In Proc. of IEEE Conf. on CVPR'03, Madison, Wisconsin, 2003

Tracking Appearances with Occlusions

Ying Wu, Ting Yu, Gang Hua
Department of Electrical & Computer Engineering
Northwestern University
2145 Sheridan Road, Evanston, IL 60208
{yingwu,tingyu,ganghua}@ece.nwu.edu

Abstract

Occlusion is a difficult problem for appearance-based target tracking, especially when we need to track multiple targets simultaneously and maintain the target identities during tracking. To cope with the occlusion problem explicitly, this paper proposes a dynamic Bayesian network which accommodates an extra hidden process for occlusion and stipulates the conditions on which the image observation likelihood is calculated. The statistical inference of such a hidden process can reveal the occlusion relations among different targets, which makes the tracker more robust against partial even complete occlusions. In addition, considering the fact that target appearances change with views, another generative model for multiple view representation is proposed by adding a switching variable to select from different view templates. The integration of the occlusion model and multiple view model results in a complex dynamic Bayesian network, where extra hidden processes describe the switch of targets' templates, the targets' dynamics, and the occlusions among different targets. The tracking and inferencing algorithms are implemented by the sampling-based sequential Monte Carlo strategies. Our experiments show the effectiveness of the proposed probabilistic models and the algorithms.

1 Introduction

Tracking targets based on their appearances play an important role in many applications such as intelligent human computer interaction and video surveillance. For example, before the detailed facial motion can be recovered and before the human identities can be recognized, we need to locate and track faces in video sequences. An effective way is through matching and tracking face appearances. Since image appearances provide more comprehensive visual information to represent the targets, e.g., the faces, appearance-based tracking methods receive more and more attention.

However, if a target is partially or completely occluded, its visual appearance would dramatically deviate from its appearance template as we set for tracking. Thus, occlusion becomes a special and difficult problem for appearance-based tracking. In addition, if we are concerned about multiple targets, explicit handling of occlusion is indispensable for tracking, since occlusion would probably occur when different targets interact.

This paper addresses the occlusion problem in the multiple target tracking scenario. Different from other works on tracking multiple targets, this paper aims at solving the occlusion relationships besides keeping the trajectories. Our method is based on a dynamic Bayesian network which models the occlusion process explicitly. This model consists of multiple hidden Markov processes: the dynamics of each individual target, and the process of the occlusion relation. In addition, the model describes the formation (or generation) of the image observations, jointly conditioned on the targets states and their occlusion relations. Then, tracking is to infer the states of all these hidden Markov chains based on the sequence of image observations.

In addition, we investigate two representations for the appearances: i.e., single view and multiple views. The single view appearance is represented by an appearance template associated with a transformation that depicts the motion and deformation of the template. Since the appearances change with views, we extend this "view+transformation" representation to the multiple view case, by switching among a set of templates and transformations. This mechanism is also modelled by a generative model which contains a hidden switching process.

The combination of the occlusion modelling and the multiple view representation results in a multilevel dynamic Bayesian network. Due to the complexity in the structure of the generative model, the inference of the model is approximated by the sampling-based sequential Monte Carlo strategies. Various test sequences showed the effectiveness of this approach to handle the occlusion situations.

The proposed approach accommodates the inference of the occlusion relations of multiple targets and the switch of multiple views into a probabilistic tracking framework. Not limited to multiple face tracking, the proposed generative model is general and valid for many tracking scenarios which need to handle occlusion explicitly.

The paper is organized as follows. Section 3 presents the

dynamic Bayesian network for occlusion. The sequential Monte Carlo strategy is described in Section 4. Section 5 presents the multiple view appearance model. The generative model that combines the occlusion model and multiple view switching can also be found in Section 5. Experiments are given in Section 6 and conclusions are in Section 7.

2 Previous Work

The target representations affect the effectiveness and efficiency of tracking algorithms. Many approaches have been studied based on different target representations, e.g., image appearances [2, 3, 6, 9, 17] and geometrical shapes [1, 7, 15]. Shape-based approaches are concerned about the matching between shape models and image features. They need to deal with more ambiguities in tracking but are less sensitive to lighting. On the other hand, since massive image appearance data contain very rich information for characterizing targets, appearance-based methods would not be sensitive to image resolutions, but special attention needs to be taken for deformation and lighting.

