Adaptive Visual Tracking and Recognition
using Particle Filters

Shaouha Zhou [*]
Rama Chellappa [*]
Baback Moghaddam [†]

Abstract

This paper presents an improved method for simultaneous tracking and recognition of human faces from video where a time series model is used to resolve the uncertainties in tracking and recognition. The improvements mainly arise from three aspects: (i) modeling the inter-frame appearance changes within the video sequence using an adaptive appearance model and an adaptive-velocity motion model; (ii) modeling the appearance changes between the video frames and gallery images by constructing intra- and extra-personal spaces; and (iii) utilization of the fact that the gallery images are in frontal views. By embedding them in a particle filter, we are able to achieve a stabilized tracker and an accurate recognizer when confronted by pose and illumination variations.

[*] University of Maryland
[†] MERL - Research Laboratory

ADAPTIVE VISUAL TRACKING AND RECOGNITION USING PARTICLE FILTERS

Shaohua Zhou, Rama Chellappa, and Baback Moghaddam

Center for Automation Research (CfAR)
University of Maryland, College Park, MD 20742
{shaohua, rama}@cfar.umd.edu

Mitsubishi Electric Research Laboratories (MERL)
201 Broadway, Cambridge, MA 02139
baback@merl.com

ABSTRACT

This paper presents an improved method for simultaneous tracking and recognition of human faces from video [1], where a time series model is used to resolve the uncertainties in tracking and recognition. The improvements mainly arise from three aspects: (i) modeling the inter-frame appearance changes within the video sequence using an adaptive appearance model and an adaptive-velocity motion model; (ii) modeling the appearance changes between the video frames and gallery images by constructing intra- and extra-personal spaces; and (iii) utilization of the fact that the gallery images are in frontal views. By embedding them in a particle filter, we are able to achieve a stabilized tracker and an accurate recognizer when confronted by pose and illumination variations.

1. INTRODUCTION

Video-based face recognition entails disambiguating uncertainties in both tracking and recognition. While conventional methods [2] resolve both uncertainties separately, i.e. after tracking is accomplished, recognition is applied, we have proposed in [1] a framework to model both uncertainties in a unified way to realize simultaneous tracking and recognition. As evidenced by the empirical results (on a modest databases) in [1], this algorithm improves its recognition rate over the conventional ones without sacrificing accuracy in tracking.

Though the time series formulation allows very general models, our earlier effort invoked rather simple models. For example, only a simple constant-velocity motion model with fixed noise variance was used; with a fixed noise variance it is hard to reach a compromise between rapid movement (favoring large variance) and attaining computational efficiency (against large variance). Also, a simple Laplacian density model based on the distance to a fixed template was used to deal with appearance changes between the frames. Secondly, modeling appearance changes between probe video and gallery set could have been more accurate. Finally, prior knowledge that all gallery images are in frontal views was not used. All these factors may yield unsatisfactory results in both tracking and recognition when confronted by pose and illumination variations.

This paper attempts to improve our previous approach in the following three aspects. (i) Modeling the inter-frame appearance changes within the video sequence using an adaptive appearance model [3] and an adaptive-velocity motion model, both adaptive to the observations. (iii) Modeling the appearance changes between the video frames and gallery images by constructing intra-

and extra-personal spaces which can be treated as a 'generalized' version of discriminative analysis [4]. (iii) Utilization of the fact that the gallery images are in frontal views. By embedding them in a particle filter, we are able to achieve a stabilized tracker and an accurate recognizer when confronted by pose and illumination variations.

The rest of the paper is organized as follows. After a brief review of the time series model for recognition in Sec. 2, we describe in Sec. 3 the three features that improve the performance. Experimental results and discussions are then presented in Sec. 4. Sec. 5 concludes the paper.

2. REVIEW OF SIMULTANEOUS TRACKING AND RECOGNITION

In this section, we briefly present the propagation model for recognition, consisting of the following three components, namely the motion transition equation, the identity equation, and the observation likelihood. and define the recognition task as a statistical inference problem, which can be solved using particle filters.

