Probabilistic Human Recognition from Video*

Shaohua Zhou and Rama Chellappa

Center for Automation Research (CfAR)
Department of Electrical and Computer Engineering
University of Maryland, College Park, MD 20740
{shaohua, rama}@cfar.umd.edu

Abstract. This paper presents a method for incorporating temporal information in a video sequence for the task of human recognition. A time series state space model, parameterized by a *tracking state vector* and a *recognizing identity variable*, is proposed to simultaneously characterize the kinematics and identity. Two *sequential importance sampling* (SIS) methods, a brute-force version and an efficient version, are developed to provide numerical solutions to the model. The joint distribution of both state vector and identity variable is estimated at each time instant and then propagated to the next time instant. *Marginalization* over the state vector yields a robust estimate of the posterior distribution of the identity variable. Due to the propagation of identity and kinematics, a *degeneracy* in posterior probability of the identity variable is achieved to give improved recognition. This evolving behavior is characterized using changes in *entropy*. The effectiveness of this approach is illustrated using experimental results on low resolution face data and upper body data.

1 Introduction

Human recognition has been an extensive research area for a long time, using a variety of human biometrics like fingerprint, retinal or iris scan, face, and human body to name a few. The problem of human recognition system can be stated as follows: given a gallery of still templates and a probe, either a still image or a video sequence containing a certain biometric, determine the closest template in the gallery to the one in the probe, or rank the templates in the gallery according to a similarity measure.

Significant research has been conducted on still-to-still human recognition [12,7] where the probe is a still image. Usually, an abstract representation of an image after a suitable geometric and photometric registration is formed using the various approaches and then recognition is performed based on this new representation.

However, research efforts using video as the probe, or still-to-video recognition, are relatively few due to the following challenges [16] in typical applications like surveillance and access control: poor video quality, large illumination and

* This work was completed with the support of the DARPA HumanID Grant N00014-00-1-0908. All correspondences are addressed to shaohua@cfar.umd.edu.

A. Heyden et al. (Eds.): ECCV 2002, LNCS 2352, pp. 681–697, 2002.

pose variations, and low image resolution. Most video-based recognition systems [2] do the following: the face or the body is first detected and then tracked over time. Only when a frame satisfying certain criteria (size, pose, etc) is acquired, recognition is performed using a still-to-still recognition technique. For this, the biometric part (face, whole body, etc) is cropped from the frame and transformed or registered with appropriate parameters.

There are several unresolved issues in the above recognition systems: 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 that selects several good frames to perform recognition per frame, then votes on these recognition results for a final solution is rather *ad hoc*.

In this paper, we attempt to overcome these difficulties with a unified probabilistic framework. We formulate human recognition as a hypothesis-testing problem and derive the posterior probability of each hypothesis given the observation, which contains a transformed and noise-corrupted version of one hypothesis. To accommodate the above requirement and video sequences, we propose a time series state space model that is parameterized by a tracking state vector (e.g. affine transform parameter) and an identity variable. The model consists of three equations:

- a *state equation* governing the temporal behavior of the kinematics vector,
- an *identity equation* governing temporal evolution of the identity variable,
- an *observation equation* at each time instant.

Using the SIS technique, the joint distribution of state vector and identity variable is estimated at each time instant and then propagated to the next time instant governed by the state and identity equations. The marginal distribution of the identity variable is estimated to provide a recognition result.

There is no need for selecting good video frames in this framework. Ultimately, the two tasks, namely discriminating the identity and determining the transformation parameter, are unified and solved here. However, a face detector is still needed to provide the prior distribution for the state vector.

In the following, Section 2 summarizes some related work in the literature. Section 3 introduces the mathematical framework and establishes the time-evolving behavior of posterior probability of identity variable based on some minor assumptions. Section 4 presents the SIS technique and the actual recognition algorithms. Section 5 presents and discusses experimental results, and Section 6 presents conclusions.

2 Related Work

There are numerous papers in the literature on recognizing human using biometrics. Interested readers may refer to [7] for a general review of statistical pattern recognition, and [16] for a recent survey on face recognition.

Probabilistic visual tracking in video sequences has recently gained significant attention. Generally, a state space model is applied to accommodate the

dynamics of a video sequence. The task of visual tracking is reduced to solving the posterior distribution of the state vector given an observation. Isard and Blake [6] proposed the CONDENSATION algorithm to track an object in a cluttered environment. In their work, the object is represented by a robust active contour. A near real-time performance and a high tracking accuracy are reported. However, only the tracking problem is considered.

