

BAYESIAN METHODS FOR FACE RECOGNITION FROM VIDEO

Rama Chellappa, Shaohua Zhou *

Center for Automation Research
EE Department, University of Maryland
College Park, MD 20742

Baoxin Li

Sharp Laboratories of America
5750 NW Pacific Rim Blvd.
Camas, WA 98607

ABSTRACT

Face recognition (FR) from video necessitates simultaneously solving two tasks, recognition and tracking. To accommodate the video, a time series state space model is introduced in a Bayesian approach. Given this model, the goal reduces to estimating the posterior distribution of the state vector given the observations up to the present. The *Sequential Importance Sampling* (SIS) technique is invoked to generate a numerical solution to this model. However, the ultimate goal is to estimate the posterior distribution of the identity of humans for recognition purposes. Presented here are two methods to approximate the above distribution under different experimental scenarios.

1. INTRODUCTION

Bayesian analysis of video has recently gained significant attention in the computer vision community since the seminal work by 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 state vector (e.g. affine parameters) and developed the CONDENSATION algorithm to provide a numerical approximation to the posterior distribution of the state vector, and to propagate it over time according to the state equation. This has been extended to many areas [2, 3], including face recognition [4, 5, 6]. Refer to [7, 8] for surveys and [9] for experiments on face recognition.

Experiments reported in [9] evaluate still-to-still scenarios, where the gallery and the probe consist of both still facial images. Some well-known still-to-still FR approaches include Principal Component Analysis (PCA) [10], Linear Discriminant Analysis (LDA) [11, 12], and Elastic Graph Matching (EGM)[13]. Typically, recognition is performed based on an abstract representation of an image after suitable geometric and photometric normalizations are performed.

Following [9], we define the gallery and probe as follows: the gallery consists of still facial templates and the probe consists of video sequences containing the facial region. There are many instances where still-to-video algorithms are useful. Denote the gallery set as $H = \{I_1, I_2, \dots, I_N\}$, indexed by the identity variable n , which lies in a finite sample space $\mathcal{N} = \{1, 2, \dots, N\}$. We also adopt the time series state space model to characterize the evolving dynamics of and identity in the probe video. Let x_t be the state vector and y_t be the observation respectively at time t . Given this model, the goal reduces to computing the posterior distribution of the state vector given the observations up to time

t , denoted by $\pi_t(x_t) = p_t(x_t|y_{0:t})$ with $y_{0:t} = \{y_0, y_1, \dots, y_t\}$. The SIS technique can be invoked to generate a numerical solution. Ultimately, we need to estimate the posterior distribution of the identity, $\pi_t(n_t) = p_t(n_t|y_{0:t})$, where n_t is the human identity variable at time t .

Presented here are two methods for approximating the distribution $\pi_t(n_t)$ under different experimental scenarios but same still-to-video setup. Method I [5] parameterizes the model with only an affine tracking state, denoted by θ_t , and approximates and propagates $\pi_t(\theta_t)$ using the SIS algorithm. The distribution $\pi_t(n_t)$ is estimated by marginalizing $\pi_t(\theta_t)$ over a proper affine region around the posterior mean $E_{\pi}(x_t)$. Method II [6] parameterizes the model with the affine tracking state θ_t and the recognizing identity variable n_t , approximates and propagates the joint distribution $\pi_t(\theta_t, n_t)$ using the SIS algorithm. The distribution $\pi_t(n_t)$ is a free estimate from $\pi_t(\theta_t, n_t)$, i.e., the true marginal distribution of $\pi_t(\theta_t, n_t)$.

Section 2 introduces a general time series state space model and briefly reviews the SIS algorithm that approximates its solution. Sections 3 and 4 respectively describe the experimental scenarios and presents the two aforementioned methods and their results. Section 5 concludes the paper.

2. SIS ALGORITHM

A general time series state space model consists of the following three components:

1. State equation governing the state evolution:

$$x_t = g_t(x_{t-1}, u_t); t \geq 1, \quad (1)$$

where u_t is the state noise and $g_t(\cdot, \cdot)$ the state evolving function. Denote the state transition probability as $p_t(x_t|x_{t-1})$.