Many different types of appearance models have been investigated, such as color appearances [3], eigen appearances [2], texture appearances [9], layered image template appearances [17], and the appearances combining image template and lighting [6]. All of these models parameterize the appearances for target representations. Tracking targets includes the estimation of these parameters.

There are two methodologies to this problem: *bottom-up* and *top-down*. The bottom-up approaches generally formulate the problem as nonlinear optimization problems which minimize some error functions, e.g., flow residue [2, 6] and color discrepancy [3]. On the other hand, the top-down approaches adopt the idea of *analysis-by-synthesis*, by directly verifying plenty of hypotheses [7, 15].

Most bottom-up algorithms are computationally more efficient, but they are subject to the validation of the small motion assumption, and it is hard for them to cope with occlusions unless the appearance model itself is robust against occlusions. On the other hand, most top-down algorithms involve more computation, but the motion estimation tends to be more accurate and more robust. In addition, occlusion can be modelled from top-down in the same framework.

The generative model approaches take a top-down methodology, by modelling the hidden factors that would affect the observed data [10]. Once the structure and the parameters of the model are set, those hidden factors can be inferred and the parameters can be learnt from the data. As a special case, dynamic Bayesian networks model dynamic systems and temporal signals [16]. The inference of the networks provides tracking results directly.

To track multiple appearances with occlusions, this paper describes a class of dynamic Bayesian networks that accommodates the hidden process of occlusion and model the switching of the appearance templates of multiple views.

3 A Generative Model for Occlusion

We take a "view+transformation" approach to represent the state of a target, which consist of an appearance template T and a transformation H. The template T can be any kind of templates, such as an image template, an edge map template, or a texture template. The transformation H can be an affine transformation or a homography transformation.

To make the description clearer, we limit the description to the situation of tracking two targets (i.e., A and B). We denote the *target state* of target k at time t by $\mathbf{X}_t^k$. The tracking task is to infer $\mathbf{X}_t^A$ and $\mathbf{X}_t^B$ based on all the observed image evidence $\underline{\mathbf{Z}}_t = \{\mathbf{Z}_1, \cdots, \mathbf{Z}_t\}$, where $\mathbf{Z}_t$ is the image *measurement* (or *observation*) at time t, i.e., to estimate $p(\mathbf{X}_t|\underline{\mathbf{Z}}_t) = p((\mathbf{X}_t^A, \mathbf{X}_t^B)|\underline{\mathbf{Z}}_t)$, where $\mathbf{X}_t = (\mathbf{X}_t^A, \mathbf{X}_t^B)$.

We are concerned about the occlusions between these targets, i.e., a target is occluded by a known object. This paper does not investigate a more challenging situation where the target is occluded by a completely unknown object, since no clue from the occluding object can be used for occlusion detection. But it will be part of our future work.

The tracking process can be viewed as the density propagation [7] from $p(\mathbf{X}_{t-1}|\underline{\mathbf{Z}}_{t-1})$ to $p(\mathbf{X}_t|\underline{\mathbf{Z}}_t)$, and it is governed by the dynamic model $p(\mathbf{X}_{t+1}|\mathbf{X}_t)$ and the observation model $p(\mathbf{Z}_t|\mathbf{X}_t)$, since we have

$$p(\mathbf{X}_t|\underline{\mathbf{Z}}_t) \propto p(\mathbf{Z}_t|\mathbf{X}_t) \int p(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\underline{\mathbf{Z}}_{t-1}) d\mathbf{X}_{t-1}$$

In addition, since the motion of two targets are independent, we have $p(\mathbf{X}_t|\mathbf{X}_{t-1}) = p(\mathbf{X}_t^A|\mathbf{X}_{t-1}^A) p(\mathbf{X^B}_t|\mathbf{X}_{t-1}^B)$. Then we have

$$p(\mathbf{X}_t|\underline{\mathbf{Z}}_t) \propto p(\mathbf{Z}_t|\mathbf{X}_t^A, \mathbf{X}_t^B) \int p(\mathbf{X}_t^A|\mathbf{X}_{t-1}^A) \times p(\mathbf{X}_t^B|\mathbf{X}_{t-1}^B) p(\mathbf{X}_{t-1}|\underline{\mathbf{Z}}_{t-1}) d\mathbf{X}_{t-1}$$

