CAR-TR-976
CS-TR-4359
UMIACS-TR-2002-39

April 2002
N00014-00-1-0908

Probabilistic Recognition of Human Faces from Video

Shaohua Zhou, Volker Krüger, and Rama Chellappa

Center for Automation Research
University of Maryland
College Park, MD 20742-3275

Abstract

Recognition of human faces using a gallery of still or video images and a probe set of videos is systematically investigated using a probabilistic framework. In *still-to-video* recognition, where the gallery consists of still images, a time series state space model is proposed to fuse temporal information in a probe video, which simultaneously characterizes the kinematics and identity using a *motion vector* and an *identity variable*, respectively. The joint posterior distribution of the motion vector and the identity variable is estimated at each time instant and then propagated to the next time instant. *Marginalization* over the motion vector yields a robust estimate of the posterior distribution of the identity variable. A computationally efficient *sequential importance sampling* algorithm is developed to provide a numerical solution to the model. Theoretical derivations under weak assumptions demonstrate that, due to the propagation of the identity variable over time, a *degeneracy* in the posterior probability of the identity variable is exploited to give improved recognition.

The gallery is generalized to videos in order to realize *video-to-video* recognition. An *exemplar-based learning* strategy is adopted to automatically select video representatives from the gallery, serving as mixture centers in an updated likelihood measure. The SIS algorithm is applied to approximate the posterior distribution of the motion vector, the identity variable, and the exemplar index, and the marginal distribution of the identity variable produces the recognition result. The model formulation is very general and allows a variety of image representations and transformations.

Experimental results using videos collected by NIST/USF and CMU illustrate the effectiveness of this approach in both still-to-video and video-to-video scenarios with appropriate model choices.

Key Words: Face recognition (FR), still-to-video, video-to-video, time series state space model, sequential importance sampling (SIS), exemplar-based learning (EBL).

[0]This work was done with the support of the DARPA HumanID Project (Grant N00014-00-1-0908). Zhou and Chellappa are affiliated with the Center for Automation Research (CfAR) and the Department of Electrical and Computer Engineering. Krüger is affiliated with CfAR.

1. Introduction

Probabilistic video analysis has recently received significant attention in the computer vision community since the seminal work of Isard and Blake [4]. In their effort to solve the problem of visual tracking, they introduced a time series state space model parameterized by a tracking motion vector (e.g. affine transformation parameters), denoted by θ_t. The CONDENSATION algorithm was developed to provide a numerical approximation to the posterior distribution of the motion vector at time t given the observations up to t, i.e., $\pi(\theta_t) = p(\theta_t|z_{0:t})$ where $z_{0:t} = (z_0, z_1, \ldots, z_t)$ and z_t is the observation at time t, and to propagate it over time according to the kinematics. This algorithm has been extended to many areas [5, 15, 23], including human face recognition (FR) [14]. In this paper, we will systematically investigate the incorporation of temporal information in a video sequence for the FR task.

FR has been an extensive research area for a long time. Refer to [1, 2] for surveys and [3] for reports on experiments. The experiments reported in [3] evaluate *still-to-still* scenarios, where the gallery and the probe set both consist of still facial images. Some well-known still-to-still FR approaches include Principal Component Analysis (PCA) [9], Linear Discriminant Analysis (LDA) [10, 11], and Elastic Graph Matching (EGM) [12]. Typically, recognition is performed based on an abstract representation of the face image after suitable geometric and photometric registrations.

Following [3], we define a *still-to-video* scenario: the gallery consists of still facial templates and the probe set consists of video sequences containing the facial region. Denote the gallery by $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, indexed by the identity variable n, which lies in a finite sample space $\mathcal{N} = \{1, 2, \ldots, N\}$. Though significant research has been conducted on still-to-still recognition, research efforts on still-to-video recognition are less common due to the following challenges [2] in typical surveillance applications: poor video quality, significant illumination and pose variations, and low image resolution. Most video-based recognition systems [13] do the following: the face is first detected and then tracked over time. When a frame satisfying certain criteria (size, pose) is acquired, recognition is performed using a still-to-still recognition technique. To do this, the face part is cropped from the frame and transformed or registered using appropriate transformations.

There are several unresolved issues in the above approach: criteria for selecting good frames and estimation of parameters for registration. Also, still-to-still recognition does not effectively exploit temporal information. A common strategy, which selects several good frames, performs recognition on each frame and then votes on these recognition results for a final solution, is rather *ad hoc*.

Many of these difficulties can be overcome by the proposed probabilistic framework. To fuse temporal information, a time series state space model is adopted to characterize the evolving kinematics and identity in the probe video. Three basic components of the model are:

- a *motion equation* governing the kinematic behavior of the tracking motion vector,
- an *identity equation* governing the temporal evolution of the identity variable,
- an *observation equation* depicting the observing behavior by establishing a link between the motion vector and the identity variable.

Using the sequential importance sampling (SIS) [4, 6, 7, 8] technique, the joint posterior distribution of the motion vector and the identity variable, i.e., $\pi(n_t, \theta_t) = p(n_t, \theta_t|z_{0:t})$, is estimated at each time instant and then propagated to the next time instant as governed by motion and identity equations. The marginal distribution of the identity variable, $\pi(n_t) = p(n_t|z_{0:t})$, is estimated to provide a recognition result. An SIS algorithm is developed to approximate the distribution $\pi(n_t)$ in the still-to-video scenario. It achieves computational efficiency over the CONDENSATION algorithm by considering the discrete nature of the identity variable.

The still templates in the gallery can be generalized to video sequences in order to realize *video-to-video* recognition. In video-to-video recognition, exemplars and their prior probabilities are learned from the gallery videos to serve as still templates in the still-to-video scenario. Person n may have a collection

of several exemplars, say $\mathcal{C}^n = \{c_1^n, c_2^n, \ldots, c_{K_n}^n\}$ indexed by k. The likelihood is then modeled as a mixture density with exemplars as mixture centers. We first compute the joint distribution $\pi(n_t, k_t, \theta_t)$ using the SIS algorithm and marginalize it to yield $\pi(n_t)$. In the experiments reported here, the subject walks on a treadmill with his/her head moving naturally, giving rise to significant variations across poses. The proposed method successfully copes with these pose variations, as evidenced by the experimental results.

It is worth emphasizing that (i) our model can take advantage of any still-to-still recognition algorithm [9, 10, 11, 12] by embedding any distance measure used by such an algorithm into our likelihood measurement; and (ii) it allows a variety of image representations and transformations.

The organization of this paper is as follows: Section 2 reviews some related studies on video-based FR. Section 3 introduces the time series state space model for recognition and establishes the time-evolving behavior of $\pi(n_t)$ under some weak assumptions. Section 4 briefly reviews the SIS principle from the viewpoint of a general state space model and develops an SIS algorithm to solve the still-to-video recognition problem, with special emphasis on its computational efficiency. Section 5 describes the experimental scenarios for still-to-video recognition and presents results using the HID data collected by NIST/USF and the CMU MoBo database. In Section 6, Subsection 6.1 presents the exemplar-based learning algorithm, and Subsection 6.2 describes some modifications in the model and an algorithm that accommodates video-to-video recognition. Experimental results for the video-to-video recognition problem using CMU data are presented in Subsection 6.3. Section 7 concludes the paper.

2. Related Literature

Nearly all video-based recognition systems apply still-image-based recognition to selected good frames. The face images are warped into frontal views whenever pose and depth information about the faces is available [13].

In [27, 18, 26], RBF (Radial Basis Function) networks are used for tracking and recognition purposes. The system in [27] uses an RBF network for recognition. Since no warping is done, the RBF network has to learn the individual variations as well as possible transformations. The performance appears to vary widely, depending on the size of the training set. In [18], face tracking is based on an RBF network to provide feedback to a motion clustering process. Good tracking results were demonstrated, but person authentication results were referred to as future work. [26] presents a fully automatic person authentication system that uses video break, face detection, and authentication modules and cycles over successive video images until a high recognition confidence is reached. During operation, the face is tracked, face images are normalized and then used for authentication with an RBF network. This system was tested on three image sequences; the first was taken indoors with one subject, the second was taken outdoors with two subjects, and the third was taken outdoors with one subject in stormy conditions. Perfect results were reported on all three sequences, when verified against a database of 20 still face images.

In [19] a system called *PersonSpotter* is described. This system is able to capture, track and recognize a person walking toward or passing a stereo CCD camera. It has several modules, including a head tracker and a landmark finder. The landmark finder uses a dense graph consisting of 48 nodes learned from 25 example images to find landmarks such as eyes and nose tip. An elastic graph matching scheme is employed to identify the face.

A multimodal-based person recognition system is described in [13]. This system consists of a face recognition module, a speaker identification module, and a classifier fusion module. The most reliable video frames and audio clips are selected for recognition. 3D information about the head is used to detect the presence of an actual person as opposed to an image of that person. Recognition and verification rates of 100% were achieved for 26 registered clients.

In [16], recognition of faces over time is implemented by constructing a face identity surface. The face is first warped to a frontal view, and its Kernel Discriminant Analysis (KDA) features over time

form a trajectory. It is shown that the trajectory distances accumulate recognitive evidence over time.

In [14] a generic approach to simultaneous object tracking and verification is proposed. The approach is based on posterior probability density estimation using sequential Monte Carlo methods [4, 6, 7, 8]. Tracking is formulated as a probability density propagation problem and the algorithm also provides verification results. However, there was no systematic evaluation of recognition performance. There are significant differences between our approach and the algorithm described in [14]; we highlight these differences in Section 7.

Black and Jepson [15] used a CONDENSATION-based algorithm to match temporal trajectories. Models of temporal trajectories, arising from gestures and facial expressions, are trained beforehand and matched against human motions in a new image sequence. The joint posterior distribution of model selection, local stretching, scaling, and position evolves as time proceeds.

3. A Model for Recognition in Video

In this section we present details on how we establish the propagation model for recognition and discuss its impact on the posterior distribution of the identity variable under some weak assumptions.

3.1. A Time Series State Space Model for Recognition

We use the following mathematical notation:

- An image I. It could be represented using raw intensity values $I(R)$ in the image region R, or abstract features extracted from $I(R)$, such as PCA or LDA features.
- A transformed version of image I, $\mathcal{T}_\theta\{I\}$, with *transformation vector* $\theta \in \Theta$, a continuous sample space. The transformation can be either geometric, photometric, or both. If I is represented by abstract features, so is $\mathcal{T}_\theta\{I\}$.
- A gallery $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, indexed by an *identity variable* n lying in a finite sample space $\mathcal{N} = \{1, 2, \ldots N\}$. I_n is the still template for the identity n, which may or may not be registered beforehand. Here it is assumed that each identity has only one representative still facial image.
- θ_t is the *transformation vector* or the *motion vector*, n_t the identity variable, and z_t the *observation*, respectively, at time instant t. $x_t = (n_t, \theta_t)$ is the *overall state vector*. $z_{0:t} = (z_0, z_1, \ldots, z_t)$ is a collection of observations up to time t.

