A Robust Algorithm for Probabilistic Human Recognition From Video

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

Abstract

Human recognition from video requires solving the two tasks, recognition and tracking, simultaneously. This leads to a parameterized time series state space model, representing both motion and identity of the human. Sequential Monte Carlo (SMC) algorithms, like CONDENSATION [3], can be developed to offer numerical solutions to this model. However, in outdoor environments, the solution is more likely to diverge from the foreground, causing failures in both recognition and tracking. In this paper, we propose an approach for tackling this problem by incorporating the constraint of temporal continuity in the observations. Experimental results demonstrate improvements over its CONDENSATION counterpart.

1. Introduction

Bayesian analysis of video has recently gained significant attention in the computer vision community since the seminal work of Isard and Blake [3]. 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 approach has been extended to many areas [1, 6, 11], including human recognition using a variety of biometrics, e.g., face, human body. Refer to [4] for a general review of statistical pattern recognition, and to [10] for a recent survey on face recognition.

Reported in [11] is an extension to video-based human recognition, with the gallery and the probe [9] consisting of still templates and video sequences respectively. In this work, the time series state space model is re-parameterized by a tracking state vector and a recognizing identity variable, characterizing the dynamics and identity of a human, assuming temporal constancy in the identity variable. The observation is modeled as a transformed and noise-corrupted version of a template in the gallery. By employing the SMC technique, the joint distribution of the state vector and the identity variable 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 marginal distribution of the identity variable is just a free estimate, which provides the recognition result.

However, it turns out that, when confronting some nontrivial circumstances, e.g., inhomogeneous background and non-uniformly illuminated foreground, this formulation may lead to a solution that is more likely to move away from the foreground, causing failures in both tracking and recognition. In this paper, we propose a two-stage approach to tackle this problem by separating the tasks of tracking and recognition at one stage and integrating them at the other stage. This approach essentially incorporates the temporal continuity constraint in the observations [7]. The SMC framework is used to compute the posterior probability.

In the following, Sec. 2 briefly reviews the time series state space model for recognition and practical model choices. Sec. 3 first introduces the CONDENSATION like algorithm to solve the model, addresses its weakness confronting the nontrivial situations, and then proposes the two-stage SMC algorithm in order to overcome the weakness. Experimental results are included, and Sec. 4 concludes the paper.

2 State Space Model for Recognition

Let I denote an image represented using raw intensity values on the image region R, i.e. $I(R)$, and let $f(I;\theta)$ denote the transformed version of image I undergoing a photometric transformation (e.g. histogram equalization) and a geometric transformation, parameterized by an affine state

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

vector $\theta = (a_1, a_2, a_3, a_4, t_x, t_y)$ where $\{a_1, a_2, a_3, a_4\}$ are deformation parameters and $\{t_x, t_y\}$ are 2-D translation parameters . A hypothesis gallery H consists of still template images, i.e., $H = \{I_1, \ldots, I_N\}$ which is indexed by an identity variable n belonging to a finite sample space $\mathcal{N} = \{1, \ldots N\}$. The time series state space model for recognition consists of the following components.

1. State equation. Let θ_t be the *state vector* and u_t the *state noise* respectively at time t,

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

Assume that $p(u_t)$ or $p(\theta_t|\theta_{t-1})$ is time-invariant and Gaussian, with its mean and covariance matrix manually set.

2. Identity equation. Assuming that identity does not change as time proceeds,

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

3. Observation equation. Let y_t be the *observation* and v_t the *observation noise* at time t respectively,

$$f(y_t; \theta_t) = I_{n_t} + v_t; \qquad t \geq 1, \tag{3}$$

Assume $p(v_t)$ or the likelihood $p(y_t|n_t, \theta_t)$ to be time-invariant and truncated-Laplacian:

$$p(y_t|n_t, \theta_t) \propto \begin{cases} \exp(-\|v_t\|/\sigma) & \text{if } \|v_t\| \leq \lambda\sigma \\ \exp(-\lambda) & \text{if } \|v_t\| > \lambda\sigma, \end{cases} \tag{4}$$

where $\|I(R)\| = \sum_{r \in R} |I(r)|$, σ and λ are manually specified.

4. Prior distributions. $p(\theta_0|y_0)$ is assumed to be Gaussian, whose mean and covariance matrix are manually set. $p(n_0|y_0)$ is assumed to be uniform over $\mathcal{N}$, i.e.,

$$p(n_0 = n|y_0) = 1/N. \tag{5}$$

5. Statistical independence:

$$\begin{aligned} n_0 \perp \theta_0, u_t \perp v_s; \quad t, s \geq 1 \\ u_t \perp u_s, v_t \perp v_s; \quad t, s \geq 1 \,\&\, t \neq s, \end{aligned} \tag{6}$$

where $\perp$ implies statistical independence.