If there is no occlusion between A and B, the observation likelihood $p(\mathbf{Z}_t|\mathbf{X}_t^A, \mathbf{X}_t^B)$ can be uniquely determined. However, when one target occludes the other, the occlusion relation has to be known before the likelihood can be uniquely calculated, i.e., the likelihood should be conditioned on the occlusion relations additionally. Let $\alpha_t \in \{0, 1, 2\}$ denote the occlusion relation, i.e., $\alpha_t = 0$ indicates no occlusion, $\alpha_t = 1$ indicates A $\bigwedge$ B, and $\alpha_t = 2$ indicates B $\bigwedge$ A, where $\bigwedge$ means "occludes". Then based on the joint likelihood $p(\mathbf{Z}_t|\mathbf{X}_t^A, \mathbf{X}_t^B, \alpha_t)$, we have

$$p(\mathbf{X}_t, \alpha_t|\underline{\mathbf{Z}}_t) \propto p(\mathbf{Z}_t|\mathbf{X}_t^A, \mathbf{X}_t^B, \alpha_t) \int p(\mathbf{X}_t^A|\mathbf{X}_{t-1}^A) \times p(\mathbf{X}_t^B|\mathbf{X}_{t-1}^B) p(\alpha_t|\alpha_{t-1}) p(\mathbf{X}_{t-1}, \alpha_{t-1}|\underline{\mathbf{Z}}_{t-1}) d\mathbf{X}_{t-1} \quad (1)$$

where $p(\alpha_t|\alpha_{t-1})$ describes the transition of occlusion relation. Thus, based on Equation 1, the probabilistic dynamic system can be illustrated by a factorized graphical model (a factorized dynamic Bayesian network) in Figure 1.

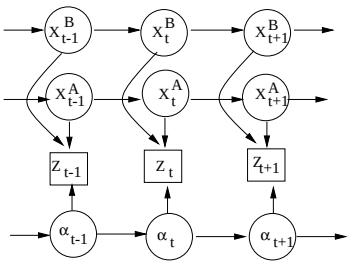

Figure 1: A hidden process $\{\alpha_t\}$ is accommodated in the dynamic Bayesian network to present the occlusion relationships.

The posterior density of occlusion can be obtained through integrating out $\mathbf{X}_t^A$ and $\mathbf{X}_t^B$ from the joint posterior probability, i.e.,

$$p(\alpha_t|\underline{\mathbf{Z}}_t) = \int\int p(\mathbf{X}_t^A, \mathbf{X}_t^B, \alpha_t|\underline{\mathbf{Z}}_t) d\mathbf{X}_t^A d\mathbf{X}_t^B \quad (2)$$

As a generative model, this dynamic Bayesian network models the forwarding process of image formation. In the graphical model, there are three hidden Markov processes, $\{\mathbf{X}_t^A\}$, $\{\mathbf{X}_t^B\}$ and $\{\alpha_t\}$, which are to be inferred from the observation data $\underline{\mathbf{Z}}_t$, based on all the conditional probabilities as illustrated by arrows in the graph. Specifically, to characterize the model, we need to model the dynamics of the two targets $p(\mathbf{X}_t^A|\mathbf{X}_{t-1}^A)$ and $p(\mathbf{X}_t^B|\mathbf{X}_{t-1}^B)$, the transition model $p(\alpha_t|\alpha_{t-1})$ of the occlusion process $\{\alpha_t\}$, and the observation likelihood $p(\mathbf{Z}_t|\mathbf{X}_t^A, \mathbf{X}_t^B, \alpha_t)$.

We employ a constant velocity model for the target dynamics $p(\mathbf{X}_t^k|\mathbf{X}_{t-1}^k), k \in \{A, B\}$. In addition, the transition $p(\alpha_t|\alpha_{t-1})$ of the occlusion process is described by a finite state machine, i.e.,

$$\mathbf{T}_\alpha = [T_\alpha(i,j)] = [p(\alpha_t = j|\alpha_{t-1} = i)].$$

The observation likelihood $p(\mathbf{Z}_t|\mathbf{X}_t^A, \mathbf{X}_t^B, \alpha_t)$ is modelled based on the innovations, i.e, the discrepancies between the predicted appearance and the actual image observations. Denote the predicted region of the k-th target at time t by $R_t^k = R(\mathbf{X}_t^k)$. Then, the predicted region of $\mathbf{X}_t$ is the union of two targets', i.e.,

$$R_t = R(\mathbf{X}_t) = R((\mathbf{X}_t^A, \mathbf{X}_t^B)) = R_t^A \bigcup R_t^B$$

