Switching Hypothesized Measurements: A Dynamic Model with Applications to Occlusion Adaptive Joint Tracking

Yang Wang **Tele Tan**
Institute for Infocomm Research, Singapore
{ywang, teletan}@i2r.a-star.edu.sg

Kia-Fock Loe
Dept. Computer Science, National Univ. Singapore
loekf@comp.nus.edu.sg

Abstract

This paper proposes a dynamic model supporting multimodal state space probability distributions and presents the application of the model in dealing with visual occlusions when tracking multiple objects jointly. For a set of hypotheses, multiple measurements are acquired at each time instant. The model switches among a set of hypothesized measurements during the propagation. Two computationally efficient filtering algorithms are derived for online joint tracking. Both the occlusion relationship and state of the objects are recursively estimated from the history of measurement data. The switching hypothesized measurements (SHM) model is generally applicable to describe various dynamic processes with multiple alternative measurement methods.

1 Introduction

Visual tracking is important in such application areas as human-computer interaction, surveillance, and visual reconstruction. Given a sequence of images containing the objects that are represented with a parametric motion model, parameters of the motion model are required to be estimated in successive frames. Tracking could be difficult due to the potential variability such as partial or full occlusions of objects, appearance changes caused by the variation of object poses or illumination conditions, as well as distractions from background clutter.

One principle challenge for visual tracking is to develop an accurate and effective model representation. The variability in visual environments usually results in a multimodal state space probability distribution. The Kalman filter [Brown, 1983; Rohr, 1994], a classical choice employed in tracking work, is restricted to representing unimodal probability distributions. Switching linear dynamic systems (SLDS) [Pavlovic and Rehg, 2000] and their equivalents [Shumway and Stoffer, 1991; Kim, 1994] have been used to describe dynamic processes. Intuitively, a complex dynamic system is represented with a set of linear models controlled by a switching variable. Joint probabilistic data association (JPDA) [Bar-Shalom and Fortmann, 1988] and multiple hypothesis tracking (MHT) [Cox and Hingorani, 1996] techniques, which represent multimodal distributions by constructing data association hypotheses, can be cast in the framework of SLDS as well. Moreover, Monte Carlo methods such as the Condensation algorithm [Isard and Blake, 1996] support multimodal probability densities with sample based representation. By retaining only the peaks of the probability density, relatively fewer samples are required in the work of Cham and Rehg [1999]. A switching model framework of the Condensation algorithm is also proposed by Isard and Blake [1998].

On the other hand, the measurement process is another essential issue to deal with the potential variability. Measurements are not readily available from image sequences in visual tracking. Even an accurate tracking model may have a poor performance if the measurements are too noisy. Parametric models can be used to characterize appearance changes of target regions [Hager and Belhumeur, 1998]. In the work of Galvin et al. [1999], two virtual snakes, a background and a foreground snake for each object, are generated to resolve the occlusion when two objects intersect. Rasmussen and Hager [2001] describe a joint measurement process for tracking multiple objects enumerating all possible occlusion relationships. The measurement with respect to the most possible occlusion relationship is determined from the current frame. Moreover, layered approach [Wang and Adelson, 1994; Ayer and Sawhney, 1995; Jojic and Frey, 2001; Tao et al., 2002] is an efficient way to represent multiple moving objects. A moving object is characterized by a coherent motion model over its support region.

In this paper, the idea of switching hypothesized measurements (SHM), which results in a SHM model supporting multimodal distributions, is proposed to handle the potential variability in visual tracking. The approach acquires a set of hypothesized measurements for different occlusion hypotheses at each time instant. Comparing with the above mentioned state space models, the SHM approach switches among a set of hypothesized measurements rather than switches among a set of models. Two computationally efficient filtering algorithms are derived for jointly tracking multiple objects. Both the occlusion relationship and state of the objects are estimated from the history of measurements.

2 Model

2.1 Hypothesized Measurement

For a hidden state sequence $\{\mathbf{z}_k\}$ ($k \in \mathbf{N}$), the objective of online tracking is to recursively estimate $\mathbf{z}_k$ from the set of all available measurements $\mathbf{y}_{1:k} = \{\mathbf{y}_i\}_{1 \le i \le k}$ up to time k. For a certain complex system, the estimation may be influenced by a mode or switching state sequence $\{s_k\}$ as well, with $s_k \in \{1, 2, \ldots, L\}$ ($L \in \mathbf{N}$). Specifically, the mode switching originates from the measurement process in our work. The notion of a measurement is extended to a set of L hypothesized measurements $\mathbf{y}_k = (\mathbf{y}_{k,1}, \mathbf{y}_{k,2}, \ldots, \mathbf{y}_{k,L})$ at each time instant. Each $\mathbf{y}_{k,j}$ ($1 \le j \le L$) is called a hypothesized measurement since it is obtained by assuming that the switching state s_k is j at time k.

