PARTICLE FILTER WITH MODE TRACKER(PF-MT) FOR VISUAL TRACKING ACROSS ILLUMINATION CHANGE

Amit Kale, Namrata Vaswani** and Christopher Jaynes**

* Ctr. for Visualization and Virtual Environments and Dept. of Computer Science
University of Kentucky, $\{amit, jaynes\}$@$cs.uky.edu$
**Dept. of ECE, Iowa State University, Ames, IA 50011, namrata@iastate.edu

Index Terms— Lighting,Tracking,Monte Carlo methods

ABSTRACT

In recent work, the authors introduced a multiplicative, low dimensional model of illumination that is computed as a linear combination of a set of simple-to-compute Legendre basis functions. The basis coefficients describing illumination change, are can be combined with the "shape" vector to define a joint 'shape"-illumination space for tracking. The increased dimensionality of the state vector necessitates an increase in the number of particles required to maintain tracking accuracy. In this paper, we utilize the recently proposed PF-MT algorithm to estimate the illumination vector. This is motivated by the fact that, except in case of occlusions, multimodality of the state posterior is usually due to multimodality in the "shape" vector (e.g. there may be multiple objects in the scene that roughly match the template). In other words, given the "shape" vector at time t, the posterior of the illumination (probability distribution of illumination conditioned on the "shape" and illumination at previous time) is unimodal. In addition, it is also true that this posterior is usually quite narrow since illumination changes over time are slow. The choice of the illumination model permits the illumination coefficients to be solved in closed form as a solution of a regularized least squares problem. We demonstrate the use of our method for the problem of face tracking under variable lighting conditions existing in the scene.

1. INTRODUCTION

Visual tracking involves generating an inference about the motion of an object from measured image locations in a video sequence. Unfortunately, this goal is confounded by sources of image appearance change that are only partly related to the position of the object in the scene. For example, changes in pose of the object or illumination can cause a template to change appearance over time and lead to tracking failure.

For situations where the weak perspective assumptions hold, shape change for rigid objects can be captured by a low dimensional "shape" vector(here "shape" refers to location and scale change, in general can also be affine). Tracking is the problem of causally estimating a hidden state sequence corresponding to this "shape" vector, $\{X_t\}$ (that is Markovian with state transition pdf $p(X_t|X_{t-1})$, from a sequence of observations, $\{Y_t\}$, that satisfy the Hidden Markov Model (HMM) assumption ($X_t \rightarrow Y_t$ is a Markov chain for each t, with observation likelihood denoted $p(Y_t|X_t)$). This interpretation forms the basis of several tracking algorithms including the well-known Condensation algorithm [6] and its variants. A similarly concise model is required if we are to robustly estimate illumination changes in a statistical tracking framework while avoiding undue increase in the dimensionality of the problem. The study of appearance change as a function of illumination is a widely studied area in computer vision [2, 1]. These methods focus on accurate models of appearance under varying illumination and their utility for object recognition. However they typically require an explicit 3-D model of the object. This limits their use in surveillance where a 3-D model or a large number of images of every object to be tracked under different illumination conditions is unavailable [5]. Examples of such tasks that involve tracking objects through simultaneous illumination and "shape" change are shown in Figure 2. Note that features that are considered to be invariant to illumination could be unreliable [2] in such situations.