The actual image appearance observation is collected on the predicted region R_t and denoted by $I(R_t)$. Denote the predicated appearance by $T_t = T(\mathbf{X}_t^A, \mathbf{X}_t^B, \alpha_t)$ which depends on the value of α_t. As illustrated in Figure 2, we denote the overlapping region of the two targets by

$$O_t = O(\mathbf{X}_t) = O((\mathbf{X}_t^A, \mathbf{X}_t^B)) = R_t^A \bigcap R_t^B$$

which is independent of α_t. Then, $\forall u \in R_t$,

$$T_t(u) = \begin{cases} T^A(\mathbf{X}_t^A(u)), & u \in R_t^A - O_t \\ T^B(\mathbf{X}_t^B(u)), & u \in R_t^B - O_t \\ T^C(\mathbf{X}_t^C(u)), & u \in O_t \end{cases}$$

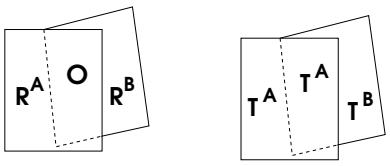

Figure 2: The occlusion relations of $\alpha = 1$.

where u is a pixel location in a region, and C indicates the occluding target, i.e.,

$$C = C(\alpha_t) = \begin{cases} \phi, & \alpha_t = 0 \\ A, & \alpha_t = 1 \\ B, & \alpha_t = 2 \end{cases}$$

Then, the observation likelihood is modelled by:

$$p(\mathbf{Z}_t|\mathbf{X}_t, \alpha_t) \propto \exp\left[-\frac{\sum_{u \in R_t} \mathcal{D}(T_t(u), I_t(u))}{M(R_t)}\right] \quad (3)$$

where $M(R_t)$ is the number of pixels in the region R_t, and $\mathcal{D}(T_t(u), I_t(u)) = |T_t(u) - I_t(u)|^2$.

Specially attention should be taken for the case where one target is fully occluded by the other one as illustrated in Figure 3, since no image evidence can be used to support the existence of the fully occluded target. Consequently, the tracker would not be able to follow the occluded target again. Under this circumstance, the regain of tracking the

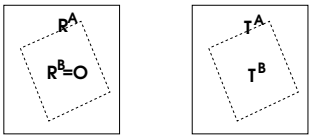

Figure 3: Target B is fully occluded by A.

fully occluded target would depends on motion prediction of target and the detection around the border of the occluding target. Such a mechanism can be implemented by reducing the likelihood of the full occlusion events. Then, we have $p(\mathbf{Z}_t|\mathbf{X}_t, \alpha_t) \propto \exp[-H(\mathbf{Z}_t, \mathbf{X}_t, \alpha_t)]$, where $H(\mathbf{Z}_t, \mathbf{X}_t, \alpha_t) =$

$$\frac{\sum_{u \in R_t^A} \mathcal{D}(T_t^A(u), I_t(u)) + \sum_{u \in R_t^B} \mathcal{D}(T_t^B(u), I_t(u))}{M(R_t^A) + M(R_t^B)} \quad (4)$$

4 Sequential Monte Carlo Tracking

The densely-connected structure of the factorized graphical model as shown in Figure 1 is complex. The structure variational analysis can be taken to analyze the graphical model [11]. Analytical results of a set of fixed-point equations were obtained based on some simplifications such as

linear observation likelihood [5, 11]. In addition, the fixed-point equations reveal a co-inference phenomenon [19]. However, in general, the exact probabilistic inference of the hidden processes would be very difficult especially when the observation likelihood is complicated.

On the other hand, statistical sequential Monte Carlo strategies provide a computational approach to this problem [4, 13, 14], in which a probability density is approximated by a set of weighted particles. The evolution of the set of particles according to the dynamic Bayesian network characterizes the behavior of the dynamic system, and the hidden processes can be recovered from the set of particles. Many particle-based algorithms have been studied for visual tracking [7, 15, 19].