A time series state space model for recognition can now be described as follows:

1. Motion equation:
$$\theta_t = g(\theta_{t-1}, u_t); \quad t \geq 1, \tag{1}$$
where u_t is *noise* in the motion model and $g(.,.)$ characterizes the evolving motion, both as functions of t.

2. Identity equation:
$$n_t = n_{t-1}; \quad t \geq 1, \tag{2}$$
assuming that the identity does not change over time.

3. Observation equation:
$$\mathcal{T}_{\theta_t}\{z_t\} = I_{n_t} + v_t; \quad t \geq 1, \tag{3}$$
where v_t is *observation noise* at time instant t, and assuming that the transformed observation is a noise-corrupted version of some still template in the gallery.

4. Prior distributions:
$$p(\theta_0|z_0), p(n_0|z_0). \tag{4}$$

5. Noise distributions:
$$p(u_t), p(v_t); \quad t \geq 1. \tag{5}$$
The noise distribution $p(u_t)$ determines the motion state transition probability $p(\theta_t|\theta_{t-1})$, and $p(v_t)$ determines the observation likelihood $p(z_t|n_t, \theta_t)$.

6. Statistical independence:
$$\begin{aligned} n_0 \perp \theta_0, u_t \perp v_s; \quad t, s \geq 1 \\ u_t \perp u_s, v_t \perp v_s; \quad t, s \geq 1 \ \& \ t \neq s \end{aligned} \tag{6}$$
where $\perp$ indicates statistical independence.

Using the overall state vector x_t, Eqs. (1) and (2) can be combined into one state equation (in a normal sense) which is completely described by the overall state transition probability
$$p(x_t|x_{t-1}) = p(n_t|n_{t-1})p(\theta_t|\theta_{t-1}) \ . \tag{7}$$

Given this model, our goal is to compute the posterior probability $\pi(n_t)$. It is in fact a probability mass function (PMF), as well as a marginal probability of $\pi(n_t, \theta_t)$, which is a mixture distribution. Therefore, the problem is reduced to computing the posterior probability.

3.2. The Posterior Probability of the Identity Variable

It is very interesting to study the evolution of the posterior probability $\pi(n_t)$ as time proceeds since the identity variable does not change by assumption, i.e., $p(n_t|n_{t-1}) = \delta(n_t - n_{t-1})$, where $\delta(.)$ is a discrete impulse function at the origin.

Using time recursion, the Markov properties, and the statistical independence embedded in the model, we can easily derive
$$\begin{aligned} p(n_{0:t}, \theta_{0:t}|z_{0:t}) &= p(n_{0:t-1}, \theta_{0:t-1}|z_{0:t-1}) \frac{p(z_t|n_t, \theta_t)p(n_t|n_{t-1})p(\theta_t|\theta_{t-1})}{p(z_t|z_{0:t-1})} \\ &= p(n_0, \theta_0|z_0) \prod_{s=1}^{t} \frac{p(z_s|n_s, \theta_s)p(n_s|n_{s-1})p(\theta_s|\theta_{s-1})}{p(z_s|z_{0:s-1})} \\ &= p(n_0|z_0)p(\theta_0|z_0) \prod_{s=1}^{t} \frac{p(z_s|n_s, \theta_s)\delta(n_s - n_{s-1})p(\theta_s|\theta_{s-1})}{p(z_s|z_{0:s-1})}. \end{aligned} \tag{8}$$

Therefore, by marginalizing over $\theta_{0:t}$ and $n_{0:t-1}$, we get
$$p(n_t = l|z_{0:t}) = p(l|z_0) \int_{\theta_0} \cdots \int_{\theta_t} p(\theta_0|z_0) \prod_{s=1}^{t} \frac{p(z_s|l, \theta_s)p(\theta_s|\theta_{s-1})}{p(z_s|z_{0:s-1})} d\theta_t \ldots d\theta_0. \tag{9}$$

Thus $\pi(n_t = l)$ is determined by the prior distribution $p(n_0 = l|z_0)$ and the product of the likelihood functions, $\prod_{s=1}^{t} p(z_s|l, \theta_s)$. If a uniform prior is assumed, then $\prod_{s=1}^{t} p(z_s|l, \theta_s)$ is the only determining factor.

Suppose that the following two conditions hold:

- (A) The prior probability for each identity is the same,
$$p(n_0|z_0) = 1/N; \quad n_0 \in \mathcal{N}, \tag{10}$$

- (B) For the correct identity $l \in \mathcal{N}$, there exists a constant $\eta > 1$ such that

$$p(z_t|l, \theta_t) \geq \eta p(z_t|n_t, \theta_t); \quad t \geq 1, n_t \in \mathcal{N}, n_t \neq l. \tag{11}$$

A trivial choice of η is the lower bound on the likelihood ratio, i.e.

$$\eta = \inf_{t \geq 1, n_t \neq l, \theta_t \in \Theta} \frac{p(z_t|l, \theta_t)}{p(z_t|n_t, \theta_t)}. \tag{12}$$

Section 7 discusses the empirical choice of η.

Substitution of Eqs. (10) and (11) into Eq. (9) gives

$$p(l|z_{0:t}) \geq \eta^t p(n_t|z_{0:t}); \quad n_t \in \mathcal{N}, n_t \neq l, \tag{13}$$

where $\eta^t = \prod_{s=1}^{t} \eta$. Since $\sum_{n_t=1}^{N} p(n_t|z_{0:t}) = 1$, it is easy to see that for the correct identity,

$$p(l|z_{0:t}) \geq p(l|z_{0:t-1}). \tag{14}$$

In other words, the posterior probability of the correct hypothesis is non-decreasing as time proceeds.

More interestingly, from Eq. (13) we have

$$(N-1)p(l|z_{0:t}) \geq \eta^t \sum_{n_t=1, n_t \neq l}^{N} p(n_t|z_{0:t}) = \eta^t (1 - p(l|z_{0:t})), \tag{15}$$

i.e.

$$p(l|z_{0:t}) \geq \frac{\eta^t}{\eta^t + N - 1}. \tag{16}$$

Since $\eta > 1$ and $p(l|z_{0:t}) \leq 1$,

$$\lim_{t \to \infty} p(l|z_{0:t}) = 1, \tag{17}$$

implying that $\pi(n_t)$ degenerates to the identity l for some sufficiently large t.

However, all these derivations are based on conditions (A) and (B). Though it is easy to satisfy condition (A), difficulty arises in practice in satisfying condition (B) for all the frames in the sequence. Fortunately, as we will see in the experiment in Section 5, numerically this degeneracy is still reached even if condition (B) is satisfied only for most but not all frames in the sequence. This issue is also addressed in Section 7.

To measure the evolving uncertainty remaining in the identity variable as observations accumulate, we use the notion of entropy [17]. Given a PMF $p(x); x \in \mathcal{N}$, the entropy is defined as

$$H(x) = -\sum_{x \in \mathcal{N}} p(x) \log_2 p(x). \tag{18}$$

Entropy essentially measures the average uncertainty about the random variable x with PMF $p(x)$. It is well known that among all distributions taking values on $\{1, \ldots, N\}$, the uniform distribution yields a maximum $\log_2 N$ and the degenerate case yields the minimum 0, i.e., $0 \leq H \leq \log_2 N$. Similarly, conditional entropy is defined as

$$H(x|y) = -\sum_{y} p(y) \sum_{x} p(x|y) \log_2 p(x|y). \tag{19}$$

In the context of this problem, conditional entropy $H(n_t|z_{0:t})$ captures the evolving uncertainty of the identity variable given observations $z_{0:t}$. However, knowledge of $p(z_{0:t})$ is needed to compute $H(n_t|z_{0:t})$. We assume it degenerates to an impulse around the actual observations $\tilde{z}_{0:t}$ since we observe only this particular sequence, i.e., $p(z_{0:t}) = \delta(z_{0:t} - \tilde{z}_{0:t})$. Now

$$H(n_t|z_{0:t}) = -\sum_{n_t \in \mathcal{N}} p(n_t|\tilde{z}_{0:t}) \log_2 p(n_t|\tilde{z}_{0:t}). \tag{20}$$

Under the conditions mentioned above, we expect $H(n_t|z_{0:t})$ to decrease as time proceeds.

4. Sequential Importance Sampling Algorithm

Consider a general time series state space model fully determined by (i) the overall state transition probability $p(x_t|x_{t-1})$, (ii) the observation likelihood $p(z_t|x_t)$, and (iii) the prior probability $p(x_0)$ and statistical independence: $u_t \perp u_s, v_t \perp v_s; t \neq s$. We wish to compute the posterior probability $\pi(x_t) = p(x_t|z_{0:t})$.

If the model is linear with Gaussian noise, it is analytically solvable by a Kalman filter which essentially propagates the mean and variance of a Gaussian distribution over time. For nonlinear and non-Gaussian cases, an extended Kalman filter (EKF) and its variants have been proposed to arrive at an approximate analytic solution [29]. Recently, the SIS technique, a special case of the Monte Carlo method [4, 8, 6, 7], has been used to provide a numerical solution and to propagate an arbitrary distribution over time.

4.1. Importance Sampling

The essence of the Monte Carlo method is to represent an arbitrary probability distribution $\pi(x)$ closely by a set of discrete samples. It would be ideal to draw i.i.d. samples $\{x^{(m)}\}_{m=1}^{M}$ from $\pi(x)$. However, this is often difficult to implement, especially for non-trivial distributions. Instead, a set of samples $\{x^{(m)}\}_{m=1}^{M}$ is drawn from an *importance function* $g(x)$ which is easy to sample from, and a weight

$$w^{(m)} = \pi(x^{(m)})/g(x^{(m)}) \tag{21}$$

is then assigned to each sample. This technique is called *Importance Sampling* (IS). It can be shown [7] that the *importance sample set* $\mathcal{S} = \{(x^{(m)}, w^{(m)})\}_{m=1}^{M}$ is *properly weighted* to the target distribution $\pi(x)$. To accommodate a video, importance sampling is used in a sequential fashion, which leads to SIS. SIS propagates $\mathcal{S}_{t-1}$ according to the *sequential importance function*, say $g(x_t|x_{t-1})$, and calculates the weight using

$$w_t = w_{t-1}p(z_t|x_t)p(x_t|x_{t-1})/g(x_t|x_{t-1}). \tag{22}$$

In the CONDENSATION algorithm [4], $g(x_t|x_{t-1})$ is taken to be $p(x_t|x_{t-1})$ and Eq. (22) becomes

$$w_t = w_{t-1}p(z_t|x_t), \tag{23}$$

In fact, Eq. (23) is implemented by first resampling the sample set $\mathcal{S}_{t-1}$ according to w_{t-1} and then updating the weight w_t using $p(z_t|x_t)$. For a complete description of the SIS method, refer to [7, 8].