2. Observation equation depicting the observational behavior:

$$y_t = h_t(x_t, v_t); t \geq 1, \quad (2)$$

where v_t is the observation noise and $h_t(\cdot, \cdot)$ the observation function. Denote the likelihood as $p_t(y_t|x_t)$.

3. Prior probability $p_0(x_0)$ and statistical independence:

$$u_t \perp u_s, v_t \perp v_s; t, s \geq 1 \text{ \& } t \neq s. \quad (3)$$

Using this model, we attempt to compute the filtering posterior probability $\pi_t(x_t) = p(x_t|y_{0:t})$. If the model is linear with Gaussian noise, it is analytically solvable by a Kalman filter which essentially propagates the mean and variance of a Gaussian distribution over time. For nonlinear and non-Gaussian cases, an extended Kalman filter (EKF) is proposed to arrive at an approximate

*Partially supported by the DARPA Grant N00014-00-1-0908.

analytic solution. Recently, the SIS technique, a special case of Monte Carlo method, [1, 14, 15, 16] has been used to provide a numerical solution and to propagate an arbitrary distribution over time.

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

$$w^{(m)} = \pi(x^{(m)})/g(x^{(m)}) \quad (4)$$

is assigned to each sample. This technique is called *Importance Sampling* (IS). It can be shown[16] that the *importance sample set* $S = \{(x^{(m)}, w^{(m)})\}_{m=1}^M$ 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 S_{t-1} according to the *sequential importance function* $g_t(x_t|x_{t-1})$, and calculates the weight using

$$w_t = w_{t-1} p_t(y_t|x_t) p_t(x_t|x_{t-1}) / g_t(x_t|x_{t-1}). \quad (5)$$

For a complete description of the SIS method, refer to [14, 16].

3. METHOD I

This method[5] has been tested on a database containing 19 subjects. In building the database, each person was asked to sit on a chair at a fixed distance from the camera so that the scale was approximately the same for all persons, and to move his/her head and make any desired facial expression, simulating an automatic teller machine or access control scenario. Fig. 1 shows some sample frames from a probe video and some templates in the gallery.

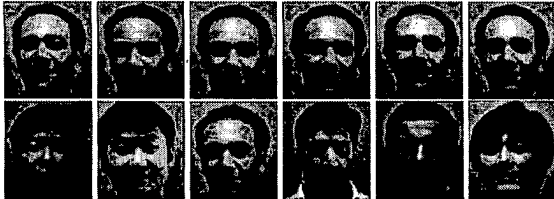


Fig. 1. I. Sample frames from the video (top) and templates (bottom). The templates will be referred to as face 1, face 2, ..., etc., counting from left, and obviously face 3 in the bottom row is the true hypothesis.

The state vector x_t is taken to be affine tracking parameter θ_t , which obeys a first-order Gaussian-Markov model, i.e., the transition probability, assumed time-invariant $p(\theta_t|\theta_{t-1})$, obeys a Gaussian distribution. In addition, a local deformation is introduced to account for the residual motion due to inaccuracies in affine modeling and other factors such as facial expressions. The observation y_t is taken to be Gabor-filtered jets [13] defined on a sparse grid, shown in Fig. 2. Note that it is the grid in the template image that undergoes the affine motion and local deformation. The local deformation is implemented by performing a local search around each grid point for its best match when updating the likelihood

measurement. The likelihood is assumed to be time-invariant and modeled as a truncated Gaussian:

$$p(y_t|\theta_t) = \begin{cases} (\sqrt{2\pi}\sigma_0)^{-1} \exp\{-(e_t)^2/(2\sigma_0^2)\} & \text{if } |e_t| < \delta \\ K & \text{otherwise,} \end{cases} \quad (6)$$

where δ is a threshold and K a constant. The error e_t is computed as

$$e_t = \frac{1}{N_J} \sum_k \frac{J_m^{(j)} \cdot J_s^{(j)}}{\|J_m^{(j)}\| \cdot \|J_s^{(j)}\|}, \quad (7)$$

where N_J is the number of jets, $J_m^{(j)}$ the jet for the j -th grid point in the template and $J_s^{(j)}$ its counterpart in the current frame. It needs to be emphasized that $J_s^{(j)}$ is found by first applying to the grid the affine motion with θ_t followed by a local search.

