Stan Z. Li
Anil K. Jain (Eds.)

Handbook of Face Recognition

Springer

Contents

Chapter 8. Face Tracking and Recognition from Video
Rama Chellappa, Shaohua Kevin Zhou 1

Chapter 8. Face Tracking and Recognition from Video[⋆]

Rama Chellappa and Shaohua Kevin Zhou

Department of Electrical and Computer Engineering and
Center for Automation Research
University of Maryland
College Park, MD 20742
{rama, shaohua}@cfar.umd.edu

While most face recognition algorithms take still images as probe inputs, this chapter presents a video-based face recognition approach that takes video sequences as inputs. Since the detected face might be moving in the video sequence, we inevitably have to deal with uncertainty in tracking as well as that in recognition. Rather than resolving these two uncertainties separately, our strategy is to perform simultaneous tracking and recognition of human faces from a video sequence.

In general, a video sequence is a collection of still images; so still-image-based recognition algorithms can always be applied. An important property of a video sequence, however, is its temporal continuity. While this property has been exploited for tracking, it has not been used for recognition. In this chapter, we systematically investigate how temporal continuity can be incorporated for video-based recognition.

Our probabilistic approach solves still-to-video recognition, where the gallery consists of still images and the probes are video sequences. 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 (SIS) algorithm is used to estimate the posterior distribution. Empirical results demonstrate that, due to the propagation of the identity variable over time, a degeneracy in posterior probability of the identity variable is achieved.

The organization of the chapter is as follows: Section 1 sets the framework for face recognition in video. Section 2 covers in detail all the components of the si-

[⋆] Partially supported by the DARPA/ONR Grant N00014-00-1-0908. We also thank Dr. Baback Moghaddam at Mitsubishi Electric Research Labs.

multaneous tracking and recognition approach. Section 3 presents some techniques for enhancing tracking and recognition accuracy via modeling inter-frame appearance and appearance changes between video frames and gallery images. Section 4 addresses future research issues and discussions.

1 Review

Probabilistic video analysis has recently gained significant attention in the computer vision community since the seminal work of Isard and Blake [1]. 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., $p(\theta_t|z_{0:t})$ 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. The CONDENSATION algorithm, also known as the particle filter, was originally proposed [2] in the signal processing literature and has been used to solve many other vision tasks [3, 4], including human face recognition [5]. In this chapter, we will systematically investigate the application of particle filter for face recognition in a video sequence.

Face recognition has been an active research area for a long time. Refer to [6, 7] for surveys and [8] for reports on experiments. Experiments reported in [8] evaluate still-to-still scenarios, where the gallery and the probe set consist of both still facial images. Some well-known still-to-still face recognition approaches include Principal Component Analysis [9], Linear Discriminant Analysis [10, 11], and Elastic Graph Matching [12]. Typically, recognition is performed based on an abstract representation of the face image after suitable geometric and photometric transformations and corrections.

Following [8], 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 as $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$, 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 relatively fewer due to the following challenges [7] in typical surveillance applications: poor video quality, significant illumination and pose variations, and low image resolution. Most existing video-based recognition systems ([13] and references in [7]) attempt the following: the face is first detected and then tracked over time. Only when a frame satisfying certain criteria (size, pose) is acquired, recognition is performed using still-to-still recognition technique. For this, the face part is cropped from the frame and transformed or registered using appropriate transformations. This approach attempts to resolve uncertainties in tracking and recognition sequentially and separately and needs a criteria for selecting good frames and estimation of parameters for registration. Also, still-to-still recognition does not effectively exploit temporal information.

To overcome these difficulties, we propose a tracking-and-recognition approach, which attempts to resolve uncertainties in tracking and recognition simultaneously in a unified probabilistic framework. To fuse temporal information, the 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 establishing a link between the motion vector and the identity variable.

Using the SIS [14, 15, 16] technique, the joint posterior distribution of the motion vector and the identity variable, i.e., $p(n_t, \theta_t|z_{0:t})$ is estimated at each time instant and then propagated to the next time instant governed by motion and identity equations. The marginal distribution of the identity variable, i.e., $p(n_t|z_{0:t})$, is estimated to provide the recognition result. An SIS algorithm is presented to approximate the distribution $p(n_t|z_{0:t})$ in the still-to-video scenario. It achieves computational efficiency over its CONDENSATION counterpart by considering the discrete nature of the identity variable.

It is worth emphasizing that (i) our model can take advantage of any still-to-still recognition algorithm [9, 10, 11, 12] by embedding distance measures used therein in our likelihood measurement; and (ii) it allows a variety of image representations and transformations. Section 3 presents an enhancement technique by incorporating more sophisticated appearance-based models. The appearance models are used for tracking (modeling inter-frame appearance changes) and recognition (modeling appearance changes between video frames and gallery images), respectively. Table 1 summarizes the proposed approach and others, in term of using temporal information.

Process	Operation	Use of temporal information
Visual tracking	Modeling the inter-frame differences	In tracking
Visual recognition	Modeling the difference between probe and gallery images	Not related
Tracking-then-recognition	Combining tracking and recognition sequentially	Only in tracking
Tracking-and-recognition	Unifying tracking and recognition	In both tracking and recognition.

Table 1. A summary of the proposed approach and others.

2 Simultaneous Tracking and Recognition from Video

In this section, we first present the details on the propagation model for recognition and discuss its impact on the posterior distribution of identity variable. We then proceed to solve the model using the SIS algorithms.

2.1 A Time Series State Space Model for Recognition

The recognition model consists of the following components:

- *Motion equation*
  In its most general form, the motion model can be written as

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

  where u_t is noise in the motion model, whose distribution determines the motion state transition probability $p(\theta_t|\theta_{t-1})$. The function $g(.,.)$ characterizes the evolving motion and it could be a function learned offline or given a priori. One of the simplest choice is an additive function, i.e., $\theta_t = \theta_{t-1} + u_t$, which leads to a first-order Markov chain.
  The choice of θ_t is dependent on the application. Affine motion parameters are often used when there is no significant pose variation available in the video sequence. However, if a 3-D face model is used, 3-D motion parameters should be used accordingly.
- *Identity equation*
  Assuming that the identity does not change as time proceeds, we have

  $$n_t = n_{t-1}; \quad t \geq 1, \tag{2}$$

  In practice, one may assume a small transition probability between identity variables to increase the robustness.
- *Observation equation*
  By assuming that the transformed observation is a noise-corrupted version of some still template in the gallery, the observation equation can be written as

  $$\mathcal{T}_{\theta_t}\{z_t\} = I_{n_t} + v_t; \quad t \geq 1, \tag{3}$$

  where v_t is observation noise at time t, whose distribution determines the observation likelihood $p(z_t|n_t, \theta_t)$, and $\mathcal{T}_{\theta_t}\{z_t\}$ is a transformed version of the observation z_t. This transformation could be either geometric or photometric or both. However, when confronting sophisticated scenarios, this model is far from sufficient. One should use a more realistic likelihood function as shown in Section 3.
- *Statistical independence*
  We assume statistical independence between all noise variables u_t's and v_t's.

- *Prior distribution*
  The prior distribution $p(n_0|z_0)$ is assumed to be uniform, i.e.,

$$p(n_0|z_0) = \frac{1}{N}; \quad n_0 = 1, 2, \ldots, N. \tag{4}$$

  In our experiments, $p(\theta_0|z_0)$ is assumed to be Gaussian, whose mean comes from an initial detector or a manual input and whose covariance matrix is manually specified.