In [7], the authors introduced a multiplicative, low dimensional model of illumination that is computed as a linear combination of a set of simple-to-compute Legendre basis functions. Such a multiplicative model can be interpreted as an approximation of the illumination image as discussed in Weiss [12]. The basis coefficients describing illumination change can be combined with the "shape" vector (i.e. affine or similarity group) to define a joint "shape-illumination" space for tracking. Assuming that a "shape space" of dimension $N_u = 3$ corresponding to x, y translation and scale and the number of illumination coefficients, $N_\Lambda = 7$ is sufficient to capture a significant variability from the initial template to its repositioned and re-lit counterparts in successive frames, we need to sample a 10 dimensional space. It is a well known fact that as state dimension increases, the effective particle size reduces and hence more particles are needed for a certain accuracy. The question is can we do better than brute force PF on a 10 dim space? We can utilize the fact that, except in case of occlusions, multimodality of the state posterior is usually due to multimodality in the "shape" vector (e.g. there may be multiple objects in the scene that roughly match the template). Or in other words, given the "shape" vector at time t, the posterior of the illumination (probability distribution of illumination conditioned on the "shape", the image and illumination at the previous time instant is unimodal. In addition, it is also true that this posterior is usually quite narrow since illumination changes over time are slow.

Under these two assumptions, we can utilize the PF-MT algorithm proposed in [11, 10]. The main idea is to split the entire state vector into "effective basis" and "residual space". We run the SIR PF (sample from state transition pdf) [3] on the effective basis, but approximate importance sampling from the residual posterior by its

This work was funded by NSF CAREER Award IIS-0092874 and by Department of Homeland Security

mode. For our problem, we run SIR PF on[1] "shape" and we compute the mode of the posterior of illumination conditioned on the "shape", previous illumination vector and the current image. The mode computation turns out to be a regularized least squares problem in our case and hence can actually be done in closed form. This idea can also be understood as an approximation of the Rao Blackwellized PF (RB-PF) [9], but is more general since it only requires the subsystem to have a unimodal posterior (need not be linear Gaussian). We would like to point out though that for the specific observation model considered in this paper, RB-PF can also be used. However, if the observation noise is non-Gaussian or the illumination model is nonlinear, RB-PF will not be applicable. In order to further reduce the number of particles required, we also effectively use the Aux PF [8] to improve resampling efficiency.

2. STATE SPACE MODEL

2.1. Illumination model

The image template throughout the tracking sequence can be expressed as:

$$T_t(x,y) = L_t(x,y)R(x,y) \tag{1}$$

where $L_t(x,y)$ denotes the illumination image in frame t and $R(x,y)$ denotes a fixed reflectance image [12]. Thus if the R is known, tracking becomes the problem of estimating the illumination image and a "shape"-vector. Of course, R is typically unavailable and the illumination image can only be computed modulo the illumination contained in the image template T_0,

$$L_t(x,y) = \tilde{L_t}(x,y)L_0(x,y)R(x,y) = \tilde{L_t}(x,y)T_0(x,y) \tag{2}$$

where L_0 is the initial illumination image and $\tilde{L}_t$ is the unknown illumination image for frame t.

Our proposed model of appearance change [7], then, is the product of T_0 with an approximation of L_t which is constructed using a linear combination of a set of N_Λ Legendre basis functions defined over the template of size,M. Let $p_k(x)$ denote the k th Legendre basis function. Then, for $N_\Lambda = 2k+1$, $\mathbf{\Lambda} = [\lambda_0, \cdots, \lambda_{N_\Lambda}]^T$, the scaled intensity value at a pixel of the template T_t is computed as:

$$\begin{aligned}\hat{T_t}(x,y) = (\frac{1}{N_\Lambda}(\lambda_0 + \lambda_1 p_1(x) + \cdots + \lambda_k p_k(x) + \\ \lambda_{k+1}p_1(y) + \cdots + \lambda_{N_\Lambda} p_k(y)) + 1)T_0(x,y)\end{aligned} \tag{3}$$

so that when $\Lambda \equiv 0$ $\hat{T_t} = T_0$. For purposes of notation, we will denote the effect of Λ on T_0 as

$$\mathbf{\Delta}\Lambda T_0 \equiv T_0 \otimes \mathbf{P}\Lambda + T_0 \tag{4}$$

where

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2n+1}p_0 & \cdots & \frac{1}{2n+1}p_n(y_1) \\ \vdots & \vdots & \vdots \\ \frac{1}{2n+1}p_0 & \cdots & \frac{1}{2n+1}p_n(y_M) \end{bmatrix}. \tag{5}$$

