Face Recognition from Video: A CONDENSATION Approach

Shaohua Zhou, Volker Krueger, and Rama Chellappa
Center for Automation Research (CfAR)
Department of Electrical & Computer Engineering
University of Maryland, College Park, MD 20740
Emails: {shaohua, vok, rama}@cfar.umd.edu

Abstract

The aim of this work is to investigate how to exploit the temporal information in a video sequence for the task of face recognition. Following the approach in [11], we propose a probabilistic model parameterized by a tracking state vector and a recognizing identity variable, simultaneously characterizing the kinematics and identity of humans. We then invoke a CONDENSATION [8] approach to provide a numerical solution to the model. Once the joint posterior distribution of the state vector and the identity variable is estimated, we marginalize it over the state vector to yield a robust estimate of the posterior distribution of the identity variable. Due to the propagation of identity and dynamics, a degeneracy in the posterior distribution 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 video data.

1 Introduction

Face recognition (FR) has been an extensive research area for over 10 years. See [3, 16] for surveys and [14] for experiments. Experiments reported in [14] are still-image-based approaches, including Principal Component Analysis (PCA) [15], Linear Discriminant Analysis (LDA) [7, 1], Elastic Graph Matching [10], and so on. Usually, an abstract representation of an image after a suitable geometric and photometric registration is formed and then recognition is performed based on this new representation.

Research efforts on video data are relatively fewer due to the following challenges [16] in typical applications like surveillance and access control: poor video quality, low image resolution, and large illumination and pose variations. This requires simultaneous solutions to tracking and recognition. Most video-based FR systems [4] split two tasks separately by performing the following: the face is first detected and then tracked over time. Only when a 'best' frame satisfying certain criteria is acquired, recognition is performed using still-image-based recognition technique. For this, the face is cropped from the frame and transformed or registered with appropriate parameters. The right choice of criteria for selecting good frames and the estimation of parameters for registration are often determined in an *ad hoc* manner.

However, one could solve two tasks simultaneously by probabilistic reasoning. Following [14], we define the gallery and probe as follows: the gallery consists of still facial templates and and the probe consists of video sequences containing the facial region. Denote the gallery set as $H = \{I_1, I_2, \ldots, I_N\}$, indexed by the identity variable n, which lies in a finite sample space $\mathcal{N} = \{1, 2, \ldots, N\}$. We also adopt the time series state space model to characterize the evolving kinematics and/or 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 $y_{0:t} = \{y_0, y_1, \ldots, y_t\}$. The CONDENSATION [8] algorithm, or in general the Sequential Importance Sampling (SIS) algorithm, 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.

In the scheme proposed by Li and Chellappa [11], the model is parameterized with an affine tracking state, denoted by θ_t, and $\pi_t(\theta_t)$ is approximated and propagated 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 $E_\pi(\theta_t)$. We present it in detail in Section 2.

Following [11], we also propose a probabilistic model in this paper. We parameterize this model with the affine

*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. Volker Krueger is affiliated with CfAR only.

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

tracking state θ_t and the recognizing identity variable n_t. The joint distribution $\pi_t(n_t, \theta_t)$ is approximated and propagated using the CONDENSATION algorithm. The distribution $\pi_t(n_t)$ is a free estimate from $\pi_t(n_t, \theta_t)$, i.e., the true marginal distribution of $\pi_t(n_t, \theta_t)$.

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

In the following, Section 2 summarizes some related work in the literature. Section 3 starts with some basics of SIS algorithm and then presents the proposed recognition algorithm. Section 4 introduces the practical choices in modeling. Section 5 presents and discusses experimental results, and Section 6 concludes with final remarks.

2 Related Literature

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 [8] proposed the CONDENSATION algorithm to deal with the difficult problem of tracking an object in a cluttered environment. In [8], the object is represented by a robust active contour. Near real-time performance with high tracking accuracy is achieved. However, only the tracking problem is considered in [8].

Black and Jepson [2] used 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 human motion in a new image sequence. The joint posterior distribution of model selection, local stretching, scaling, and position evolves as time proceeds.