To illustrate the idea of hypothesized measurement, consider the measurement process for jointly tracking two objects, e.g. a rectangle and a circle, in an image sequence $\{g_k\}$. To deal with occlusions between the two objects when measuring the kth frame g_k, the switching state s_k is introduced to describe the depth ordering at time k. $s_k \in \{1, 2\}$, where s_k equals 1 if the rectangle is in front of the circle, and 2 if the circle is in front of the rectangle. The hypothesized measurement $\mathbf{y}_{k,j}$ ($1 \le j \le 2$) is denoted as $(\mathbf{y}_{k,j}^{(1)}, \mathbf{y}_{k,j}^{(2)})^T$, where $\mathbf{y}_{k,j}^{(1)}$ is the measurement for the rectangle, and $\mathbf{y}_{k,j}^{(2)}$ is the measurement for the circle under the jth hypothesis.

Under the hypothesis of $s_k = 1$, i.e. the circle is occluded by the rectangle at time k, the rectangle should be measured first to acquire $\mathbf{y}_{k,1}^{(1)}$. Then the observed rectangle is masked in the image. The occluded area of the circle is ignored and only the visible region is matched normally to get $\mathbf{y}_{k,1}^{(2)}$. Similarly, under the hypothesis of $s_k = 2$, i.e. the rectangle is occluded by the circle, the circle should be matched first to get $\mathbf{y}_{k,2}^{(2)}$, then the masked image is used to measure $\mathbf{y}_{k,2}^{(1)}$. Thus, the occlusion will not affect the measurement result. It is obvious that both hypothesized measurements support the condition of nonocclusion since different depth orderings of nonoverlapping objects are visually equivalent. The probabilities of the hypotheses should be equal in the case of nonocclusion.

Unfortunately, the occlusion relationship is not given before hand. The objective of our SHM approach is to estimate both the switching state and the hidden state from the history of the hypothesized measurements.

2.2 Linear SHM Model for Joint Tracking

For joint tracking of M ($M \in \mathbf{N}$) objects in the scene, the switching state s_k represents the occlusion relationship at time k. $s_k \in \{1, 2, \ldots, L\}$, and $L = M!$. We assume the switching state follows a first order Markov chain with the following transition probability,

$$p(s_{k+1} = i \mid s_k = j) = \alpha_{i,j}, \tag{1}$$

with $\sum_i \alpha_{i,j} = 1$. The hidden state $\mathbf{z}_k$ is denoted as $(\mathbf{z}_k^{(1)}, \mathbf{z}_k^{(2)}, \ldots, \mathbf{z}_k^{(M)})^T$, with $\mathbf{z}_k^{(m)}$ ($1 \le m \le M$) being the state for the mth object at time k. For a linear process with Gaussian noise, the hidden state transition model becomes

$$\mathbf{z}_{k+1} = \mathbf{F}\mathbf{z}_k + \mathbf{n}, \tag{2a}$$

$$p(\mathbf{z}_{k+1} \mid \mathbf{z}_k) = N(\mathbf{z}_{k+1}; \mathbf{F}\mathbf{z}_k, \mathbf{Q}), \tag{2b}$$

where $\mathbf{F} \in \mathbf{R}^{n_z} \times \mathbf{R}^{n_z}$ is the state transition matrix, $\mathbf{n}$ is a zero-mean Gaussian noise with covariance matrix $\mathbf{Q}$, and $N(\mathbf{z}; \mathbf{m}, \Sigma)$ is a Gaussian density with argument $\mathbf{z}$, mean $\mathbf{m}$, and covariance Σ.

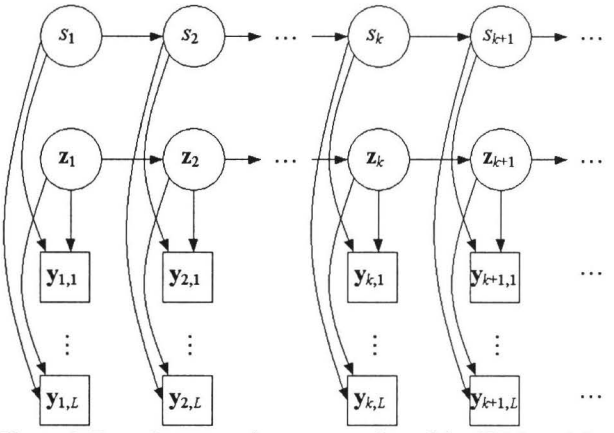

Figure 1: Bayesian network representation of the SHM model

Given the switching state s_k at time k, the corresponding hypothesized measurement $\mathbf{y}_{k,s_k}$ could be considered as a proper measurement centering on the hidden state, while every other $\mathbf{y}_{k,j}$ for $j \ne s_k$ is an improper measurement generated under a wrong assumption. The improper measurement should be weakly influenced by the hidden state $\mathbf{z}_k$ and have a large variance. To simplify the computation, we assume a normal distribution for a proper measurement and a uniform distribution for an improper measurement. The measurement model is simplified as

$$\mathbf{y}_{k,j} = \begin{cases} \mathbf{z}_k + \mathbf{v}_{k,j}, \text{if } j = s_k, \\ \mathbf{w}, \text{otherwise}, \end{cases} \tag{3a}$$