We take a sequential Monte Carlo approach to inferencing the factorized dynamic Bayesian network in Figure 1. The posterior density $p(\mathbf{X}_t^A, \mathbf{X}_t^B, \alpha_t | \underline{\mathbf{Z}}_t)$ is represented by a set of weighted particles $\{x_t^{A,(n)}, x_t^{B,(n)}, \alpha_t^{(n)}, \pi^{(n)}\}$. The sampling-based algorithm is summarized in Figure 4.

Generate $\{x_{t+1}^{A,(n)}, x_{t+1}^{B,(n)}, \alpha_{t+1}^{(n)}, \pi_{t+1}^{(n)}\}$ from $\{x_t^{A,(n)}, x_t^{A,(n)}, \alpha_t^{(n)}, \pi_t^{(n)}\}$.

1. `Re-sampling.`
   Resample the particle set $\{x_t^{A,(n)}, x_t^{B,(n)}, \alpha_t^{(n)}\}$ to produce $\{x_t'^{A,(n)}, x_t'^{B,(n)}, \alpha_t'^{(n)}\}$ based on $\{\pi_t^{(n)}\}$.
2. `Prediction.` For each $(x_t'^{A,(n)}, x_t'^{B,(n)}, \alpha_t'^{(n)})$:
   (a) sample the density of the target dynamics $p(x_{t+1}^A | x_t^A)$ to produce $x_{t+1}^{A,(n)}$ from $x_t'^{A,(n)}$;
   (b) sample the target dynamics $p(x_{t+1}^B | x_t^B)$ to produce $x_{t+1}^{B,(n)}$ from $x_t'^{B,(n)}$;
   (c) sample the finite state machine $\mathbf{T}_\alpha$ of $p(\alpha_{t+1}|\alpha_t)$ to produce $\alpha_{t+1}^{(n)}$ from $\alpha_t'^{(n)}$.
3. `Correction.` Re-weight each particle by calculating the likelihood

$$\pi_{t+1}^{(n)} = p(\mathbf{Z}_{t+1} | x_{t+1}^{A,(n)}, x_{t+1}^{B,(n)}, \alpha_{t+1}^{(n)}).$$

Then normalize all the new weights to 1.

Figure 4: The sequential Monte Carlo algorithm for the factorized dynamic Bayesian network in Figure 1.

Based on the weighted particle set at each time instant, we obtain the estimation of the hidden states:

$$\hat{\mathbf{X}}_t^k = \sum_n x_t^{k,(n)} \pi_t^{(n)}, \qquad k = \{A, B\},$$

$$\hat{\alpha}_t = arg\max_\alpha \sum_{\alpha_t^{(n)} = \alpha} \pi_t^{(n)}, \qquad \alpha = \{0, 1, 2\}.$$

5 Switching Multiple Views

Most appearance-based methods are sensitive to view changes and large deformations, since appearances are view-based. Subspace-based techniques can be employed to learn the appearance-based representations which are robust to views [12] and large appearance changes [2]. These representations are suitable for target detection and recognition, but the dimensionality of the subspace is high for the tracking tasks.

To model view changes, we simplify the subspace-based approaches, and represent a target by maintaining a finite set of examplar view templates, each of which is associated with a transformation, i.e.,$\{(T_1, H_1), \cdots, (T_V, H_V)\}$. Denote an indicator variable by $\beta \in \{1, \cdots, V\}$. Our representation stipulates that the whole set of appearances under different views can be divided into a set of non-overlapped subsets represented by (T_β, H_β). In other words, for any appearance, a unique view template T_β and a suitable transformation exist. This method extends the "view+transoformation" approach to a "switch view+transformation" representation in the spirit of the Toyama and Blake's examplar-based tracking [18].

This representation is different from subspace representations. In subspace methods, since an appearance is modelled by a linear/nonlinear combination of a set of appearance basis, the methods are global. On the other hand, our "switch view+transformation" approach identifies a specific "mode" (although it is a special case of linear combination), and it is local, like a piece-wise spline in the appearance space. Thus, our approach uses a switch β to switch among different "modes" or views templates.

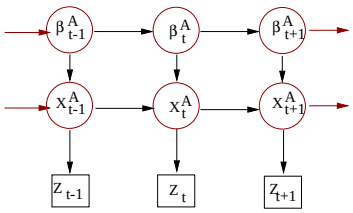

Figure 5: A discrete hidden process $\{\beta_t^A\}$ is used to switch among different views of the target A.