In [12], 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 over time form a trajectory. It is shown that the trajectory distance accumulates recognition evidence over time. However, this recognition is still deterministic.

Li and Chellappa [11] performed simultaneous tracking and verification via sequential posterior estimation. For each template k in the gallery, they rectifies it onto the first frame of the query video. Then, they invoke the SIS algorithm to obtain an updated set of samples for $\pi_t(\theta_t)$. To compute $\pi_t(n_t = k)$, they first evaluate the mean value $E_\pi(\theta_t)$ of θ_t, then marginalize $\pi_t(\theta_t)$ over a proper region A, i.e.,

$$\pi_t(n_t = k) = \int_A \pi_t(\theta_t) d\theta_t \tag{1}$$

where A is a hypercube around $E_\pi(\theta_t)$:

$$A = [E_\pi(\theta_t) - \Sigma, E_\pi(\theta_t) + \Sigma]. \tag{2}$$

They choose as the hypothesis k giving rise to the maximum probability $\pi_t(n_t = k)$. Experimental results on both synthetic data and real sequences (some using face information as well) are presented in [11]. Our method detailed in Section 4 is somewhat similar to this approach, but there are significant differences from it. We discuss them in Section 5.

3 SIS and Proposed Algorithm

In this section, we first introduce how to numerically solve a general time series state space model using SIS algorithm, then propose our algorithm as a special case.

3.1 SIS

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

1. State equation governing the state evolution:

$$x_t = g_t(x_{t-1}, u_t); t \geq 1, \tag{3}$$

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

2. Observation equation depicting the observed behavior:

$$y_t = h_t(x_t, v_t); t \geq 1, \tag{4}$$

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

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

$$\begin{aligned} u_t \perp v_s; \quad t, s \geq 1 \\ u_t \perp u_s, v_t \perp v_s; \quad t, s \geq 1 \ \& \ t \neq s \end{aligned} \tag{5}$$

where $\perp$ implies statistical independence.

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 over time the mean and variance which completely determines a Gaussian distribution. For nonlinear and non-Gaussian cases, extended Kalman filter (EKF) and its variants such as the iterated extended Kalman filter have been used to arrive at an approximate solution. Recently, the SIS technique, a special case of Monte Carlo method, [8, 6, 9, 13] 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 $\pi(x)$ 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

$$w^{(m)} = \pi(x^{(m)})/g(x^{(m)}) \tag{6}$$

is assigned to each sample. This technique is called *Importance Sampling* (IS). It can be shown[13] 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* $g_t(x_t|x_{t-1})$, and calculates the weight using

$$w_t = w_{t-1} p_t(y_t|x_t) p_t(x_t|x_{t-1}) / g_t(x_t|x_{t-1}). \tag{7}$$

In the CONDENSATION algorithm [8], $g_t(x_t|x_{t-1})$ is taken to be $p_t(x_t|x_{t-1})$ and Eq. (7) becomes

$$w_t = w_{t-1} p_t(y_t|x_t), \tag{8}$$

In fact, Eq. (8) 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 $p_t(y_t|x_t)$. For a complete description of the SIS method, refer to [6, 13].

3.2 Proposed Algorithm

In the context of this problem, the posterior probability distribution $\pi_t(n_t, \theta_t)$ is represented by a set of **indexed and weighted** samples $\mathcal{S}_t = \{(n_t^{(m)}, \theta_t^{(m)}, w_t^{(m)})\}_{m=1}^{M}$ with n_t as the **index**. It can be easily shown [17] that the sample set $\{n_t, \beta_{n_t}\}_{n_t=1}^{N}$ representing the distribution $\pi_t(n_t)$ can be constructed by summing the weights of the samples belonging to the same index n_t, i.e.,

$$\beta_{n_t} = \sum_{m=1, n_t^{(m)}=n_t}^{M} w_t^{(m)}. \tag{9}$$

The algorithm shown in Fig. 1 is an extension to CONDENSATION [8] for computing the joint distribution $\pi_t(n_t, \theta_t)$.