$$p(\mathbf{y}_{k,j} \mid s_k, \mathbf{z}_k) = \begin{cases} N(\mathbf{y}_{k,j}; \mathbf{z}_k, \mathbf{R}_{k,j}), \text{if } j = s_k, \\ \text{a constant, otherwise}, \end{cases} \tag{3b}$$

where $\mathbf{v}_{k,j}$ is a zero-mean Gaussian noise with covariance matrix $\mathbf{R}_{k,j}$, and $\mathbf{w}$ is a uniformly distributed noise, whose density is a small positive constant. For the measurement of M objects, $\mathbf{y}_{k,j}$ is denoted as $(\mathbf{y}_{k,j}^{(1)}, \mathbf{y}_{k,j}^{(2)}, \ldots, \mathbf{y}_{k,j}^{(M)})^T$, and $\mathbf{v}_{k,j}$ is written as $(\mathbf{v}_{k,j}^{(1)}, \mathbf{v}_{k,j}^{(2)}, \ldots, \mathbf{v}_{k,j}^{(M)})^T$. Given the current state, the conditional independence among the hypothesized measurements is assumed to make the model computationally efficient.

$$p(\mathbf{y}_k \mid s_k = j, \mathbf{z}_k) = p(\mathbf{y}_{k,1}, \mathbf{y}_{k,2}, \ldots, \mathbf{y}_{k,L} \mid s_k = j, \mathbf{z}_k)$$
$$= p(\mathbf{y}_{k,j} \mid s_k = j, \mathbf{z}_k) \prod_{l \neq j} p(\mathbf{y}_{k,l} \mid s_k = j, \mathbf{z}_k)$$
$$\propto N(\mathbf{y}_{k,j}; \mathbf{z}_k, \mathbf{R}_{k,j}). \quad (4)$$

The SHM model can be represented by a dynamic Bayesian network shown in figure 1.

3 Method

3.1 Measurement

Multiple, occluding objects are modeled using layer representation. Layers are indexed by $m = 1, 2, \ldots, M$, with layer 1 being the layer that is closest to the camera and layer m being behind layer 1, 2, …, m–1. There is one object in each layer. The number of all occlusion relationship hypotheses (or depth ordering permutations) is $L = M!$. Each permutation is tagged with a index j ($1 \le j \le L$).

Under each permutation hypothesis, the object in the front layer 1 should be measured first from the image g_k at time k. Then the object in layer 2 can be matched from the masked image, and so on. At last, the object in layer M can be measured. Occluded points are not matched when measuring the objects. Measurement results of nonoverlapping objects should be equivalent for different depth ordering permutations. Given the reference image g_r ($r < k$), the measurement is based on minimizing the mean of squared intensity differences between the current image and the reference region. Under the jth hypothesis, $\mathbf{y}_{k,j}^{(m)}$ is the hypothesized measurement of the mth object, and $e_{k,j}^{(m)}$ is the corresponding squared difference mean at time k. The vector $(e_{k,j}^{(1)}, e_{k,j}^{(2)}, \ldots, e_{k,j}^{(M)})^T$ is written as $\mathbf{e}_{k,j}$. The covariance matrix $\mathbf{R}_{k,j}$ is obtained by assuming that the components of the measurement noise are uncorrelated to each other, and the variances is proportional to the corresponding squared difference mean.

3.2 SHM Filter

From a Bayesian perspective, the online tracking problem is to recursively calculate the posterior state space distribution. Given the measurement data up to time k, the probability density function (pdf) of the state is expressed as

$$p(s_k = j, \mathbf{z}_k \mid \mathbf{y}_{1:k})$$
$$= p(s_k = j \mid \mathbf{y}_{1:k})\, p(\mathbf{z}_k \mid s_k = j, \mathbf{y}_{1:k})$$
$$= \beta_{k,j}\, N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j}), \quad (5)$$

where $p(s_k = j \mid \mathbf{y}_{1:k})$ is denoted as $\beta_{k,j}$, with $\sum_j \beta_{k,j} = 1$, and the conditional density $p(\mathbf{z}_k \mid s_k = j, \mathbf{y}_{1:k})$ is modeled as a normal distribution $N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j})$ under each switching state hypothesis. Thus the pdf $p(\mathbf{z}_k \mid \mathbf{y}_{1:k})$ is a mixture of L Gaussians.