Accommodating this switching view representation in the generative model, the dynamic Bayesian net for a signal target can be illustrated in Figure 5, where $\{\beta_t^A\}$ is the hidden process, and we have

$$p(\mathbf{X}_{t+1}^A, \beta_{t+1}^A | \mathbf{X}_t^A, \beta_t^A) = p(\mathbf{X}_{t+1}^A | \mathbf{X}_t^A, \beta_{t+1}^A) p(\beta_{t+1}^A | \beta_t^A)$$

where $p(\mathbf{X}_{t+1}^A | \mathbf{X}_t^A, \beta_{t+1}^A)$ describes the switch of view templates and its dynamics, and $p(\beta_{t+1}^A | \beta_t^A)$ models the transition of the switch event which is stipulated by a finite state machine:

$$\mathbf{T}_\beta^A = [T_\beta^A(i,j)] = \left[p(\beta_t^A = j | \beta_{t-1}^A = i)\right].$$

Although we can perform the structure variational analysis on this graphical model in Figure 5 (see [16]), a more flexible approach for inferencing is again the sequential Monte Carlo strategies. Similar to the mixed-state CONDENSATION [8], a particle for the target A is represented as $\{x_t^{A,(n)}, \beta_t^{A,(n)}, \pi_t^{(n)}\}$. The evolution of the set of particles is generated by the dynamic Bayesian net model. The estimate of the view is given by:

$$\hat{\beta}_t^A = \arg\max_{\beta} \sum_{\beta_t^{(n)}=\beta} \pi_t^{(n)}; \quad (5)$$

$$\hat{\mathbf{X}}_t^A = \sum_{\beta_t^{(n)}=\hat{\beta}_t^A} \pi_t^{(n)} x_t^{A,(n)} \pi_t^{(n)}. \quad (6)$$

Naturally, the combination of the occlusion model in Figure 1 and the model for switching views in Figure 5 results in a new dynamic Bayesian network as illustrated in Figure 6, which models the occlusion of multiple targets as well as multiple views. Taking the sequential Monte Carol

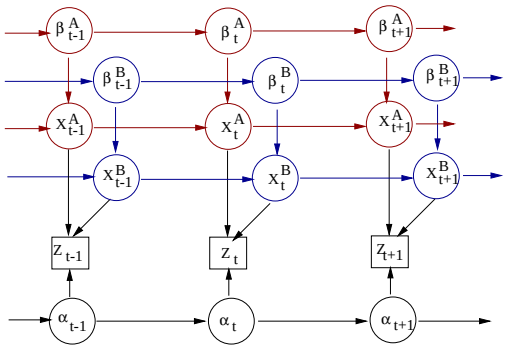

Figure 6: A hidden process $\{\alpha_t\}$ controls the occlusion relations among different targets and $\{\beta_t^k\}$ switches among different views for the k-th target, where $k \in \{A, B\}$.

methods similar as those in previous sections, the inference of this dynamic Bayesian net is straightforward.

6 Experiments

The proposed methods have been applied to the task of tracking two moving and occluding faces. We report the experiments in three tracking scenarios including occlusion, view changes and the combination of the two.

Our first experiment was concerned about the inference of occlusions induced by the interaction of two targets, and the generative model in Figure 1 applied. In this case, the appearance of a face was represented by a single pre-trained view template of the face and an affine transformation. The tracking task was to estimate the affine parameters for both templates as well as the occlusion relation when the two faces crossed. We employed two types of view templates: one was the image template, and the other was the texture template based on wavelet transformations.

Since the overlapping can be directly calculated once $\mathbf{X}_t^{A,(n)}$ and $\mathbf{X}_t^{B,(n)}$ are given, the uncertainty remained for occlusion variable α_t is either $\alpha_t = 1$ or $\alpha = 2$. Then the transition of $\{\alpha_t\}$ is reduced to a two-state machine. In the experiment, we set

$$\mathbf{T}_\alpha = p(\alpha_j|\alpha_i) = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}, \quad i, j \in \{1, 2\}. \quad (7)$$

because we found the results were not sensitive to $\mathbf{T}_\alpha$. The tracking results can be seen in "`occlusion.mpg`"[1]. Some sample frames of the tracking results are shown in Figure 7. In this experiment, the size of the particle set was 2000. When the two faces crossed, the tracker proved to keep locking on the two faces with the right identities, because the occlusion relation was recovered during tracking, which greatly helped to maintain the identities of different targets. The occlusion was estimated by maximizing the *a posteriori* in Equation 5. The recovered occlusion process $\{\alpha_t\}$ is shown in Figure 8. The estimates of the occlusion