The following two propositions are useful for guiding the development of the SIS algorithm.

PROPOSITION 1. *When $\pi(x)$ is a PMF defined on a finite sample space, the proper sample set should consist of all samples in the sample space.*

PROPOSITION 2. *If a set of weighted random samples $\{(x^{(m)}, y^{(m)}, w^{(m)})\}_{m=1}^{M}$ is proper with respect to $\pi(x, y)$, then a new set of weighted random samples $\{(y'^{(k)}, w'^{(k)})\}_{k=1}^{K}$, which is proper with respect to $\pi(y)$, the marginal of $\pi(x, y)$, can be constructed as follows:*
1) Remove repetitive samples from $\{y^{(m)}\}_{m=1}^{M}$ to obtain $\{y'^{(k)}\}_{k=1}^{K}$, where all $y'^{(k)}$'s are distinct;
2) Sum the weights $w^{(m)}$ belonging to the same sample $y'^{(k)}$ to obtain the weight $w'^{(k)}$, i.e.,

$$w'^{(k)} = \sum_{m=1, y^{(m)}=y'^{(k)}}^{M} w^{(m)}. \tag{24}$$

4.2. Algorithms and Computational Efficiency

In this framework, the posterior probability $\pi(n_t, \theta_t)$ is represented by a set of *indexed and weighted* samples

$$\mathcal{S}_t = \{(n_t^{(m)}, \theta_t^{(m)}, w_t^{(m)})\}_{m=1}^{M} \tag{25}$$

where n_t is the above index. By Proposition 2, we can sum the weights of the samples belonging to the same index n_t to obtain a proper sample set $\{n_t, \beta_{n_t}\}_{n_t=1}^{N}$ with respect to the posterior PMF $\pi(n_t)$.

A straightforward implementation of the CONDENSATION algorithm (Fig. 1) for simultaneous tracking and recognition is not efficient in terms of its computational load. Since $\mathcal{N} = \{1, 2, \ldots, N\}$ is a countable sample space, we need N samples for the identity variable n_t according to Proposition 1. Assume that, for each identity variable n_t, J samples are needed to represent θ_t. Then we need $M = J * N$ samples in total. Further assume that one resampling step takes T_r units of time (ut), one predicting step takes T_p ut, one updating step takes T_u ut, the normalizing step takes T_n ut, and the marginalizing step takes T_m ut. In the updating step, there are two substeps: computing the transformed image $\mathcal{T}_{\theta_t}\{z_t\}$, and evaluating the likelihood $p(z_t|n_t, \theta_t)$, taking T_t and T_l ut respectively. Obviously, the total computation time to deal with one video frame is $J * N * (T_r + T_p + T_t + T_l) + T_n + T_m$ ut. It is well known that computing the transformed image is much more expensive than the other operations, i.e., $T_t >> \max(T_r, T_p, T_l)$. Therefore, as the number of templates N grows, the computational load increases dramatically.

Since the sample space $\mathcal{N}$ is countable, an exhaustive search is possible. Mathematically, we relax the random sampling in the identity variable n_t by constructing samples as follows: for each $\theta_t^{(j)}$,

$$(1, \theta_t^{(j)}, w_{t,1}^{(j)}), (2, \theta_t^{(j)}, w_{t,2}^{(j)}), \ldots, (N, \theta_t^{(j)}, w_{t,N}^{(j)}).$$

We in fact use the following notation for the sample set:

$$\mathcal{S}_t = \{(\theta_t^{(j)}, w_t^{(j)}, w_{t,1}^{(j)}, w_{t,2}^{(j)}, \ldots, w_{t,N}^{(j)})\}_{j=1}^{J}, \tag{26}$$

with $w_t^{(j)} = \sum_{n=1}^{N} w_{t,n}^{(j)}$.

Based on this, we can first resample the 'marginal' sample set $\{(\theta_{t-1}^{(j)}, w_{t-1}^{(j)})\}_{j=1}^{J}$ to arrive at $\theta_{t-1}^{'(j)}$'s. After resampling, we set $w_{t-1}^{'(j)} = 1$ and $w_{t-1,n}^{'(j)} = w_{t-1,n}^{(j)} / w_{t-1}^{(j)}$ for all $j = 1, 2, \ldots, N$. We then predict the $\theta_t^{(j)}$'s using $p(\theta_t|\theta_{t-1}^{'(j)})$. Finally, the propagation of the joint distribution requires the following weighting step:

$$w_{t,n}^{(j)} = w_{t-1,n}^{'(j)} * p(z_t|n, \theta_t^{(j)}). \tag{27}$$

The proposed algorithm is summarized in Fig. 2.

The crux of this algorithm lies in that, instead of propagating random samples on both the motion vector and the identity variable, we can keep the samples of the identity variable fixed and let those of the motion vector be random. Also, we propagate only the marginal distribution for motion tracking, but we still propagate the joint distribution for recognition purposes.

Compared to the CONDENSATION algorithm, the proposed algorithm is more efficient and accurate since its total computation time is $J*(T_r+T_p+T_t)+J*N*T_l+T_n+T_m$ ut, a tremendous improvement over CONDENSATION when dealing with a large database since the majority computational time $J * T_t$ does not depend on N. To appreciate its accuracy, consider the following case for the CONDENSATION algorithm: all samples belonging to identity $n_t = l$, where l is the correct hypothesis, are drawn with a bias in θ_t. More specifically, though $p(z_t|l, \theta_t) > p(z_t|n_t, \theta_t)$ holds provided that the same θ_t is applied on both sides, different θ_t's can be chosen in a biased way to disobey this inequality. The proposed algorithm eliminates such a bias.

5. Still-to-Video based Face Recognition

In this section we describe the still-to-video scenarios actually used in our experiments and their practical model choices, followed by the results obtained. Two databases were used in the still-to-video experiments.

Database-1 contains video sequences with subjects walking towards a camera in order to simulate scenarios typical of visual surveillance. There are 30 subjects, each having one face template. The face gallery is shown in Fig. 3. The probe set contains 30 video sequences, one for each subject. Fig. 4 shows some example frames in one probe video and Fig. 5 shows a set of face images cropped by hand from probe videos and then normalized. Note that the gallery is captured under different circumstances from the probe and that the probe has considerable variation in scale. These images were collected, as part of the HumanID project, by researchers at the National Institute of Standards and Technology and the University of South Florida.

Database-2 comes from the Carnegie-Mellon University Motion of the Body (MoBo) database [28], which was collected under the HumanID project. There are 25 subjects. The video sequences show the individuals walking on a treadmill so that they move their heads naturally. Different walking styles have been simulated to assure a variety of conditions that are likely to appear in real life: *walking slowly*, *walking fast*, *inclining*, and *carrying an object*. Therefore, four videos per person and 99 videos in total (with one *carrying* video missing) are available. During recording of the videos, illumination conditions do not alter. Each video consists of 300 frames (480×640 pixels per frame) captured at 30 Hz. The inner face regions in these videos contain between 30×30 and 40×40 pixels.

Only 25 *slowWalk* videos are used as probes in the still-to-video experiment while all 99 videos are in used in the video-to-video experiment. Some example images from the videos (*slowWalk*) are shown in Fig. 10. Fig. 3 shows the face gallery in Database-2 with facial images in almost frontal views cropped from the probe videos and then normalized using their eye positions.

5.1. Model Choices

The model parameters used in the experiments are chosen as follows.

1. Image representation. For Database-1, an image reconstructed from the top 300 principal components or eigenfaces [9] is used to represent the face. Fig. 3 shows the top 10 eigenfaces. For Database-2, the raw intensity values are used.

2. Geometric transformation is only affine. Specifically, $\theta = (a_1, a_2, a_3, a_4, t_x, t_y)$ where $\{a_1, a_2, a_3, a_4\}$ are deformation parameters and $\{t_x, t_y\}$ are 2-D translation parameters. This is a reasonable approximation since there is no significant out-of-plane motion as the subject walks towards the camera. Regarding the photometric transformation, only histogram equalization is performed to partially compensate for contrast variations. The complete transformation $\mathcal{T}_\theta\{z\}$ is as follows: affine transform z using $\{a_1, a_2, a_3, a_4\}$, crop out the region of interest at position $\{t_x, t_y\}$ with the same size as the still template in the gallery, and perform histogram equalization.

3. Prior distribution $p(\theta_0|z_0)$ is a Gaussian whose mean comes from the initial face detector and whose covariance matrix is manually specified.

4. Noise distribution $p(u_t)$ is a Gaussian whose mean and covariance matrix are manually specified. Also function $g(.,.)$ in Eq. (1) is assumed to be an addition function. Furthermore, $p(u_t)$ is not time-varying. This is essentially a first-order Markov Gaussian model with constant velocity. Given the scenario that the subject is walking towards the camera, the scale increases with time. However, under perspective projection, this increase is no longer linear so that the constant-velocity model is not optimal. However, experimental results show that as long as the samples of θ can cover the motion, this model is sufficient.

5. The noise distribution $p(v_t)$ or the likelihood $p(z_t|n_t, \theta_t)$ could in general be functions of distance between $\mathcal{T}_{\theta_t}\{z_t\}$ and I_{n_t}, say $d(\mathcal{T}_{\theta_t}\{z_t\}, I_{n_t})$. This can be any distance metric used in still-to-still

recognition algorithms. Here, we take it to be a 'truncated' Laplacian:

$$p(z_t|n_t,\theta_t) \propto \begin{cases} \sigma^{-1}\exp(-\|v_t\|/\sigma) & \text{if } \|v_t\| \leq \lambda\sigma \\ \sigma^{-1}\exp(-\lambda) & \text{otherwise} \end{cases} \tag{28}$$

where $\|I(R)\| = \sum_{r\in R}|I(r)|$, and σ and λ are manually specified. Furthermore, $p(v_t)$ is not time-varying. The Gaussian distribution is widely used as a noise model, accounting for sensor noise, digitization noise, etc. However, given the observation equation $v_t = \mathcal{T}_{\theta_t}\{z_t\} - I_{n_t}$, the dominant part of v_t becomes the high-frequency residual if θ_t is not proper, and it is well known that the high-frequency residuals of natural images are more Laplacian-like. The 'truncated' Laplacian is used to provide a chance for the samples to accommodate abrupt motion changes.