At time k+1, the set of hypothesized measurements $\mathbf{y}_{k+1}$ becomes available, and it is used to update $\{\beta_{k,j}, \mathbf{m}_{k,j}, \mathbf{P}_{k,j}\}_{1 \le j \le L}$ to $\{\beta_{k+1,i}, \mathbf{m}_{k+1,i}, \mathbf{P}_{k+1,i}\}_{1 \le i \le L}$ via Bayes' rule.

$$p(s_{k+1}, s_k, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k+1})$$
$$= \frac{1}{p(\mathbf{y}_{k+1} \mid \mathbf{y}_{1:k})} p(\mathbf{y}_{k+1} \mid s_{k+1}, \mathbf{z}_{k+1})\, p(s_{k+1}, s_k, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k})$$
$$\propto p(\mathbf{y}_{k+1} \mid s_{k+1}, \mathbf{z}_{k+1})\, p(s_{k+1}, s_k, \mathbf{z}_{k+1} \mid \mathbf{y}_{1:k}). \quad (6)$$

In principle, the filtering process has three stages: prediction, update, and collapsing [Murphy, 1998]. From (1), (2), (4) and (6), the approximate inference algorithm can be derived in a similar way as that in a Gaussian sum filter [Anderson and Moore, 1979].

$$\beta_{k+1,i} = p(s_{k+1} = i \mid \mathbf{y}_{1:k+1})$$
$$\approx \frac{\sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}{\sum_i \sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}, \quad (7a)$$

$$p(\mathbf{z}_{k+1} \mid s_{k+1} = i, s_k = j, \mathbf{y}_{1:k+1}) \approx N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i|j}, \mathbf{P}_{k+1,i|j}), \quad (7b)$$

$$p(\mathbf{z}_{k+1} \mid s_{k+1} = i, \mathbf{y}_{1:k+1})$$
$$\approx \sum_j \beta_{k+1,i|j} N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i|j}, \mathbf{P}_{k+1,i|j})$$
$$\approx N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i}, \mathbf{P}_{k+1,i}), \quad (7c)$$

where

$$\mathbf{m}_{k+1|k,j} = \mathbf{F}\mathbf{m}_{k,j},$$
$$\mathbf{P}_{k+1|k,j} = \mathbf{F}\mathbf{P}_{k,j}\mathbf{F}^T + \mathbf{Q},$$
$$\mathbf{S}_{k+1,i|j} = \mathbf{P}_{k+1|k,j} + \mathbf{R}_{k+1,i},$$
$$\mathbf{K}_{k+1,i|j} = \mathbf{P}_{k+1|k,j}\mathbf{S}_{k+1,i|j}^{-1},$$
$$\mathbf{m}_{k+1,i|j} = \mathbf{m}_{k+1|k,j} + \mathbf{K}_{k+1,i|j}(\mathbf{y}_{k+1,i} - \mathbf{m}_{k+1|k,j}),$$
$$\mathbf{P}_{k+1,i|j} = \mathbf{P}_{k+1|k,j} - \mathbf{K}_{k+1,i|j}\mathbf{P}_{k+1|k,j},$$
$$\beta_{k+1,i|j} = \frac{\alpha_{i,j}\beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}{\sum_j \alpha_{i,j}\beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})},$$
$$\mathbf{m}_{k+1,i} = \sum_j \beta_{k+1,i|j}\mathbf{m}_{k+1,i|j},$$
$$\mathbf{P}_{k+1,i} = \sum_j \beta_{k+1,i|j}[\mathbf{P}_{k+1,i|j} +$$
$$(\mathbf{m}_{k+1,i|j} - \mathbf{m}_{k+1,i})(\mathbf{m}_{k+1,i|j} - \mathbf{m}_{k+1,i})^T].$$

The state at time k+1 is estimated as

$$\hat{s}_{k+1} = \arg\max_i p(s_{k+1} = i \mid \mathbf{y}_{1:k+1}) = \arg\max_i \beta_{k+1,i}, \quad (8a)$$
$$\hat{\mathbf{z}}_{k+1} = \arg\max_{\mathbf{z}_{k+1}} p(\mathbf{z}_{k+1} \mid s_{k+1} = \hat{s}_{k+1}, s_k = \hat{s}_k, \mathbf{y}_{1:k+1})$$
$$= \mathbf{m}_{k+1,\hat{s}_{k+1}|\hat{s}_k}. \quad (8b)$$

From (7) it can be seen that the computation of the SHM filter is slightly more complex than that of multiple Kalman filters or Gaussian sum filters.

3.3 Fast SHM Filter

When occlusion is the main factor in the potential variability of joint tracking, we can assume that the measurement noise under the true occlusion hypothesis is small. When the noise becomes zero, the measurement model can be simplified as

$$p(\mathbf{y}_{k,j} \mid s_k, \mathbf{z}_k) = \begin{cases} \delta(\mathbf{y}_{k,j} - \mathbf{z}_k), \text{if } j = s_k, \\ \text{a constant, otherwise,} \end{cases} \quad (9)$$

where $\delta(\cdot)$ is the Dirac delta function. Consider the minimized mean of squared differences as a part of the hypothesized measurement, so that the definition of measurement can be generalized as

$$\mathbf{y}'_{k,j} = (\mathbf{y}_{k,j}, \mathbf{e}_{k,j}), \ \mathbf{y}'_k = (\mathbf{y}'_{k,1}, \ldots, \mathbf{y}'_{k,L}). \quad (10)$$

