(\beta|\mathbf{x}_{1:T})$ is computed as

$$\mathrm{p}(\beta|\mathbf{x}_{1:T}) \propto \pi(\beta)\mathrm{p}(\mathbf{x}_{1:T}|\beta) = \pi(\beta)\prod_{t=1}^{T}\mathrm{p}(\mathbf{x}_t|\beta) \tag{8}$$

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 $\mathrm{p}(\mathbf{x}_t|\beta)$ for different t's. However, each $\mathbf{y}_t$ is associated 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 to be $\mathrm{p}(\mathbf{y}|\theta, \alpha, \beta)$. Equation (1) then becomes

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

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

2.1 Recognition Setting and Issues

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

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

$$\hat{\alpha} = \arg\min_{\{1,2,\ldots,N\}}\mathrm{p}(\alpha|\mathbf{y}_{1:T}). \tag{10}$$

Usually $\mathrm{p}(\mathbf{y}|\theta, \alpha)$ is a class-dependent density, either pre-specified or learned.

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 $\mathrm{p}(\alpha|\mathbf{y}_{1:T})$ to represent the identity of the image group $\mathbf{y}_{1:T}$. Recognition is performed by matching $\hat{\alpha}$'s belonging to different groups of images using a metric $\mathrm{k}(.,.)$. Say, $\hat{\alpha}_1$ is for group 1 and $\hat{\alpha}_2$ for group 2, the point distance

$$\hat{\mathrm{k}}_{1,2} \doteq \mathrm{k}(\hat{\alpha}_1, \hat{\alpha}_2) \tag{11}$$

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 $\mathrm{p}(\alpha_1)$ and $\mathrm{p}(\alpha_2)$, we use the following expected distance [58]:

$$\bar{\mathrm{k}}_{1,2} \doteq \int_{\alpha_1}\int_{\alpha_2}\mathrm{k}(\alpha_1,\alpha_2)\mathrm{p}(\alpha_1)\mathrm{p}(\alpha_2)\mathrm{d}\alpha_1\mathrm{d}\alpha_2. \tag{12}$$

Ideally, we wish to compare the two probability distributions using quantities such as the Kullback-Leibler distance [9]. However, computing such quantities is numerically prohibitive when α is of high dimensionality.

The second scheme is preferred as it uses complete statistical information, whereas the first one based on point estimates uses partial information. For example, if only the conditional mean is used, the covariance structure or higher-order statistics is not used. However, there are circumstances when the first scheme is appropriate: the posterior distribution $\mathrm{p}(\alpha|\mathbf{y}_{1:T})$ is highly peaked or even degenerate at $\hat{\alpha}$. This might occur when (a) the variance parameters are taken to be very small; or (b) we let T go to ∞ (i.e., keep observing the same object for a long time).

To evaluate the expected distance $\bar{\mathrm{k}}$, we resort to importance sampling [47]. Other sampling techniques such as Monte Carlo Markov chain [47] can also be applied. Suppose that, say for group 1, the importance function is $\mathrm{q}_1(\alpha_1)$, and weighted sample set is $\{\alpha_1^{(i)}, w_1^{(i)}\}_{i=1}^{I}$, the expected distance is approximated as

$$\bar{\mathrm{k}}_{1,2} \simeq \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}w_1^{(i)}w_2^{(j)}\mathrm{k}\left(\alpha_1^{(i)},\alpha_2^{(j)}\right)}{\sum_{i=1}^{I}w_1^{(i)}\sum_{j=1}^{J}w_2^{(j)}}. \tag{13}$$

The point distance is approximated as

$$\hat{\mathrm{k}}_{1,2} \simeq \mathrm{k}\frac{\left(\sum_{i=1}^{I}w_1^{(i)}\alpha_1^{(i)}\sum_{i=1}^{I}w_1^{(i)},\right.}{\left.\sum_{j=1}^{J}w_2^{(j)}\alpha_2^{(j)}\sum_{j=1}^{J}w_2^{(j)}\right)}. \tag{14}$$

3. *The effects of the transformation:* Even though recognition based on a single image has been attempted for a long time,

most efforts assume only one alignment parameter $\hat{\theta}$ and compute the probability $\mathrm{p}(\mathbf{y}|\hat{\theta}, \alpha)$. Any recognition algorithm computing some distance measure can be thought of as using a properly defined Gibbs distribution. The underlying assumption is that

$$\mathrm{p}(\theta) = \delta(\theta - \hat{\theta}), \tag{15}$$

where $\delta(.)$ is an impulse function. Using (15), (5) becomes

$$\mathrm{p}(\alpha|\mathbf{y}) \propto \pi(\alpha) \int_{\theta} \mathrm{p}(\mathbf{y}|\theta, \alpha)\delta(\theta - \hat{\theta})\mathrm{d}\theta = \pi(\alpha)\mathrm{p}(\mathbf{y}|\hat{\theta}, \alpha). \tag{16}$$

Incidentally, if the Laplace method [47] is used to approximate the integral $\int_{\theta} \mathrm{p}(\mathbf{y}|\theta, \alpha)\mathrm{p}(\theta)\mathrm{d}\theta$ and the maximizer $\hat{\theta}_{\alpha} = \arg\min_{\theta} \mathrm{p}(\mathbf{y}|\theta, \alpha)\mathrm{p}(\theta)$ does not depend on α, say $\hat{\theta}_{\alpha} = \hat{\theta}$, then

$$\begin{aligned} \mathrm{p}(\alpha|\mathbf{y}) &\propto \pi(\alpha) \int_{\theta} \mathrm{p}(\mathbf{y}|\theta, \alpha)\mathrm{p}(\theta)\mathrm{d}\theta \\ &\simeq \pi(\alpha)\mathrm{p}(\mathbf{y}|\hat{\theta}, \alpha)\mathrm{p}(\hat{\theta})\sqrt{(2\pi)^r/|\mathtt{I}(\hat{\theta})|} \\ &\propto \pi(\alpha)\mathrm{p}(\mathbf{y}|\hat{\theta}, \alpha), \end{aligned} \tag{17}$$

where $\mathtt{I}(\theta)$ is an $r \times r$ matrix (r is the dimension of *theta*) whose ij-th element is

$$\mathtt{I}_{ij}(\theta) = -\frac{\partial^2 \log \mathrm{p}(\theta)}{\partial\theta_i \partial\theta_j}. \tag{18}$$

This gives rise to the same decision rule as implied by (16) and also partly explains why the simple assumption (15) can work in practice.