4 Model Choices

In this section, we specify the practical model choices used in this paper.

1. State equation (3) consists of two subequations.

- Motion subequation:

$$\theta_t = \theta_{t-1} + u_t; \quad t \geq 1, \tag{10}$$

Initialize *a sample set* $\mathcal{S}_0 = \{(n_0^{(m)}, \theta_0^{(m)}, 1)\}_{m=1}^{M}$
according to prior distributions $p(n_0|y_0)$ *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* $\alpha_t^{(m)} = p(y_t|n_t^{(m)}, \theta_t^{(m)})$.
  End
  Normalize *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

Figure 1. The proposed algorithm.

where u_t is the *motion noise* at time t. It is assumed that u_t is time-invariant, Gaussian with its mean and covariance matrix manually specified. This is a first-order Gaussian-Markov motion model.

- Identity subequation:

$$n_t = n_{t-1}; \quad t \geq 1, \tag{11}$$

assuming that identity does not change as time proceeds.

2. Observation equation:

$$f(y_t; \theta_t) = I_{n_t} + v_t; \quad t \geq 1, \tag{12}$$

where v_t is the *observation noise* at time t. It is assumed that $p(v_t)$ or the likelihood $p(y_t|n_t, \theta_t)$ is a 'truncated' Laplacian:

$$p(y_t|n_t, \theta_t) = \begin{cases} L \exp\left(-\|I_{n_t} - f(y_t; \theta_t)\|/\sigma\right) \\ \quad \text{if } \|I_{n_t} - f(y_t; \theta_t)\| \leq \lambda\sigma \\ L \exp\left(-\lambda\right) \\ \quad \text{if } \|I_{n_t} - f(y_t; \theta_t)\| > \lambda\sigma, \end{cases}$$

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.

Also, $\theta = (a_{11}, a_{12}, a_{21}, a_{22}, t_x, t_y)$ where $\{a_{11} - a_{22}\}$ are deformation parameters and $\{t_x, t_y\}$ are the 2-D translation parameters. $f(y_t, \theta_t)$ is obtained by the following procedures: firstly, an affine geometric transformation is applied to the whole image with $\{a_{11} - a_{22}\}$; then we crop the image part centered at $\{t_x, t_y\}$, with its size same as the template; finally, histogram equalization is applied for enhancement.

Figure 2. The image database used in experiment. The image size is 60x60.

3. Prior distribution $p(\theta_0)$ is Gaussian, whose mean and covariance matrix are manually specified. Prior distribution $p(n_0)$ is uniform on $\mathcal{N}$, i.e.,

$$p(n_0) = 1/N; \quad n_0 = 1, 2, \ldots, N. \tag{13}$$

4. Statistical independence other than those established in equation (5):

$$n_0 \perp \theta_0. \tag{14}$$

Obviously, this model is nonlinear by the nature of the nonlinear transformation $f(y_t, \theta_t)$, and non-Gaussian by the nature of the observation noise v_t. Therefore, we proceed to compute the joint posterior distribution $\pi_t(n_t, \theta_t)$ using the CONDENSATION algorithm proposed in Section 3. $\pi_t(n_t)$ is simply the marginal distribution of $\pi_t(n_t, \theta_t)$.

5 Experiments and Discussions

In this section, we first present the experimental results using these choices, followed by discussions.

5.1 Experimental Results

In our experiment, we captured video sequences with subjects walking towards a camera in order to simulate typical scenarios like in visual surveillance. Using the terminology of the FERET test [14], the gallery set as shown in Fig. 2 contains 12 still images, one for each subject, and the probe set contains 12 query video sequences, one for each subject. Fig. 3 gives some example frames in one query. Note the considerable change in scale.