Using an overall state vector $x_t = (n_t, \theta_t)$, Eq. (1) and (2) can be combined into one state equation (in a normal sense) which is completely described by the overall state transition probability

$$p(x_t|x_{t-1}) = p(n_t|n_{t-1})p(\theta_t|\theta_{t-1}) \ . \tag{5}$$

Given this model, our goal is to compute the posterior probability $p(n_t|z_{0:t})$. It is in fact a probability mass function (PMF) since n_t only takes values from $\mathcal{N} = \{1, 2, ..., N\}$, as well as a marginal probability of $p(n_t, \theta_t|z_{0:t})$, which is a mixed distribution. Therefore, the problem is reduced to computing the posterior probability.

The Posterior Probability of Identity Variable

The evolution of the posterior probability $p(n_t|z_{0:t})$ as time proceeds is very interesting to study as 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 zero, i.e. $\delta(x) = 1$ if $x = 0$; otherwise, $\delta(x) = 0$.

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

$$\begin{aligned} p(n_{0:t}, \theta_{0:t}|z_{0:t}) &= p(n_{0:t-1}, \theta_{0:t-1}|z_{0:t-1}) \frac{p(z_t|n_t, \theta_t)p(n_t|n_{t-1})p(\theta_t|\theta_{t-1})}{p(z_t|z_{0:t-1})} \\ &= p(n_0, \theta_0|z_0) \prod_{i=1}^{t} \frac{p(z_i|n_i, \theta_i)p(n_i|n_{i-1})p(\theta_i|\theta_{i-1})}{p(z_i|z_{0:i-1})} \\ &= p(n_0|z_0)p(\theta_0|z_0) \prod_{i=1}^{t} \frac{p(z_i|n_i, \theta_i)\delta(n_i - n_{i-1})p(\theta_i|\theta_{i-1})}{p(z_i|z_{0:i-1})}. \end{aligned} \tag{6}$$

Therefore, by marginalizing over $\theta_{0:t}$ and $n_{0:t-1}$, we obtain the marginal posterior distribution for the identity j,

$$p(n_t = j|z_{0:t}) = p(n_0 = j|z_0) \int_{\theta_0} \cdots \int_{\theta_t} p(\theta_0|z_0) \prod_{i=1}^{t} \frac{p(z_i|j, \theta_i)p(\theta_i|\theta_{i-1})}{p(z_i|z_{0:i-1})} d\theta_t \ldots d\theta_0. \tag{7}$$

Thus $p(n_t = j|z_{0:t})$ is determined by the prior distribution $p(n_0 = j|z_0)$ and the product of the likelihood functions, $\prod_{i=1}^{t} p(z_i|j, \theta_i)$. If a uniform prior is assumed, then $\prod_{i=1}^{t} p(z_i|j, \theta_i)$ is the only determining factor.

If we further assume that, for the correct identity $l \in \mathcal{N}$, there exists a constant $\eta > 1$ such that,

$$p(z_t|n_t = l, \theta_t) \geq \eta p(z_t|n_t = j, \theta_t); \quad t \geq 1, j \in \mathcal{N}, j \neq l, \tag{8}$$

we have been able to show [30] that the posterior probability for the correct identity l, $p(n_t = l|z_{0:t})$, is lower-bounded by an increasing curve which converges to 1.

To measure the evolving uncertainty remaining in the identity variable as observations accumulate, we use the notion of entropy [18]. In the context of this problem, conditional entropy $H(n_t|z_{0:t})$ is used. However, the knowledge of $p(z_{0:t})$ is needed to compute $H(n_t|z_{0:t})$. We assume that it degenerates to an impulse at 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,

$$H(n_t|z_{0:t}) = -\sum_{n_t \in \mathcal{N}} p(n_t|\tilde{z}_{0:t}) \log_2 p(n_t|\tilde{z}_{0:t}). \tag{9}$$

We expect that $H(n_t|z_{0:t})$ decreases as time proceeds since we start from an equiprobable distribution to a degenerate one.

2.2 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) prior probability $p(x_0)$ and statistical independence among all noise variables. We wish to compute the posterior probability $p(x_t|z_{0:t})$.

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, extended Kalman filter and its variants have been used to arrive at an approximate analytic solution [19]. Recently, the SIS technique, a special case of Monte Carlo method, [16, 14, 15] has been used to provide a numerical solution and propagate an arbitrary distribution over time.

Importance Sampling

The essence of Monte Carlo method is to represent an arbitrary probability distribution $\pi(x)$ 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

$$w^{(m)} = \pi(x^{(m)})/g(x^{(m)}) \tag{10}$$

is assigned to each sample. This technique is called Importance Sampling. It can be shown[15] 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

$$w_t = w_{t-1}p(z_t|x_t)p(x_t|x_{t-1})/g(x_t|x_{t-1}). \tag{11}$$

In the CONDENSATION algorithm, $g(x_t|x_{t-1})$ is taken to be $p(x_t|x_{t-1})$ and Eq. (11) becomes

$$w_t = w_{t-1}p(z_t|x_t), \tag{12}$$

In fact, Eq. (12) 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 [15, 16].

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 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.,

$$w'^{(k)} = \sum_{m=1}^{M} w^{(m)} \quad \delta(y^{(m)} - y'^{(k)}). \tag{13}$$

Algorithms and Computational Efficiency

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

$$\mathcal{S}_t = \{(n_t^{(m)}, \theta_t^{(m)}, w_t^{(m)})\}_{m=1}^{M} \tag{14}$$

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 $p(n_t|z_{0:t})$.

A straightforward implementation of the CONDENSATION algorithm 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. Hence, we need $M = J * N$ samples in total. Further assume that one resampling step takes T_r seconds (s), one predicting step T_p s, computing one transformed image T_t s, evaluating likelihood once T_l s, one updating step T_u s. Obviously, the bulk of computation is $J * N * (T_r + T_p + T_t + T_l)$ s to deal with one video frame as the computational time for the normalizing step and the marginalizing step is negligible. 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 increases dramatically.

There are various approaches in the literature for reducing the computational complexity of the CONDENSATION algorithm. In [20], random particles are guided by deterministic search. Assumed density filtering approach [21], different from CONDENSATION, is even more efficient. Those approaches are general and do not explicitly exploit the special structure of the distribution in this setting: a mixed distribution of continuous and discrete variables. To this end, we propose the following algorithm.

As the sample space $\mathcal{N}$ is countable, 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 $\theta_t^{(j)}$,

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

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

$$\mathcal{S}_t = \{(\theta_t^{(j)}, w_t^{(j)}, w_{t,1}^{(j)}, w_{t,2}^{(j)}, \ldots, w_{t,N}^{(j)})\}_{j=1}^{J}, \tag{15}$$

with $w_t^{(j)} = \sum_{n=1}^{N} w_{t,n}^{(j)}$. The proposed algorithm is summarized in Fig. 1.

Thus, instead of propagating random samples on both motion vector and identity variable, we can keep the samples on the identity variable fixed and let those on the motion vector be random. Although we propagate only the marginal distribution for motion tracking, we still propagate the joint distribution for recognition purposes.

The bulk of computation of the proposed algorithm is $J*(T_r+T_p+T_t)+J*N*T_l$ s, a tremendous improvement over the traditional CONDENSATION algorithm when dealing with a large database since the majority computational time $J * T_t$ does not depend on N.