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

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

$$\mathrm{p}(\theta|\mathbf{y}) \propto \int_{\alpha} \mathrm{p}(\mathbf{y}|\theta, \alpha)\pi(\alpha)\mathrm{d}\alpha. \tag{19}$$

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 $\mathrm{p}(\alpha|\mathbf{y}_{1:T})$ when T is large. In [64], the discrete case of α in a video sequence is studied. However, it is very challenging to extend this study to the continuous case. Experimentally (refer to Section 3.1.2), we find that $\mathrm{p}(\alpha|\mathbf{y}_{1:T})$ becomes more and more peaked as T increase, which seems to suggest a degenerancy in the true value α_{true}.

3 Instances of Probabilistic Identity Characterization

3.1 Face Recognition from a Group of Still Images

The main challenge is to specify the likelihood $\mathrm{p}(\mathbf{y}|\theta, \alpha)$. Practical considerations require that (a) the identity encoding coefficient α should be compact so that our target space where α resides is low dimensional, and (b) 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 $\mathbf{y}$ can be well explained by a subspace, whose basis vectors are encoded in a matrix denoted by $\mathbf{B}$ (i.e., there exist linear coefficients α such that $\mathbf{y} \approx \mathbf{B}\alpha$). 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 $\mathbf{B}$ 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 $\mathbf{y}$ from the transformation condition θ to the canonical condition $\bar{\theta}$ and the other way is to warp the basis matrix $\mathbf{B}$ 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 $\mathbf{B}_{\theta}$.

1. *Subspace identity encoding—Invariant to localization, illumination, and pose:* The 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 $\mathcal{T}\{\mathbf{y}; \varepsilon\}$, where the $\mathcal{T}\{.; \varepsilon\}$ is a localization operator, cropping the region of interest and normalizing it to 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 [8, 17] often discretize this due to the difficulty in handling 3D to 2D projection. Suppose the quantized pose set is $\{1, 2, \ldots, V\}$. To achieve pose invariance, we concatenate all the images [17, 67] $\{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^V\}$ under all the views and a fixed illumination λ to form a very long vector $\mathbf{Y}^{\lambda} = [\mathbf{y}^{1,\lambda}, \mathbf{y}^{2,\lambda}, \ldots, \mathbf{y}^{V,\lambda}]^{\mathrm{T}}$. To

further achieve invariance to illuminations, we invoke the Lambertian reflectance model, ignoring the shadow pixels. Now, λ is actually a 3D vector describing the illuminant. We now follow [67] to derive a bilinear algorithm that is summarized as follows.

Because all $\mathbf{y}^v$'s are illuminated by the same λ, the Lambertian model gives,

$$\mathbf{Y}^\lambda = \mathbf{W}\lambda. \tag{20}$$

Following [66], we assume that

$$\mathbf{W} = \sum_{i=1}^{m} \alpha^i \mathbf{W}_i, \tag{21}$$

and we have

$$\mathbf{Y}^\lambda = \sum_{i=1}^{m} \alpha^i \mathbf{W}_i \lambda, \tag{22}$$

where $\mathbf{W}_i$'s are illumination-invariant bilinear basis and $\alpha = [\alpha^1, \alpha^2, \ldots, \alpha^m]^{\mathrm{T}}$ provides an illuminant-invariant identity signature. The bilinear basis can be easily learned as shown in [15, 66]. Thus, α is also pose-invariant because, for a given view v, we take the part in $\mathbf{Y}$ corresponding to this view and still have

$$\mathbf{y}^{\lambda,v} = \sum_{i=1}^{m} \alpha^i \mathbf{W}_i^v \lambda, \tag{23}$$

where $\mathbf{W}_i^v$ take the part in $\mathbf{W}_i$ corresponding to view v.

In summary, the basis matrix $\mathbf{B}_\theta$ for $\theta = (\varepsilon, \lambda, v)$ with ε absorbed in $\mathbf{y}$ is expressed as $\mathbf{B}_{\lambda,v} = [\mathbf{W}_1^v\lambda, \mathbf{W}_2^v\lambda, \ldots, \mathbf{W}_m^v\lambda]$.

We focus on the following likelihood:

$$\mathrm{p}(\mathbf{y}|\theta) = \mathrm{p}(\mathbf{y}|\varepsilon, \lambda, v, \alpha) = \mathrm{Z}_{\lambda,v,\alpha}^{-1} \exp\{-\mathrm{D}(\mathcal{T}\{\mathbf{y};\varepsilon\}, \mathbf{B}_{\lambda,v}\alpha)\}, \tag{24}$$

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

$$\mathrm{D}(\mathcal{T}\{\mathbf{y};\varepsilon\}, \mathbf{B}_{\lambda,v}\alpha) = (\mathcal{T}\{\mathbf{y};\varepsilon\} - \mathbf{B}_{\lambda,v}\alpha)^{\mathrm{T}} \Sigma^{-1} (\mathcal{T}\{\mathbf{y};\varepsilon\} - \mathbf{B}_{\lambda,v}\alpha)/2, \tag{25}$$

with a given Σ (say $\Sigma = \sigma^2\mathrm{I}$ where I is an identity matrix), then (24) becomes a multivariate Gaussian and the partition function $\mathrm{Z}_{\lambda,v,\alpha}$ does not depend on the parameters any more. However, even though (24) is a multivariate Gaussian, the posterior distribution $\mathrm{p}(\alpha|\mathbf{y}_{1:N})$ is no longer Gaussian.

2. *Experimental results:* We use a portion of the "illum" subset of the PIE database [52]. This part includes 68 subject under 12 lighting sources and at nine poses. In total, we have $68 \times 12 \times 9 = 7344$ images. Figure 3 displays one PIE object under 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 among the training, gallery, and probe sets.