Using SIS, the tracking problem can be numerically solved by approximating $\pi_t(\theta_t)$. For pure tracking, we let the template be the facial part in the first frame. Fig. 2 shows some pure tracking results. For both tracking and recognition, we use templates in the gallery set. In order to evaluate $\pi_t(n_t = n)$ for template n in the gallery, we first invoke SIS to obtain $\pi_t(\theta_t)$, then compute it as follows:

$$\pi_t(n_t = n) = \int_A \pi_t(\theta_t) d\theta_t, \quad (8)$$

where A is a proper region interval around the posterior mean $E_\pi(\theta_t)$. The complete algorithm I is summarized below.

Algorithm I

Initialization: Rectify the template grid onto the first frame using EGM. Draw M random samples from $p_0(\theta_0)$.

Tracking and Recognition:

Tracking: at time $t > 0$, invoke the SIS algorithm to obtain an updated set of samples for $\pi_t(\theta_t)$. To compute the likelihood of each sample, a local search around each node is performed to account for the deformation before computing the matching error.

Recognition and Mean Shape Evaluation: at any time $t > 0$, the tracked set of jets is given by $E_\pi(\theta_t)$, plus a local search; and a final matching score is calculated using the mean shape; the posterior probability is computed in an interval around $E_\pi(\theta_t)$.

Fig. 2 shows pure tracking results with tracked grid points superimposed on the image. Even under difficult situations as shown in Fig. 2, tracking is successfully maintained. Fig. 3 shows the posterior probability $\pi_t(n_t)$ and the matching scores computed using the mean shape.

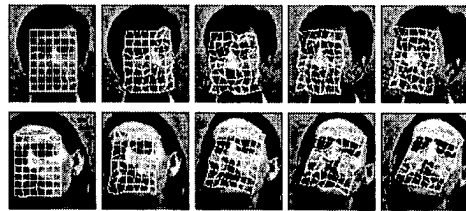


Fig. 2. I. Tracking results. Note the rotation in depth in the upper row and the large in-plane rotation in the lower row.

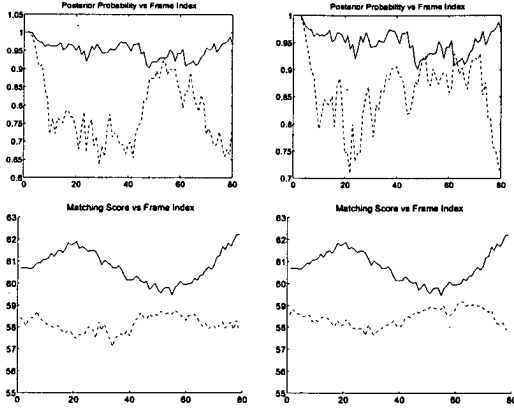


Fig. 3. I. Posterior probabilities and matching scores. Solid line is from the true hypothesis (face 3 in Fig.2). Dashed lines are from face 1 (left) and face 5 (right) corresponding to wrong hypotheses, respectively.

4. METHOD II

This method[6] has been tested on a database collected as part of the HumanID project by National Institute of Standards and Technology and University of South Florida researchers. It contains 30 subjects walking towards a camera in order to simulate typical scenarios in visual surveillance. There are 30 subjects, each having one face template in the gallery and one video in the probe. The complete face gallery is shown in Fig. 4. Fig. 5 gives some example frames in one probe video. Note that the gallery was captured under different lighting circumstances from the probe and that the face in the probe is of low resolution, small size, and considerable scale change.



Fig. 4. II. The face gallery (upper) with image size 48x42. The top 10 eigenfaces (lower).