We define $\otimes$ as an operator that scales the rows of P with the corresponding element of T_0 written as a vector. Given a proposal image region G and T_0, the Legendre coefficients that relight T_0 to resemble G can be computed by solving the least squares problem:

$$T_0 \otimes \mathbf{P}\Lambda = A_{T_0}\Lambda \approx T_0 - G \tag{6}$$

where $A_{T_0} \triangleq T_0 \otimes \mathbf{P}$.

[1] in case of occlusions one can also treat the mean intensity (0th order Legendre coefficient) as part of the effective basis

2.2. System and Observation Model

The new illumination model can be combined with "shape" to define a joint "shape-illumination" vector

$$X_t = \begin{bmatrix} u_t \\ \Lambda_t \end{bmatrix} \tag{7}$$

where $u_t = [s \quad t_x \quad t_y]'$ corresponds to a three dimensional "shape" space encompassing scale (s) and translations t_x and t_y and $\Lambda \in \mathbb{R}^{N_\Lambda}$ corresponds to coefficients of the Legendre polynomials of order k.

The system dynamics is assumed to be a random walk model on object "shape", u_t and on illumination coefficients, Λ_t i.e.

$$u_{t+1} = u_t + \nu_{ut}, \;\; \nu_{ut} \sim h(.) \tag{8}$$

$$\Lambda_{t+1} = \Lambda_t + \nu_{\Lambda t}, \;\; \nu_{\Lambda t} \sim \mathcal{N}(0, \Pi) \tag{9}$$

where $\Pi_{N_\Lambda \times N_\Lambda}$ is a diagonal covariance matrix (variance of individual components of Λ) and $h(.)$ denotes the pdf of ν_{ut} which is described in Section (4.1).

Let T_0 denote the original template and let M denote the number of pixels in it. The observation at time t, Y_t is the image at time t. We assume the following image formation process: the image intensities of the image region that contains the object are illumination scaled versions of the intensities of the original template, T_0, plus Gaussian noise. Also, the proposals of the image region that contains the object are obtained by applying the dynamics of the object's "shape" to each point of the template. Also, the rest of the image (which does not contain the object) is independent of the object intensity or "shape"(and hence can be thrown away). Thus we have the following observation model:

$$Y_t\left(\mathbf{J}u_t + \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{Y}_0 \end{bmatrix}\right) = \Delta\Lambda_t T_0 + \psi_t \tag{10}$$

where $\psi_t \sim \mathcal{N}(0, V)$ where $V_{M\times M}$ is a diagonal covariance matrix (variance of individual pixel noise) and $\mathbf{J}$ is

$$\mathbf{J} = \begin{bmatrix} \mathbf{X}_0 - \bar{x}_0\mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{Y}_0 - \bar{y}_0\mathbf{1} & \mathbf{0} & \mathbf{1} \end{bmatrix}. \tag{11}$$

where $\mathbf{X}_0$ and $\mathbf{Y}_0$ denote the x and y coordinates of each point on the template and $\bar{x}_0$ and $\bar{y}_0$ denote the corresponding means. $\mathbf{1}$ and $\mathbf{0}$ denote a vector of ones and zeros of size M respectively. $\mathbf{J}$ can be easily modified for the affine case as described in [6]. For brevity we will denote the image region in Y_t indicated by a vector u as:

$$G_t^u = Y_t\left(\mathbf{J}u + \begin{bmatrix} \mathbf{X}_0 \\ \mathbf{Y}_0 \end{bmatrix}.\right) \tag{12}$$

Thus the observation likelihood can be written as:

$$p(Y_t|X_t) = p(Y_t|u_t, \Lambda_t) = \exp[-\frac{||G_t^{u_t} - \Delta\Lambda_t T_0||^2}{v}] \tag{13}$$

where $(V)_{i,i} = v$.