Given this model, the goal is to first evaluate the joint posterior distribution $p(n_t, \theta_t|y_{0:t})$, where $y_{0:t} = \{y_0, \ldots, y_n\}$, then to marginalize it over θ_t to yield $p(n_t|y_{0:t})$. Note that $p(n_t|y_{0:t})$ is in fact a probability mass function (*pmf*) defined on the sample space $\mathcal{N}$.

3 Algorithms and Experimental Results

This model is nonlinear, due to the transformation part, and non-Gaussian, due to the observation noise, and has no analytic solution. We use SMC techniques [2, 5, 8] to provide a numerical approximation to the solution. The essence of SMC is to represent an arbitrary probability distribution by a set of weighted samples.

Algorithm I
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* $p(\theta_0|y_0)$.
For $t = 1, 2, \ldots$
  For $m = 1, 2, \ldots, M$
    Resample $\mathcal{S}_{t-1} = \{(n_{t-1}^{(m)}, \theta_{t-1}^{(m)}, \alpha_{t-2}^{(m)})\}_{m=1}^M$ *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* $\alpha_t^{(m)} = \alpha_t^{(m)} / \sum_{m=1}^M \alpha_t^{(m)}$.
  Marginalize *over* θ_t *to obtain weight* β_{n_t} *for* n_t.
End

Figure 1. The CONDENSATION-like algorithm I

One popular SMC algorithm is the CONDENSATION[3]. Fig. 1 describes the Algorithm I, a CONDENSATION-like algorithm, for solving our model. In the context of this problem, the posterior distribution $p(n_t, \theta_t|y_t)$ is represented by the sample set $\mathcal{S}_t = \{(n_t^{(m)}, \theta_t^{(m)}, \alpha_t^{(m)})\}_{m=1}^M$, and the posterior distribution $p(n_t|y_t)$ is represented by the sample set $\{n_t, \beta_{n_t}\}_{n_t=1}^N$ where β_{nt} is a marginal weight obtained using the following equation: $\beta_{n_t} = \sum_{m=1, n_t^{(m)}=n_t}^M \alpha_t^{(m)}$. Theoretical discussions on the evolution of the posterior probability $p(n_t|y_{0:t})$ in this framework and some experimental results are presented in [11].

The success of Algorithm I relies on that fact that likelihood maximizing pair $(\bar{n}_t, \bar{\theta}_t)$ among the generated samples $\{(n_t^{(m)}, \theta_t^{(m)}, \alpha_t^{(m)})\}_{m=1}^M$ is the actual value for tracking and recognition. There are two issues involved here. One is how good the state equation characterizes the dynamics is and the other is the observation equation. We focus on the latter in this paper.

A detailed examination of $p(y_t|n_t, \theta_t)$ shows that the key is the distance measure $D = \|f(y_t; \theta_t) - I_{n_t}\|$. The likelihood maximizing pair $(\bar{n}_t, \bar{\theta}_t)$ possesses the smallest distance D. However, this likelihood function may not behave well under certain nontrivial circumstances, e.g., inhomogeneous background and non-uniformly illuminated foreground because (i) the non-uniform illumination causes the foreground to deviate from the face subspace; (ii) the inhomogeneous background tends to absorb samples, and (iii) time propagation based on these samples yields inferior solution. Fig. 2 presents such an example with some frames extracted from a probe video and its MAP tracking results (dashed bounding box) obtained by Algorithm I.

In our experiment, we use video sequences with subjects walking towards a camera in order to simulate typi-

cal scenarios in visual surveillance. There are 30 subjects, each having one face template and one upper body template. The face gallery and the upper body gallery are as shown in Fig. 3. The probe set contains 30 video sequences, one for each subject. Fig. 2 gives some example frames in one probe video. These images and videos were collected, as part of the HumanID project, by National Institute of Standards and Technology and University of South Florida researchers. Note (i) that the probe set is taken outdoors, with sunshine casting strong shadow on the face in some video sequence (see Fig. 2), however, this non-uniform illumination does not vary significantly throughout one probe video; (ii) that both gallery sets are captured under different circumstances from the probe and (iii) that the probe has considerable variations in scale.

In order to overcome these unfavorable conditions, we propose the following. Since the illumination does not vary significantly from frame to frame, we first compute the tracking distribution $p(\theta_t|y_{0:t})$ by employing this temporal continuity in the observation, where a tracking template T is other than those in the gallery, say a cut version in some frame of the same video. A second observation equation is introduced here:

$$f(y_t, \theta_t) = T + r(t), \tag{7}$$

where $r(t)$ is the observation noise obeying a truncated-Laplacian distribution $q(y_t|\theta_t)$ with different parameters σ and λ, originally defined in Eqn. 4. By doing this, the produced affine samples will gather around the biometric part. Then, we use Eqn. 3 in order to compute $p(n_t, \theta_t|y_{0:t})$ by binding n_t and θ_t together and obtain $p(n_t|y_{0:t})$ by marginalizing over θ_t. This leads to Algorithm II, a two-stage algorithm, detailed in Fig. 4.