2.1. Motion Transition Equation

Denote the motion parameter by θ_t. 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 [5]. However such a model may overfit the training data and may not necessarily succeed with the testing video where objects can move arbitrarily 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 [1], i.e., $\theta_t = \theta_{t-1} + u_t$, where u_t has a fixed noise variance, say $u_t = r_0 * u_0$ and r_0 is a fixed constant measuring the noisy extent and u_0 is a 'standardized' random variable/vector. If r_0 is small, it is very hard to model the rapid movement; if r_0 is large, many more particles are needed to accommodate the large noise variance, yielding computational inefficiency. All these factors make the use of such a model ineffective. In this paper, we propose a generalization using an adaptive-velocity model. Our strategy is, at time t, to propagate only the point estimate $\hat{\theta}_{t-1}$ (we use the MAP estimate in the experiment) and predict its motion velocity ν_t using a first-order approximation, and diffuse it using additive noise $u_t = r_t * u_0$ with adaptive noise variance r_t. Sec. 3.1 presents a method for computing ν_t and r_t. In summary, we have

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

2.2. Identity Equation

Denoting the identity variable by $n_t \in \mathcal{N} = \{1, 2, ..., N\}$, indexing the gallery set $\{I_1, ..., I_N\}$ with each individual n possessing one facial image I_n in frontal view, and assuming that the identity does not change as time proceeds, we have

$$n_t = n_{t-1}, \quad t \geq 1. \qquad (2)$$

In practice, one may assume a slight transition probability between identity variables for increasing the robustness.

2.3. Observation Likelihood

In [1], our empirical results show that combining contributions (or scores) from both tracking and recognition in the likelihood yields the best performance in both tracking and recognition. We continue our effort along this line.

To compute the tracking score which measures the inter-frame appearance changes, we introduce an appearance model a_t for tracking [1], i.e.,

$$y_t \equiv \mathcal{T}\{z_t; \theta_t\} = a_t + v_t, \quad t \geq 1, \qquad (3)$$

where y_t is the image patch of interest in the video frame z_t, parameterized by θ_t, and noise component v_t determines the tracking score $p_a(z_t|\theta_t)$. Note that y_t is a transformed version of the observation z_t and this transformation could be either geometric or photometric or both.

In [6], a fixed template, $a_t \equiv a_0$, is matched with observations to minimize a cost function in the form of sum of squared distance (SSD). This is equivalent to assuming that 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, say $a_t = \hat{y}_{t-1}$, i.e., the 'best' patch of interest in the previous frame. A fixed template cannot handle appearance changes in the video, while a rapidly changing model is susceptible to drift. In [3], Jepson *et. al.* proposed an online appearance model (OAM) to realize a robust visual tracker, which is a mixture of three components. We use a modified version of the OAM model detailed in Sec. 3.1 and the actual calculation of $p_a(z_t|\theta_t)$ is also detailed there.

To compute the recognition score which measures the appearance changes between probe videos and gallery images, we assume that the transformed observation is a noise-corrupted version of some still template in the gallery, i.e.,

$$y_t = I_{n_t} + w_t, \quad t \geq 1, \qquad (4)$$

where w_t is the *observation noise* at time t, whose distribution determines the recognition score $p_n(z_t|n_t, \theta_t)$. We will physically define this quantity in Sec. 3.3.

To fully exploit the fact that all gallery images are in frontal view, we also compute in Sec. 3.2 how likely the patch y_t is in frontal view and denote this score by $p_f(z_t|\theta_t)$. If the patch is in frontal view, we believe in the 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

$$p(z_t|n_t, \theta_t) = p_a \{p_f \, p_n + (1 - p_f) \, N^{-1}\}. \qquad (5)$$

---

[1] We denote: $y_t = \mathcal{T}\{z_t; \theta_t\}$, $y_t^{(j)} = \mathcal{T}\{z_t; \theta_t^{(j)}\}$, $\hat{y}_t = \mathcal{T}\{z_t; \hat{\theta}_t\}$.