2.3 Experimental Results

In this section we describe the still-to-video scenarios used in our experiments and model choices, followed by a discussion of results. Two databases are used in the still-to-video experiments.

Database-0 was collected outside a building. We mounted a video camera on a tripod and requested subjects to walk straight towards the camera in order to simulate typical scenarios in visual surveillance. Database-0 includes one face gallery,

```
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 the sample by drawing $(\theta_t^{(j)})$ from $p(\theta_t|\theta_{t-1}^{'(j)})$.
    Compute the transformed image $\mathcal{T}_{\theta_t^{(j)}}\{z_t\}$.
    For $n = 1, \ldots, N$
      Update the 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 $\theta_t$ to obtain weight $\beta_{n_t}$ for $n_t$.
End
```

Fig. 1. The computationally efficient SIS algorithm.

and one probe set. The images in the gallery are listed in Fig. 2. The probe contains 12 videos, one for each individual. Fig. 2 gives some frames in a probe video.

In Database-1, we have video sequences with subjects walking in a slant path towards the camera. There are 30 subjects, each having one face template. The face gallery is shown in Fig. 3. The probe contains 30 video sequences, one for each subject. Fig. 3 gives some example frames extracted from one probe video. As far as imaging conditions are concerned, the gallery is very different from the probe, especially in lighting. This is similar to the 'FC' test protocol of the FERET test [8]. These images/videos were collected, as part of the HumanID project, by National Institute of Standards and Technology and University of South Florida researchers.

Table 2 summaries the features of the two databases.

Database	Database-0	Database-1
No. of subjects	12	30
Gallery	Frontal face	Frontal face
Motion in probe	Walking straight towards the camera	Walking straight towards the camera
Illumination variation	No	Large
Pose variation	No	Slight

Table 2. Summary of the two databases.

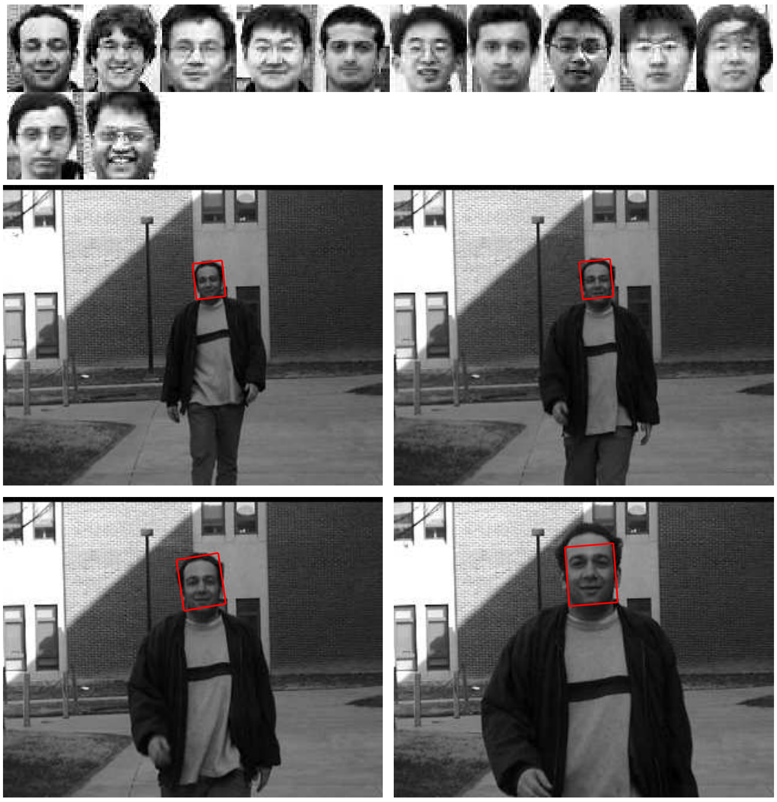


Fig. 2. Database-0. The 1st row: the face gallery with image size being 30x26. The 2nd and 3rd rows: four example frames in one probe video with image size being 320x240 while the actual face size ranges approximately from 30x30 in the first frame to 50x50 in the last frame. Notice that the sequence is taken under a well-controlled condition so that there are no illumination or pose variations between the gallery and the probe.

Results for Database-0

We consider affine transformation. Specifically, the motion is characterized by $\theta = (a_1, a_2, a_3, a_4, t_x, t_y)$ 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 as the subjects walk towards the camera. Regarding the photometric transformation, only zero-mean-unit-variance operator is performed to partially compensate for contrast variations. The complete transformation $\mathcal{T}_\theta\{z\}$ is processed as follows: affine transform z using $\{a_1, a_2, a_3, a_4\}$, crop out the interested region at position $\{t_x, t_y\}$ with the same size as the still template in the gallery, and perform zero-mean-unit-variance operation.

A time-invariant first-order Markov Gaussian model with constant velocity is used for modeling motion transition. Given that the subject is walking towards the

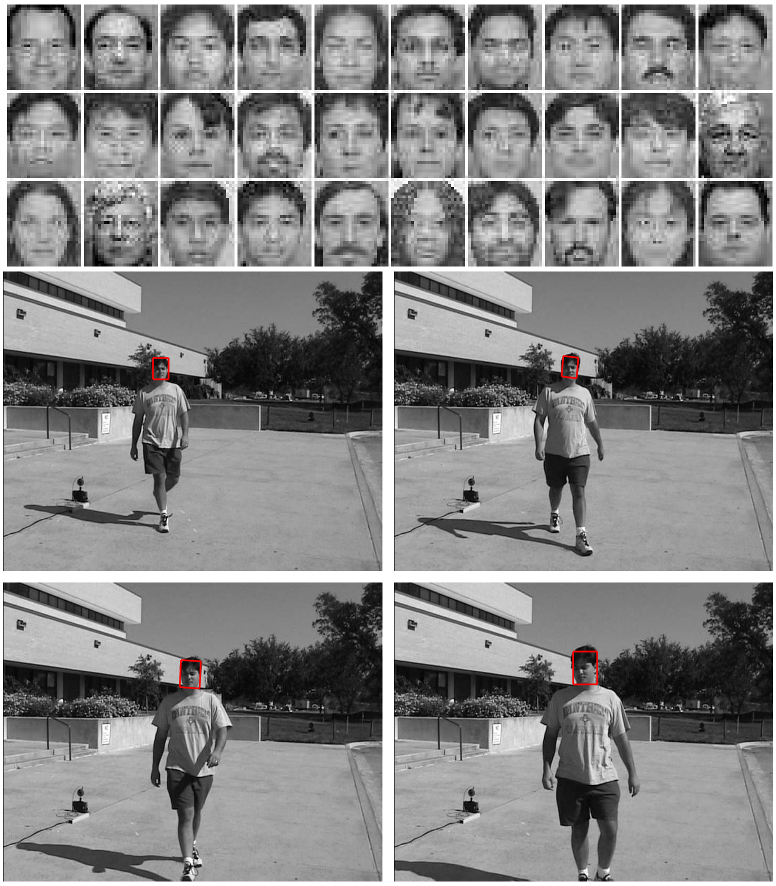


Fig. 3. Database-1. The 1st row: the face gallery with image size being 30x26. The 2nd and 3rd rows: 4 example frames in one probe video with image size being 720x480 while the actual face size ranges approximately from 20x20 in the first frame to 60x60 in the last frame. Notice the significant illumination variations between the probe and the gallery.

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 is sufficient.