This technique has been extended to perform other tasks beside tracking. Li and Chellappa [9] performed simultaneous tracking and verification via sequential posterior estimation. At each time instant, the posterior probability of the state vector is evaluated using the SIS technique. The verification probability is computed on a proper region of the state space. They presented experimental results on both synthetic data and real sequences (some using face information as well). Our method is somewhat similar to this approach, but there are significant differences from it. We will highlight them in the discussion part in Sec. 5.

Black and Jepson [1] use a CONDENSATION-based algorithm to match temporal trajectories. Models of temporal trajectories, such as gestures and facial expressions, are trained beforehand, and are gradually matched against the human motion in a new image sequence. The joint posterior distribution of model selection, local stretching, scaling, and position evolves as time proceeds.

Edwards et al [5] learn how individual faces vary through video sequences by decoupling sources of image variations such as expression, lighting and pose. The assumption that the identity remains the same throughout the video sequence is used.

Torre et al [14] propose a probabilistic framework to perform tracking and activity recognition using video. They first track the motion of rigid and non-rigid appearance, constrained with a mixture of pre-trained Hidden Markov Models (HMM) [13] with one model for each activity. After tracking, activity recognition is performed using standard Viterbi algorithm [13].

In [10], recognition of face over time is implemented by constructing a face identity surface. The face is first warped to a frontal view, and its KDA (Kernel Discriminant Analysis) features are used to form a trajectory. It is shown that the recognitive evidence accumulates over time along the trajectory. However, this recognition algorithm is deterministic.

3 Mathematical Framework

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

3.1 A Time Series State Space Model for Recognition

We will use the following mathematical notations:

- An image I. It could be represented using raw intensity values on the image region R, i.e. $I(R)$, or abstract features extracted from $I(R)$.

- A transformed version of image I, $f(I;\theta)$, with *state vector* $\theta \in \Theta$, a continuous *state space*. The transform can be either geometric or photometric or both.
- A gallery set $H = \{I_1, I_2, \ldots, I_N)\}$, indexed by an *identity variable* $n \in \mathcal{N} = \{1, 2, \ldots N\}$. I_n is the still template for the identity n, which may be registered beforehand or not.

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

1. State equation:
$$\theta_t = g(\theta_{t-1}) + u_t; \quad t \geq 1, \tag{1}$$
where θ_t is the *state vector* and u_t the *state noise* respectively at time instant t. If $g(.)$ is an identity function, it is a Brownian motion model.
2. Identity equation:
$$n_t = n_{t-1}; \quad t \geq 1, \tag{2}$$
assuming that identity does not change as time proceeds.
3. Observation equation:
$$f(y_t;\theta_t) = I_{n_t} + v_t; \quad t \geq 1, \tag{3}$$
where y_t is the *observation* and v_t the *observation noise* respectively at time instant t.
4. Prior distributions:
$$p(\theta_0|y_0), p(n_0|y_0). \tag{4}$$
5. Noise distributions:
$$p(u_t) \text{ or } p(\theta_t|\theta_{t-1}), p(v_t) \text{ or } p(y_t|n_t,\theta_t); \quad t \geq 1, \tag{5}$$
where $p(y_t|n_t,\theta_t)$ is the likelihood.
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 $\perp$ implies statistical independence.

Given this model, our goal is to compute the posterior probability $p(n_t|y_{0:t})$, where $y_{0:t} = \{y_0, \ldots, y_n\}$. It is in fact a probability mass function (PMF), as well as a marginal probability of $p(n_t, \theta_t|y_{0:t})$. Therefore, the problem is reduced to computing the posterior probability.

3.2 The Posterior Probability of Identity Variable

The evolution of the posterior probability $p(n_t|y_{0:t})$ as time proceeds is very interesting since the identity variable does not change by assumption, i.e., $p(n_t|n_{t-1}) = \delta(n_t - n_{t-1})$.