Assume that $\mathbf{e}_{k,j}$ is independent on $\mathbf{y}_{k,j}$ and $\mathbf{z}_k$, and the posterior density of the squared difference mean is of exponential distribution for each object (More accurate expression could be derived using the χ^2 distribution.) under the true hypothesis, the pdf of $\mathbf{e}_{k,j}$ is factorized as

$$p(\mathbf{e}_{k,j} \mid s_k, \mathbf{z}_k) = p(\mathbf{e}_{k,j} \mid s_k) = \prod_m p(e_{k,j}^{(m)} \mid s_k), \quad (11a)$$

$$p(e_{k,j}^{(m)} \mid s_k) \begin{cases} \propto \exp[-e_{k,j}^{(m)}], \text{if } j = s_k. \\ \text{is a constant, otherwise.} \end{cases} \quad (11b)$$

The generalized measurement model now becomes

$$\begin{aligned} & p(\mathbf{y}'_k \mid s_k = j, \mathbf{z}_k) \\ &= p(\mathbf{y}'_{k,j} \mid s_k = j, \mathbf{z}_k) \prod_{l \neq j} p(\mathbf{y}'_{k,l} \mid s_k = j, \mathbf{z}_k) \\ &\propto p(\mathbf{y}_{k,j} \mid s_k = j, \mathbf{z}_k)\, p(\mathbf{e}_{k,j} \mid s_k = j, \mathbf{z}_k) \\ &\propto \eta_{k,j} \delta(\mathbf{y}_{k,j} - \mathbf{z}_k), \end{aligned} \quad (12)$$

where $\eta_{k,j} = \prod_m \exp[-e_{k,j}^{(m)}]$.

At time k, the conditional pdf of the state is expressed as

$$p(s_k = j, \mathbf{z}_k \mid \mathbf{y}'_{1:k}) = \beta_{k,j} \delta(\mathbf{z}_k - \mathbf{y}_{k,j}), \quad (13)$$

where $\beta_{k,j}$ is $p(s_k = j \mid \mathbf{y}'_{1:k})$, and $p(\mathbf{z}_k \mid s_k = j, \mathbf{y}'_{1:k})$ equals $\delta(\mathbf{z}_k - \mathbf{y}_{k,j})$. From appendix A, the filtering algorithm is

$$\begin{aligned} \beta_{k+1,i} &= p(s_{k+1} = i \mid \mathbf{y}'_{1:k+1}) \\ &= \frac{\sum_j \alpha_{i,j} \beta_{k,j} \eta_{k+1,i} N(\mathbf{y}_{k+1,i}; \mathbf{F}\mathbf{y}_{k,j}, \mathbf{Q})}{\sum_i \sum_j \alpha_{i,j} \beta_{k,j} \eta_{k+1,i} N(\mathbf{y}_{k+1,i}; \mathbf{F}\mathbf{y}_{k,j}, \mathbf{Q})}. \end{aligned} \quad (14)$$

The state at time k is estimated as

$$\hat{s}_k = \arg\max_j p(s_k = j \mid \mathbf{y}'_{1:k}) = \arg\max_j \beta_{k,j}, \quad (15a)$$

$$\hat{\mathbf{z}}_k = \arg\max_{\mathbf{z}_k} p(\mathbf{z}_k \mid s_k = \hat{s}_k, \mathbf{y}'_{1:k}) = \mathbf{y}_{k,\hat{s}_k}. \quad (15b)$$

The result $\{\mathbf{y}_{k,\hat{s}_k}\}$ can then be plugged into a Kalman filter to achieve improved performance. Such a SHM-Kalman filter keeps the multimodality of the SHM model and has attractive computation requirement. In addition, the collapsing stage is not necessary in the fast SHM filter.

4 Implementation

In practice, we use the second order (constant velocity) model. The hidden state transition function is

$$\begin{pmatrix} \mathbf{z}_{k+2} \\ \mathbf{z}_{k+1} \end{pmatrix} = \begin{pmatrix} 2\mathbf{I} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{z}_{k+1} \\ \mathbf{z}_k \end{pmatrix} + \begin{pmatrix} \mathbf{n} \\ \mathbf{0} \end{pmatrix}, \quad (16)$$

where $\mathbf{z}_k$ is the tracked entity (e.g. position and orientation). The hidden state and state transition matrix can be correspondingly defined. The switching state transition probability is set as

$$\alpha_{i,j} = \begin{cases} 1 - \lambda, \text{if } i = j, \\ \dfrac{\lambda}{L-1}, \text{otherwise,} \end{cases} \quad (17)$$

where λ is a small positive value (0.1 in this paper) so that two successive switching states are more likely to be of the same label. At the beginning, the reference image g_r is set as the initial image g_0. When there is a high confidence in nonocclusion, the reference image can be adaptively updated. The objects are assumed to be separated from each other in g_0. The initial $\beta_{0,j}$, $\mathbf{m}_{0,j}$, and $\mathbf{P}_{0,j}$ should be equal for different j because of nonocclusion. $\beta_{0,j} = p(s_0 = j) = \frac{1}{L}$. The initial mean $\mathbf{m}_{0,j}$ is set as a zero vector. The initial covariance matrix $\mathbf{P}_{0,j}$ is set as diagonal with small variances since the initialization is assumed to be accurate. When an object is totally occluded by the other objects, no points of the target region will be matched. The estimation is based on the result of time k using the state transition function when no visible region of the object is expected at time k+1.