5.2. Experimental Results

For Database-1, Fig. 6 presents the plot of the posterior probability $\pi(n_t)$, and Fig. 7 the conditional entropy $H(n_t|z_{0:t})$ and the MMSE estimate of the scale parameter a_1, all against t. In Fig. 4, the tracked face is superimposed on the image using a bounding box.

Suppose the correct identity for Fig. 4 is l. From Fig. 6, we can easily observe that the posterior probability $\pi(n_t = l)$ increases as time proceeds and eventually approaches 1, and all other π's for $n_t \neq l$ go to 0. Fig. 7 plots the decrease in the conditional entropy $H(n_t|z_{0:n})$ and the increase in the scale parameter, which matches with the scenario: a subject walking towards a camera.

Table 1 summarizes the average recognition performance and computational time of CONDENSATION and the proposed algorithm when applied to Database-1. As far as performance is concerned, there is no significant difference between the two algorithms, with the proposed one giving slightly better results. Obviously, the proposed algorithm is more efficient than the CONDENSATION algorithm. It is about 10 times faster, as shown in Table I. This experiment was implemented in C++ on a PC with a P-III650 CPU and 512M RAM, with the number of motion samples J chosen to be 200 and the number of templates in the gallery N to be 30.

The relatively low recognition rate is due to the fact that the probe video was taken outdoors with the sun casting a strong shadow on the face, while the gallery was taken indoors. Compare the images in Figs. 3 and 5. This is somewhat similar to the FC test reports in [3]. Histogram equalization is not able to account for illumination changes effectively. As a comparison, we used the images shown in Fig. 5 as probes to perform still-to-still face recognition using the eigenface approach [9]. The recognition result is less than 30% for the top match.

Algorithm	CONDENSATION	Proposed
Recognition rate within top 1 match	47%	50%
Recognition rate within top 3 matches	70%	77%
Time per frame	22s	2.1s

TABLE 1
Performances of algorithms when applied to Database-1.

The recognition result for Database-2 is presented in Fig. 8. Compared to Database-1, better performance is achieved because there is no illumination change. However, due to the pose variations present in the database, using one frontal view is not sufficient to represent one face class. This motivates us to obtain more representative views for one face class, leading to the discussions in the following sections.

6. Video-to-Video based Face Recognition

In this section we introduce our *video-to-video* based face recognition approach. It enhances the *still-to-video* approach by taking an entire video, instead of a single image, to represent the face of an individual. The *video-to-video* based approach has two stages. In the first learning stage, *exemplars*, which are selected representatives from the raw video, are automatically extracted from gallery videos. The exemplars are used to summarize the gallery video information. In the second part, exemplars are then used as centers for probabilistic mixture distributions for tracking and recognition processes. Probabilistic methods are attractive in this context as they allow systematic handling of uncertainty and an elegant way of fusing temporal information.

In Subsection 6.1 we present the learning stage and explain, how the exemplars are generated from a gallery video. In Subsection 6.2 we explain how tracking and recognition is performed. In Subsection 6.3 we present experimental results on the 99 video sequences of 25 individuals in Database-2.

6.1. Exemplar-based Learning

In order to realize *video-to-video* recognition, a probabilistic model needs to be learned from each gallery video V. Denote the gallery by $\mathcal{V} = \{V_1, V_2, \ldots, \mathcal{V}_N\}$. We take an approach which is similar to the ones proposed in [20, 21]. These two approaches have in common that they try to find a set of exemplars that best describe the set of training images, i.e., that minimize the expected distance between the given set of images $\mathcal{Z} = \{z_1, z_2, \ldots, z_J\}$ and a set of exemplars (cluster centers) $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$:

$$E\left\{d(\mathcal{Z}, \mathcal{C})\right\} \quad . \tag{29}$$

In other words, let $\mathcal{Z} = \{z_1, z_2, \ldots, z_J\}$ again be the sequence of video images. It is being searched for a set of exemplars $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ such that

$$p(z_t) = \sum_{c \in \mathcal{C}} \int_{\Theta} p(z_t|\theta, c) p(\theta|c) p(c) d\theta \tag{30}$$

is maximal for all t. Here, $p(z_t|\theta, c)$ is the observation equation, given by

$$\begin{aligned} p(z_t|x) &\equiv p(z_t|\theta, c) \\ &\propto \exp\left[-\frac{1}{2\sigma^2} d(\mathcal{T}_\theta\{z_t\}, c)\right] \quad , \end{aligned} \tag{31}$$

where the choice of σ depends on the choice of d.

In [20], a K-means clustering technique is applied to minimize Eq. (30), and in [21] an EM approach is used. In spite of the essential differences between these approaches, they suffer from the following drawbacks:

- They find K exemplars, where K has to be given in advance. For face recognition this drawback is restrictive: Clearly, in Eq. (29) the distance measure d may be chosen arbitrarily and for face recognition it is preferable to choose one of the well-evaluated ones (PCA, LDA, etc.) [3]. Thresholds and variances for each of these measures that minimize misclassification are known and considering them requires a dynamic choice of the number of clusters rather than a static one.
- They have to store the training data in order to compute the clusters, which becomes difficult for long video streams.

Inspired by the probabilistic interpretation of the RBF neural network [22], we developed an online technique to learn the exemplars: At each time step t, $p(z_t|\theta, c)$ of Eq. (30) is maximized. If $p(z_t|\theta, c) < \rho$ for some ρ (which depends on the choice of d) then $\mathcal{T}_\theta\{z_t\}$ is added to the set of exemplars.

The details of the learning algorithm are as follows:

1. The first step is the alignment- or tracking-step: a cluster k and a deformation $\theta \in \Theta$ are found such that $d(\mathcal{T}_\theta\{z_t\}, c_k)$ is minimized:

$$\begin{aligned} \theta_t &\longleftarrow \arg\min_\theta \min_k d(\mathcal{T}_\theta\{z_t\}, c_k) \text{ and} \\ k_t &\longleftarrow \arg\min_k d(\mathcal{T}_{\theta_t}\{z_t\}, c_k) \end{aligned} \tag{32}$$

2. The second step generates a new cluster center, if necessary: if

$$p(z_t|\theta_t, c_{k_t}) < \rho$$

then

$$\mathcal{C} \longleftarrow \mathcal{C} \cup \{\mathcal{T}_{\theta_t}\{z_t\}\},$$

Count the number of times, $count(k_t) = count(k_t) + 1$, that cluster c_{k_t} approximated image z_t best.

3. Repeat steps 1 and 2 until all the frames in the video are processed.

4. Compute the mixture weights $\mu_k \propto count(k)$.

The result of this learning procedure is

1. a set $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ of aligned exemplars c_k

2. a prior μ_k for each of the exemplars c_k.

Clearly, the more carefully the set Θ is chosen, the fewer exemplars are generated. Allowing Θ, e.g., to compensate only for translations, exemplars are generated to compensate for scale changes and rotation.

Given a gallery $\mathcal{V}$ of videos, the above has to be carried out for each video. During recognition, as will be explained in the next section, the exemplars are used as centers of mixture models.

The above learning algorithm is motivated by the online learning approaches to artificial neural networks (ANNs) [24, 25]. Clearly, many kinds of enhancements can be attempted (topology preserving maps, neighborhood relations, etc.). An online learning algorithm for exemplars used during testing could allow, in a bootstrapping manner, learning new exemplars from probe videos.

In [26] a similar learning approach was presented. Unlike our approach, face images are not normalized with respect to Θ, which results in a far larger number of clusters. In [27] a 'Unit Face' RBF model is proposed where for each individual, a single RBF network is trained. The authors have also investigated different geometrical normalizations and have experimented with a preprocessing step, such as the application of a 'difference of Gaussians' or Gabor wavelets.

The goal of both of [26] and [27] was to build a representation of face intensity by using an RBF network. We want to make it clear that this is exactly what we do not want! Our intention is to choose a well-known face representation (such as PCA) *in advance*. *Then* we learn the different exemplars of a single face. The advantage is that in this way we inherit the "face recognition capabilities" of the PCA, LDA, etc. techniques.

6.2. Tracking and Recognition with Multiple Exemplars

After applying the learning algorithm we have for each individual n a set of exemplars $\mathcal{C}^n = \{c_1^n, c_2^n, \ldots, c_{K_n}^n\}$. In order to recognize individuals with multiple exemplars, our SIS approach has to be developed further.

Exemplars as Mixture Centers

To take into account a set of exemplars $\mathcal{C}^n = \{c_1^n, c_2^n, \ldots, c_{K_n}^n\}$ for individual n, we refine the likelihood computation of Eq. (31) as follows:

$$\begin{aligned} p(z_t|x) &\equiv p(z_t|n,\theta) \\ &\propto \sum_{c\in\mathcal{C}^n} p(z_t|\theta,c)p^n(c) \qquad (33) \\ &\propto \sum_{c\in\mathcal{C}^n} \exp\left[-\frac{1}{2\sigma^2}d(\mathcal{T}_\theta\{z\},c)\right]\mu_c^n \ . \qquad (34) \end{aligned}$$

The exemplars in $\mathcal{C}^{n_t}$ are used as the mixture centers of a joint distribution; $p^{n_t}(c) = \mu_c^{n_t}$ is the prior for mixture center c of individual n_t.

Dynamic Model

In Section 5, a dynamic model for Eq. (1) has to be given in advance. Here, however, while learning exemplars from video sequences, a dynamic model can also be learned. In Eq. (7)

$$p(x_t|x_{t-1}) \equiv p(\theta_t, n_t|\theta_{t-1}, n_{t-1})$$

defines the probability of the state variable changing from x_{t-1} to x_t. Learning of a dynamic model has been discussed in [20].