By time recursion, Markov properties and statistical independence embedded in the model, we can easily derive:

$$\begin{aligned} p(n_{0:t}, \theta_{0:t}|y_{0:t}) &= p(n_{0:t-1}, \theta_{0:t-1}|y_{0:t-1}) \frac{p(y_t|n_t, \theta_t)p(n_t|n_{t-1})p(\theta_t|\theta_{t-1})}{p(y_t|y_{0:t-1})} \\ &= p(n_0, \theta_0|y_0) \prod_{s=1}^{t} \frac{p(y_s|n_s, \theta_s)p(n_s|n_{s-1})p(\theta_s|\theta_{s-1})}{p(y_s|y_{0:s-1})} \\ &= p(n_0|y_0)p(\theta_0|y_0) \prod_{s=1}^{t} \frac{p(y_s|n_s, \theta_s)\delta(n_s - n_{s-1})p(\theta_s|\theta_{s-1})}{p(y_s|y_{0:s-1})}. \end{aligned} \quad (7)$$

Therefore, by marginalizing over $\theta_{0:t}$ and $n_{0:t-1}$, we get

$$p(n_t = n|y_{0:t}) = p(n|y_0) \int_{\theta_0} \dots \int_{\theta_t} p(\theta_0|y_0) \prod_{s=1}^{t} \frac{p(y_s|n, \theta_s)p(\theta_s|\theta_{s-1})}{p(y_s|y_{0:s-1})} d\theta_t \dots d\theta_0. \quad (8)$$

Thus $p(n_t = n|y_{0:t})$ is determined by the prior distribution $p(n_0 = n|y_0)$ and the product of the likelihood functions, $\prod_{s=1}^{t} p(y_s|n, \theta_s)$. If a uniform prior is assumed as below, then $\prod_{s=1}^{t} p(y_s|n, \theta_s)$ is the only determining factor.

Suppose that, (A) the prior probability for each hypothesis is same,

$$p(n_0|y_0) = 1/N; \quad n_0 \in \mathcal{N}, \quad (9)$$

and (B) for the correct hypothesis $c \in \mathcal{N}$, there exists a constant $K > 1$ such that,

$$p(y_t|c, \theta_t) \geq K * p(y_t|n_t, \theta_t); \quad t \geq 1, n_t \in \mathcal{N}, \ \& \ n_t \neq c, \quad (10)$$

where K is a lower bound on the likelihood ratio.

Substitution of Eqns. 9 and 10 into Eqn. 8 gives

$$p(c|y_{0:t}) \geq K^t * p(n_t|y_{0:t}); \quad n_t \in \mathcal{N}, \ \& \ n_t \neq c, \quad (11)$$

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

$$p(c|y_{0:t}) \geq p(c|y_{0:t-1}). \quad (12)$$

In other words, the posterior probability of the correct hypothesis is nondecreasing as time proceeds.

More interestingly, from Eqn. 11, we have

$$(N-1)p(c|y_{0:t}) \geq K^t * \sum_{n_t=1, n_t \neq c}^{N} p(n_t|y_{0:t}) = K^t * (1 - p(c|y_{0:t})), \quad (13)$$

i.e.,

$$p(c|y_{0:t}) \geq \frac{K^t}{K^t + N - 1}. \quad (14)$$

Since $K > 1$ and $p(c|y_{0:t}) \leq 1$,

$$\lim_{t\to\infty} p(c|y_{0:t}) = 1, \tag{15}$$

implying that degeneracy will be reached in the correct identity variable 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 order to satisfy condition (B) for all the frames in the sequence. Fortunately, as we will see in the experiment in Sec. 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 Sec. 5.

To change a viewing angle, we use the notation of entropy [3]. Given a PMF $p(x); x \in \mathcal{N}$, the entropy is defined as:

$$H(x) = -\sum_{x\in\mathcal{N}} p(x)\log_2 p(x). \tag{16}$$

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:

$$H(x|y) = -\sum_{y} p(y)\sum_{x} p(x|y)\log_2 p(x|y). \tag{17}$$

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

$$H(n_t|y_{0:t}) = -\sum_{n_t\in\mathcal{N}} p(n_t|\tilde{y}_{0:t})\log_2 p(n_t|\tilde{y}_{0:t}). \tag{18}$$

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

4 SIS Algorithms

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 are then proposed to give an approximate analytic solution. Recently, a sequential Monte Carlo method [6,4,8,11] has been used to provide a numerical solution for propagating an arbitrary distribution.

4.1 Importance Sampling

The key idea of Monte Carlo method is that an arbitrary probability distribution $\pi(x)$ can be represented closely by a set of discrete samples. It is ideal to draw i.i.d. samples $\{x^{(m)}\}_{m=1}^{M}$ from $\pi(x)$. However it 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, then a weight $w^{(m)} = \pi(x^{(m)})/g(x^{(m)})$ is assigned to each sample. In the ideal case, the importance function $g(x)$ is $\pi(x)$ itself, each sample has its weight 1. This technique is called *Importance Sampling* (IS). To accommodate a video, importance sampling is used in a sequential fashion, leading to *Sequential Importance Sampling* (SIS). For a complete SIS recipe, refer to [4,11]. We now introduce the following definition and two propositions.