The likelihood measurement is simply set as a 'truncated' Laplacian:

$$p_1(z_t|n_t,\theta_t) = \mathsf{L}(\|\mathcal{T}_{\theta_t}\{z_t\} - I_{n_t}\|;\sigma_1,\tau_1) \tag{16}$$

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

$$\mathsf{L}(x;\sigma,\tau) = \begin{cases} \sigma^{-1}\exp(-x/\sigma) & \text{if } \ x \le \tau\sigma \\ \sigma^{-1}\exp(-\tau) & \text{otherwise} \end{cases} \tag{17}$$

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 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. 4 presents the plot of the posterior probability $p(n_t|z_{0:t})$, the conditional entropy $H(n_t|z_{0:t})$ and the minimum mean square error (MMSE) estimate of the scale parameter $sc = \sqrt{(a_1^2 + a_2^2 + a_3^2 + a_4^2)/2}$, all against t. In Fig. 2, the tracked face is superimposed on the image using a bounding box.

Suppose the correct identity for Fig. 2 is l. From Fig. 4, we can easily observe that the posterior probability $p(n_t = l|z_{0:t})$ increases as time proceeds and eventually approaches 1, and all others $p(n_t = j|z_{0:t})$ for $j \neq l$ go to 0. Fig. 4 also plots the decrease in conditional entropy $H(n_t|z_{0:t})$ and the increase in scale parameter, which matches with the scenario of a subject walking towards a camera.

Table 3 summarizes the average recognition performance and computational time of the CONDENSATION and the proposed algorithm when applied to Database-0. Both algorithms achieved 100% recognition rate with top match. The proposed algorithm is much more efficient than the CONDENSATION algorithm. It is more than 10 times faster as shown in Table I. This experiment was implemented in C++ on a PC with P-III 1G 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 12.

Algorithm	CONDENSATION	Proposed
Recognition rate within top 1 match	100%	100%
Time per frame	7s	0.5s

Table 3. Recognition performance of algorithms when applied to Database-0.

Results on Database-1

Case 1: Tracking and Recognition using Laplacian Density

We first investigate the performance using the same setting as described in Section 2.3. Table 4 shows that the recognition rate is very poor, only 13% are correctly identified using top match. The main reason is that the 'truncated' Laplacian density is not able to capture the appearance difference between the probe and the gallery, thereby indicating a need for more effective appearance modeling. Nevertheless, the tracking accuracy [2] is reasonable with 83% successfully tracked because

[2] We manually inspect the tracking results by imposing the MMSE motion estimate on the final frame as shown in Figs. 2 and 3 and determine if tracking is successful or not for this sequence. This is done for all sequences and the tracking accuracy is defined as the ratio of the number of sequences successfully tracked to the total number of all sequences.

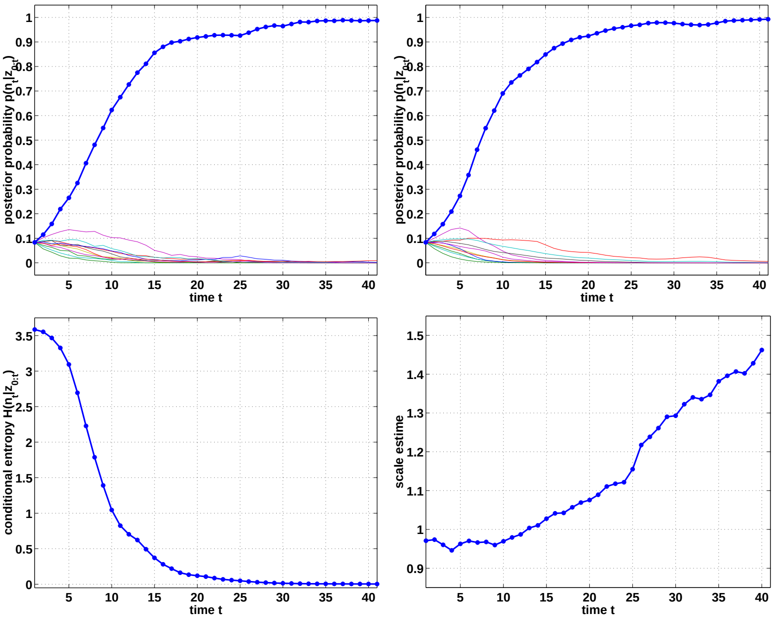


Fig. 4. Posterior probability $p(n_t|z_{0:t})$ against time t, obtained by the CONDENSATION algorithm (top left) and the proposed algorithm (top right). Conditional entropy $H(n_t|z_{0:t})$ (bottom left) and MMSE estimate of scale parameter sc (bottom right) against time t. The conditional entropy and the MMSE estimate are obtained using the proposed algorithm.

we are using multiple face templates in the gallery to track the specific face in the probe video. After all, faces in both the gallery and the probe belong to the same class of human face and it seems that the appearance change is within the class range.

Case 2: Pure Tracking using Laplacian Density

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

$$q(z_t|\theta_t) = \mathsf{L}(\|\mathcal{T}_{\theta_t}\{z_t\} - T_0\|; \sigma_2, \tau_2), \tag{18}$$

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

Table 4 shows that 87% are successfully tracked by this simple tracking model, which implies that the appearance within the video remains similar.

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

Table 4. Performances of algorithms when applied to Database-1.

Case 3: Tracking and Recognition using Probabilistic Subspace Density

As mentioned in Case 1, we need a new appearance model to improve the recognition accuracy. Of the many approaches suggested in the literature, we decided to use the approach suggested by Moghaddam et. al. [17] due to its computational efficiency and high recognition accuracy. However, here we model only the intra-personal variations. Modeling both intra/extra-personal variations is considered in Section 3.2.

We need at least two facial images for one identity to construct the intra-personal space (IPS). Apart from the available gallery, we crop out the second image from the video ensuring no overlap with the frames actually used in probe videos. Fig. 5 (top row) shows a list of such images. Compare with Fig. 3 to see how the illumination varies between the gallery and the probe.

We then fit a probabilistic subspace density [26] on top of the IPS. It proceeds as follows: a regular PCA is performed for the IPS. Suppose the eigensystem for the IPS is $\{(\lambda_i, e_i)\}_{i=1}^d$, where d is the number of pixels and $\lambda_1 \geq ... \geq \lambda_d$. Only top r principal components corresponding to top s eigenvalues are then kept while the residual components are considered as isotropic. We refer the reader to [26] for full details. Fig. 5 (middle row) show the eigenvectors for the IPS. The density is written as follows:

$$\mathsf{Q}(x) = \{\frac{exp(-\frac{1}{2}\sum_{i=1}^{r}\frac{y_i^2}{\lambda_i})}{(2\pi)^{r/2}\prod_{i=1}^{r}\lambda_i^{1/2}}\}\{\frac{exp(-\frac{\epsilon^2}{2\rho})}{(2\pi\rho)^{(d-r)/2}}\}, \tag{19}$$

where the principal components y_i's, the reconstruction error ϵ^2, and the isotropic noise variance ρ are defined as:

$$y_i = e_i^{\mathsf{T}} x, \quad \epsilon^2 = \|x\|^2 - \sum_{i=1}^{r} y_i^2, \quad \rho = (d-r)^{-1}\sum_{i=r+1}^{d}\lambda_i. \tag{20}$$

It is easy to write the likelihood as follows:

$$p_2(z_t|n_t, \theta_t) = \mathsf{Q}_{IPS}(\mathcal{T}_{\theta_t}\{z_t\} - I_{n_t}). \tag{21}$$

Table 4 lists the performance by using this new likelihood measurement. It turns out that the performance is significantly better that in Case 1, with 93% tracked successfully and 83% correctly recognized within top 1 match. If we consider the top 3 matches, 97% are correctly identified.