3. PARTICLE FILTER WITH MODE TRACKER(PF-MT) ALGORITHM

A naive approach would be to simply apply the SIR PF [4] to the system (8),(9) and observation model (10). In (9) we use a $k = 3$ order Legendre basis and hence $N_\Lambda = 2k+1 = 7$. Thus the total state

Algorithm 1 PF-MT. Going from π^N_{t-1} to $\pi^N_t(X_t) = \sum_{i=1}^N w_n^{(i)}\delta(X_t - X_t^i)$, $X_t^i = [u^i, \Lambda^i]$

1. *Auxiliary Resampling:* Compute g_t^i using (17) and resample X_{t-1}^i according to it. Reset the weights of the resampled to $(w_{t-1}^i)^{new}$ defined in (18).
2. *Importance Sample (IS) on effective basis:* $\forall i$, sample $\nu_{ut} \sim h(u)$ and compute $u_t^i = u_{t-1}^i + \nu_{ut}$
3. *Mode Tracking (MT) in residual space:* $\forall i$, compute m_t^i using (15) and set $\Lambda_t^i = m_t^i$
4. *Weighting:* Compute w_t^i using (16).

space dimension is 10. It is a well known fact that the number of particles required for a certain accuracy increases with state dimension [4], making the PF very expensive to run. But notice that conditioned on u_t, the posterior of Λ_t is unimodal. Also, we observed in expts that covariance of change of Λ_t was small. Thus we can use the recently proposed PF-MT idea [10] for this problem. The main idea is to importance sample (IS) "shape", u_t^i from its state transition model (8), but replace IS by posterior Mode Tracking (MT) for illumination, Λ_t, i.e. we compute the mode (denote it as m_t^i) of $p(\Lambda_t|u_t^i, \Lambda_{t-1}^i, Y_t)$ and set $\Lambda_t^i = m_t^i$. In exact PF, one would compute m_t^i and use a Gaussian about m_t^i as the IS density. Replacing IS by MT is a valid approximation when the the covariance is small[10] which is true in our case.

Now, it is easy to see that

$$p(\Lambda_t|u_t^i, \Lambda_{t-1}^i, Y_t) \propto p(Y_t|u_t^i, \Lambda_t)p(\Lambda_t|\Lambda_{t-1}^i) \quad (14)$$

where the first term is defined in (13) and the second term is given by (10). Thus m_t^i can be computed as the minimizer of the $-\log[.]$ of (14) and this turns out to be a nice regularized least squares problem (regularization term is the weighted distance from Λ_{t-1}) with a closed form solution given by

$$m_t^i = \Lambda_{t-1}^i + (\Pi^{-1} + A_{T_0}^T V^{-1} A_{T_0})^{-1} A_{T_0}^T V^{-1} (G_t^{u_t^i} - \Delta\Lambda_t T_0) \quad (15)$$

Note all the multipliers can be pre-computed, making this a very fast computation. With the above importance sampling strategy, the weighting term will be [10]

$$w_t^i \propto w_{t-1}^i p(Y_t|u_t^i, \Lambda_t^i)p(\Lambda_t^i|\Lambda_{t-1}^i), \;\; \Lambda_t^i = m_t^i \quad (16)$$

Using this method greatly reduces the weight variance, thus reducing the number of particles required for a certain accuracy (or improving tracking accuracy when number of particles available is small). We have shown the comparison of PF-MT with other existing methods - SIR PF (called FULLPF), Auxiliary PF(called FULLPFWAP) and PF without tracking illumination (called NOILLUM) in Figure 2.