Definition 1: A set of weighted random samples $\mathcal{S} = \{(x^{(m)}, w^{(m)})\}_{m=1}^{M}$ is **proper** with respect to $\pi(x)$ if for any integrable function $h(x)$,

$$\lim_{M\to\infty} \frac{\sum_{m=1}^{M} h(x^{(m)})w^{(m)}}{\sum_{m=1}^{M} w^{(m)}} = E_{\pi}[h(x)]. \tag{19}$$

Proposition 1: When $\pi(x)$ is a PMF defined on a finite sample space, the proper sample set should exactly include 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 the 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 weight $w^{(m)}$ belonging to the same sample $y^{'(k)}$ to obtain the weight $w^{'(k)}$, i.e.,

$$w^{'(k)} = \sum_{m=1, y^{(m)}=y^{'(k)}}^{M} w^{(m)}. \tag{20}$$

Proof:
By Definition 1, for any integrable function $h(x, y)$, we have

$$\lim_{M\to\infty} \frac{\sum_{m=1}^{M} h(x^{(m)}, y^{(m)})w^{(m)}}{\sum_{m=1}^{M} w^{(m)}} = E_{\pi}[h(x, y)]. \tag{21}$$

By taking the above $h(x, y)$ as a function of only y, $g(y) = \int h(x, y)dx$, we get

$$\lim_{M\to\infty} \frac{\sum_{m=1}^{M} g(y^{(m)})w^{(m)}}{\sum_{m=1}^{M} w^{(m)}} = E_{\pi}[g(y)]. \tag{22}$$

However, in the above equation, some $y^{(m)}$ might be of same value. It is necessary to merge them to arrive at a new sample set $\{(y^{'(k)}, w^{'(k)})\}_{k=1}^{K}$, where $w^{'(k)}$ representing the weights belonging to $y^{'(k)}$, i.e., using Eqn. 20.

4.2 Algorithms and Their Efficiency

In the context of this framework, the posterior probability $p(n_t, \theta_t|y_{0:t})$ is represented by a set of *indexed and weighted* samples

$$\mathcal{S}_t = \{(n_t^{(m)}, \theta_t^{(m)}, w_t^{(m)})\}_{m=1}^{M} \tag{23}$$

with n_t as 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 $p(n_t|y_{0:t})$. We apply the following algorithm to solve this problem. Algorithm I shown in Fig. 1 is similar to CONDENSATION [6] for computing the joint distribution $p(n_t, \theta_t|y_{0:t})$.

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. Algorithm I

Algorithm I 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. Hence, we need $M = J * N$ samples in total. Further assume that one resampling step takes T_r unit time (ut), one predicting step T_p ut, one updating step T_u ut, the normalizing step T_n ut, and the marginalizing step T_m ut. In the updating step, there are two substeps: computing the transformed image $f(y_t; \theta_t)$ and evaluating the likelihood $p(y_t|n_t, \theta_t)$, taking T_t and T_l ut respectively . Obviously, the total computation is $J * N * (T_r + T_p + T_t + T_l) + T_n + T_m$ ut to deal with one video frame. It is well known that computing the transformed image is much more expensive than other operations, i.e., $T_t >> max(T_r, T_p, T_l)$. Therefore, as the number of templates N grows, the computational load increase dramatically.

To avoid the 'brute force' method in Algorithm I, we develop an efficient Algorithm II, which is motivated by the fact that the sample space $\mathcal{N}$ is countable.

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

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

In Algorithm II, we in fact use the following notation for the sample set,

$$\mathcal{S}_t = \{(\theta_t^{(j)}, w_t^{(j)}, w_{t,1}^{(j)}, w_{t,2}^{(j)}, \ldots, w_{t,N}^{(j)})\}_{j=1}^{J}, \tag{24}$$

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, 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 $\theta_t^{(j)}$'s, governed by $p(\theta_t | \theta_{t-1}^{'(j)})$. Finally, the propagation of the joint distribution requires the weighting step as follows:

$$w_{t,n}^{(j)} = w_{t-1,n}^{'(j)} * p(y_t | n, \theta_t^{(j)}). \tag{25}$$

The core of Algorithm II lies in that, instead of propagating random samples on both state vector and identity variable, we can keep the samples on the identity variable fixed and let those on the state vector random. Also, we propagate the marginal distribution for tracking the state vector, but we still propagate the joint distribution for recognition purposes. Algorithm II is summarized in Fig. 2.