During training, the images are prepreprocessed by aligning the eyes and mouth to desired positions. No flow computation is carried on for further alignment. After the preprocessing step, the face image is of size 48×40, (i.e., $d = 48 \times 40 = 1920$). 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 $\mathbf{B}_\theta$ or the bilinear basis $\mathbf{W}_i$'s. As mentioned before, θ includes the illumination direction λ and the view pose v, where λ is a continuous-valued random vector and v is a discrete random variable taking values in $\{1, 2, \ldots, V\}$ with $p = 9$.

The images belonging to the remaining 34 subjects are used as gallery and probe sets. To form a gallery set of the 34 subjects, for each subject, we use an I-group of 12 images under all the illuminations under one pose v_p (e.g., one row of Figure 3); to form a probe set, we use I-groups under the other pose v_g. We mainly concentrate on the case with $v_p \neq v_g$. Thus, we have $9 \times 8 = 72$ 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 preprocessed images and therefore the unknown transformation parameter includes the affine localization parameter, the light direction, and the discrete pose. The prior distribution $\mathrm{p}(\varepsilon_t)$ is assumed to be Gaussian, whose mean is found by a background subtraction algorithm and whose covariance matrix is manually specified. Numeric computation is conducted as in [68]. The metric $\mathrm{k}(.,.)$ actually used in our experiments is the correlation coefficient:

$$\mathrm{k}(\mathbf{x}, \mathbf{y}) = \{(\mathbf{x}^{\mathrm{T}}\mathbf{y})^2\}/\{(\mathbf{x}^{\mathrm{T}}\mathbf{x})(\mathbf{y}^{\mathrm{T}}\mathbf{y})\}. \tag{26}$$

Figure 5 shows the marginal posterior distribution of the first element α^1 of the identity variable α, i.e., $\mathrm{p}(\alpha^1|\mathbf{y}_{1:T})$, with different T's. From Fig. 5, we notice that (a) the posterior probability $\mathrm{p}(\alpha^1|\mathbf{y}_{1:T})$ has two modes, which might fail those algorithms using the point estimate, and (b) it becomes more peaked and tightly supported as N increases, which empirically supports the asymptotic behavior mentioned in Section 2.1.

Figure 6 shows the recognition rates for all the 72 tests. In general, when the poses of the gallery and probe sets are far apart, the recognition rates decrease. The best gallery sets for

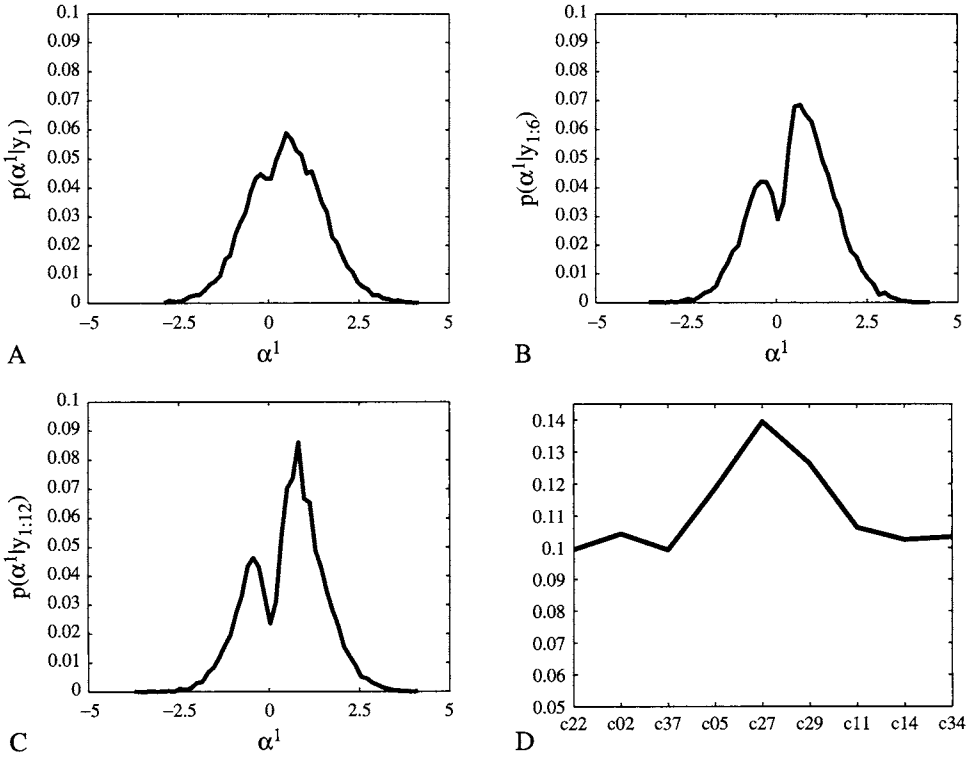

FIGURE 5 The posterior distributions $p(\alpha^1|y_{1:T})$ with different T's: **A:** $p(\alpha^1|y_1)$; **B:** $p(\alpha^1|y_{1:6})$; and **C:** $p(\alpha^1|y_{1:12})$, and **D:** the posterior distribution $p(v|y_{1:12})$. Notice that $p(\alpha^1|y_{1:T})$ has two modes and becomes more peaked as T increases.

recognition are those in frontal poses and the worst gallery sets are those in profile views.

For comparison, Table 2 shows the average recognition rates for four different methods: our two probabilistic approaches using $\bar{k}$ and $\hat{k}$, respectively, the PCA approach [54], and the statistical approach [48] using the Kullback (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 significantly outperform the other two approaches due to transformation-invariant subspace modeling. The KL approach [48] 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 illuminations and poses (though this specific information is stripped).

As mentioned in Section 2.1, we can infer the transformation parameters using the posterior probability $p(\theta|y_{1:T})$. Figure 5 also shows the obtained $p(v|y_{1:12})$ for one probe I-group. In this case, the actual pose is $v = 5$ (i.e., camera c_{27}), which has the maximum probability in Fig. 5(D). Similarly, we can find an estimate for ε, which is quite accurate as the background subtraction algorithm already provides a clean position.