To improve resampling efficiency, we used the look-ahead resampling idea of Auxiliary PF [8]. This performs resampling of the past particles when the current observation, Y_t comes in, and uses the likelihood of X_{t-1}^i generating Y_t to resample, i.e. it resamples according to

$$g_t^i = w_{t-1}^i p(Y_t|X_{t-1}^i) \quad (17)$$

After resampling, the weights of the resampled particles are set to

$$(w_{t-1}^i)^{new} = \frac{w_{t-1}^i}{N g_t^i} = \frac{p(Y_t|X_{t-1}^i)}{N} \quad (18)$$

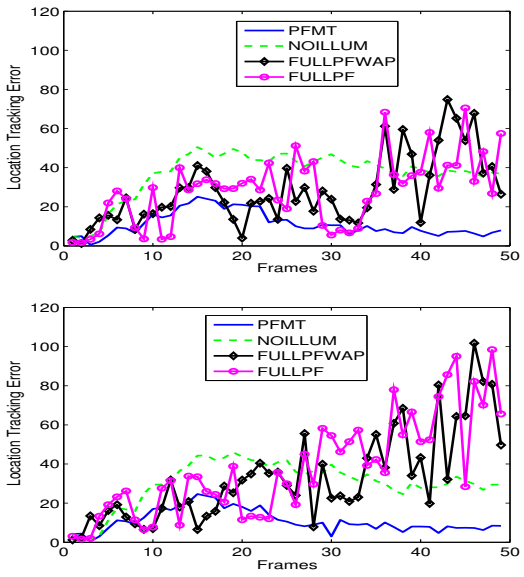

Fig. 1. Comparison of errors from ground truth of location using different particle filters using 300 particles (top) and 100 particles(bottom).

The complete algorithm is summarized in Algorithm 1.

Note, for the particular form of the observation model that we use (additive Gaussian noise), one can also use Rao-Blackwellized PF [9]. But if the observation noise were non-Gaussian (e.g. in order to model occlusions or outliers), RB-PF cannot be used. Also, in case of occlusions, one may want to use the mean intensity (λ_0) also as part of the effective basis (importance sample for it) while mode tracking for the rest.

4. MODEL PARAMETER ESTIMATION AND RESULTS

4.1. Learning the Model Parameters

We need to learn the noise models for "shape"($h(.)$), of illumination (Π) and the observation noise covariance (V). We assume that we have a static camera acquiring images and that the illumination conditions, although variable within the scene, do not change significantly over time. Ground truth video sequences consisting of a starting template T_0 and its location and shape in subsequent frames, $G_t, t = 1, \cdots, N_f$ are used to compute state-vectors $X_t = [u_t \quad \Lambda_t]^T, t = 1, \cdots, N_f$ using (12) and (6)for this motion with the corresponding approximations $\{\hat{G}_t = \Delta\Lambda_t T_0, t = 1, \cdots, N_f\}$

We consider the "shape" difference vectors $du_t = u_t - u_{t-1}$ for $t = 1, \cdots, N_f$. Assuming that the individual components of du_t are independent, we build a "shape" sampling distribution $h(u)$ as follows. Given $du_t, t = 1 \ldots N$ the dynamic model was estimated for the shape vector. The horizontal displacement dt_x and scale change ds were modeled as Gaussian random vectors whose parameters were computed using standard MLE techniques. In order to take into account the nature of human gait the vertical displacement dt_y was modeled as a mixture of two Gaussians whose parameters were estimated using EM. The "shape" sampling distri-

bution is given by

$$h(u) = [\mathcal{N}(\mu_s, \sigma_s^2) \quad \mathcal{N}(\mu_{t_x}, \sigma_{t_x}^2) \quad \sum_{i=1}^{2} \alpha_i \mathcal{N}(\mu_i, \sigma_i)]$$

A third order Legendre polynomial ($N_\Lambda = 2 * 3 + 1 = 7$) was used to represent the illumination effects. Given $d\Lambda_t = \Lambda_t - \Lambda_{t-1}$ for $t = 1, \cdots, N_f$ we estimate Π as

$$\Pi = \frac{1}{N_f - 1} \sum_{t=2}^{N_f} (\Lambda_t - \Lambda_{t-1})(\Lambda_t - \Lambda_{t-1})^T \tag{19}$$