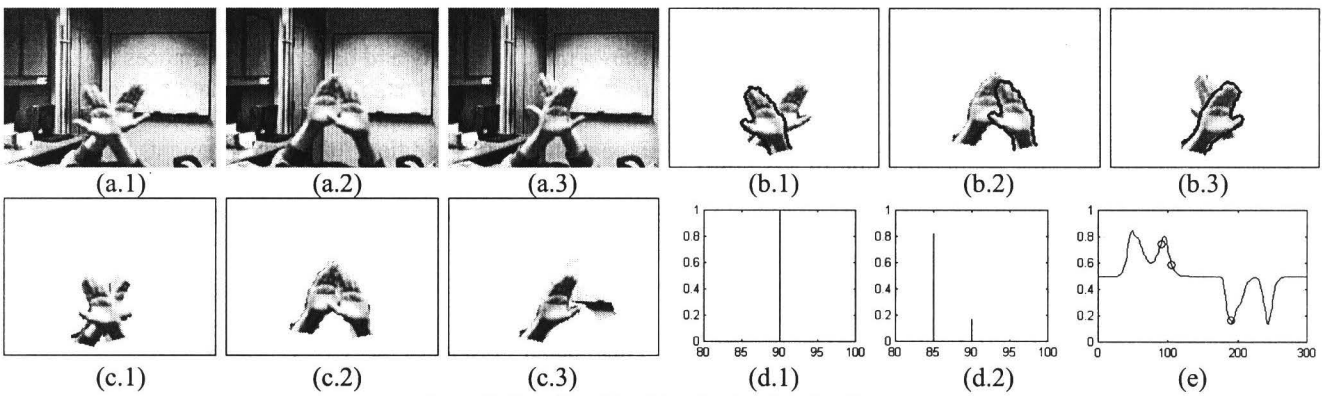

Figure 2: Results of tracking the two hands of a person

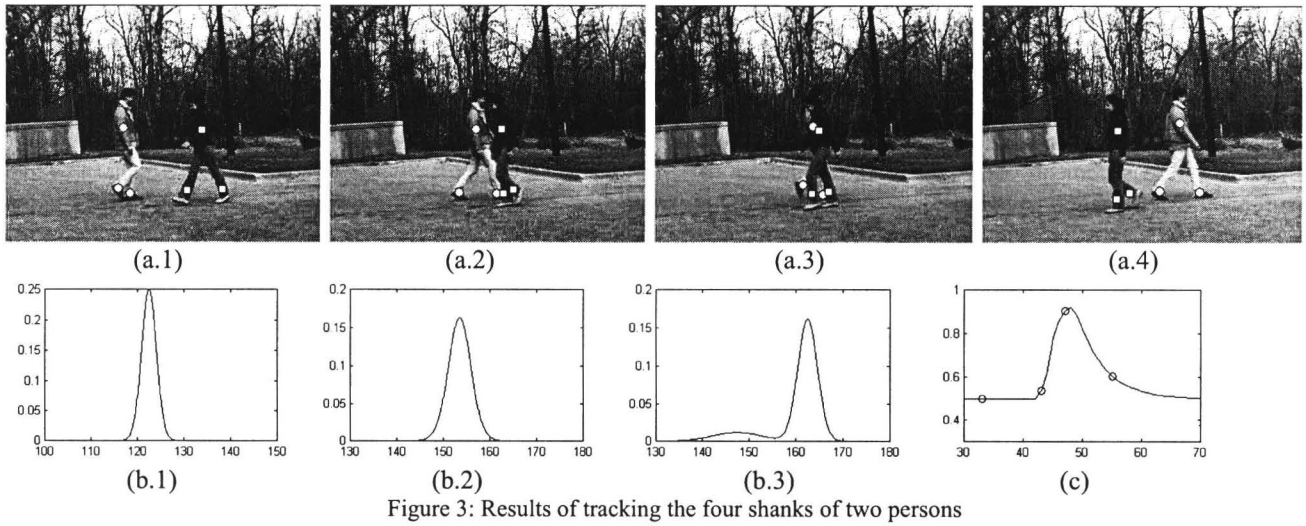

Figure 3: Results of tracking the four shanks of two persons

5 Results and Discussion

Test results of two video sequences are shown in our experiments. The state is the position and orientation, and the second order model is employed. Each measurement is a translation and rotation.