In fact, Eqns. 3 and 7 can be combined to get a new likelihood: $p(y_t|n_t, \theta_t)q(y_t|\theta t)$. We can then (i) marginalize over n_t to get weights on θ_t for tracking and resampling; and (ii) marginalize over θ_t to get weights for n_t for recognition. This will have the same effect as Algorithm II. In Fig. 4, the sample set $\{\theta_t^{(j)}, w_t^{(j)})\}_{j=1}^J$ characterizes the tracking distribution $p(\theta_t|y_{0:t})$, the sample set

$$\{(1, \theta_t^{(j)}, \alpha_{t,1}^{(j)}), (2, \theta_t^{(j)}, \alpha_{t,2}^{(j)}), \ldots, (N, \theta_t^{(j)}, \alpha_{t,N}^{(j)})\}_{j=1}^J,$$

or $\{(\theta_t^{(j)}, \alpha_{t,1}^{(j)}, \alpha_{t,2}^{(j)}, \ldots, \alpha_{t,N}^{(j)})\}_{j=1}^J$ in short, characterizes the joint distribution $p(n_t, \theta_t|y_{0:t})$ since $\mathcal{N}$ is a finite sample space to be enumerated, and $\{n_t, \beta_{n_t}\}_{n_t=1}^N$ characterizes the recognizing distribution $p(n_t|y_{0:t})$, where β_{n_t} is the marginal weight.

Fig. 2 shows the MAP results (solid bounding box) obtained by Algorithm II. Fig. 5 presents the posterior probability $p(n_t|y_{0:t})$. We can observe similar convergence to the correct identity over time as those in [11] for Algorithm II, while Algorithm I converges to the wrong person.

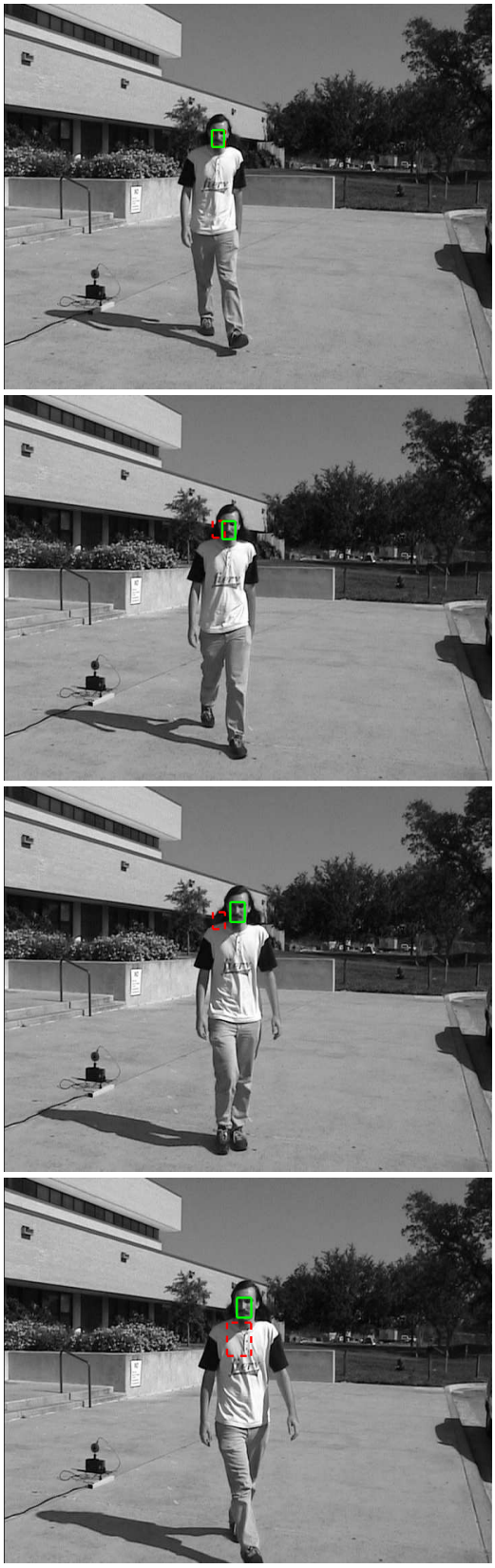

Figure 2. Sample frames in one probe video. The image size is 720x480 with the actual face size ranging roughly from 25x25 in to 40x40.