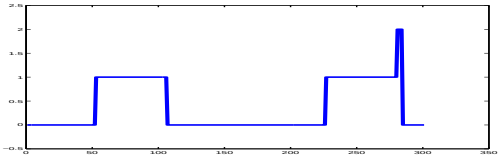

Figure 8: The recovered occlusion process $\{\alpha_t\}$.

relations were quite accurate, except for the frames where the occlusion was about to occur or about to finish. But this phenomenon was reasonable since the occlusion relations were weak and uncertain at those time instants. Since a face went back and forth in front of the other face in the sequence, the occlusion events $\alpha = 1$ occurred in two time intervals. This is clearly indicated in Figure 8.

Figure 9: The three view templates used for the multiple appearances switching.

The second experiment was about the multiple view model, and the generative model in Figure 5 applied. The task was to track a single face but the motion of the face contains out-plane rotations, which resulted in multiple distinguishable views. In this experiment, we exploited three view templates: one front view, and two profile views, with three homography transformations associated with each

[1] All results can be accessed from http://www.ece.nwu.edu/~yingwu

Figure 7: Two faces are tracked (in red or green) during the occlusion. One becomes dark if occluded. Their occlusion relations are inferred and the identities of the two faces are maintained. (See "`occlusion.mpg`" for detail.)

template. The three templates are shown in Figure 9. Here, $\beta = 1/2/3$ denotes left profile, front and right profile views, respectively. The transition of the view switching process $\{\beta_t\}$ was a three-state FSM:

$$\mathbf{T}_\beta = p(\beta_j|\beta_i) = \begin{bmatrix} 0.8 & 0.15 & 0.05 \\ 0.1 & 0.8 & 0.1 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}. \quad (8)$$

The result for the single face with multiple views is shown in the sequence "`multiview.mpg`". Some sample frames are shown in Figure 10. The size of the particle set in the sequential Monte Carlo inference was 1000. When the face turned, the correct view template was automatically selected and the tracker switched to this view template and kept tracking. Since the particle set represents the density, it implicitly keeps all the view hypotheses and the priors of these hypotheses are propagated from previous time instants. The calculation of the likelihood of the image observation given these view hypotheses can strengthen or weaken these hypotheses. The one with the maximum posterior probability was selected as the estimation of the view template "mode" at each time instant. The recovered process of mode switching is shown in Figure 11. We see

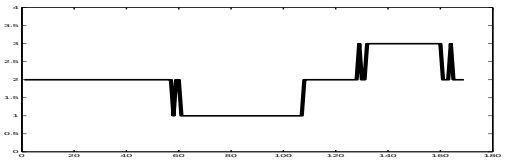

Figure 11: The recovered switching process $\{\beta_t\}$.

clearly from this figure that the person turns his head around when he moves.

In the third experiment, we tracked two faces under occlusion and multiple views, and the method in Figure 6 applied. The same as the second experiment, we used a three-view templates with homography transformations. And $\mathbf{T}_\alpha$ used Equation 7, and $\mathbf{T}_\beta^A$ and $\mathbf{T}_\beta^B$ used Equation 8.

The sequence "`occlu_multiview.mpg`" demonstrates the tracking result for the two faces with multiple views. Some sample frames are shown in Figure 12. Due to the complexity of the dynamic Bayesian net in Figure 6 used in this experiment, more particles are needed for effective Monte Carlo. We used 4000 particles to obtain the result. By accommodating the processes of occlusion and view switching, the tracker needs to infer more hidden factors based on the image observations, thus more computation is involved. But the payoff is huge: the tracker becomes more robust and the recovered hidden factors provide quantitative clues for evaluating the tracking performance online.