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. Algorithm II

Compared with Algorithm I, Algorithm II is more efficient and accurate. The total computation of Algorithm II is $J*(T_r+T_p+T_t)+J*N*T_l+T_n+T_m\ ut$, a tremendous saving over Algorithm I 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 Algorithm I: all samples belonging to identity $n_t = c$, where c is the correct hypothesis, are drawn with a bias in θ_t. More specifically, though $p(y_t|c,\theta_t) > p(y_t|n_t,\theta_t)$ holds provided that the same θ_t is applied in both sides, different θ_t's can be chosen in a bias to disobey this inequality. However, Algorithm II elegantly eliminates such a bias.

5 Experiments and Discussions

5.1 Experimental Results

We use video sequences with subjects walking towards a camera in order to simulate typical scenarios like in visual surveillance. There are 30 subjects, each having one face template and one upper body template. The face gallery and the upper body gallery are as shown in Fig. 3. The probe set contains 30 video sequences, one for each subject. Fig. 4 gives some example frames in one probe video. Note that both galleries are captured under different circumstances from the probe and that the probe has considerable change in scale. These images were collected, as part of the HumanID project, by National Institute of Standards and Technology and University of South Florida researchers.

Model parameters used in the experiments are chosen as follows.

1. Image representation. A reconstructed image from top 300 principal components or eigenfaces [15] is used to represent the face, while the raw intensity array is used to represent the upper body. Fig. 5 shows the top 10 eigenfaces.

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. It is a reasonable approximation since there is no significant out-of-plane motion in the scenario that the subject walks towards the camera. Regarding photometric transformation, only histogram equalization, a typical but coarse method dealing for enhancement, is adopted for face gallery and no preprocessing operation is performed for body gallery.

3. Prior distribution $p(\theta_0|y_0)$ is Gaussian, whose mean comes from the initial detector and whose covariance matrix is manually specified.

4. Noise distribution $p(u_t)$ or $p(\theta_t|\theta_{t-1})$ is Gaussian, whose mean and covariance matrix are manually specified. Also function $g(.)$ in Eqn. 1 is assumed to be an identity function. Furthermore, $p(u_t)$ is not time-varying. This is essentially a constant-velocity model. 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, causing the constant-velocity model to be not optimal. However, experimental results show that as long as the samples of θ can cover the motion, this model can be applied for simplicity.

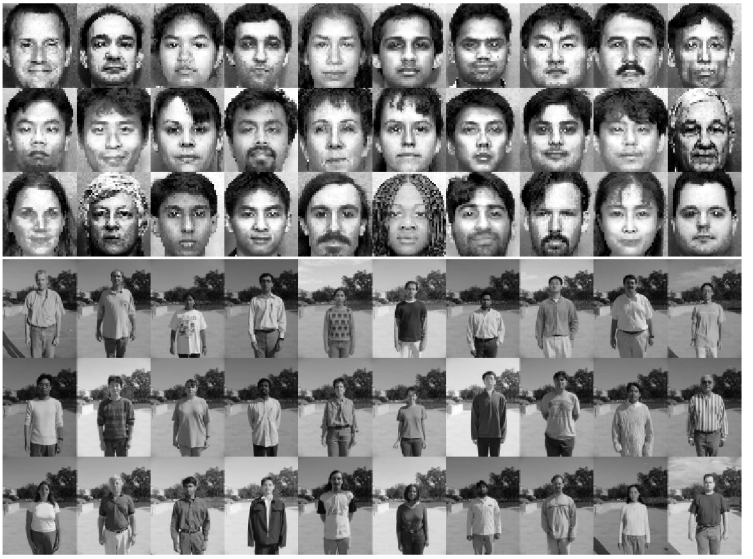

Fig. 3. The face and body galleries. The face image size is 48x42 and the body image size is 100x75.