Figure 3. The face and body galleries. The face size: 70x60 and the body size: 100x75.

Algorithm II
Initialize $\{(\theta_0^{(j)}, 1)\}_{j=1}^{J}$ *according to prior distributions* $p(\theta_0|y_0)$ *and form* $\{(\theta_0^{(j)}, 1, \ldots, 1)\}_{j=1}^{J}$ *for* $p(n_0, \theta_0|y_0)$.
For $t = 1, \ldots$
  For $j = 1, 2, \ldots, J$
    Resample $\{(\theta_{t-1}^{(j)}, w_{t-1}^{(j)})\}_{j=1}^{J}$ *to obtain a new sample* $(\theta_{t-1}^{'(j)}, 1)$. *Set* $w_{t-1,n}^{'(j)} = w_{t-1,n}^{(j)} / w_{t-1}^{(j)}$ *for* $n \in \mathcal{N}$.
    Predict *sample by drawing* $\theta_t^{(j)}$ *from* $p(\theta_t|\theta_{t-1}^{'(j)})$.
      Update *weight using* $w_t^{(j)} = q(y_t|\theta_{t-1}^{'(j)})$.
    For $n = 1, \ldots, N$
      Update *weight using*
      $\alpha_{t,n}^{(j)} = p(y_t|n, \theta_t^{(j)}) * \alpha_{t-1,n}^{'(j)} * w_t^{(j)}$.
    End
  End
  Normalize *weight using* $w_t^{(j)} = w_t^{(j)} / \sum_{j=1}^{J} w_t^{(j)}$ *and* $\alpha_{t,n}^{(j)} = \alpha_{t,n}^{(j)} / \sum_{j=1,n=1}^{J,N} \alpha_{t,n}^{(j)}$
  Marginalize *over* θ_t *to obtain weight* β_{n_t} *for* n_t.
End

Figure 4. The two-stage algorithm II

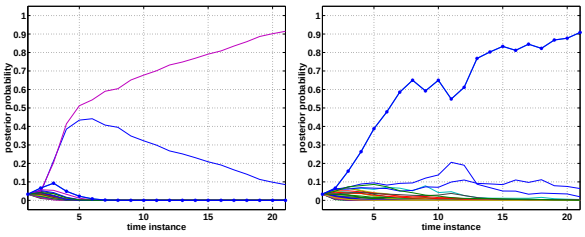

Figure 5. The distribution $p(n_t|y_{0:t})$ obtained by Algorithm I (left) and II (right). The '*' denotes the correct identity.

We sum all the weights over time to rank all the templates in the gallery. It turns out that Algorithm II has an improved performance (roughly 10% higher) over Algorithm I. However, using the body gallery produces a higher performance than using the face gallery. The reasons might be: (i) that the body template is bigger than the face template; (ii) that the non-uniform illumination is prominent only on the face region; and (iii) that the subjects are wearing the same dress for the probe and body galleries.

4 Conclusion

We have introduced a time series state space model for recognition, assuming temporal constancy in the identity variable. The CONDENSATION-like algorithm is not robust against certain outdoor imaging conditions, causing failures in both tracking and recognition. To tackle them, we propose a two-stage approach that essentially constructs a new likelihood measurement by using the constraint of temporal continuity in the observations. It turns out that the improved tracking and recognition are achieved, compared to its CONDENSATION counterpart.

References

[1] M. J. Black and A. D. Jepson. A probabilistic frameworkfor matching temporal trajectories. *Proc. of ICCV*, 1999.
[2] 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.
[3] M. Isard and A. Blake. Contour tracking by stochatic propagation of conditional density. *Proc. of ECCV*, 1996.
[4] A. K. Jain, R. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. PAMI*, 22:4–37, 2000.
[5] G. Kitagawa. Monte carlo filter and smoother for non-gaussian nonlinear state space models. *J. Computational and Graphical Statistics*, 5:1–25, 1996.
[6] B. Li and R. Chellappa. Simultaneous tracking and verification via sequential posterior estimation. *Proc. of CVPR*, pages 110–117, 2000.
[7] Y. Li, S. Gong, and H. Liddell. Constucting structures of facial identities on the view sphere using kernel discriminant analysis. *Proc. of the 2rd Intl. Workshop on SCTV*, 2001.
[8] J. S. Liu and R. Chen. Sequential monte carlo for dynamic systems. *JASA*, 93:1031–1041, 1998.
[9] P. J. Philipps, H. Moon, S. Rivzi, and P. Ross. The FERET testing protocol. *Face Recognition: From Theory to Applications*, 83:244–261, 1998.
[10] W. Y. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. *UMD CAR-TR-948*, 2000.
[11] S. Zhou and R. Chellappa. Probabilistic human recognition from video. *Proc. of ECCV*, 2002.