Computation of Posterior Distribution

The posterior probability distribution $\pi(n_t, k_t, \theta_t)$ (where n refers to the individual and k to the exemplar index) is represented by a set of M indexed and weighted particles

$$\left\{\left(n^{(m)}, k^{(m)}, \theta^{(m)}, w^{(m)}\right)\right\}_{m=1\ldots M}^{t} \ . \qquad (35)$$

Note that we have, for better readability, indexed the entire set with t, instead of each component. Since all exemplars for each person are aligned, we do not have to treat the different exemplars for a single person separately. We can therefore increase efficiency if we rewrite (35) as:

$$\left\{\begin{bmatrix} n^{(m)}, 1, \theta^{(m)}, w_1^{(m)} \\ \vdots \\ n^{(m)}, K_{n^{(m)}}, \theta^{(m)}, w_{K_{n^{(m)}}}^{(m)} \end{bmatrix}\right\}_{m=1\ldots M'}^{t} \ . \qquad (36)$$

Set (36) is a set of $K_{n^{(m)}} \times 4$ dimensional matrices, and each matrix represents one particle, where $K_{n^{(m)}} = \left|\mathcal{C}^{n^{(m)}}\right|$. We can now easily marginalize over $\mathcal{C}^{n^{(m)}}$ to compute the posterior probability $\pi(n_t, \theta_t)$: We get with

$$\hat{w}^{(m)} = \sum_{k=1}^{K_{n^{(m)}}} \mu_k^{n^{(m)}} w_k{}^{(m)} \qquad (37)$$

a new set of weighted sample vectors:

$$\left\{\left(n^{(m)}, \theta^{(m)}, \hat{w}^{(m)}\right)\right\}_{m=1\ldots M'}^{t} \ . \qquad (38)$$

In Eq. (37), $\mu_k^{n^{(m)}}$ is the prior of exemplar k of person $n^{(m)}$.

To compute the identity from the particle set (38) we marginalize over θ as in Eq. (24).

6.3. Experimental Results

We have tested the video-to-video-based recognition algorithm on 99 video sequences of 25 different individuals in Database-2. As mentioned before, there are four walking styles: *walking slowly*, *walking fast*, *inclining* and *carrying an object*. Therefore, four videos per person are available. In the experiments we used one or two of the video types as gallery videos for training, while the remaining ones were used as probes for testing.

From each gallery video, a first face sample was cropped by hand. Based on this sample, the training process was initiated. Four examples of automatically extracted exemplar sets are shown in Fig. 11 (extracted from the *slowWalk* videos). The top row shows the exemplars of subjects 04006 and 04079 (six exemplars each). The leftmost exemplar in each of the two sets is the hand-extracted one. The second and third rows of Fig. 11 show the exemplars of subject 04015, and the fourth and fifth rows the exemplars of subject 04022. The top left exemplar of each of the two sets is again the hand-extracted one. Clearly, the number of generated exemplars depends on the variety of different views that appear in the gallery video. To generate these exemplars, we set $\rho = 0.65$ and the standard deviation per pixel to $\sigma = 0.4$. Increasing σ to $\sigma = 0.5$ decreases the number of exemplars by roughly a factor of two.

During testing, these exemplar galleries were used to compute, over time, the posterior probabilities $\pi(n_t)$. It is interesting to see how the posterior probabilities develop over time. Examples of this can be seen in Fig. 12. The dashed line refers to the correct hypothesized identity, and the other five curves refer to the probabilities of the top matching identities other than the true one. One can see in the left and right plots that the dashed line (true hypothesis) increases quickly to 1. In order to consider *all* the frames of the video, we restart the algorithm after convergence. Recognition is established by that identity to which the SIS converged most often.

Examples illustrating the robustness as well as the limits of our approach are shown in Figs. 11, 12, 13 and 14: Due to the severe differences between the gallery exemplars (derived from "slowWalk") in Fig. 11 (4th and 5th rows) and the probe video (see sample images from the probe video in Fig. 13), the recognition of subject 04022 was not successful (Fig. 12, top). On the other hand, in spite of the differences between the gallery exemplars and the probe video, subject 04079 was always recognized successfully (Fig. 12, bottom). The major problems that we encountered during our experiments were:

1. Some subjects appeared very different in the gallery video and in the probe videos: This was the case for about 50% of the failed experiments.

2. Some subjects looked down while walking: This was the case for roughly 10 subjects (Fig. 15). In some cases, where the subject looked down in the gallery as well as in the probe, this wasn't necessarily a problem, but in other cases (see Fig. 15, left), it led to misclassification.

Clearly, both problems can be solved by using more gallery videos. We therefore did a second set of experiments, where two videos were used as a gallery video, while the testing was carried out on the remaining two videos. The overall recognition results for one and two gallery videos are summarized in Table 2. The 'g' indicates which videos were used in the gallery. The gallery contained 25 different individuals; however, for the "carrying" video set, only 24 different individuals were available.

In [20] it is demonstrated that the dynamic information can also be learned. We have done extensive experiments to combine facial and dynamic information for recognition. However, we have observed that the dynamic information of persons can change severely with walking speed. We have therefore not used that information for recognition.

Video images from our test data were converted from color to gray value images, but no further processing was done. Throughout our experiments we used the Euclidean distance measure. The set of deformations Θ included scale and translation. Shear and rotation were not considered.

slow	fast	incline	carrying
g	100%	96%	92%
92%	g	100%	96%
100%	96%	g	96%
88%	96%	92%	g
g	g	100%	96%
g	100%	g	100%
g	100%	96%	g
100%	g	g	96%
100%	g	100%	g
100%	100%	g	g

slow	fast	incline	carrying
g	96%	92%	88%
92%	g	92%	92%
96%	96%	g	96%
88%	88%	83%	g
g	g	96%	92%
g	100%	g	100%
g	96%	96%	g
100%	g	g	96%
92%	g	96%	g
100%	96%	g	g

TABLE 2
Overall recognition rates in percent for $\sigma = 0.4$ (left), and $\sigma = 0.5$ (right). The 'g' indicates the video used as gallery.

7. Discussion and Conclusions

We have presented a systematic method for face recognition from a probe video, compared with a gallery of still templates. A time series state space model is used to accommodate the video, and SIS algorithms provide the numerical solutions to the model. This probabilistic framework, which overcomes many difficulties arising in conventional recognition approaches using video, is registration-free and avoids the need for selecting good frames. It turns out that an immediate recognition decision can be made in our framework due to the degeneracy of the posterior probability of the identity variable. The conditional entropy can also serve as a good indication of the convergence. In addition, the still templates in the gallery are generalized to videos by learning exemplars from the gallery video. The still-to-video framework then performs recognition from probe videos. It is shown by our experiments that this video-to-video algorithm achieves better performance than the still-to-video algorithm. However, in order to show that our approach is capable of recognizing faces in practice, one needs to work with much larger face databases.

Once the exemplars are generated, they could theoretically be used as gallery images in a *multiple-still-to-multiple-still* (or *still-to-multiple-still*) face recognition approach. Here, recognition would be based on pairwise computation of the distance measure $d(\cdot, \cdot)$. If we use one of the image representations that are already well known for face recognition (such as PCA) and an appropriate distance measure, we can consider the techniques that were tested in the FERET test [3] as a "baseline" for our method. This means that we can predict the recognition results for our still-to-video and video-to-video recognition methods, and the FERET performances are lower bounds.

The following issues are worthy of investigation in the future.

1. The assumption that identity does not change as time proceeds, i.e., $p(n_t|n_{t-1}) = \delta(n_t - n_{t-1})$, could be relaxed by having nonzero transition probabilities between different identity variables. Using nonzero transition probabilities will enable an easier transition to the correct choice if the initial choice is incorrect, making the algorithm more robust.

2. Choice of the likelihood distribution $p(z_t|n_t, \theta_t)$ and condition (B). In general, the smaller $d(\mathcal{T}_{\theta_t}\{z_t\}, I_{n_t})$ is, the higher the likelihood $p(z_t|n_t, \theta_t)$ and the higher the posterior $p(n_t|z_{0:t})$. In this sense, an accurate solution to this problem is determined by the basic problem: how can we find an efficient distance metric d or how well is condition (B) satisfied? Fig. 9 plots, against the logarithm of

the scale parameter, the 'average' likelihood of the correct identity

$$\frac{1}{N}\sum_{n\in\mathcal{N}} p(I_n|n,\theta),$$

and of the incorrect identities

$$\frac{1}{N(N-1)}\sum_{m\in\mathcal{N},n\in\mathcal{N},m\neq n} p(I_m|n,\theta),$$

of the face gallery as well as the 'average' likelihood ratio, i.e., the ratio between the above two quantities. The observation is that only within a narrow 'band' is condition (B) well satisfied. Therefore, the success of the SIS algorithm depends on how well the samples lie in a similar 'band' in the high-dimensional affine space. Also, the lower bound η in condition (B) is too strict. If we take the mean of the 'average' likelihood ratio shown in Fig. 9 as an estimate of η (roughly 1.5), Eq. (16) tells us that after 20 frames, the probability $p(l|y_{0:t})$ reaches 0.99! However, this does not happen in the experiments due to noise in the observations and incomplete parameterization of the transformations.

3. Resampling. In the recognition algorithm, the marginal distribution $\{(\theta_{t-1}^{(j)}, w_{t-1}^{'(j)})\}_{j=1}^{J}$ is sampled to obtain the sample set $\{(\theta_t^{(j)}, 1)\}_{j=1}^{J}$. This may cause problems in principle since there is no conditional independence between θ_t and n_t given $y_{0:t}$. However, in a practical sense, this is not a big disadvantage because the purpose of resampling is to "provide chances for the good streams (samples) to amplify themselves and hence rejuvenate the sampler to produce a better result for future states as the system evolves" [7]. The resampling scheme can be simple random sampling with weights (as in CONDENSATION), residual sampling, or a local Monte Carlo method.

4. Computational load. As mentioned earlier, two important numbers affecting the computation are J, the number of motion samples, and N, the size of the database. (i) The choice of J is an open question in the statistics literature. In general, larger J's produce more accurate results. (ii) The choice of N depends on the application. Since a small database was used in this experiment, this is not a big issue here. However, the computational burden may be excessive if N is large. One possibility is to use a continuous parameterized representation, say γ, instead of a discrete identity variable n. The task then reduces to computing $p(\gamma_t, \theta_t|y_{0:t})$. We then can easily rank the gallery using the estimated γ_t.

5. We highlight the differences from Li and Chellappa's approach [14]. In [14], basically only the tracking motion vector is parameterized in the state-space model. The identity is used only in the initialization step to rectify the template onto the first frame of the sequence. In our approach, however, both the tracking motion vector and the identity variable are parameterized in the state-space model, which offers us one more degree of freedom and leads to a different approach to deriving the solution. The SIS technique is used in both approaches to numerically approximate the posterior probability given the observations. Again, in [14], the posterior probability of the motion vector and the verification probability are estimated by marginalizing on a proper region of state space, redefined at each time instant. However, we always compute the joint density, i.e., the posterior probability of the motion vector and the identity variable, and the posterior probability of the identity variable is just a free estimate by marginalizing over the motion vector. Note that there is no time propagation of the verification probability in [14], while we always propagate the joint density. One consequence is that we guarantee that $\sum_{n_t\in\mathcal{N}} p(n_t|z_{0:t}) = 1$, but there is no such guarantee in [14]. The approach of [14] in some sense is more like a batch method by running the algorithm for different templates, while ours is truly recursive. Another important consequence is that in our approach the degeneracy in the correct identity eventually leads to immediate decisions, while no such decision can be readily made from the verification probability in [14]. In addition, in terms of tracking accuracy, if the wrong template is rectified on the first frame in the initialization step, the tracking is more likely to be attracted to the noisy background, while our approach is more robust since we consider all templates at the same time.