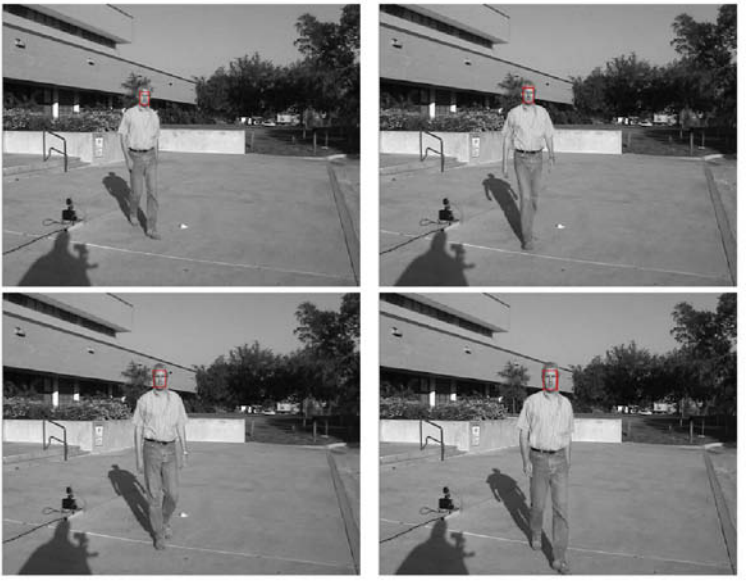

Fig. 4. Example frames in one probe video. The image size is 720x480 while the actual face size ranges approximately from 20x20 in the first frame to 60x60 in the last frame.

Fig. 5. The top 10 eigenfaces.

5. Noise distribution $p(v_t)$ or the likelihood $p(y_t|n_t, \theta_t)$ is a 'truncated' Laplacian:

$$p(y_t|n_t, \theta_t) = \begin{cases} L * \exp(-\|v_t\|/\sigma) & \text{if } \|v_t\| \leq \lambda\sigma \\ L * \exp(-\lambda) & \text{if } \|v_t\| > \lambda\sigma \end{cases} \tag{26}$$

where $\|I(R)\| = \sum_{r \in R} |I(r)|$, σ and λ are manually specified, and L is a normalizing constant. Furthermore, $p(v_t)$ is not time-varying. Gaussian distribution is widely used as a noise model, accounting for sensor noise, digitization noise, etc. However, given the observation equation: $v_t = f(y_t; \theta_t) - I_{n_t}$, 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 residual of natural images is more Laplacian-like. The 'truncated' Laplacian is used to give a 'surviving' chance for samples to accommodate abrupt motion changes.

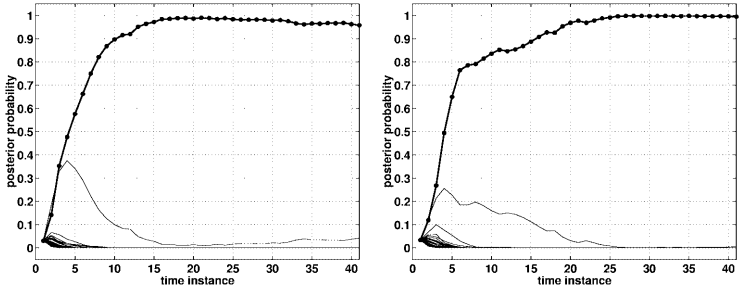

Fig. 6. Posterior probability $p(n_t|y_{0:t})$ against time instant t. Left: Algorithm I. Right: Algorithm II.

For Algorithm I and II, Fig. 6 presents the plot of the posterior probability $p(n_t|y_{0:t})$ against frame index t. Fig. 7 presents the conditional entropy $H(n_t|y_{0:t})$ against t and the MMSE estimate of the scale parameter a_1 against t, both obtained using Algorithm II. In Fig. 4, the tracked face is superimposed on the image using a bounding box.

Suppose the correct identity for Fig. 4 is c. From Fig. 6, we can easily observe that the posterior probability $p(c|y_{0:t})$ increases as time proceeds and eventually approaches 1, and all others $p(n_t \neq c|y_{0;t})$ go to 0 finally. Evidenced in Fig. 7 is the decreasing conditional entropy $H(n_t|y_{0:n})$ and the increasing scale parameter, which matches with the scenario: a subject walking towards a camera.

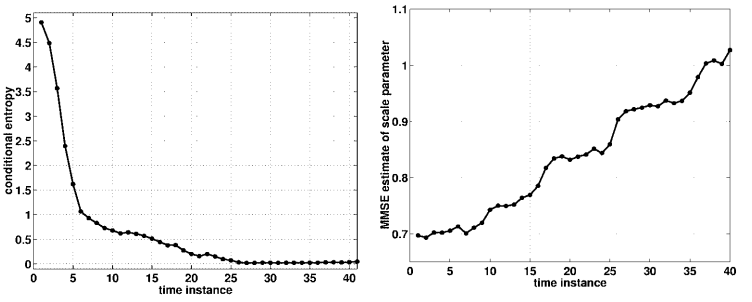

Fig. 7. Left: conditional entropy $H(n_t|y_{0:t})$ against t. Right: MMSE estimate of a_1 against t. Both are obtained using Algorithm II.