3.2 Face Recognition from a Video Sequence

Face recognition from a video inevitably requires solving tracking and recognition tasks. Visual tracking models the interframe appearance differences and visual recognition models the appearance differences between the video frames and gallery images. Simultaneous tracking and recognition [64, 65] provide a mechanism of jointly modeling interframe appearance differences and the appearance differences between video frames and gallery images. Here we focus on the case when the identity variable is discrete. In addition, we assume that the gallery set consists of images $\{h_1, h_2, \ldots, h_N\}$, with each h_α for the αth individual.

1. *Simultaneous tracking and recognition:* It is easy to show that the derivation of probabilistic identity characterization is equivalent to a time series state-space model consisting of the following three components: the motion transition equation, the identity equation, and the observation likelihood. This defines the recognition task as a statistical inference problem, which can be solved using particle filters. We briefly review the state-space model.

Motion transition equation: In its most general form, the motion model can be written as

$$\theta_t = g(\theta_{t-1}, u_t); \quad t \geq 1, \tag{27}$$

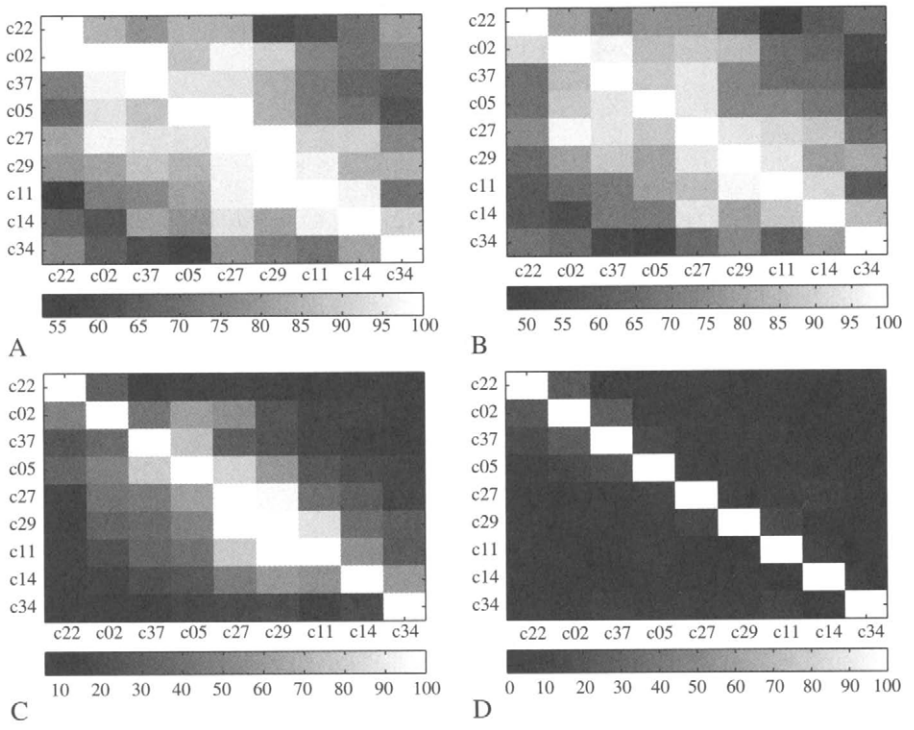

FIGURE 6 The recognition rates of all tests. **A:** Our method based on $\bar{k}$. **B:** Our method based on $\hat{k}$. **C:** The principal component approach [54]. **D:** The KL approach. Notice the different ranges of values for different methods and the diagonal entries should be ignored.

TABLE 2 Recognition rates of different methods

Method	$\bar{k}$	$\hat{k}$	PCA	KL [48]
Rec. Rate (top 1)	82%	76%	36%	6%
Rec. Rate (top 3)	94%	91%	56%	25%

KL, . . . ; PCA, principal component approach; Rec., recognition.

where u_t is *noise* in the motion model, whose distribution determines the motion state transition probability $\mathrm{p}(\theta_t|\theta_{t-1})$. The function $\mathrm{g}(.,.)$ characterizes the evolving motion and it could be a function learned offline or given a priori. One of the simplest choice is an additive function (i.e., $\theta_t = \theta_{t-1} + \mathsf{u}_t$).

The choice of θ_t is application dependent. The affine motion parameter is often used when there is no significant pose variation in the video sequence. However, if a 3D face model is used, 3D motion parameters should be used accordingly. In the experiments presented below, we set θ_t as the affine motion parameter and $\mathrm{g}(.,.)$ as the additive function.

Identity equation: Assuming that the identity does not change as time proceeds, we have

$$\alpha_t = \alpha_{t-1}, \quad t \geq 1. \tag{28}$$

In practice, one may assume a small transition probability between identity variables to increase the robustness.

Observation likelihood: In the simplest form, we assume that the transformed observation is a noise-corrupted version of some still template in the gallery, for example

$$\mathsf{z}_t = \mathcal{T}\{\mathsf{y}_t; \theta_t\} = \mathsf{h}_{\alpha_t} + \mathsf{v}_t, \quad t \geq 1, \tag{29}$$

where v_t is the *observation noise* at time t, whose distribution determines the likelihood $\mathrm{p}(\mathsf{y}_t|\alpha_t, \theta_t)$.

Particle filter for solving the model: We assume statistical independence between all noise variables and prior knowledge on the distributions $p(\theta_0)$ and $\mathrm{p}(\alpha_0)$ (uniform prior in fact). Given this model, our goal is to compute the posterior probability $\mathrm{p}(\alpha_t|\mathsf{y}_{1:t})$. It is in fact a probability mass function (PMF) because α_t only takes values from $\mathcal{N} = \{1, 2, \ldots, N\}$, as well as a marginal probability of $\mathrm{p}(\alpha_t, \theta_t|\mathsf{y}_{1:t})$, which is a mixed-type distribution. Therefore, the problem is reduced to computing the posterior probability.