REFERENCES

[1] Chellappa, R., Wilson, C. L., and Sirohey, S. (1995), Human and Machine Recognition of Faces, A Survey, *Proc. of IEEE* **83**, 705-740.

[2] Zhao, W. Y., Chellappa, R., Rosenfeld, A.. and Phillips, P. J. (2000), Face Recognition: A Literature Survey, *UMD CfAR Technical Report CAR-TR-948*.

[3] Phillips, P. J., Moon, H., Rivzi, S. A., and Rauss, P. J. (2000), The FERET Evaluation Metholodogy for Face-Recognition Algorithms, *IEEE Trans. PAMI* **22**, 1090-1104.

[4] Isard, M. and Blake, A. (1996), Contour Tracking by Stochatic Propagation of Conditional Density, *Proc. of ECCV*, 343-356.

[5] Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993), Novel Approach to Nonlinear/non-Gaussian Bayesian State Estimation, *IEE Proceedings on Radar and Signal Processing* **140**, 107-113.

[6] Kitagawa, G. (1996), Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models, *J. Computational and Graphical Statistics* **5**, 1-25.

[7] Liu, J. S. and Chen, R. (1998), Sequential Monte Carlo for Dynamic Systems, *Journal of the American Statistical Association* **93**, 1031-1041.

[8] Doucet, A., Godsill,S. J., and Andrieu, C. (2000), On Sequential Monte Carlo Sampling Methods for Bayesian Filtering, *Statistics and Computing* **10**, 197-209.

[9] Turk, M. and Pentland, A. (1991), Eigenfaces for Recognition, *Journal of Cognitive Neutoscience* **3**, 72-86.

[10] Etemad, K. and Chellappa, R. (1997), Discriminant Analysis for Recognition of Human Face Images, *Journal of Optical Society of America A* **14**, 1724-1733.

[11] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997), Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, *IEEE Trans. PAMI* **19**, 711-720.

[12] Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., Malsburg, C von der, Wurtz, R. P., and Konen, W. (1993), Distortion Invariant Object Recognition in the Dynamic Link Architecture, *IEEE Trans. Computers* **42**, 300-311.

[13] Choudhury, T., Clarkson, B., Jebara, T., and Pentland, A. (1999), Multimodal Person Recognition Using Unconstrained Audio and Video, *Proc. of Intl. Conf. on Audio- and Video-Based Biometric Person Authentication*, 176-181.

[14] Li, B. and Chellappa, R. (2000), Simultaneous Tracking and Verification via Sequential Posterior Estimation, *Proc. of CVPR*, 110-117.

[15] Black, M. J., and Jepson,A. D. (1999), A Probabilistic Framework for Matching Temporal Trajectories, *Proc. of ICCV*, 176-181.

[16] Li, Y., Gong, S., and Liddell, H. (2001), Constucting Structures of Facial Identities on the View Sphere Using Kernel Discriminant Analysis, *Proc. of Intl. Workshop on SCTV*.

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

[18] McKenna, S. and Gong, S. (1997), Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recogntion, *Proc. Intl. Conf. on Audio- and Video-based Biometric Person Authentication*, 177-183.

[19] Steffens, J., Elagin, E., and Neven, H. (1998), Personspotter - Fast and Robust System for Human Detection, Tracking, and Recognition, *Proc. Intl. Conf. on Automatic Face and Gesture Recognition*, 516-521.

[20] Toyama, K. and Blake, A. (2001), Probabilistic Tracking in a Metric Space, *Proc. of ICCV*, **2**, 50-59.

[21] Frey, B. and Jojic, N. (2000), Learning Graphical Models in Images, Videos, and their Spatial Transformations, *Proc. Conf. on Uncertainty in AI*.

[22] Lowe D. (1995), Radial Basis Function Networks, in "The Handbook of Brain Theory and Neural Networks" (M. Arbib, Ed.), pp. 779-782, MIT Press.

[23] Qian, G. and Chellappa, R. (2001), Structure from Motion Using Sequential Monte Carlo Methods, *Proc. of ICCV*, 614-621.

[24] Fritzke, B. (1995), Growling Cell Structures - a Self-organizing Network for Unsupervised and Supervised Learning, *Neural Learning* **7**, 1441-1460.

[25] Martinez, T. and Schulten, K. (1994), Topology Representing Networks, *Neural Learning* **7**, 505-522.

[26] Wechsler, H., Kakkad, V., Huang, J., Gutta, S., and Chen, V. (1997), Automatic Video-based Person Authentication Using the RBF Network, *Proc. Intl. Conf. on Audio- and Video-based Biometric Person Authentication*, 85-92.

[27] Howell, A. and Buxton H. (1996), Face Rcognition Using Radial Basis Function Neural Networks, *Proc. British Machine Vision Conference*, 455-464.

[28] Gross, R. and Shi, J. (2001) The CMU Motion of Body (MoBo) Database, *Technical Report CMU-RI-TR-01-18*, Robotics Institute, Carnegie Mellon University.

[29] Anderson, B. and Moore, J. (1979), "Optimal Fitering", Prentice Hall, Englewood Cliffs, New Jersey.

Initialize *a sample set* $\mathcal{S}_0 = \{(n_0^{(m)}, \theta_0^{(m)}, 1)\}_{m=1}^{M}$ *according to prior distributions* $p(n_0|z_0)$ *and* $p(\theta_0|z_0)$.
For $t = 1, 2, \ldots$
  For $m = 1, 2, \ldots, M$
    Resample $\mathcal{S}_{t-1} = \{(n_{t-1}^{(m)}, \theta_{t-1}^{(m)}, w_{t-1}^{(m)})\}_{m=1}^{M}$ *to obtain a new sample* $(n_{t-1}^{'(m)}, \theta_{t-1}^{'(m)}, 1)$.
    Predict *sample by drawing* $(n_t^{(m)}, \theta_t^{(m)})$ *from* $p(n_t|n_{t-1}^{'(m)})$ *and* $p(\theta_t|\theta_{t-1}^{'(m)})$.
    Update *weight using* $\alpha_t^{(m)} = p(z_t|n_t^{(m)}, \theta_t^{(m)})$.
  End
  Normalize *each weight using* $w_t^{(m)} = \alpha_t^{(m)} / \sum_{m=1}^{M} \alpha_t^{(m)}$.
  Marginalize *over* θ_t *to obtain weight* β_{n_t} *for* n_t.
End

FIG. 1 The CONDENSATION algorithm.

Initialize *a sample set* $\mathcal{S}_0 = \{(\theta_0^{(j)}, N, 1, ..., 1)\}_{j=1}^{J}$ *according to prior distribution* $p(\theta_0|z_0)$.
For $t = 1, 2, \ldots$
  For $j = 1, 2, \ldots, J$
    Resample $\mathcal{S}_{t-1} = \{(\theta_{t-1}^{(j)}, w_{t-1}^{(j)})\}_{j=1}^{J}$ *to obtain a new sample* $(\theta_{t-1}^{'(j)}, 1, w_{t-1,1}^{'(j)}, \ldots, w_{t-1,N}^{'(j)})$, *where* $w_{t-1,n}^{'(j)} = w_{t-1,n}^{(j)} / w_{t-1}^{(j)}$ *for* $n = 1, 2, \ldots, N$.
    Predict *sample by drawing* $(\theta_t^{(j)})$ *from* $p(\theta_t|\theta_{t-1}^{'(j)})$.
    For $n = 1, \ldots, N$
      Update *weight using* $\alpha_{t,n}^{(j)} = w_{t-1,n}^{'(j)} * p(z_t|n, \theta_t^{(j)})$.
    End
  End
  Normalize *each weight using* $w_{t,n}^{(j)} = \alpha_{t,n}^{(j)} / \sum_{n=1}^{N} \sum_{j=1}^{J} \alpha_{t,n}^{(j)}$ *and* $w_t^{(j)} = \sum_{n=1}^{N} w_{t,n}^{(j)}$.
  Marginalize *over* θ_t *to obtain weight* β_{n_t} *for* n_t.
End

FIG. 2 The proposed algorithm.