Figure 2 shows the tracking of two hands of a person as they cross several times in an image sequence. Figure 2a shows the three frames of the sequence. Appearance variation of the hands due to pose changes can be seen. Figure 2b and 2c demonstrate the tracking efficacy of the fast SHM filter versus the Kalman filter. The fast SHM filter successfully tracks both hands under different occlusion relationships. In figure 2b, one hand is drawn in black contour when the detected depth order indicates that it is in front of the other hand. The Kalman filter has a similar performance when occlusions are not severe, but poor under heavy occlusions. In figure 2c.3, the distraction from background clutter causes the Kalman tracker to fail. The normalized posterior distributions for the vertical position of the left hand in figure 2a.2 and 2a.3 are shown in figure 2d.1 and 2d.2. When the occlusion is not severe, measurements under the two hypotheses are the same, and the distribution is unimodal. Under heavy occlusions, the distribution becomes multimodal since the two hypothesized measurements tend to be different. The measurement under true hypothesis matches the hand correctly, while the measurement under false hypothesis is distracted by background clutter. Figure 2e shows the probabilities of the first occlusion hypothesis (the left hand being in the front) over the first 300 frames. The probabilities for the three frames in figure 2a are circled. The probabilities of the two hypotheses are equal in the nonoverlapping cases, while the probability of the true hypothesis becomes dominant under occlusions.

Figure 3 shows the results of jointly tracking the four shanks of a man and a woman as they cross. The man's right shank has been totally occluded in the sequence. There should be totally 4! = 24 hypotheses if we directly apply the SHM filter in section 3.2. To reduce the computation, two reasonable assumptions are made to prune less plausible hypotheses. Firstly, one's legs can not simultaneously occlude and be occluded by the other's legs. Secondly, the occlusion relationship between the man and woman can be determined from their bodies. Thus the whole tracking procedure is divided into three trackers. The first one tracks the two bodies of the walkers. According to the detected occlusion relationship, the two shanks of the person in the front are then tracked. At last, the shanks of the other person are tracked in the masked image. Figure 3a shows the results for the four frames of the sequence (circles are marked on the man's body and shanks, and rectangles are marked on the woman). Figure 3b shows the posterior distributions for the horizontal position of the occluded body in figure 3a.1-3a.3. When the two bodies are separated, the density is unimodal and of a small variance. The density variance increases when occlusions occur. It becomes multimodal under heavy occlusions. Figure 3c shows the probabilities of the woman's body being in the front. The probabilities for the four frames in figure 3a are circled.

Under realistic environments, it is understandable that comparing with the other hypothesized measurements, the measurement under the true occlusion hypothesis usually shows more regularity and has smaller variances (or squared difference means). Thus, the true information (the switching state and the hidden state) could be enhanced through the propagation. In addition, the acquirement of multiple measurements helps decrease the information loss (e.g. caused by background clutter) during the measurement process.

6 Conclusion

This paper proposes a SHM model for state space representation of dynamic systems and derives two efficient filtering algorithms. Our joint tracking approach explicitly

reasons about the occlusion relationships. The occlusion relationship is quantitatively analyzed throughout the propagation. The information can be used for reference update and further analysis. Moreover, experimental results show that our method helps handle appearance changes and distractions.

The SHM model discusses the measurement switching in dynamic systems. It is complementary to the idea of model switching in [Ghahramani and Hinton, 1998]. Our future study is to effectively combine these two ideas in visual tracking. Furthermore, the SHM model is generally applicable to various dynamic processes in which there are multiple alternative measurement methods.

A Inference in Fast SHM Filter

The predictive density of the state is

$$\begin{aligned}
&p(s_{k+1}=i,s_k=j,\mathbf{z}_{k+1}\mid \mathbf{y}'_{1:k})\\
&=\int p(s_{k+1}=i,\mathbf{z}_{k+1}\mid s_k=j,\mathbf{z}_k)p(s_k=j,\mathbf{z}_k\mid \mathbf{y}'_{1:k})d\mathbf{z}_k\\
&=p(s_{k+1}=i\mid s_k=j)p(s_k=j\mid \mathbf{y}'_{1:k})\cdot\\
&\quad\int p(\mathbf{z}_{k+1}\mid \mathbf{z}_k)p(\mathbf{z}_k\mid s_k=j,\mathbf{y}'_{1:k})d\mathbf{z}_k\\
&=\alpha_{i,j}\beta_{k,j}\int N(\mathbf{z}_{k+1};\mathbf{F}\mathbf{z}_k,\mathbf{Q})\delta(\mathbf{z}_k-\mathbf{y}_{k,j})d\mathbf{z}_k\\
&=\alpha_{i,j}\beta_{k,j}N(\mathbf{z}_{k+1};\mathbf{F}\mathbf{y}_{k,j},\mathbf{Q}).
\end{aligned}\tag{18}$$

After receiving the measurement $\mathbf{y}'_{k+1}$, the posterior density is updated using (6).

$$\begin{aligned}
&p(s_{k+1}=i,s_k=j,\mathbf{z}_{k+1}\mid \mathbf{y}'_{1:k+1})\\
&\propto p(\mathbf{y}'_{k+1}\mid s_{k+1}=i,\mathbf{z}_{k+1})\ p(s_{k+1}=i,s_k=j,\mathbf{z}_{k+1}\mid \mathbf{y}'_{1:k})\\
&\propto \alpha_{i,j}\beta_{k,j}\eta_{k+1,i}\delta(\mathbf{y}_{k+1,i}-\mathbf{z}_{k+1})N(\mathbf{z}_{k+1};\mathbf{F}\mathbf{y}_{k,j},\mathbf{Q})\\
&=\alpha_{i,j}\beta_{k,j}\eta_{k+1,i}\delta(\mathbf{z}_{k+1}-\mathbf{y}_{k+1,i})N(\mathbf{y}_{k+1,i};\mathbf{F}\mathbf{y}_{k,j},\mathbf{Q}).
\end{aligned}\tag{19}$$