Since the model is nonlinear and non-Gaussian in nature, there is no analytic solution. We invoke a particle filter [10, 20, 24, 34] to provide numeric approximations to the posterior distribution $\mathrm{p}(\alpha_t, \theta_t|\mathsf{y}_{1:t})$. Also, for this mixed-type

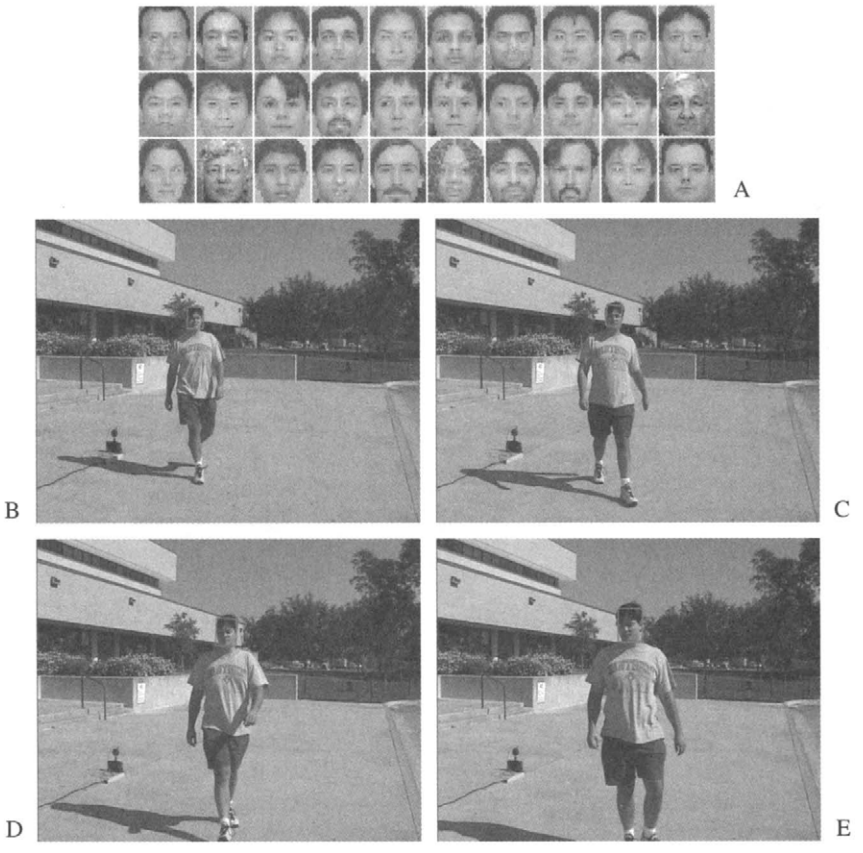

FIGURE 7 **A:** The face gallery with image size being 30 × 26. **B–E:** Four example frames in one probe video with image size being 720 × 480 while the actual face size ranges approximately from 20 × 20 in the first frame to 60 × 60 in the last frame. Notice the significant illumination variations between the probe and the gallery.

distribution, we can greatly improve the computational load by judiciously utilizing the discrete nature of the identity variable as in [64]. We [64] also theoretically justified the evolving behavior of the recognition density $p(\alpha_t|\mathbf{y}_{1:t})$ under a weak assumption.

2. *Experimental Results:* In the experiments presented in the following section, we use video sequences with subjects walking in a slant path toward the camera. There is 30 subjects, each having one face template. There is one face gallery and one probe set. The face gallery is shown in Fig. 7. The probe contains 30 video sequences, one for each subject. Figure 7 gives some example frames extracted from one probe video. As far as the imaging conditions are concerned, the gallery is very different from the probe, especially in lighting. This is similar to the "fc" test protocol of the FERET test [45]. These images/videos were collected, as part of the HumanID project, by researchers in National Institute of Standards and Technology and University of South Florida.

Case 1: Tracking and recognition using Laplacian density: We first investigate the performance under the following setting: We use an affine motion parameter, a time-invariant first-order Markov Gaussian motion transition model, and a "truncated" Laplacian observation likelihood as follows.

$$p_1(\mathbf{y}_t|\alpha_t, \theta_t) = \mathrm{LAP}(\|\mathcal{T}\{\mathbf{y}_t; \theta_t\} - \mathbf{h}_{\alpha_t}\|; \sigma_1, \tau_1) \tag{30}$$

where, $\|.\|$ is sum of absolute distance, σ_1 and λ_1 are manually specified, and

$$\mathrm{LAP}(x; \sigma, \tau) = \begin{cases} l\sigma^{-1}\exp(-x/\sigma) & \text{if } x \leq \tau\sigma \\ \sigma^{-1}\exp(-\tau) & \text{otherwise} \end{cases} \tag{31}$$

The recognition decision is based on (10). Table 3 shows that the recognition rate is very poor, only 13% of the time the top match is the correct match. The main reason is that the "truncated" Laplacian density is far from sufficient to capture the appearance difference between the probe and

TABLE 3 Algorithm performances for five cases

Case	Case 1	Case 2	Case 3	Case 4	Case 5
Tracking accuracy	83%	87%	93%	100%	NA
Recognition w/in top 1 match	13%	NA	83%	93%	57%
Recognition w/in top 3 matches	43%	NA	97%	100%	83%

NA, Not applicable.

the gallery, thereby indicating a need for improved appearance modeling. Nevertheless, the tracking accuracy[1] is reasonable with 83% successfully tracked because we are using multiple face templates in the gallery to track the specific face in the probe video. After all, faces in both the gallery and the probe sets belong to the same class of human face and it seems that the appearance change is within the class range.

Case 2: Pure tracking using Laplacian density: In Case 2, we measure the appearance change within the probe video as well as the noise in the background. To this end, we introduce a dummy template T_0, a cut version in the first frame of the video. Define the observation likelihood for tracking as

$$\mathrm{p}_2(\mathsf{y}_t|\theta_t) = \mathrm{LAP}(\|\mathcal{T}\{\mathsf{y}_t;\theta_t\} - \mathsf{T}_0\|;\sigma_2,\tau_2), \tag{32}$$

where σ_2 and τ_2 are set manually. The other setting, such as motion parameter and model, is the same as in Case 1. We still can run the particle filter algorithm to perform pure tracking.