2.4. Particle Filter: Solving the Model

We assume statistical independence between all noise variables and prior knowledge on the distributions $p(\theta_0|z_0)$ and $p(n_0|z_0)$ (uniform prior in fact). 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-type distribution. Therefore, the problem is reduced to computing the posterior probability.

Since the model is nonlinear and non-Gaussian in nature, there is no analytic solution. We invoke a particle filter [5], a special case of Sequential Monte Carlo (SMC) methods [7], to provide numerical approximations to the posterior distribution $p(n_t, \theta_t|z_{0:t})$. Also, for this mixed-type distribution, we can greatly improve the computational load by judiciously utilizing the discrete nature of the identity variable as in [1]. We [1] also theoretically justified the evolving behavior of the recognition density $p(n_t|z_{0:t})$ under a weak assumption.

3. MODEL COMPONENTS IN DETAIL

As mentioned in Sec. 1, the proposed algorithm incorporates three components which improve our previous approach. We will now examine each of these components in greater detail. The proposed algorithm is then summarized.

3.1. Modeling Inter-Frame Appearance Changes

Inter-frame appearance changes are related to the motion transition model and the appearance model for tracking. Our attempt is to make them both adaptive to the incoming frames.

3.1.1. Adaptive Appearance Model for Tracking

The OAM assumes that the observations are explained by different causes, thereby indicating the use of a mixture density of components. In the original OAM presented in [3], 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. In our implementation, we have not incorporated the L-component because there is no occlusion in our test video. Instead, to further stabilize our tracker, we have used an F-component which is a fixed template that we are expecting to observe most often. For example, in our experiment this could be just the facial image as seen from a frontal view.

We assume that the observation at time t is generated by the appearance model at time t, $a_t = \{W_t, S_t, F_t\}$, obeying a mixture of Gaussians, with W_t, S_t, F_t as mixture centers $\mu_{i,t}; i = w, s, f$. Notice that a_t only models the appearances present in all observations up to time $t - 1$. The tracking score is written as

$$p_a(z_t|\theta_t) = \sum_{i=w,s,f} m_{i,t}\mathbf{N}(y_t; \mu_{i,t}, \sigma_{i,t}^2), \qquad (6)$$

where $\{m_{i,t}; i = w, s, f\}$ are the mixing probabilities, $\{\sigma_{i,t}^2; i = w, s, f\}$ are the variances for corresponding components.

It remains to show how to update the current appearance model a_t to a_{t+1} after frame t has been tracked, i.e., having $\hat{y}_t$ available, we want to compute the new mixing probabilities, mixture centers,

and variances for time $t+1$, $\{m_{i,t+1}, u_{i,t+1}, \sigma^2_{i,t+1}; i = w, s, f\}$. We just sketch the updating equations here and refer interested readers to [3] for technical details and justifications. With a pre-defined 'updating factor' α, the updating equations are:

$$o_{i,t} = m_{i,t}\mathbf{N}(\hat{y}_t; \mu_{i,t}, \sigma^2_{w,t})/p_a(z_t|\hat{\theta}_t); \quad i = w, s, f. \quad (7)$$

$$m_{i,t+1} = \alpha o_{i,t} + (1-\alpha)m_{i,t}; \quad i = w, s, f. \quad (8)$$

$$M_{p,t+1} = \alpha \hat{y}^p_t o_{s,t} + (1-\alpha)M_{p,t}; \quad p = 1, 2. \quad (9)$$

$$S_{t+1} = \mu_{s,t+1} = \frac{M_{1,t+1}}{m_{s,t+1}}, \quad \sigma^2_{s,t+1} = \frac{M_{1,t+1}}{m_{s,t+1}} - \mu^2_{s,t+1}. \quad (10)$$

$$W_{t+1} = \hat{y}_t, \quad F_{t+1} = F_1, \quad \sigma^2_{i,t+1} = \sigma^2_{i,1}; \quad i = w, f. \quad (11)$$

To initialize a_1, we manually set $W_1 = S_1 = F_1 = F$ with F given, $\{m_{i,1}, \sigma^2_{i,1}\}; i = w, s, f$, and $M_{1,1} = m_{s,1}F$ and $M_{2,1} = m_{s,1}\sigma^2_{s,1} + F^2$. Notice that the representation for an OAM is quite general in the sense that it can be based on pixel intensity values, or other features extracted from the intensity values such as phase information derived from steerable filters as described in [3].