The time series state space model is now parameterized by both affine tracking parameters and identity variable, respectively characterizing the dynamics and identity of human, i.e., $x_t = (\theta_t, n_t)$. So, $p_t(x_t|x_{t-1}) = p_t(\theta_t|\theta_{t-1})p_t(n_t|n_{t-1})$. We assume that $p_t(\theta_t|\theta_{t-1})$ is a time-invariant Gaussian, and that there is temporal invariance in the identity, i.e., $p_t(n_t|n_{t-1}) = i(n_t - n_{t-1})$, where $i(\cdot)$ is an indicator function. The observation y_t is taken to be a reconstructed image from top 300 principal components

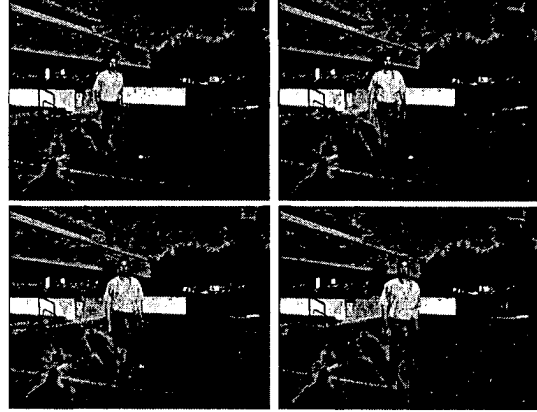


Fig. 5. II. Example frames in one probe video. The image size is 720x480 while the actual face size ranges approximately from 20x20 in the first frame to 60x60 in the last frame.

or eigenfaces (see Fig. 4 for the top 10 eigenfaces) and it is modeled as a transformed, noise-corrupted version of some template in the gallery, i.e., $f(y_t, \theta_t) = I_{n_t} + v_t$, where $f(\cdot, \cdot)$ is time-invariant image transformation function. Assume the likelihood to be a time-invariant truncated Laplacian.

$$p(y_t|\theta_t, n_t) = \begin{cases} \lambda^{-1} \exp(-\|v_t\|/\lambda) & \text{if } \|v_t\| \leq \delta \\ K & \text{otherwise,} \end{cases} \quad (9)$$

where δ is a threshold and K a constant.

By employing the SIS technique, the joint distribution of the state vector and the identity variable, $\pi_t(\theta_t, n_t)$, is estimated at current time and then propagated to the next, governed by the evolving equations for the state vector and the identity variable. The posterior distribution of the identity variable, $\pi_t(n_t)$, is just a free estimate, i.e., the marginal of $\pi_t(\theta_t, n_t)$. Algorithm II is summarized below. We have worked with two versions of Algorithm II. Algorithm IIa is a brute-force implementation; Algorithms IIb is a more efficient implementation. Details are in [6].

Algorithm II

Initialization: Draw M random samples jointly from $p_0(\theta_0)$ and the uniform prior $p_0(n_0)$.

Tracking and Recognition:

Tracking: at time $t > 0$, invoke the SIS algorithm to obtain an updated set of samples for $\pi_t(\theta_t, n_t)$.

Recognition: at any time $t > 0$, marginalizing $\pi_t(\theta_t, n_t)$ over θ_t gives rise to $\pi_t(n_t)$. Conditional entropy $H(n_t|y_{0:t})$ and MMSE estimate of θ_t are computed accordingly.

Fig. 6 presents the plot of the posterior probability $\pi_t(n_t)$ against frame instance for probe video shown in Fig. 5. Suppose that the correct identity is c . From Fig. 6, we can easily observe that the posterior probability $\pi_t(c)$ increases as time proceeds and eventually approaches 1, and all others $\pi_t(n_t \neq c)$ go to 0 finally. Refer to [6] for a justification for such convergence and more detailed discussions on the evolution of $\pi_t(n_t)$.