The per-pixel observation noise V is estimated by averaging the SSD between the corresponding pixels of $\hat{G}_t$ and G_t as

$$V = \frac{1}{N_f} \sum_{t=1}^{N_f} (\hat{G}_t - G_t) \otimes (\hat{G}_t - G_t)] \tag{20}$$

4.2. Results

Our test dataset contained several different subjects moving through challenging illumination conditions in an indoor environment(see Figure 2)including overhead, side-lit and partially shaded regions as they approach a surveillance camera. Ground truth was generated from one sequence. Figure 2 shows the face tracking results using the PF-MT algorithm using 100 particles. These sequences are typical for this setup and only three frames are shown in the interest of space. The box corresponds to the MMSE estimate of the state vector u computed as $\tilde{u} = \frac{1}{N} \sum_{i=1}^{N} u^i$. Figure 1 shows the location error from the ground truth for different particle filters. The same dynamic model was used for all the PFs. Full PF represents the case where instead of importance sampling u_t alone Λ_t is importance sampled from $\mathcal{N}(\Lambda_{t-1}^i, \Pi)$ and SIR is used. FullPFWAP represents FULLPF with SIR replaced by the Auxiliary PF. NOILLUM represents the case where no illumination model is used while PF-MT represents the the case using our algorithm. As can be seen, the estimated "shape" using PF-MT has much lower error than all the other algorithms. Also it remains in track even with just 100 particles (bottom row of Figure 1).

5. CONCLUSIONS

In this paper we studied the problem of visual tracking as an inference problem in a joint "shape-illumination" space introduced in [7]. We used the PF-MT idea to exploit the fact that, except in case of occlusions, multimodality of the state posterior is usually due to multimodality in the "motion" vector and that given the "motion" vector at time t, the posterior of the illumination (probability distribution of illumination conditioned on the "motion", the image and illumination at time $t - 1$) is unimodal. We demonstrated the use of our method for tracking faces under variable lighting conditions existing in the scene without requiring an increase in the number of particles despite the higher dimensionality of the state vector.

6. REFERENCES

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. PAMI*, 25(2):218–233, 2003.

[2] P. Belhumeur and D.J.Kriegman. What is the set of images of an object under all possible illumination conditions. *IJCV*, 28(3):1–16, 1998.

Fig. 2. Face tracking using PF-MT as the individuals walk through different lighting conditions. The white box shows the image region corresponding to MMSE estimate of the "shape vector".

[3] A. Doucet, N. deFreitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[4] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/nongaussian bayesian state estimation. *IEE Proceedings-F (Radar and Signal Processing)*, pages 140(2):107–113, 1993.

[5] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. PAMI*, 20(10):1025–1039, 1998.

[6] M. Isard and A. Blake. Condensation - conditional density propogation for visual tracking. *IJCV*, 21(1):695–709, 1998.

[7] A. Kale and C. Jaynes. A joint illumination and shape model for visual tracking. *Proceedings of IEEE CVPR*, pages 602–609, 2006.

[8] M. K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446), 1999.

[9] T. Schn, F. Gustafsson, and P. Nordlund. Marginalized particle filters for nonlinear state-space models. *IEEE Trans. Sig. Proc.*, 2005.

[10] N. Vaswani, A. Yezzi, Y. Rathi, and A. Tannenbaum. Particle filters for infinite (or large) dimensional state spaces. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2006.

[11] N. Vaswani, A. Yezzi, Y. Rathi, and A. Tannenbaum. Time-varying finite dimensional basis for tracking contour deformations. In *IEEE Conf. Decision and Control (CDC)*, 2006.

[12] Y. Weiss. Deriving intrinsic images from image sequences. *Proc of ICCV*, 2001.