3.1.2. Adaptive Motion Transition Model

With the availability of the sample set $\mathcal{S}_{t-1} = \{\theta^{(j)}_{t-1}\}^J_{j=1}$ and the image patches of interest $\mathcal{Y}_{t-1} = \{y^{(j)}_{t-1}\}^J_{j=1}$, for a new observation z_t, we can predict the shift ν_t in the motion parameter using a first-order linear approximation [6], which essentially comes from the constant brightness constraint. It reads as

$$\nu_t = B_t * [\mathcal{T}\{z_t; \hat{\theta}_{t-1}\} - \hat{y}_{t-1}], \quad (12)$$

where the B_t matrix can be estimated from available data.

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: $X = [\theta^{(1)}_{t-1}, \ldots, \theta^{(J)}_{t-1}] - \hat{\theta}_{t-1} * \mathbf{1}^{tr}$, $Y = [y^{(1)}_{t-1}, \ldots, y^{(J)}_{t-1}] - \hat{y}_{t-1} * \mathbf{1}^{tr}$, where $\mathbf{1}$ is a column vector of 1's. The least square (LS) solution for B_t is $B_t = (XY^{tr}) * (YY^{tr})^{-1}$, where $(.)^{tr}$ denotes matrix transposition. To avoid the explicit inversion of the matrix YY^{tr}, we use the singular value decomposition (SVD), say $Y = USV^{tr}$. It can be easily shown that $B_t = XVS^{-1}U^{tr}$. Also, we gain some computational efficiency at the cost of some accuracy by retaining the top q components of the SVD decomposition, i.e., $B_t = XV_qS^{-1}_qU^{tr}_q$.

In practice, one may have to run several iterations till the patch $\mathcal{T}\{z_t; \hat{\theta}_{t-1}+\nu_t\}$ stabilizes, i.e., till the error $\epsilon_t = \phi(\mathcal{T}\{z_t; \hat{\theta}_{t-1}+\nu_t\}, a_t)$, which measures the distance between $\mathcal{T}\{z_t; \hat{\theta}_{t-1}+\nu_t\}$ and the updated appearance model a_t, reaches below a threshold.

The value of ϵ_t determines the quality of prediction. Therefore, if ϵ_t is small, which implies a good prediction, we only need tightly-supported noise to absorb the residual motion; if ϵ_t is large, which implies a poor prediction, we then need widely-dispersed noise 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. However, we keep lower and upper bounds on r_t. We use the following form:

$$r_t/r_0 = \eta(\epsilon_t; a, b, c) = a + (b-a)tanh(\epsilon_t/c), \quad (13)$$

where a, b are lower and upper bounds respectively, and c is the rate. Fig. 2(a) shows the function $\eta(x; 0.5, 2, 1)$.

Initialize *a sample set* $\mathcal{S}_0 = \{\theta^{(j)}_0, w^{(j)}_0 = 1/J_0\}^{J_0}_{j=1}$ *according to prior distribution* $p(\theta_0|z_0)$. *Set* $\beta_{0,l} = 1/N$.
For $t = 1, 2, \ldots$
    Calculate *the MAP estimate* $\hat{\theta}_{t-1}$, *the adaptive motion shift* ν_t *by Eq. (12), the noise variance* r_t *by Eq. (13), and the particle number* J_t *by Eq. (17).*
    For $j = 1, 2, \ldots, J_t$
    Draw *the sample* $u^{(j)}_t$ *for* u_t *with variance* r_t. *Construct the sample* $\theta^{(j)}_t$ *by Eq. (1). Compute transformed image* $y^{(j)}_t$.
        For $l = 1, 2, \ldots, N$
        Update *the weight using* $\alpha^{(j)}_{t,l} = \beta_{t,l}p(z_t|l, \theta^{(j)}_t)$ *by Eq. (5).*
        End
    End
    Normalize *the weight using* $w^{(j)}_{t,l} = \alpha^{(j)}_{t,l}/\sum_{j,l}\alpha^{(j)}_{t,l}$ *and compute* $w^{(j)}_t = \sum_l w^{(j)}_{t,l}$ *and* $\beta_{t,l} = \sum_j w^{(j)}_{t,l}$.
End