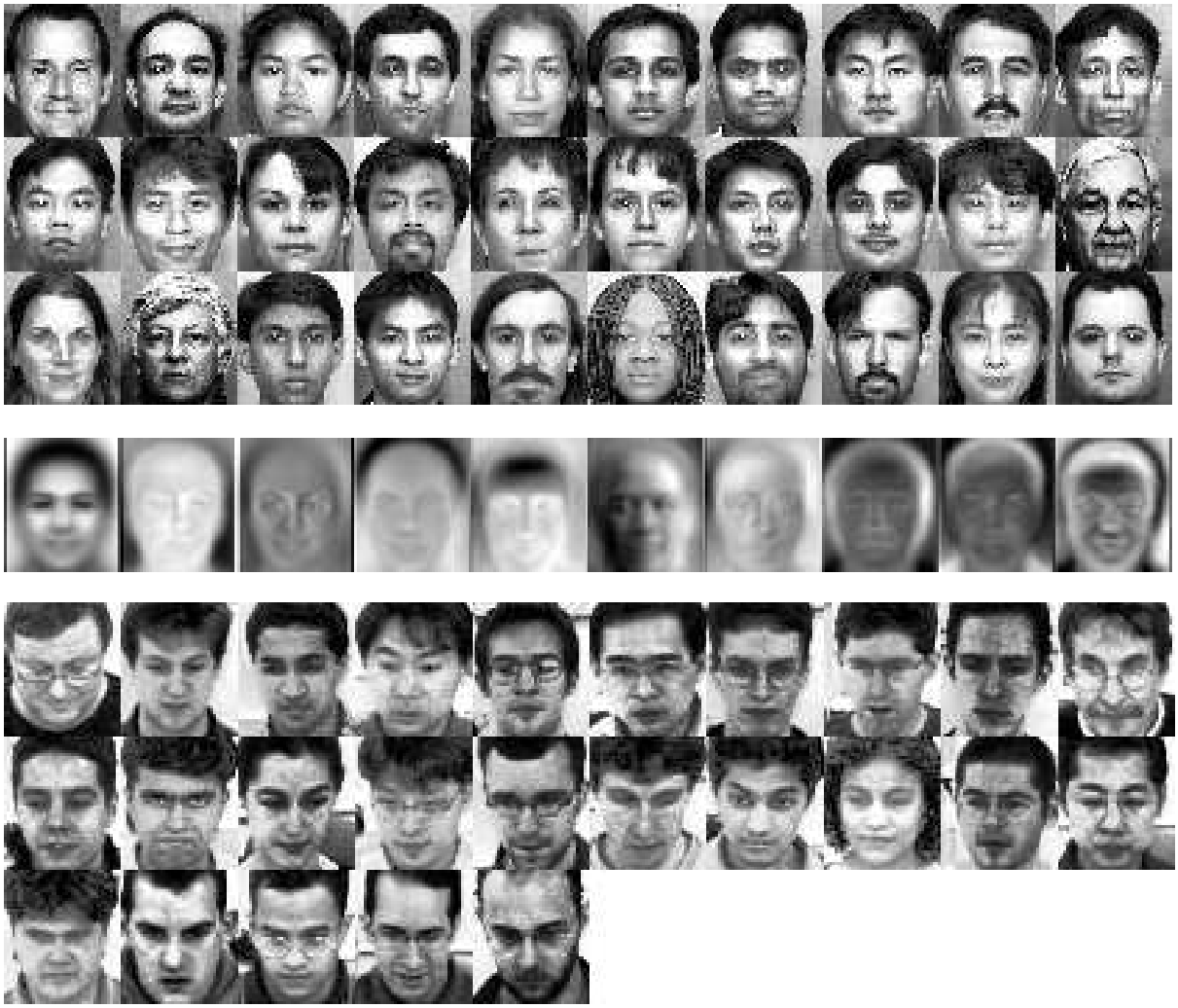

FIG. 3 Upper row: The face gallery in Database-1. Middle row: The top 10 eigenfaces for Database-1. Bottom row: The face gallery in Database-2. All images are normalized to 48x42.

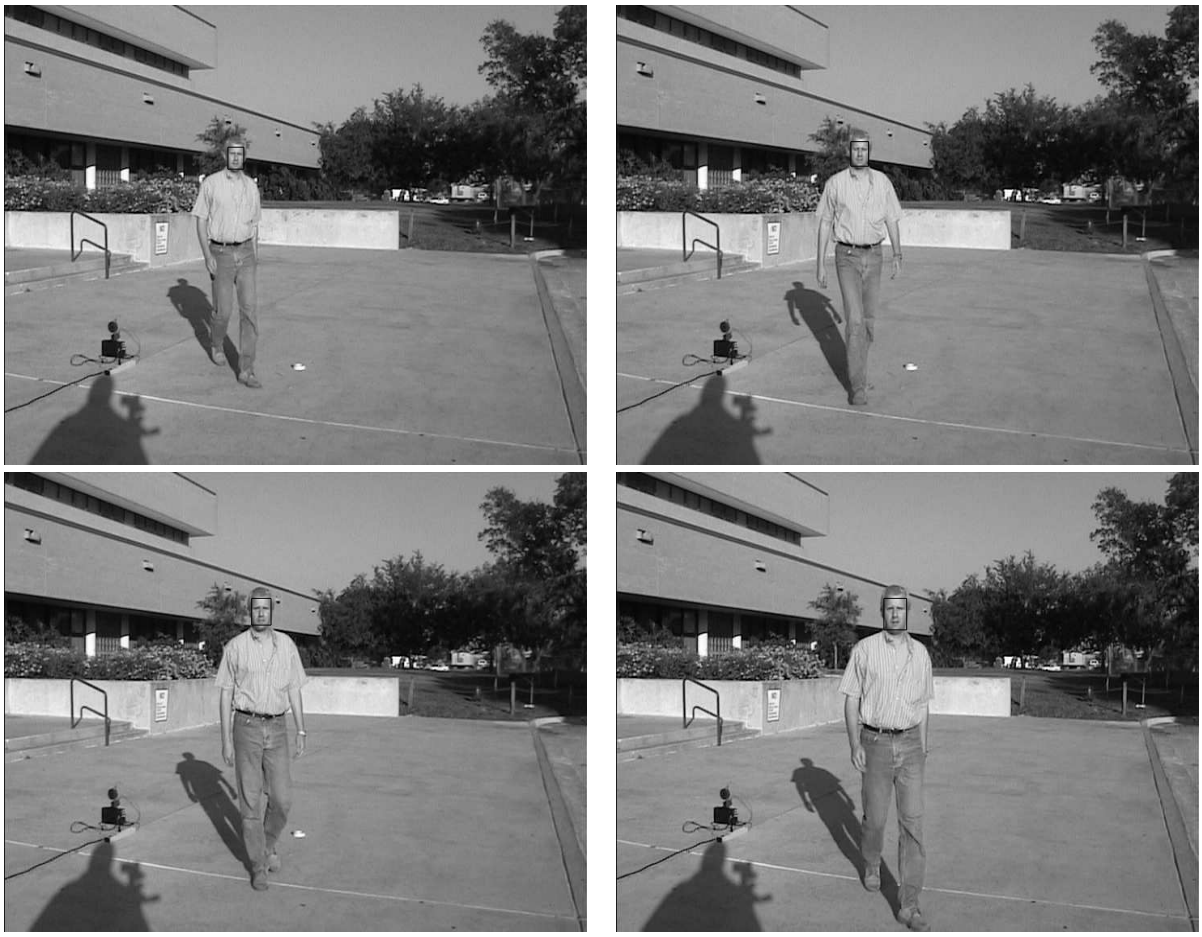

FIG. 4 Example frames in one probe video in Database-1. The image size is 720x480 while the actual face size ranges from approximately 20x20 in the first frame to 60x60 in the last frame.

FIG. 5 Facial images cropped from probe videos in Database-1 and then normalized.

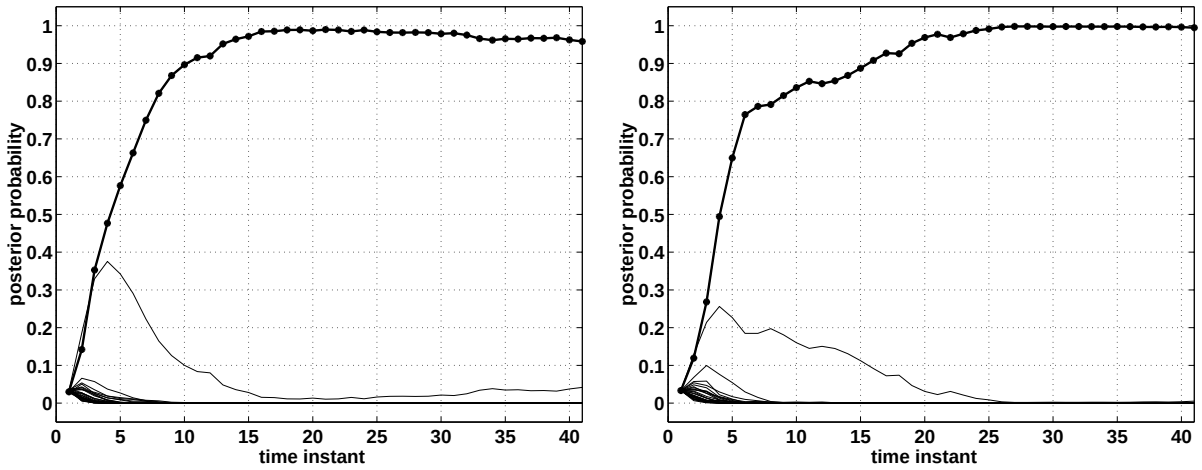

FIG. 6 Posterior probability $\pi(n_t)$ against time instant, obtained by the the CONDENSATION algorithm (left) and the proposed algorithm (right).

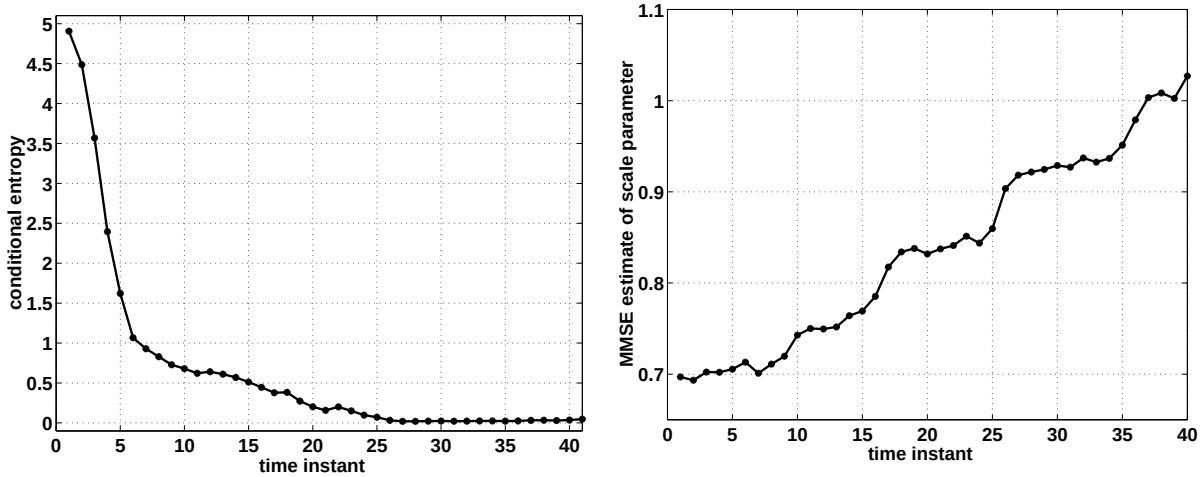

FIG. 7 Conditional entropy $H(n_t|z_{0:t})$ (left) and MMSE estimate of scale parameter (right) against time instant. Both are obtained using the proposed algorithm.

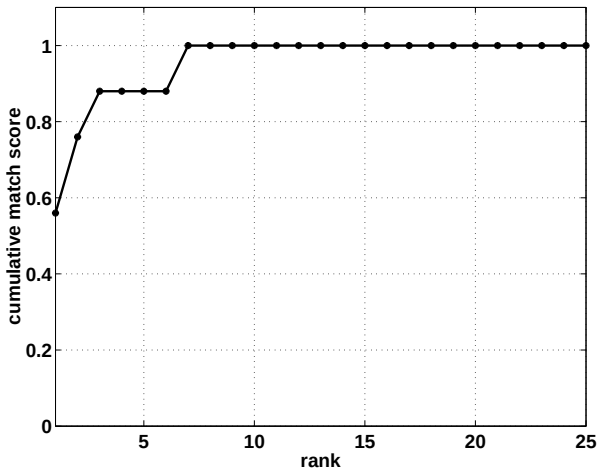

FIG. 8 Cumulative match curve for Database-2.