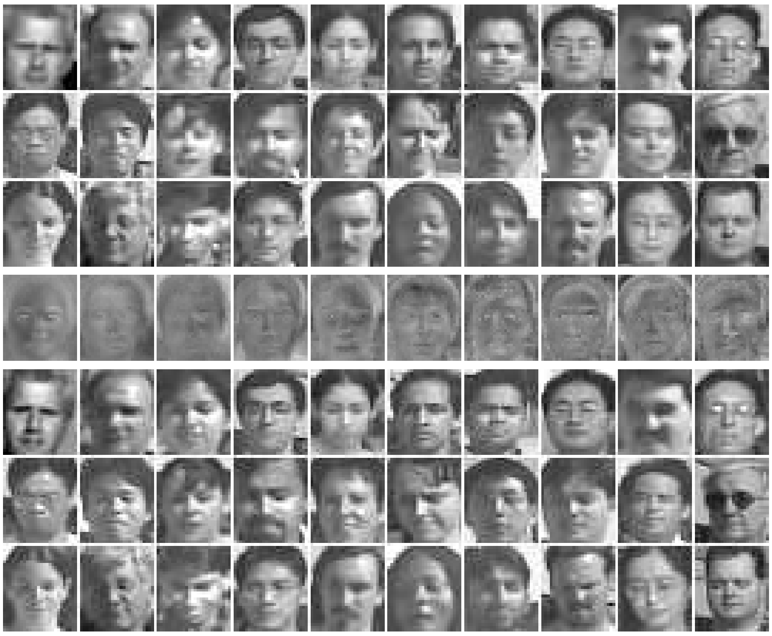


Fig. 5. Database-1. Top row: the second facial images for training probabilistic density. Middle row: top 10 eigenvectors for the IPS. Bottom row: the facial images cropped out from the largest frontal view.

Case 4: Tracking and Recognition using Combined Density

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

$$p_3(z_t|n_t, \theta_t) = p_2(z_t|n_t, \theta_t)q(z_t|\theta_t) \tag{22}$$

Again, all other setting is the same as in Case 1. We now obtain the best performance so far: no tracking error, 93% are correctly recognized as the first match, and no error in recognition when top 3 matches are considered.

Case 5: Still-to-still Face Recognition

We also performed an experiment for still-to-still face recognition. We selected the probe video frames with the best frontal face view (i.e. biggest frontal view) and cropped out the facial region by normalizing with respect to the eye coordinates manually specified. This collection of images is shown in Fig. 5 (bottom row) and it is fed as probes into a still-to-still face recognition system with the learned probabilistic subspace as in Case 3. It turns out that the recognition result is 57% correct for the top one match, and 83% for the top 3 matches. Clearly, Case 4 is the best among all.

3 Enhancing Tracking and Recognition Accuracy

The general formulation of our recognition model leaves room for enhancing tracking and recognition accuracy. In other words, one can employ different model choices tuned for specific scenarios. However, here we present rather general appearance-based models. This technique includes enhanced approaches to modeling inter-frame appearance changes for accurate tracking and modeling appearance changes between probe video frames and gallery images for accurate recognition.

3.1 Modeling Inter-frame Appearance Changes

To model inter-frame appearance changes, a certain appearance model A_t is needed. In [28] and Section 2, a fixed template, $A_t = T_0$, is matched with observations to minimize a cost function in the form of sum of squared distance (SSD). This is equivalent to assuming that the noise V_t is a normal random vector with zero mean and a diagonal (isotropic) covariance matrix. At the other extreme, one could use a rapidly changing model, for example, taking A_t as the 'best' patch of interest in the previous frame. However, a fixed template cannot handle appearance changes in the video, while a rapidly changing model is susceptible to drift. Thus, it is necessary to have a model which is a compromise between these two cases, which leads to an online appearance model [22].

Inter-frame appearance changes are also related to the motion transition model. In a visual tracking problem, it is ideal to have an exact motion model governing the kinematics of the object. In practice, however, approximate models are used. There are two types of approximations commonly found in the literature. (i) One is to learn a motion model directly from a training video [1]. However such a model may overfit the training data and may not necessarily succeed when presented with testing videos containing objects arbitrarily moving at different times and places. Also one cannot always rely on the availability of training data in the first place. (ii) Secondly, a fixed constant-velocity model with fixed noise variance is fitted for simplicity as in Section 2. Let r_0 be a fixed constant measuring the extent of noise. If r_0 is small, it is very hard to model rapid movements; if r_0 is large, it is computationally inefficient since many more particles are needed to accommodate large noise variance. All these factors make the use of such a model ineffective. In our work, we overcome this by introducing an adaptive-velocity model.

Adaptive Appearance Model

The adaptive appearance model assumes that the observations are explained by different causes, thereby indicating the use of a mixture density of components. In [22], three components are used, namely the W-component characterizing two-frame variations, the S-component depicting the stable structure within all past observations (though it is slowly-varying), and the L-component accounting for outliers such as occluded pixels. However, in our implementation, we have incorporated only the S, and W-components.

As an option, in order to further stabilize our tracker one could use an F-component which is a fixed template that one is expecting to observe most often. For example, in face tracking this could be just the facial image as seen from a frontal view. In the sequel, we derive the equations as if there is an F-component. However, the effect of this component can be ignored by setting its initial mixing probability to zero.

We now describe our mixture appearance model. The appearance model at time t, $A_t = \{W_t, S_t, F_t\}$, is a time-varying one that models the appearances present in all observations up to time $t-1$. It obeys a mixture of Gaussians, with W_t, S_t, F_t as mixture centers $\{\mu_{i,t};\ i = w, s, f\}$ and their corresponding variances $\{\sigma^2_{i,t};\ i = w, s, f\}$ and mixing probabilities $\{m_{i,t};\ i = w, s, f\}$. Notice that $\{m_{i,t}, \mu_{i,t}, \sigma^2_{i,t};\ i = w, s, f\}$ are 'images' consisting of d pixels that are assumed to be is independent of each other.

In summary, the observation likelihood is written as

$$p(Y_t|\theta_t) = p(Z_t|\theta_t) = \prod_{j=1}^{d}\{ \sum_{i=w,s,f} m_{i,t}(j)\mathtt{N}(Z_t(j); \mu_{i,t}(j), \sigma^2_{i,t}(j))\}, \tag{23}$$

where $\mathtt{N}(x; \mu, \sigma^2)$ is a normal density

$$\mathtt{N}(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\{-\rho(\frac{x-\mu}{\sigma})\}, \quad \rho(x) = \frac{1}{2}x^2. \tag{24}$$

Model Update

To keep our paper self-contained, we show how to update the current appearance model A_t to A_{t+1} after $\hat{Z}_t$ becomes available, i.e., we want to compute the new mixing probabilities, mixture centers, and variances for time $t+1$, $\{m_{i,t+1}, \mu_{i,t+1}, \sigma^2_{i,t+1};\ i = w, s, f\}$.

It is assumed that the past observations are exponentially 'forgotten' with respect to their contributions to the current appearance model. Denote the exponential envelop by $\mathcal{E}_t(k) = \alpha \exp(-\tau^{-1}(t-k))$ for $k \le t$, where $\tau = n_h/\log 2$, n_h is the half-life of the envelope in frames, and $\alpha = 1 - \exp(-\tau^{-1})$ to guarantee that the area under the envelope is 1. We just sketch the updating equations as follows and refer the interested readers to [22] for technical details and justifications.