Table 3 shows that 87% are successfully tracked by this simple tracking model, which implies that the appearance within the video remains similar. Figure 8A shows the posterior probability $\mathrm{p}(\alpha_t|\mathsf{y}_{1:t})$ for the video sequence in Fig. 7. Starting from uniform $\mathrm{p}(\alpha_0) = N^{-1}$, the posterior probability for the correct identity approaches one as time proceeds, and all others decrease to zero. This evolving behavior is characterized by the notion of entropy as shown in Fig. 8B. Also, the tracking results, inferred from $p(\theta_t|\mathsf{y}_{1:t})$, are illustrated in Fig. 7.

Case 3: Tracking and recognition using probabilistic subspace density: As mentioned in Case 1, we need a new appearance model to improve the recognition accuracy. We use the approach suggested by Moghaddam et al. [39] due to its computational efficiency and high recognition accuracy. However, in our implementation, we model only intrapersonal variations instead of both intra/extra-personal variations for simplicity.

[1]We inspect the tracking results by imposing the minimum mean squared density motion estimate on the final frame as shown in Figure 7 and determine if tracking is successful or not for this sequence. This is done for all sequence and tracking accuracy is defined as the ratio of the number of sequences successfully tracked to the total of all sequence.

We need at least two facial images for one identity to construct the intrapersonal space (IPS). Apart from the available gallery, we crop out the second image from the video ensuring no overlap with the frames actually used in probe videos. Figure 9A shows a list of such images. Compare with Fig. 7 to see how the illumination varies between the gallery and the probe.

We then fit a probabilistic subspace density [39] on top of the IPS. It proceeds as follows: A regular PCA is performed for the IPS. Suppose the eigensystem for the IPS is $\{(\lambda_i, \mathsf{e}_i)\}_{i=1}^{d}$, where d is the number of pixels and $\lambda_1 \geq \ldots \geq \lambda_d$. Only top s principal components corresponding to the top s eigenvalues are then kept while the residual components are considered as isotropic. We refer the reader to the original paper [39] for full details. Figure 9(B) show the eigenvectors for the IPS. The density is written as follows:

$$\mathrm{Q}_{IPS}(\mathsf{x}) = \left\{\frac{exp(-1/2\sum_{i=1}^{s} y_i^2/\lambda_i)}{(2\pi)^{s/2}\prod_{i=1}^{s}\lambda_i^{1/2}}\right\}\left\{\frac{exp(-\epsilon^2/2\rho)}{(2\pi\rho)^{(d-s)/2}}\right\}, \tag{33}$$

where $y_i = \mathsf{e}_i^T\mathsf{x}$ for $i = 1, 2, \ldots, s$ is the ith principal component of x, $\epsilon^2 = \|\mathsf{x}\|^2 - \sum_{i=1}^{s} y_i^2$ is the reconstruction error, and $\rho = (\sum_{i=s+1}^{d}\lambda_i)/(d-q)$. It is easy to write the likelihood as follows:

$$\mathrm{p}_3(\mathsf{y}_t|\alpha_t,\theta_t) = \mathrm{Q}_{IPS}(\mathcal{T}\{\mathsf{y}_t;\theta_t\} - \mathsf{h}_{\alpha_t}). \tag{43}$$

Table 3 lists the performance by using this new likelihood measurement. It turns out that the performance is significantly better that in Case 1, with 93% tracked successfully and 83% recognized as the top 1 match. If we consider the top three matches, 97% are correctly identified.

Case 4: Tracking and recognition using combined density: In Case 2, we have studied appearance changes within a video sequence. In Case 3, we have studied the appearance change between the gallery and the probe. In Case 4, we attempt to take advantage of both cases by introducing a combined likelihood defined as follows:

$$\mathrm{p}_4(\mathsf{y}_t|\alpha_t,\theta_t) = \mathrm{p}_3(\mathsf{y}_t|\alpha_t,\theta_t)\mathrm{p}_2(\mathsf{y}_t|\theta_t) \tag{35}$$

Again, all other setting is as in Case 1. We now obtain the best performance so far: no tracking error, 93% are correctly recognized as the first match, and no error in recognition is seen when the top three matches are considered.

Case 5: Still-to-still face recognition: To make a comparison, we also performed an experiment on still-to-still face recognition. We selected the probe video frames with the best frontal face view (i.e., biggest frontal view) and cropped out the facial region by normalizing with respect to the eye coordinates manually specified. This collection of images is shown in Fig. 9C and it is fed as probes into a still-to-still

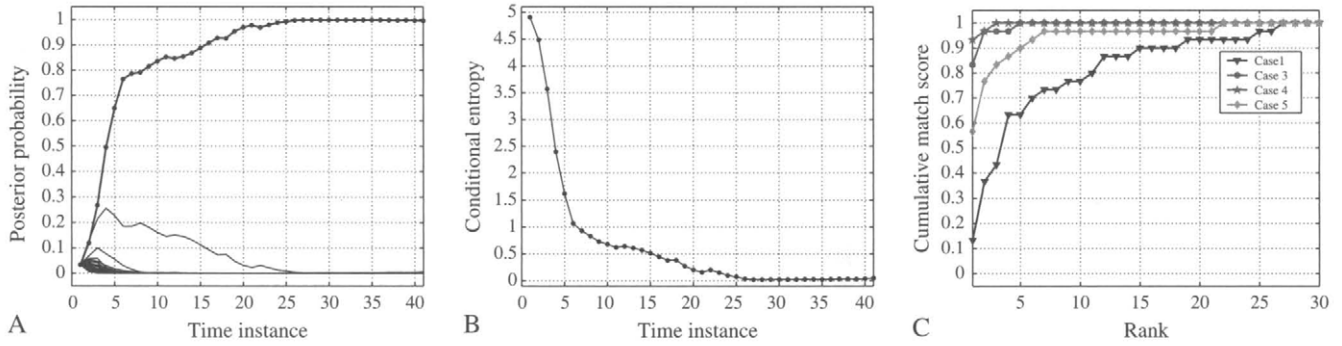

FIGURE 8 **A:** Posterior probability $p(\alpha_t|\mathbf{y}_{1:t})$. **B:** Entropy of posterior probability $p(\alpha_t|\mathbf{y}_{1:t})$. **C:** Cumulative match curves for the dataset.