Fig. 1. The proposed algorithm.

3.2. Score of Being in Frontal View

Since all gallery images are in frontal view, we simply build such a score by fitting a probabilistic subspace (PS) density on top of the gallery images [4], assuming that they are i.i.d. samples from the frontal face space (FFS). It proceeds as follows: a regular PCA is first performed (after removing the sample mean). Suppose the eigensystem for the FFS is $\{(\lambda_i, e_i)\}^d_{i=1}$, where d is the number of pixels and $\lambda_1 \geq \ldots \geq \lambda_d$. Only top s principal components corresponding to top s eigenvalues are then kept while the residual components are considered as isotropic. We refer the reader to the original paper [4] for full details. The PS density is written as follows:

$$Q(x) = \left\{ \frac{exp(-\frac{1}{2}\sum^s_{i=1}\frac{q^2_i}{\lambda_i})}{(2\pi)^{s/2}\prod^s_{i=1}\lambda^{1/2}_i} \right\} \left\{ \frac{exp(-\frac{err^2}{2\rho})}{(2\pi\rho)^{(d-s)/2}} \right\}, \quad (14)$$

where $q_i = e^T_i x$ for $i = 1, \ldots, s$ is the ith principal component of x, $err^2 = \|x\|^2 - \sum^s_{i=1}q^2_i$ is the reconstruction error, and $\rho = (\sum^d_{i=s+1}\lambda_i)/(d-q)$. It is easy to write $p_f(z_t|\theta_t)$ as follows:

$$p_f(z_t|\theta_t) = Q_{FFS}(y_t). \quad (15)$$

3.3. Modeling Appearance Changes between Probe Video Frames and Gallery Images

We adopt the MAP rule developed in [4] 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 variations in appearances belonging to the same identity while the extra-personal space (EPS) is used to cover all variations in appearances belonging to the different identities. At least two facial images for one identity are needed to construct the IPS. Apart from the available gallery, we crop out the second image from the video ensuring no overlap with frames used in probe videos. The above PS density density is fitted separately on top of the IPS and the EPS, yielding two different eigensystems. The recognition score $p_n(z_t|n_t, \theta_t)$ is finally computed as

$$p_n(z_t|n_t, \theta_t) = Q_{IPS}(y_t - I_{n_t})/Q_{EPS}(y_t - I_{n_t}). \quad (16)$$

3.4. Proposed Algorithm

We adjust the particle number J_t based on the following two heuristics. (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. (ii) As proved in [1], 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 a weak assumption). Accordingly, we increase the particle number by a fixed amount d if H_t increases; otherwise we deduct d from J_t. Combining these two, we have, with $I[.]$ being an indication function

$$J_t = J_0 \{ \eta(\epsilon_t; a, b, c) + d * (-1)^{I[H_{t-1} < H_{t-2}]} \}. \quad (17)$$

The proposed algorithm is summarized in Fig. 1, where $w_{t,l}^{(j)}$ is the weight of the particle $(n_t = l, \theta_t = \theta_t^{(j)})$ for the posterior density $p(n_t, \theta_t|z_{0:t})$; $w_t^{(j)}$ is the weight of the particle $\theta_t = \theta_t^{(j)}$ for the posterior density $p(\theta_t|z_{0:t})$; and $\beta_{t,l}$ is the weight of the particle $n_t = l$ for the posterior density $p(n_t|z_{0:t})$.