The EM algorithm [27] is invoked. Since we assume that the pixels are independent to each other, we can deal with each pixel separately. The following computation is valid for $j = 1, 2, \ldots, d$ where d is the number of pixels in the appearance model.

Firstly, the posterior responsibility probability is computed as

$$o_{i,t}(j) \propto m_{i,t}(j)\mathtt{N}(\hat{Z}_t(j); \mu_{i,t}(j), \sigma^2_{i,t}(j)); \quad i = w, s, f, \quad \& \sum_{i=w,s,f} o_{i,t}(j) = 1. \tag{25}$$

Then, the mixing probabilities are updated as

$$m_{i,t+1}(j) = \alpha\ o_{i,t}(j) + (1-\alpha)\ m_{i,t}(j); \quad i = w, s, f, \tag{26}$$

and the first- and second-moment images $\{M_{p,t+1};\ p=1,2\}$ are evaluated as

$$M_{p,t+1}(j) = \alpha\ \hat{Z}_t^p(j) o_{s,t}(j) + (1-\alpha)\ M_{p,t}(j);\quad p=1,2. \tag{27}$$

Finally, the mixture centers and the variances are updated as:

$$S_{t+1}(j) = \mu_{s,t+1}(j) = \frac{M_{1,t+1}(j)}{m_{s,t+1}(j)},\quad \sigma^2_{s,t+1}(j) = \frac{M_{2,t+1}(j)}{m_{s,t+1}(j)} - \mu^2_{s,t+1}(j). \tag{28}$$

$$W_{t+1}(j) = \mu_{w,t+1}(j) = \hat{Z}_t(j),\quad \sigma^2_{w,t+1}(j) = \sigma^2_{w,1}(j), \tag{29}$$

$$F_{t+1}(j) = \mu_{f,t+1}(j) = F_1(j),\quad \sigma^2_{f,t+1}(j) = \sigma^2_{f,1}(j). \tag{30}$$

Model Initialization

To initialize A_1, we set $W_1 = S_1 = F_1 = T_0$ (with T_0 supplied by a detection algorithm or manually), $\{m_{i,1}, \sigma^2_{i,1};\ i = w,s,f\}$, and $M_{1,1} = m_{s,1}T_0$ and $M_{2,1} = m_{s,1}\sigma^2_{s,1} + T_0^2$.

Adaptive Velocity State Model

The state transition model we use possesses a term for modeling adaptive velocity. The adaptive velocity in the state parameter is calculated using a first-order linear prediction based on the appearance difference between two successive frames. The previous particle configuration is incorporated in our prediction scheme.

Construction of the particle configuration involves the costly computation of image warping (in our experiments, it usually takes about half of the computation load). In a conventional particle filtering algorithm the particle configuration is used only to update the weight, i.e., computing weight for each particle by comparing the warped image with the online appearance model using the observation equation. But, our approach in addition uses the particle configuration in the state transition equation. In some sense, we 'maximally' utilize the information contained in the particles (without wasting the costly computation of image warping) since we use it in both state and observation models.

Adaptive Velocity

With the availability of the sample set $\Theta_{t-1} = \{\theta_{t-1}^{(j)}\}_{j=1}^J$ and the image patches of interest $\mathcal{Y}_{t-1} = \{y_{t-1}^{(j)}\}_{j=1}^J$ with $y_{t-1}^{(j)} = \mathcal{T}_{\theta_t^{(j)}}\{z_t\}$, for a new observation z_t, we can predict the shift in the motion vector (or adaptive velocity) $\nu_t = \theta_t - \hat{\theta}_{t-1}$ using a first-order linear approximation [28, 23, 29], which essentially comes from the constant brightness constraint, i.e., there exists a θ_t such that

$$\mathcal{T}_{\theta_t}\{z_t\} \simeq \hat{y}_{t-1}. \tag{31}$$

Approximating $\mathcal{T}_{\theta_t}\{z_t\}$ via a first-order Taylor series expansion around $\hat{\theta}_{t-1}$ yields

$$\mathcal{T}_{\theta_t}\{z_t\} \simeq \mathcal{T}_{\hat{\theta}_{t-1}}\{z_t\} + C_t(\theta_t - \hat{\theta}_{t-1}) = \mathcal{T}_{\hat{\theta}_{t-1}}\{z_t\} + C_t\nu_t, \tag{32}$$

where C_t is the Jacobian matrix.

Combining (31) and (32) gives

$$\hat{y}_{t-1} \simeq \mathcal{T}_{\hat{\theta}_{t-1}}\{z_t\} + C_t\nu_t, \tag{33}$$

i.e.,

$$\nu_t \simeq -B_t(\mathcal{T}_{\hat{\theta}_{t-1}}\{z_t\} - \hat{y}_{t-1}), \tag{34}$$

where B_t is the pseudo-inverse of the C_t matrix, which can be efficiently estimated from the available data Θ_{t-1} and $\mathcal{Y}_{t-1}$.

Specifically, to estimate B_t we stack into matrices the differences in motion vectors and image patches, using $\hat{\theta}_{t-1}$ and $\hat{y}_{t-1}$ as pivotal points:

$$\begin{cases} \Theta^{\delta}_{t-1} = [\theta^{(1)}_{t-1} - \hat{\theta}_{t-1}, \ \ldots, \ \theta^{(J)}_{t-1} - \hat{\theta}_{t-1}], \\ \mathcal{Y}^{\delta}_{t-1} = [y^{(1)}_{t-1} - \hat{y}_{t-1}, \ \ldots, \ y^{(J)}_{t-1} - \hat{y}_{t-1}]. \end{cases} \tag{35}$$

The least square solution for B_t is

$$B_t = (\Theta^{\delta}_{t-1}\mathcal{Y}^{\delta\,\mathrm{T}}_{t-1})(\mathcal{Y}^{\delta}_{t-1}\mathcal{Y}^{\delta\,\mathrm{T}}_{t-1})^{-1}. \tag{36}$$

However, it turns out that the matrix $\mathcal{Y}^{\delta}_{t-1}\mathcal{Y}^{\delta\,\mathrm{T}}_{t-1}$ is very often rank-deficient due to the high dimensionality of the data (unless the number of the particles exceeds the data dimension). To overcome this, we use the singular value decomposition.

$$\mathcal{Y}^{\delta}_{t-1} = USV^{\mathrm{T}} \tag{37}$$

It can be easily shown that

$$B_t = \Theta^{\delta}_{t-1}VS^{-1}U^{\mathrm{T}}. \tag{38}$$

To gain some computational efficiency, we can further approximate

$$B_t = \Theta^{\delta}_{t-1}V_qS_q^{-1}U_q^{\mathrm{T}}, \tag{39}$$

by retaining the top q components. Notice that if only a fixed template is used [23], the B matrix is fixed and pre-computable. But, in our case, the appearance is changing so that we have to compute the B_t matrix in each time step.