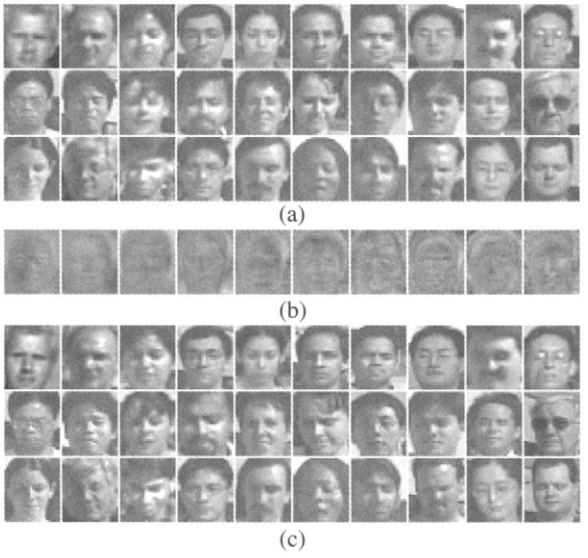

FIGURE 9 **A:** The second facial images for training probabilistic density. **B:** The top-ten eigenvectors for the intrapersonal space. **C:** The facial images cropped out from the largest frontal view.

face recognition system with the learned probabilistic subspace as in Case 3. It turns out that the recognition result is 57% correct for the top-one match, and 83% for the top three matches. The cumulative match curves for Case 1 and Cases 3–5 are presented in Fig. 8C. Clearly, Case 4 is the best. We also implemented the original algorithm by Moghaddam et al. [39] (i.e., both intra-extrapersonal variations are considered, the recognition rate is similar to that obtained in Case 5).

4 Conclusions

The chapter presented a hierarchic framework for face pattern and face recognition theory. Current face recognition approaches are classified according to their placements in this framework. We then proposed a unified framework in which many approaches can be cast. Two instances were given to as examples of the framework: one on recognition from groups of still images under variations and the other on recognition from video sequences within which tracking and recognition interacts.

Acknowledgment

Partially supported by National Science Foundation grant 03-25119.

References

[1] M. S. Barlett, H. M. Ladesand, and T. J. Sejnowski, "Independent component representations for face recognition," *Proc. SPIE 3299*, 528–539 (1998).

[2] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Patt. Anal. Mach. Intell.* 25, 218–233 (2003).

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Patt. Analy. Mach. Intell.* 19, 711–720 (1997).

[4] M. J. Black and Y. Yacoob, "Recognizing facial expressions in image sequences using local paramterized models of image motion," *Int. J. Comput. Vis.* 25, 23–48 (1997).

[5] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Patt. Anal. Mach. Intell.* 25, 1063–1074 (2003).

[6] V. Bruce, *Recognizing Faces.* (Lawrence Erlbaum Associates, London, 1988).

[7] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proc. IEEE*, 83, 705–740 (1995).

[8] T. Cootes, K. Walker, and C. Taylor, "View-based Active appearance models," *Proceedings of International Conference on Automatic Face and Gesture Recognition* (Grenoble, France, 2000).

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* (Wiley, New York, 1991).

[10] A. Doucet, N. d. Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice.* (Springer-Verlag, New York, 2001).

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* (Wiley-Interscience, New York, 2001).

[12] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *J. Opt. Soc. Am. A* 1724–1733 (1997).

[13] A. Fitzgibbon and A. Zisserman, "Joint manifold distance: a new approach to appearance based clustering," *Proc. IEEE Computer Socienty Conference on Computer Vision and Pattern Recognition* (Madison, WI, 2003).

[14] R. T. Frankot and R. Chellappa, "A method for enforcing integrability in shape from shaging problem," *IEEE Trans. Patt. Anal. Mach. Intell.* 10, 439–451 (1988).

[15] W. T. Freeman and J. B. Tenenbaum, "Learning bilinear models for two-factor problems in vision," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Puerto Rico, 1997).

[16] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Patt. Anal. Mach. Intell.* 23, 643–660 (2001).

[17] R. Gross, I. Matthews, and S. Baker, "Eigen light-fields and face recognition across pose," *Proc. Intenational Conference on Automatic Face and Gesture Recognition* (Washington, DC, 2002).

[18] R. Gross, I. Matthews, and S. Baker, "Fisher light-fields for face recognition across pose and illumination," *Proc. German Symposium on Pattern Recognition* (2002).

[19] R. Hietmeyer, "Biometric identification promises fast and secure processings of airline passengers," *Int. Civil Aviation Org. J.* 55, 10–11 (2000).

[20] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," *European Conference on Computer Vision*, (Cambridge, U.K., 1996), 343–356.

[21] T. Kanade, *Computer Recognition of Human Faces* (Birhauser, Basel, Switzerland, and Stuttgart, Germany, 1973).

[22] M. D. Kelly, "Visual identification of people by computer," *Tech. rep. AI-130* (Stanford AI project, Stanford, CA, 1970).

[23] M. Kirby and L. Sirovich, "Application of Karhunen-Loéve procedure of the characterization of human faces," *IEEE Trans. Patt. Anal. Mach. Intell.* 12, 103–108 (1990).

[24] G. Kitagawa, "Monte carlo filter and smoother for non-gaussian nonlinear state space models," *J. Computat. Graph. Stat.* 5, 1–25 (1996).

[25] B. Knight and A. Johnston, "The role of movement in face recognition," *Visual Cognition*, 4, 265–274 (1997).

[26] V. Krueger and S. Zhou, "Exemplar-based face recognition from video," *European Conference on Computer Vision* (Copenhagen, Denmark, 2002).

[27] M. Lades, J. C. Vorbruggen, J. Buhmann, *et al.*, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Comput.* 42, 300–311 (1993).

[28] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging affects on face images," *IEEE Trans. Patt. Anal. Mach. Intell.* 24, 242–455 (2002).