Table 1 summarizes the average recognition performance and computational time of two algorithms. As far as performance is concerned, there is no remarkable difference between the two algorithms, with Algorithm II giving slightly better results. However, there are big differences between two galleries. The reasons are summarized as follows: (i) the probe video is taken outside with the sunshine casting a strong shadow on the face while the face gallery is taken inside and the body gallery taken outside; (ii) the subjects are wearing the same dress for the probe and body galleries; and (iii) the body template is bigger than the face template. Fig. 8 shows a list of facial images cropped by hand from probe videos and then normalized. We then perform still-to-still face recognition using the eigenface approach [15], the recognition result is less than 30% for the top one match and less than 50% for top three matches.

Obviously, algorithm II is more efficient than Algorithm I. It is about 10 times faster than Algorithm I as shown in Table I. Note that this experiment is implemented in C++ on a PC with P-III650 CPU and 512M RAM with the number of motion samples J chosen to be 200, the number of templates in the gallery N to be 30.

Table 1. Summary of Algorithms I and II.

Algorithm	I		II	
Gallery	Face	Body	Face	Body
Probability of correct match as the top match	47%	93%	50%	93%
Probability of correct match within top 3 matches	70%	93%	77%	93%
Probability of correct match within top 5 matches	83%	100%	87%	100%
Time per frame	22s		2.1s	

Fig. 8. Facial images cropped from probe videos and then normalized.

5.2 Discussions and Future Work

The following issues are worthy of investigation in the future.

1. The assumption that the 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. Consider this example: Initially, the most likely identity might be not the desired one because of low-resolution of the data or some artifacts; then, the desired one may pop up. Under our assumption that the identity remains the same, the incorrect choice will grab most of the samples initially. But, having nonzero transition probabilities will enable us an easier transition to the correct choice, making this algorithm more robust.

2. Choice of likelihood distribution $p(y_t|n_t, \theta_t)$ and condition (B) . In general, the smaller $\|v_t\|$ is, the higher the likelihood $p(y_t|n_t, \theta_t)$ and the higher the posterior $p(n_t|y_{0:t})$. In this sense, an accurate solution to this problem is determined by the basic problem: how can we find an efficient distance metric or how good is the condition (B) satisfied? Fig. 9 plots, against the logarithm of the scale parameter, the 'average' likelihood of the correct hypothesis, $\frac{1}{N}\sum_{n\in\mathcal{N}} p(I_n|n,\theta)$, and that of the incorrect hypotheses, $\frac{1}{N(N-1)}\sum_{m\in\mathcal{N},n\in\mathcal{N},m\neq n} p(I_m|n,\theta)$, 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' that the condition (B) is well satisfied. Therefore, the success of SIS algorithm depends on how good the samples lie in a similar 'band' in the high-dimensional affine space. Also, the lower bound K in condition (B) is too strict. If we take the mean of the 'average' likelihood ratio shown in Fig. 9 as an estimate of K (roughly 1.5), Eqn. 14 tells that, after 20 frames, the probability $p(c|y_{0:t})$ reaches 0.99! However, this is not reached in the experiments due to noise in the observations and incomplete parameterization of transformations.

3. Resampling. In Algorithm II, 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 problem in principal 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

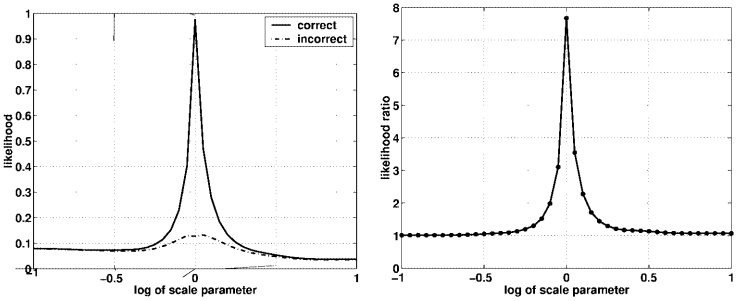

Fig. 9. Left: the 'average' likelihood of the correct hypothesis and incorrect hypotheses against the log of scale parameter. Right: The 'average' likelihood ratio against the log of scale parameter.

for future states as system evolves'[11]. Resampling scheme can be either simple random sampling with weights (like in CONDENSATION), residual sampling, or local Monte Carlo methods. Thus, this can be considered as a special resampling strategy which also amplifies samples with high weights.