7 Discussion and Conclusions

Appearance-based tracking is useful in many applications such as face tracking, but is confronted by the problem of occlusion, especially when multiple appearances are concerned. This paper presents a generative model to accommodate a hidden process of occlusion relations among multiple targets. The likelihood of the image observation is conditioned on the configuration of the states of multiple appearances as well as an occlusion relation among them. Graphically, such a generative model is a factorized dynamic Bayesian network with multiple hidden Markov chains. In addition, this paper also presents a multiple view representation for appearances by a "switch view+transformation" approach. Accommodating multiple views in the dynamic Bayesian network results in a mode-switch model. The inference of the hidden processes is made possible through particle-based sequential Monte Carlo methods, by which the the mode and transformations of different appearances as well as their occlusion relations can be recovered.

The generative models explicitly represent the hidden factors which affect the image observations, thus the recovery of these hidden factors would provide significant interpretation of the image sequences besides tracking. Since analytical results are in general hard to obtain, when more factors are included in the generative model, the computational complexity tends to be more tremendous. Thus, more efficient Monte Carlo methods should be developed to ease these computational issues. In addition, instead of presetting the parameters in the models, learning these parameters from training data would be more plausible. Our future work will include these two issues.

Acknowledgments

This work was supported in part by Northwestern startup funds for YW and Murphy Fellowships for TY and GH. We also thank the anonymous reviewers for their valuable comments.

Figure 10: Tracking one face with out-plane rotations with the switching multiple view model. A suitable appearance template is selected automatically at each time instant. (See "`multiview.mpg`" for detail.)

Figure 12: Two faces move across inducing occlusion, and the motion of the faces contains out-plane rotations. The occlusion (the occluded one is shown in dark) are inferred and the suitable view templates are switched. (See "`occlu_multiview.mpg`" for detail.)

References

[1] Stan Birchfield. Ellitical head tracking using intensity gradient and color histograms. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 232–237, Santa Barbara, California, June 1998.

[2] Michael Black and Allan Jepson. Eigentracking: Robust matching and tracking of articulated object using a view-based representation. In *Proc. European Conf. Computer Vision*, volume 1, pages 343–356, Cambridge, UK, 1996.

[3] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume II, pages 142–149, Hilton Head Island, South Carolina, 2000.

[4] Arnaud Doucet, S. J. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[5] Zoubin Ghahramani and Michael Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–275, 1997.

[6] Greg Hager and Peter Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20:1025–1039, 1998.

[7] Michael Isard and Andrew Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. of European Conf. on Computer Vision*, pages 343–356, Cambridge, UK, 1996.

[8] Michael Isard and Andrew Blake. A mixed-state condensation tracker with automatic model-switching. In *Proc. of IEEE Int'l Conf. on Computer Vision*, pages 107–112, India, 1998.

[9] Allan Jepson, David Fleet, and Thomas El-Maraghi. Robust online appearance models for visual tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume I, pages 415–422, Kauai, Hawaii, Dec. 2001.

[10] Nebojsa Jojic, Nemanja Petrovic, Brendan Frey, and Thomas S. Huang. Transformed hidden Markov models: Estimating mixture models and inferring spatial transformations in video sequences. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, June 2000.

[11] Micheal Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 2000.

[12] S. Z. Li, X G. Lv, and H. J Zhang. View-based clustering of object appearances based on independent subspace analysis. In *Proc. IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, July 2001.

[13] Jun Liu and Rong Chen. Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.*, 93:1032–1044, 1998.

[14] Jun Liu, Rong Chen, and Tanya Logvinenko. A theoretical framework for sequential importance sampling and resampling. In A. Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo in Practice*. Springer-Verlag, New York, 2000.

[15] John MacCormick and Andrew Blake. A probabilistic exclusion principle for tracking multiple objects. In *Proc. IEEE Int'l Conf. on Computer Vision*, pages 572–578, Greece, 1999.

[16] Vladimir Pavlovic, James Rehg, Tat-Jen Cham, and Kevin Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proc. IEEE Int'l Conf. on Computer Vision*, volume I, pages 94–101, Corfu, Greece, Sept. 1999.

[17] Hai Tao, Harpreet Sawhney, and Rakesh Kumar. Dynamic layer representation with applications to tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 134–141, 2000.

[18] Kentaro Toyama and Andrew Blake. Probabilistic tracking in a metric space. In *Proc. IEEE Int'l Conf. on Computer Vision*, Vancouver, Canada, July 2001.

[19] Ying Wu and Thomas S. Huang. Robust visual tracking by co-inference learning. In *Proc. IEEE Int'l Conference on Computer Vision*, volume II, pages 26–33, Vancouver, July 2001.