We also calculate the prediction error ϵ_t between $\tilde{y}_t = \mathcal{T}_{\hat{\theta}_{t-1}+\nu_t}\{z_t\}$ the updated appearance model A_t. The error ϵ_t is defined as below:

$$\epsilon_t = \phi(\tilde{y}_t, A_t) = \sum_{i=w,s,f} \frac{2}{d}\sum_{j=1}^{d} m_{i,t}(j)\rho(\frac{\tilde{y}_t(j) - \mu_{i,t}(j)}{\sigma_{i,t}(j)}). \tag{40}$$

We use the following model

$$\theta_t = \hat{\theta}_{t-1} + \nu_t + u_t, \tag{41}$$

where ν_t is the predicted shift in the motion vector. The choice of u_t is discussed below.

Adaptive Noise

The value of ϵ_t determines the quality of prediction. Therefore, if ϵ_t is small, which implies a good prediction, we only need noise with small variance to absorb the residual motion; if ϵ_t is large, which implies a poor prediction, we then need noise with large variance to cover potentially large jumps in the motion state.

To this end, we use u_t of the form $u_t = r_t * u_0$, where r_t is a function of ϵ_t and u_0 is a 'standardized' random vector. In practice, we take u_0 as a Gaussian random vector with zero mean and a given covariance matrix. Since ϵ_t is 'variance'-type measure, we use

$$r_t = \max(\min(r_0\sqrt{\epsilon_t}, r_{max}), r_{min}), \tag{42}$$

where r_{min} is the lower bound to maintain a reasonable sample coverage and r_{max} is the upper bound to constrain computational load.

3.2 Modeling Appearance Changes between Frames and Gallery Images

We adopt the maximum a posterori rule developed in [17] for the recognition score $p_n(z_t|n_t, \theta_t)$. Two subspaces are constructed to model appearance variations. The intra-personal space (IPS) is meant to cover all the variations in appearances belonging to the same person while the extra-personal space (EPS) is used to cover all the variations in appearances belonging to different people. More than one facial image per person is needed to construct the IPS. Apart from the available gallery, we crop out four images from the video ensuring no overlap with frames used in probe videos. The probabilistic subspace density (Eq. (19)) estimation method is applied separately to the IPS and the EPS, yielding two different eigensystems. The recognition score $p_n(z_t|n_t, \theta_t)$ is finally computed as, assuming equal priors on the IPS and the EPS,

$$p_n(z_t|n_t, \theta_t) = \frac{\mathsf{Q}_{IPS}(\mathcal{T}_{\theta_t}\{z_t\} - I_{n_t})}{\mathsf{Q}_{IPS}(\mathcal{T}_{\theta_t}\{z_t\} - I_{n_t}) + \mathsf{Q}_{EPS}(\mathcal{T}_{\theta_t}\{z_t\} - I_{n_t})}. \tag{43}$$

3.3 The Complete Observation Likelihood

As in Section 2.3, we can construct a combined likelihood by taking the product of $p_a(z_n|\theta_t)$ and $p_n(z_t|n_t, \theta_t)$. To fully exploit the fact that all gallery images are in frontal view, we also compute how likely the patch y_t is in frontal view and denote this score by $p_f(z_t|\theta_t)$. We simply measure this by fitting a PS density on top of the gallery images [26], assuming that they are i.i.d. samples from the frontal face space (FFS). It is easy to write $p_f(z_t|\theta_t)$ as follows:

$$p_f(z_t|\theta_t) = \mathsf{Q}_{FFS}(\mathcal{T}_{\theta_t}\{z_t\}). \tag{44}$$

If the patch is in frontal view, we accept a recognition score; otherwise, we simply set the recognition score as equiprobable among all identities, i.e., $1/N$. The complete likelihood $p(z_t|n_t, \theta_t)$ is now defined as

$$p(z_t|n_t, \theta_t) \propto p_a \; \{p_f \; p_n + (1 - p_f) \; N^{-1}\}. \tag{45}$$

We also adjust the particle number J_t based on the following two considerations. (i) If the noise variance r_t is large, we need more particles, while conversely, fewer particles are needed for noise with small variance r_t. Based on the principle of asymptotic relative efficiency [25], we should adjust the particle number J_t in a similar fashion, i.e., $J_t = J_0 r_t / r_0$. (ii) As shown in Section 2.1, the uncertainty in the identity variable n_t is characterized by an entropy measure H_t for $p(n_t|z_{0:t})$ and H_t is a non-increasing function (under one weak assumption). Accordingly we increase the number of particles by a fixed amount J_{fix} if H_t increases; otherwise we deduct J_{fix} from J_t. Combining these two, we have

$$J_t = J_0 \frac{r_t}{r_0} + J_{fix} * (-1)^{i[H_{t-1} < H_{t-2}]}, \tag{46}$$

where $i[.]$ is an indication function.

3.4 Experimental results

We have applied our algorithm to tracking and recognizing human faces captured by a hand-held video camera in office environments. There are 29 subjects in the database. Fig. 6 lists all the images in the galley set and the top 10 eigenvectors for the FFS, IPS, and EPS, respectively. Fig. 7 presents some frames (with tracking results) in the video sequence for 'Subject-2' featuring quite large pose variations, moderate illumination variations, and quick scale changes (back and forth toward the end of the sequence).

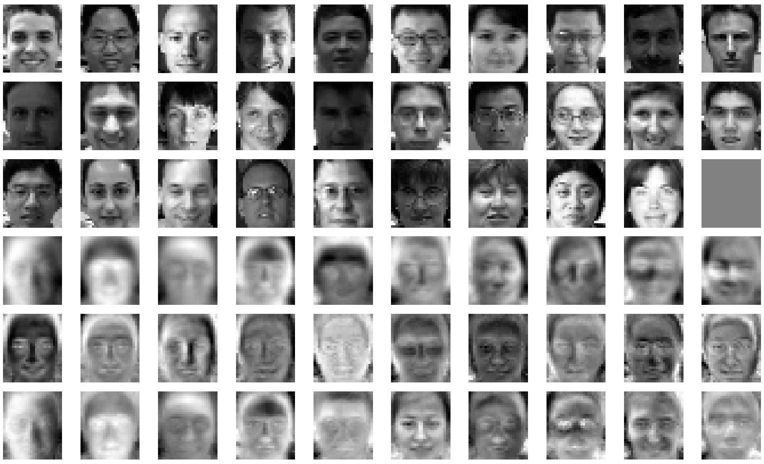
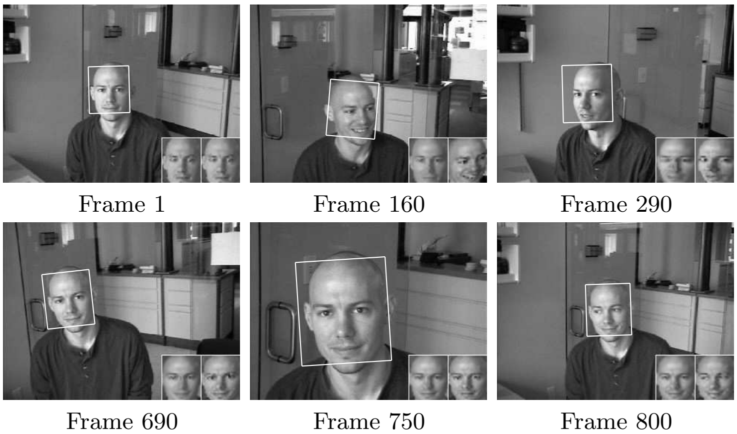


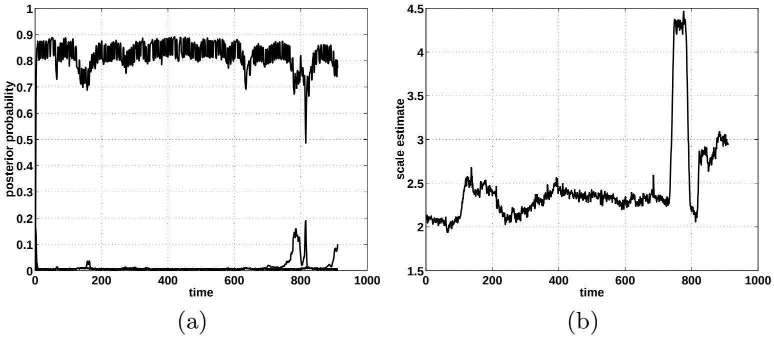
Fig. 6. Row 1-3: the gallery set with 29 subjects in frontal view. Rows 4, 5, and 6: the top ten eigenvectors for the FFS, IPS, and EPS, respectively.