4. EXPERIMENTAL RESULTS

In our implementation, we used the following practical choices. We consider affine transformations only. 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 $\{t_x, t_y\}$ denote the 2-D translation parameters. Even though significant pose/illumincation changes are present in the video, we believe that our adaptive appearance model can easily absorb them and therefore for our purposes the affine transformation is a reasonable approximation. Regarding photometric transformations, only a zero-mean-unit-variance normalization is used to partially compensate for contrast variations. The complete image transformation $\mathcal{T}\{Y; \theta\}$ is implemented as follows: affine transform z using $\{a_1, a_2, a_3, a_4\}$, crop out the region of interest at position $\{t_x, t_y\}$ with the same size as the still template in the appearance model as well as in the gallery set, and perform the zero-mean-unit-variance normalization.
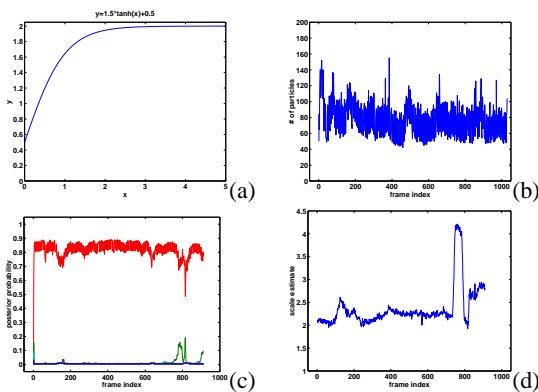
Fig. 2. (a): The function $y(x) = \eta(x; 0.5, 2, 1)$. (b) The particle number J_t *vs.* t. (c) Posterior probability $p(n_t|z_{0:t})$ *vs.* t. (d) The scale estimate for Subject-2.

We have applied our algorithm to tracking and recognizing human faces captured by a hand-held video camera in an office environments (where both camera and target motion are present).

There are 29 subjects in the database. A 100% recognition rate is achieved. Fig. 3 presents the tracking results on the video sequence for 'Subject-2' featuring quite large pose variations, moderate illumination variations, and quick scale changes (back and forth motion toward the end of the sequence). Fig. 2(b) shows the number of particles J_t against time t with $J_0 = 100$, averaging 77 particles per frame. This is much more efficient than a particle filter with fixed $J_0 = 100$. The posterior probability for 'Subject-2' is plotted in Fig. 2(c). It is very fast, taking about less than 10 frames, to reach above 0.9 level. This is mainly attributed to the discriminative power of the MAP recognition score induced by intra- and extra-personal spaces modeling. Fig. 2(d) captures the quick scale changes (a sudden increase followed by a decrease within about 50 frames) available in the video sequence by plotting the scale estimate, computed as $\sqrt{(a_1^2 + ... + a_4^2)/2}$, as a function of t.

Fig. 3. Tracking results. Frames 160, 290, 690, 750, and 800, 240x360 pixels in size, in a 890-frame-long sequence for Subject-2. The corner shows the S-component and W-component with enlarged size (the original size is 30 by 26).

5. CONCLUSION

We have improved our simultaneous tracking and recognition approach proposed in [1]. More complex models, namely adaptive appearance model, adaptive-velocity transition model, and intra- and extra-personal spaces model, are introduced to handle appearance changes between frames and between frames and gallery images. The fact that gallery images are in frontal view is enforced too. Experimental results demonstrate that tracking is stable and the recognition performance has improved.

6. REFERENCES

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

[2] T. Jebara and A. Pentland, "Parameterized structure from motion for 3D adaptive feedback tracking of faces," *Proc. of CVPR*, pp. 144–150, 1997.

[3] A. D. Jepson, D. J. Fleet, and T. El-Maraghi, "Robust online appearance model for visual tracking," *Proc. of CVPR*, vol. 1, pp. 415–422, 2001.

[4] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. PAMI*, vol. 24, pp. 780–788, 2002.

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

[6] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. on PAMI*, vol. 20, no. 10, pp. 1025–1039, 1998.

[7] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, 2001.