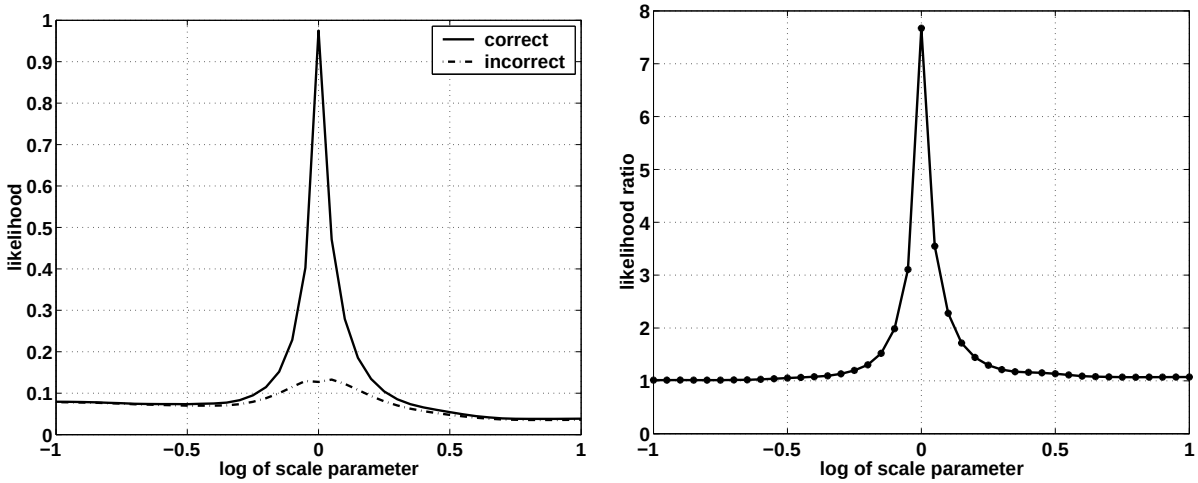

FIG. 9 Left: The 'average' likelihood of the correct hypothesis and incorrect hypotheses against the log of the scale parameter. Right: The 'average' likelihood ratio against the log of the scale parameter.

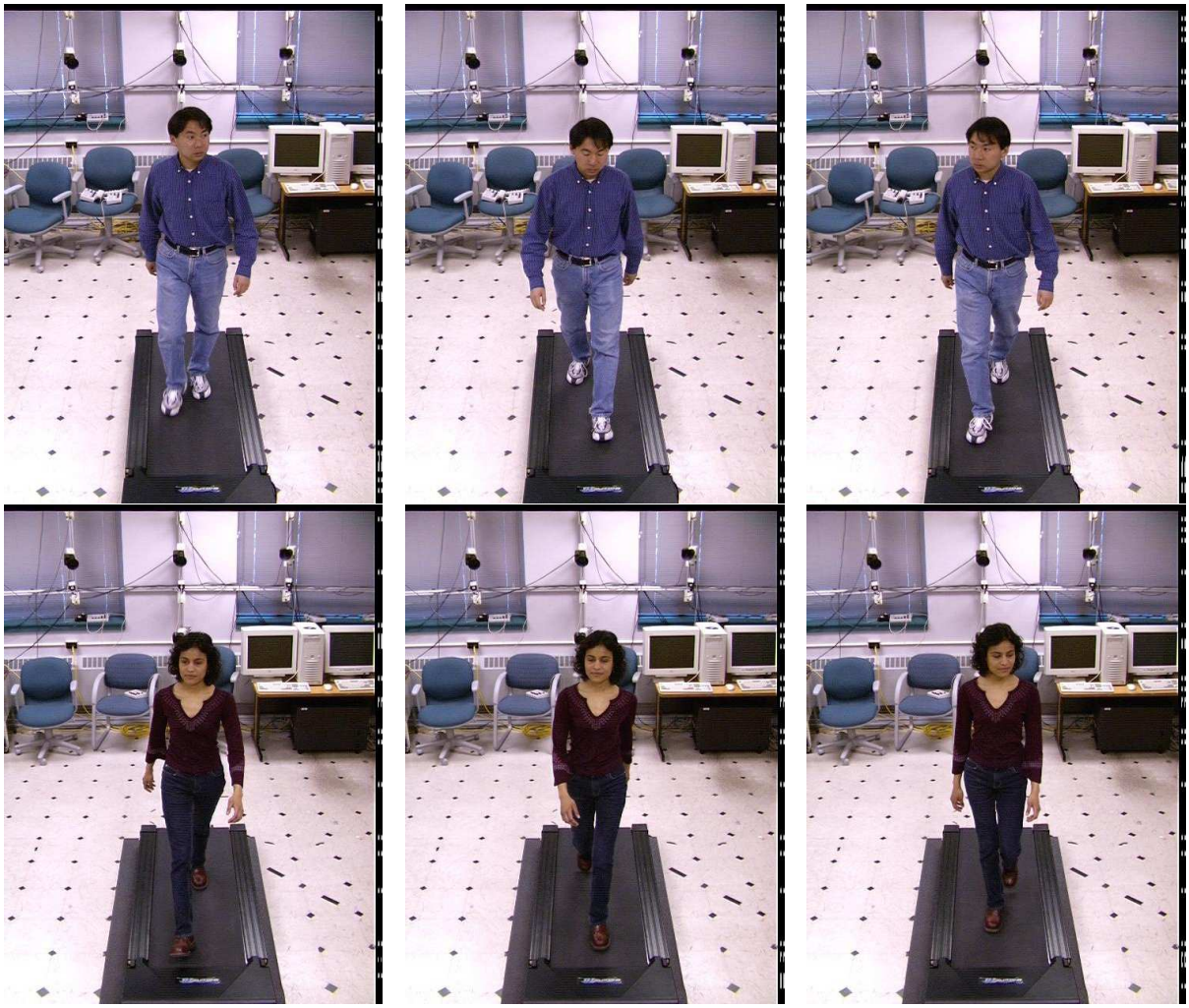

FIG. 10 Sample images from the videos (*slowWalk*) in Database-2.

FIG. 11 Exemplars of different persons in the gallery videos (*slowWalk*).

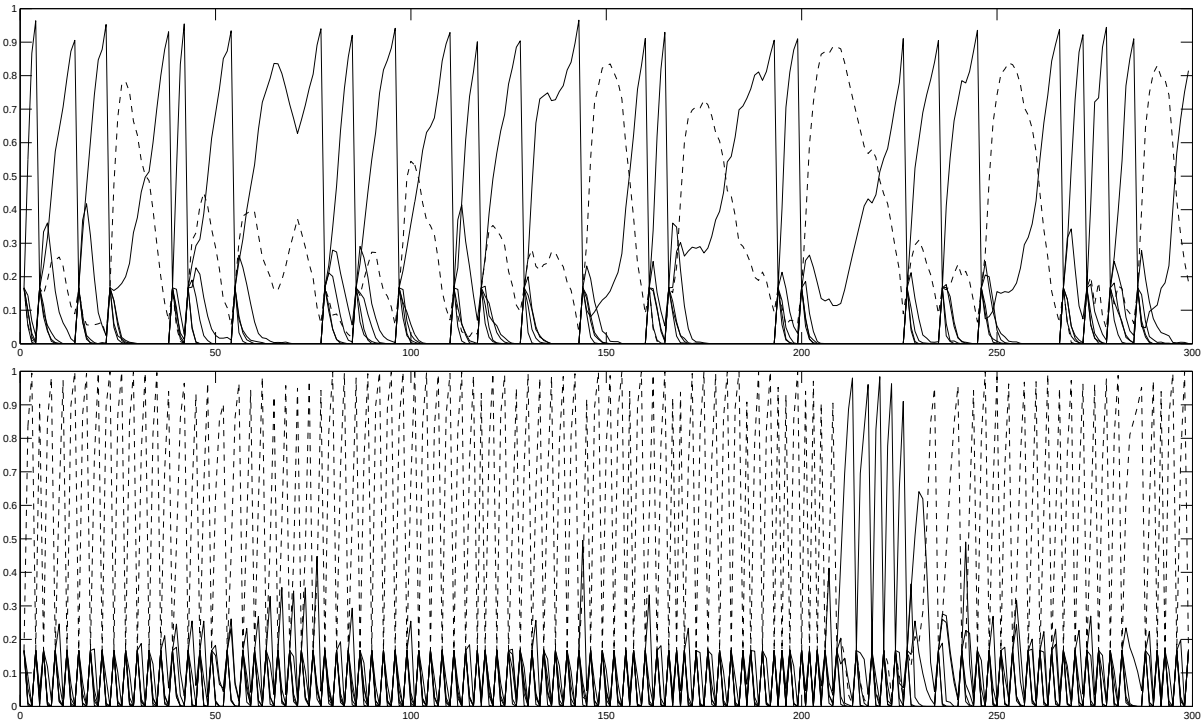

FIG. 12 Two typical probability evolutions. The top row was an unsuccessful recognition, and the bottom row was a successful one. The graphs plot the top five matches; the dashed line refers to the true hypothesis. The x-axis refers to the time t. The top graph shows the curve of subject 04022, and the bottom graph shows a typical curve (here, subject 04079). Compare the top graph with the image in Fig. 13.

FIG. 13 Sample frames 1, 35, 81, 100 of a probe video. One observes large differences from the gallery video. In this case recognition was *not* successful.

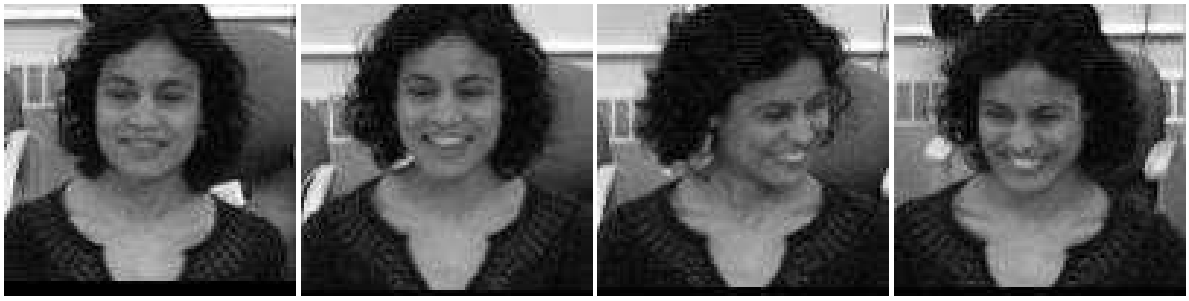

FIG. 14 Sample frames 1, 9, 40, 72 of a probe video. One observes large differences from the gallery video. In this case, however, recognition *was* successful.

FIG. 15 Failure examples, where the galleries were not sufficient to recognize the subjects.