Suppose the correct identity for Fig. 3 is c. Fig. 4 presents the posterior probability $\pi_t(n_t)$ and the number of samples sticking to the hypotheses. From Fig. 4, we can easily observe that the posterior probability $\pi_t(n_t = c)$ increase as time proceeds, which is also evidenced by the number of samples sticking to hypotheses. This is not surprising at all. The evolution of $\pi_t(n_t)$ characterizes a competition for samples among the identities of humans. Since it is assumed that the identity keeps unchanged over time, we accumulate evidence in a recursive manner such that

Figure 3. Example frames in one query. The image size is 320x240 while the actual face size ranges approximately from 20x20 in the first frame to 60x60 in the last frame.

more and more samples contribute to the identity with increasing confidence. Finally, $\pi_t(n_t)$ becomes degenerate in this identity. For an analytical derivation of the evolution of $\pi_t(n_t)$, refer to [17].

To capture the evolution of $\pi_t(n_t)$, we use the notion of entropy [5]. Given a conditional PMF $p(x|y)$ the conditional entropy is defined as:

$$\begin{aligned} H(x|y) &= -\sum_{x,y} p(x,y) \log_2 p(x|y) \\ &= -\sum_{y} p(y) \sum_{x} p(x|y) \log_2 p(x|y). \end{aligned} \tag{15}$$

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,

$$H(n_t|y_{0:t}) = -\sum_{n_t \in \mathcal{N}} p(n_t|\tilde{y}_{0:t}) \log_2 p(n_t|\tilde{y}_{0:t}). \tag{16}$$

It is well known that among all distributions taking values on $\{1, \ldots, N\}$, the uniform distribution yields a maximum of $\log_2 N$ and the degenerate case yields the minimum of 0:

$$0 \leq H(n_t|y_{0:t}) \leq \log_2 N \tag{17}$$

Fig. 5 plots the conditional entropy $H(n_t|y_{0:n})$ versus time t. As expected, $H(n_t|y_{0:n})$ starts from $\log_2(12) = 3.59$, decreases at time proceeds, and finally reaches 0.

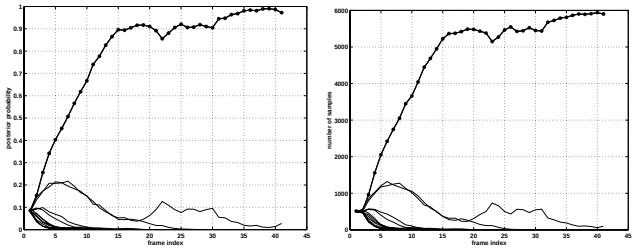

Figure 4. Left: posterior probability $p(n_t|y_{0:t})$ against t. Right: number of samples sticking to hypotheses against t.

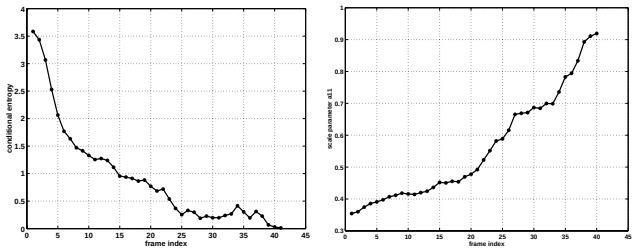

Figure 5. Left: conditional entropy $H(n_t|y_{0:t})$ against t. Right: MAP estimate of a_{11} against t.

Fig. 5 also shows the MAP estimate of the scale parameter a_{11} against frame index t. The scale increases as time proceeds, which matches the scenario where a subject is walking towards a camera. In Fig. 3, the tracked parameter is superposed on the image using a bounding box.

Table 1 summarizes the average recognition performance and computational time obtained for this database. 100% recognition is reached possibly due to the small size of this database. However, this algorithm is not so efficient in terms of computational time. Note that this experiment is implemented in C++ on a PC with P-II450 CPU and 512M RAM and the number of motion samples J is chosen to be 200.

Recognition Rate	100%
Time per Frame	16s

Table 1. A summary of the algorithm.

5.2 Discussions and Future Work

The following issues are worthy of investigation in future.

1. Modeling geometric transform as affine. Affine transform is a good approximation as long as there is no out-of-plane motion. The scenario with a subject walking towards a camera roughly satisfies this. Also histogram equalization is a typical but a coarse method to deal with changing illumination.