The probability of the switching state is updated as

$$\begin{aligned}
&\beta_{k+1,i}=p(s_{k+1}=i\mid \mathbf{y}'_{1:k+1})\\
&=\sum_j\int p(s_{k+1}=i,s_k=j,\mathbf{z}_{k+1}\mid \mathbf{y}'_{1:k+1})d\mathbf{z}_{k+1}\\
&\propto \eta_{k+1,i}\sum_j\alpha_{i,j}\beta_{k,j}N(\mathbf{y}_{k+1,i};\mathbf{F}\mathbf{y}_{k,j},\mathbf{Q}).
\end{aligned}\tag{20}$$

Since $\sum_i \beta_{k+1,i}=1$, $\beta_{k+1,i}$ can be calculated by normalizing.

$$\beta_{k+1,i}=\frac{\sum_j\alpha_{i,j}\beta_{k,j}\eta_{k+1,i}N(\mathbf{y}_{k+1,i};\mathbf{F}\mathbf{y}_{k,j},\mathbf{Q})}{\sum_i\sum_j\alpha_{i,j}\beta_{k,j}\eta_{k+1,i}N(\mathbf{y}_{k+1,i};\mathbf{F}\mathbf{y}_{k,j},\mathbf{Q})}.$$

References

[Anderson and Moore, 1979] B. D. O. Anderson and J. B. Moore, *Optimal filtering*, Prentice-Hall, 1979.

[Ayer and Sawhney, 1995] S. Ayer and H. S. Sawhney, Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding, In *Proc. International Conf. Computer Vision*, pp. 777–784, 1995.

[Bar-Shalom and Fortmann, 1988] Y. Bar-Shalom and T. E. Fortmann, *Tracking and data association*, Academic Press, 1988.

[Brown, 1983] R. G. Brown, *Introduction to random signal analysis and Kalman filtering*, John Wiley & Sons, 1983.

[Cham and Rehg, 1999] T.-J. Cham and J. M. Rehg, A multiple hypothesis approach to figure tracking, In *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 239–245, 1999.

[Cox and Hingorani, 1996] I. J. Cox and S. L. Hingorani, An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking, *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 18, pp. 138–150, 1996.

[Galvin et al., 1999] B. Galvin, B. McCane, and K. Novins, Virtual snakes for occlusion analysis, In *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 294–299, 1999.

[Ghahramani and Hinton, 1998] Z. Ghahramani and G. E. Hinton, Variational learning for switching state-space models, *Neural Computation*, vol. 12, pp. 963–996, 1998.

[Hager and Belhumeur, 1998] G. D. Hager and P. N. Belhumeur, Efficient region tracking with parametric models of geometry and illumination, *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 20, pp. 1025–1039, 1998.

[Isard and Blake, 1998] M. Isard and A. Blake, A mixed-state Condensation tracker with automatic model-switching, In *Proc. International Conf. Computer Vision*, pp. 107–112, 1998.

[Isard and Blake, 1996] M. Isard and A. Blake, Contour tracking by stochastic propagation of conditional density, In *Proc. European Conf. Computer Vision*, pp. 343–356, 1996.

[Jojic and Frey, 2001] N. Jojic and B. J. Frey, Learning flexible sprites in video layers, In *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 199–206, 2001.

[Kim, 1994] C.-J. Kim, Dynamic linear models with Markov-switching, *Journal of Econometrics*, vol. 60, pp. 1–22, 1994.

[Murphy, 1998] K. P. Murphy, Learning switching Kalman filter models, Technical Report 98-10, Compaq Cambridge Research Lab, 1998.

[Pavlovic and Rehg, 2000] V. Pavlovic and J. M. Rehg, Impact of dynamic model learning on classification of human motion, In *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 788–795, 2000.

[Rasmussen and Hager, 2001] C. Rasmussen and G. D. Hager, Probabilistic data association methods for tracking complex visual objects, *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 560–576, 2001.

[Rohr, 1994] K. Rohr, Towards model-based recognition of human movements in image sequences, *Computer Vision, Graphics, and Image Processing: Image Understanding*, vol. 59, pp. 94–115, 1994.

[Shumway and Stoffer, 1991] R. H. Shumway and D. S. Stoffer, Dynamic linear models with switching, *Journal of the American Statistical Association*, vol. 86, pp. 763–769, 1991.

[Tao et al., 2002] H. Tao, H. S. Sawhney, and R. Kumar, Object tracking with Bayesian estimation of dynamic Layer representations, *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 24, pp. 75–89, 2002.

[Wang and Adelson, 1994] J. Y. A. Wang and E. H. Adelson, Representing moving images with layers, *IEEE Trans. Image Processing*, vol. 3, pp. 625–637, 1994.