To change a viewing angle, we use the notion of entropy [17], which essentially measures the average uncertainty about a random variable. It is well known that among all distributions taking values on $\{1, \dots, N\}$, the uniform distribution yields maximum $\log_2 N$ and the degenerate case yields the minimum 0, i.e., $0 \leq H \leq \log_2 N$. 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 that 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,

$$H(n_t|y_{0:t}) = - \sum_{n_t \in \mathcal{N}} p(n_t|\tilde{y}_{0:t}) \log_2 p(n_t|\tilde{y}_{0:t}). \quad (10)$$

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 IIa. Fig. 7 shows the decreasing conditional entropy $H(n_t|y_{0:t})$ and the increasing scale parameter, which matches with the scenario: a subject walking towards a camera. In Fig. 5, the tracked face is superimposed on the image using a black bounding box.

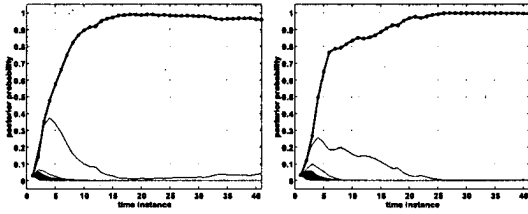


Fig. 6. II. Posterior probability $\pi_t(n_t)$ against time instance, obtained by Algorithm IIa (left) and Algorithm IIb (right).

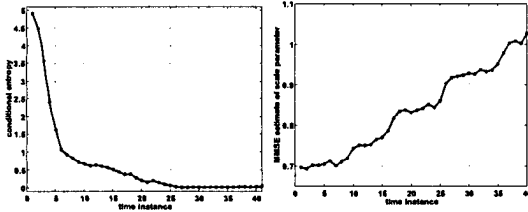


Fig. 7. II. Conditional entropy $H(n_t|y_{0:t})$ (left) and MMSE estimate of scale parameter (right) against time instance. Both are obtained using Algorithm IIb.

5. CONCLUSION

We have presented Bayesian methods for face recognition from a probe video, compared with a gallery of still templates. In both cases, a time series state space model is needed to accommodate the video and SIS algorithms provide the numerical solutions to the model. But, the posterior probability of the identity given the observations up to present, $\pi_t(n_t)$, is estimated using different strate-

gies. In addition, the still templates in the gallery can be generalized [18] to videos.

6. REFERENCES

- [1] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," *Proc. of ECCV*, 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*, vol. 140, pp. 107–113, 1993.
- [3] G. Qian and R. Chellappa, "Structure from motion using sequential monte carlo methods," *Proc. of ICCV*, pp. 614–621, 2001.
- [4] B. Li and R. Chellappa, "Simultaneous tracking and verification via sequential posterior estimation," *Proc. of CVPR*, pp. 110–117, 2000.
- [5] B. Li and R. Chellappa, "Face verification through tracking facial features," *Submitted to JOSA 2001*.
- [6] S. Zhou and R. Chellappa, "Probabilistic face recognition from video," *Submitted to ECCV 02*.
- [7] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces, a survey," *Proc. of IEEE*, vol. 83, pp. 705–740, 1995.
- [8] W. Y. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, "Face recognition: A literature survey," *UMD CAR-TR-948*, 2000.
- [9] P. J. Philipps, H. Moon, S. A. Rivzi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Trans. PAMI*, vol. 22, no. 10, pp. 1090–1104, 2000.
- [10] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 72–86, 1991.
- [11] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of Optical Society of America A*, pp. 1724–1733, 1997.
- [12] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, 1997.
- [13] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C von der Malsburg, R. P. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. Computers*, vol. 42, pp. 300–311, 1993.
- [14] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–209, 2000.
- [15] G. Kitagawa, "Monte carlo filter and smoother for non-gaussian nonlinear state space models," *J. Computational and Graphical Statistics*, vol. 5, pp. 1–25, 1996.
- [16] J. S. Liu and R. Chen, "Sequential monte carlo for dynamic systems," *Journal of the American Statistical Association*, vol. 93, pp. 1031–1041, 1998.
- [17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.
- [18] V. Krueger and S. Zhou, "Exemplar-based face recognition from video," *submitted to ECCV 02*.