2. Choice of constant-velocity dynamic 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 samples of θ can cover the motion, this model can be applied for simplicity. One future work is to train the dynamical model by examples beforehand or by sample trajectories formed by the computations up to present.

3. Choice of likelihood distribution $p(y_t|n_t, \theta_t)$. In general, $p(y_t|n_t, \theta_t)$ is a function of $\|v_t\| = \|f(y_t; \theta_t) - I_{n_t}\|$. The smaller $\|v_t\|$ is, the higher the likelihood $p(y_t|n_t, \theta_t)$ and 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?

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 high-frequency residual of natural images is more Laplacian-like. The 'truncated' Laplacian is used to give a 'surviving' chance for samples counting for abrupt motion changes. In this framework, we can easily incorporate image representations other than intensity values. We are now exploring features like PCA, and LDA, and their corresponding observation noise models.

4. Computational load. The proposed algorithm is not computationally efficient. Two important numbers affecting the computation are the number of motion samples J, and the size of database N. The actual sample number $M = J * N$. (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. An efficient algorithm without compromising the accuracy has been designed [17].

5. Co-influence of tracking and recognition. Since the joint posterior distribution is computed each time, the co-influence is obvious. If tracking fails, recognition is meaningless because we are not recognizing the face any more. If recognition is poor, for instance, some background part in the video may be more favored than the face part according to the distance measure; tracking will then stick to the background. 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 [11]. (i) Problem Formulation. In [11], 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 one more degree of freedom and leads to a different approach for deriving the solution. (ii) Solution to the problem. The SIS technique is applied in both approaches to numerically approximate the posterior probability given the observation. Again in [11], it is the posterior probability of the state vector, and the verification probability is estimated by marginalizing over a proper region of the state space redefined at each time instance. However, we always compute the joint density, i.e., the posterior probability of the state vector and the identity variable. The posterior probability of the identity variable is just a free estimate by marginalizing over the state vector. (iii) Performance. Notice that there is no time propagation of verification probability in [11] while we always propagate the joint density. One consequence is that we guarantee that $\sum_{n_t \in \mathcal{N}} \pi_t(n_t) = 1$, but there is no such guarantee in [11]. Their approach in some sense is more like a batch method, 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 [11]. 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

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

However, this model is nonlinear and non-Gaussian, leading to the possibility of an analytic solution being not available. An extension of CONDENSATION [8] is applied to provide a numerical solution. 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 be used as a good indication for convergence.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recugnition zsing class specific linear projection. *IEEE Trans. PAMI*, 19, 1997.

[2] M. J. Black and A. D. Jepson. A probabilistic frameworkfor matching temporal trajectories. *Proc. of ICCV*, pages 176–181, 1998.

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

[4] 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 Authenticamion*, pages 176–731, 1999.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[6] 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.

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

[8] M. Isard and A. Blake. Contour tracking by stochatic propagation of conditional density. *Proc. of ECCV*, 1996.

[9] G. Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *J. Computational and Graphical Statistics*, 5:1–25, 1996.

[10] J. C. Lades, M. ahd 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 Trans. Computers*, 42:300–311, 1993.

[11] B. Li and R. Chellappa. Simultaneous tracking and verification via sequential posterior estimation. *Proc. of CVPR*, pages 110–117, 2000.

[12] Y. Li, S. Gong, and H. Liddell. Conjtucting structures of facial identities on the view sphere using kernel discriminant analysis. *Proc. of the 2rd Intl. Workshop on SCTV*, 2001.

[13] J. S. Liu and R. Chen. Sequential monte carlo for dynamic systems. *Journal of the American Statistical Association*, 93:1031–1041, 1998.

[14] P. J. Philipps, H. Moon, S. Rivzi, and P. Ross. The feret testing protocol. *Face Recognition: From Theory to Applications*, 83:244–261, 1998.

[15] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neutoscience*, 3(1):71–86, 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.

[17] S. Zhou and R. Chellappa. Probabilistic human recgnition from video. *To appear in Proc. of ECCV*, 2002.