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

5. Mutual dependence of tracking and recognition. Since joint posterior distribution is computed each time, the mutual dependence is obvious. If tracking fails, the recognition is meaningless. If recognition is poor, for instant, some background region in the video might be more favored than the biometric region according to the distance measure, making the tracker stick to the background. In fact, one reason for the low performance of face gallery is this kind of tracking failure. We are now developing an algorithm which cleverly splits the tracking and recognition tasks, but still uses the idea of propagation of posterior probability for recognition.

6. Now we highlight the differences from Li and Chellappa's approach [9]. In [9], basically only the tracking state vector is parameterized in the state-space model. The identity is involved only in the initialization step to rectify the template onto the first frame of the sequence. However, in our approach both tracking state vector and identity variables are parameterized in the state-space model, which offers us one more degree of freedom and leads to a different approach for deriving the solution. The SIS technique is applied in both approaches to nu-

merically approximate the posterior probability given the observation. Again in [9], it is the posterior probability of state vector and the verification probability is 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 state vector and identity variable and the posterior probability of identity variable is just a free estimate by marginalizing over the state vector. Note that there is no time propagation of verification probability in [9] while we always propagate the joint density. One consequence is that we guarantee that $\sum_{n_t \in \mathcal{N}} p(n_t|y_{0:t}) = 1$, but there is no such guarantee in [9]. Their approach 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 indicates an immediate decision while no such decision could be readily made from the verification probability in [9]. 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 absorbed to the noisy background, while our approach is more robust since we consider all templates at the same time.

6 Conclusion

In this paper, a time series state space model is proposed to solve the two tasks of tracking and recognition simultaneously. This probabilistic framework, which overcomes many difficulties arising in conventional recognition approaches using video, is registration-free and poses no need for selecting good frames. More importantly, temporal information is elegantly exploited.

However, this model is nonlinear and non-Gaussian, with potential nonexistence of an analytic solution. Two algorithms based on the SIS technique have been applied to propagate the posterior distribution. Algorithm I, an extension of CONDENSATION, which is not so efficient in computation, while Algorithm II improves the computational efficiency and accuracy as well by incorporating the inherent property of this mixed model, namely the discreteness of identity variable. 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 for the convergence.

A final remark is that this framework is not limited to recognizing only face and upper body, and is easily extended to other recognition tasks from video.

References

1. M. J. Black and A. D. Jepson. A probabilistic frameworkfor matching temporal trajectories. *Proc. of ICCV*, pages 176–181, 1999.
2. T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. *Proc. of Intl. Conf. on Audio- and Video-Based Person Authentication*, pages 176–181, 1999.

3. T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.
4. A. Doucet, S. J. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–209, 2000.
5. T. J. Edwards, C. J. Taylor, and T. F. Cootes. Improving identification performance by intergrating evidence from sequences. *Proc. of CVPR*, pages 486–491, 1999.
6. M. Isard and A. Blake. Contour tracking by stochatic propagation of conditional density. *Proc. of ECCV*, 1996.
7. A. K. Jain, R. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. PAMI*, 22:4–37, 2000.
8. G. Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *J. Computational and Graphical Statistics*, 5:1–25, 1996.
9. Baoxin Li and R. Chellappa. Simultaneous tracking and verification via sequential posterior estimation. *Proc. of CVPR*, pages 110–117, 2000.
10. Yongmin Li, Shaogang Gong, and H. Liddell. Constucting structures of facial identities on the view sphere using kernel discriminant analysis. *Proc. of the 2rd Intl. Workshop on SCTV*, 2001.
11. J. S. Liu and R. Chen. Sequential monte carlo for dynamic systems. *Journal of the American Statistical Association*, 93:1031–1041, 1998.
12. P. J. Philipps, H. Moon, S. Rivzi, and P. Ross. The feret testing protocol. *Face Recognition: From Theory to Applications*, 83:244–261, 1998.
13. L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of IEEE*, 77(2), 1989.
14. De la Torre, F. Y. Yacoob, and L. Davis. A probabilistic framework for rigid and non-rigid appearance-based tracking and recognition. *Proc. of 4th Intl. Conf. on Auto. Face and Gesture Recognition*, 2000.
15. M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neutoscience*, 3:72–86, 1991.
16. W. Y. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. *UMD CfAR Technical Report CAR-TR-948*, 2000.