Frame 1 Frame 160 Frame 290

Frame 690 Frame 750 Frame 800

Fig. 7. Example images in 'Subject-2' probe video sequence and the tracking results.



(a) (b)

Fig. 8. Results on the 'Subject-2' sequence. (a) Posterior probabilities against time t for all identities $p(n_t|z_{0:t})$, $n_t = 1, 2, ..., N$. The line close to 1 is for the true identity. (b) Scale estimate against time t.

Tracking is successful for all the video sequences and a 100% recognition rate is achieved, while the approach in Section 2 fails to track in several video sequences due to its inability to handle significant appearance changes caused by pose and illumination variations. The posterior probabilities $p(n_t|z_{0:t})$ with $n_t = 1, 2, ...N$ obtained for the 'Subject-2' sequence are plotted in Fig. 8(a). It is very fast, taking about less than 10 frames, to reach above 0.9 level for the posterior probability corresponding to 'Subject-2', while all other posterior probabilities corresponding to other identities approach zero. This is mainly attributed to the discriminative power of the MAP recognition score induced by the IPS and EPS modeling. The approach in Section 2 usually takes about 30 frames to reach 0.9 level since only intra-personal modeling is adopted. Fig. 8(b) captures the scale change in the 'Subject-2' sequence.

4 Issues and Discussions

We have presented a probabilistic model for recognizing human faces present in a video sequence and derived several important features embedded in this framework: (i) the temporal fusion of tracking and recognition; (ii) the evolution of the posterior probability of the identity variable; (iii) the computational efficiency of the SIS algorithm for solving this model; and (iv) the generality of this model by incorporating more sophisticated appearance models.

Our current approach is appearance-based (to be more accurate, image-intensity-based). One limitation of the appearance-based learning approach is its difficulty in dealing with novel appearances not present in the training stage. Such novel appearances are easily produced by illumination, pose and facial expression variations, especially when a video sequence is used. To overcome the above limitations, the following issues are worthy of further investigations.

1. Feature-based approach. It is ideal to have features derived from the image intensities invariant to the above variations. A good example is the Elastic Graph Matching (EGM) algorithm [12]. Filter responses from Gabor wavelets are computed to form a sparse graph representation and recognition is based on graph matching results. This representation is known to be partially resistant to illumination and pose variations. It is also robust to variations in facial expression.
2. We have used only one template for each person in the gallery. Obviously we can use multiple templates per person. The still templates in the gallery can be further generalized to video sequences in order to realize *video-to-video* recognition. In [30], exemplars and their prior probabilities are learned from the gallery videos to serve as still templates in the still-to-video scenario. A person n may have a collection of K_n exemplars, say $\mathcal{C}^n = \{c_1^n, \ldots, c_k^n, \ldots, c_{K_n}^n\}$ indexed by k. The likelihood is modified as a mixture density with exemplars as mixture centers. The joint distribution $p(n_t, k_t, \theta_t | z_{0:t})$ is computed using the SIS algorithm and marginalized to yield $p(n_t | z_{0:t})$. In the experiments reported in [30], the subject walks on a tread-mill with his/her face moving naturally, giving rise to significant variations across poses. However, the proposed method successfully copes with these pose variations (using exemplars) as evidenced by the experimental results. Other learning methods can be applied to the gallery videos. For example, mixture of Gaussians can be used to replace the exemplar-learning procedure. Hidden Markov models can also be used if the person is involved in a particular activity.
3. The use of 3-D face model. There are some recent efforts on handling pose and illumination variation that invoke the 3-D face model, either explicitly or implicitly. In [24], the 3-D face model is used explicitly to recover the unknown pose. In [32], the 3-D face model is used implicitly to recover the unknown illumination. However, the above two methods are still-image-based. Incorporation of these methods into our video-based recognition framework is very appealing. If explicit 3-D models are needed, how to capture such model becomes an issue. In [24], all 3-D models are recorded separately. The obtained 3-D models

are very accurate but this need a manual operator. The alternative is to use a structure from motion algorithm [33].

References

[1] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *European Conference on Computer Vision*, 343–356, 1996.

[2] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar and Signal Processing*, 140:107–113, 1993.

[3] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories. *European Conference on Computer Vision*, 909–924, 1998.

[4] G. Qian and R. Chellappa. Structure from motion using sequential monte carlo methods. *Internation Conference on Computer Vision*, 614–621, 2001.

[5] B. Li and R. Chellappa. A generic approach to simultaneous tracking and verification in video. *IEEE Transaction on Image Processing*, 11:530–544, 2002.

[6] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of IEEE* 83:705–740, 1995.

[7] W. Y. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. *UMD CfAR Technical Report CAR-TR-948*, 2000.

[8] P. J. Phillips, H. Moon, S. Rivzi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22:1090–1104, 2000.

[9] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[10] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *Journal of Optical Society of America A* , 1724–1733, 1997.

[11] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[12] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. v. d. Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transaction on Computers*, 42:300–311, 1993.

[13] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. *International Conference on Audio- and Video-Based Person Authentication*, 176–181, 1999.

[14] G. Kitagawa. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5:1–25, 1996.

[15] J. S. Liu and R. Chen. Sequential Monte Carlo for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[16] A. Doucet, S. J. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[17] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian modeling of facial similarity. *Advances in Neural Information Processing Systems*, 11:910–916, 1999.

[18] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley, 1991.

[19] B. Anderson and J. Moore. *Optimal Filtering.* Prentice Hall, Englewood Cliffs, New Jersey, 1979.

[20] J. Sullivan and J. Rittscher. Guiding random particle by deterministic search. *International Conference on Computer Vision*, 323 –330, 2001.

[21] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. *Uncertainty in AI (UAI)*, 33 – 42, 1998.

[22] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance model for visual tracking. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:415–422, 2001.

[23] F. Jurie and M. Dhome. A simple and efficient template matching algorithm. *Internatonal Conference on Computer Vision*, 2:544–549, 2001.

[24] S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3D morphable model using linear shape and texture errror functions. *European Conference on Computer Vision*, 2002.

[25] G. Casella and R. L. Berger. *Statistical Inference.* Duxbury, 2002.

[26] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19:696–710, 1997.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society B*, 1977.

[28] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20:1025–1039, 1998.

[29] J. Bergen, P. Anadan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. European Conference on Computer Vision, 237–252, 1992.

[30] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 2003.

[31] S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance modeling in particle filters. *IEEE Conference on Multimedia and Expo*, 2003.

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

[33] A. Roy Chowdhury and R. Chellappa. Face reconstruction from video using uncertainty analysis and a generic model. *Computer Vision and Image Understanding*, 2003.