[29] K. Lee, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Madison, WI, 2003).

[30] B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans. Image Process.* 11, 530–554 (2002).

[31] S. Z. Li, A. K. Jain, eds, *Handbook of Face Recognition* (Springer-Verlag, New York, 2004).

[32] Y. Li, S. Gong, and H. Liddell, "Modelling faces dynamically across views and over time," *Proc. Int. Conf. Comput. Vis.* 554–559 (Vancouver, Canada, 2001).

[33] S. H. Lin, S. Y. Kung, and J. J. Lin, "Face recognition/detection by probabilistic decision based neural network," *IEEE Trans. Neural Networks*, 9, 114–132 (1997).

[34] J. S. Liu and R. Chen, "Sequential Monte Carlo for dynamic systems," *J. Am. Statist. Assoc.* 93, 1031–1041 (1998).

[35] C. Liu and H. Wechsler, "Evolutionary pursuit and its applications to face recognition," *IEEE Trans. Patt. Analy. Mach. Intell.* 22, 570–582 (2000).

[36] X. Liu and T. Chen, "Video-based face recognition using adaptive hidden Markov models," *Proc. IEEE Comput. Soc. Conference on Computer Vision and Pattern Recognition* (Madison, WI, 2003).

[37] M. J. Lyons, J. Biudynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Patt. Anal. Mach. Intell.* 21, 1357–1362 (1999).

[38] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Trans. Patt. Anal. Mach. Intell.* PAMI-19, 696–710 (1997).

[39] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. Patt. Anal. Mach. Intell.* 24, 780–788 (2002).

[40] A. J. O'Toole, "Psychological and neural perspectives on human faces recognition," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain (eds.), (Springer, New York, 2004).

[41] P. Penev and J. Atick, "Local feature analysis: A general statistical theory for object representation," *Networks: Comput. Neural Syst.* 7, 477–500, (1996).

[42] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Seattle, WA, 1994).

[43] P. J. Phillips, R. M. McCabe, and R. Chellappa, "Biometric image processing and recognition," *Proc. European Signal Processing Conference* (Rhodes, Greece, 1998).

[44] P. J. Phillips, "Support vector machines applied to face recognition," *Adv. Neural Inf. Process. Syst.* 11, 803–809 (1998).

[45] P. J. Phillips, H. Moon, S. Rizvi, *et al.*, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Patt. Anal. Mach. Intell.* 22, 1090–1104 (2000).

[46] P. J. Phillips, P. Grother, R. J. Micheals, *et al.*, "Face recognition vendor test 2002: evaluation report," *NISTIR 6965*, http://www.frvt.org (2003).

[47] C. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer, New York, 1999).

[48] G. Shakhnarovich, J. Fisher, and T. Darrell, "Face recognition from long-term observations," *European Conference on Computer Vision*, (Copenhagen, Denmark, 2002).

[49] A. Shashua, "On photometric issues in 3d visual recognition from a single 2d image," *Int. J. Comput. Vis.* 21, 99–122 (1997).

[50] A. Shashua and T. R. Raviv, "The quotient image: Class based re-rendering and recognition with varying illuminations," *IEEE Trans. Patt. Anal. Mach. Intell.* 23, 129–139 (2001).

[51] I. Shimshoni, Y. Moses, and M. Lindenbaum., "Shape reconstruction of 3D bilaterally symmetric surfaces," *Int. J. Comput. Vis.* 39, 97–100 (2000).

[52] T. Sim, S. Baker, and M. Bast, "The CMU pose, illumination, and expression (PIE) database," *Proc. Automatic Face and Gesture Recognition* (Washington, D.C., 2002), 53–58.

[53] Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units of facial expression analysis," *IEEE Trans. Patt. Anal. Mach. Intell.* 23, 1–19 (2001).

[54] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cog. Neurosci.* 3, 72–86 (1991).

[55] M. Vasilescu and D. Terzopoulos, "Multilinear image analysis for facial recognition," *Proc. Int. Conference on Pattern Recognition* (Quebec City, Canada, 2002).

[56] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE Trans. Patt. Anal. Mach. Intell.* 11, 733–742 (1997).

[57] L. Wolf and A. Shashua, "Kernel principal angles for classification machines with applications to image sequence interpretation," *IEEE Comput. Soc. Conference on Computer Vision and Pattern Recognition* (Madison, WI, 2003).

[58] C. Yang, R. Duraiswami, A. Elgammal, et al., "Real-time kernel-based tracking in joint feature-spatial spaces," *Tech. Report CS-TR-4567* (Univ. of Maryland, 2004).

[59] A. L. Yuille, D. Snow, E. R., et al., "Determining generative models of objects under varying illumination: shape and albedo from multiple images using svd and integrability," *Int. J. Comput. Vis.* 35, 203–222 (1999).

[60] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," *Proc. International Conference on Automatic Face and Gesture Recognition*, (Nara, Japan, 1998), 361–341.

[61] W. Zhao and R. Chellappa, "Symmetric shape from shading using self-ratio image," *Int. J. Comput. Vis.* 45, 55–752 (2001).

[62] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, 12, (2003).

[63] S. Zhou and R. Chellappa, "Probabilistic human Recognition from video," *Eur. Conf. Comput. Vis.* 3, 681–697 (2002).

[64] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Comput. Vis. Image Understand.* 91, 214–245 (2003).

[65] S. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Trans. Image Process. (to appear)* (2004).

[66] S. Zhou, R. Chellappa, and D. Jacobs, "Characterization of human faces under illumination variations using rank, integrability, and symmetry constraints," *European Conference on Computer Vision* (Prague, Czech, 2004).

[67] S. Zhou and R. Chellappa, "Illuminating light field: Image-based face recognition across illuminations and poses," *Proc. International Conference on Automatic Face and Gesture Recognition* (Seoul, Korea, 2004).

[68] S. Zhou and R. Chellappa, "Probabilistic identity characterization for face recognition," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC, June 2004).

[69] S. Zhou and R. Chellappa, "Multiple-exemplar discriminant analysis for face recognition," *Proc. International Conference on Pattern Recognition* (Cambridge, UK, August 2004).