

Joint Region Tracking with Switching Hypothesized Measurements

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

Kia-Fock Loe
Dept. CS, National University of Singapore
loekf@comp.nus.edu.sg

Abstract

This paper proposes a switching hypothesized measurements (SHM) model supporting multimodal probability distributions and presents the application of the model in handling potential variability in visual environments when tracking multiple objects jointly. For a set of occlusion hypotheses, a frame is measured once under each hypothesis, resulting in a set of measurements at each time instant. A computationally efficient SHM filter is derived for online joint region tracking. Both occlusion relationships and states of the objects are recursively estimated from the history of hypothesized measurements. The reference image is updated adaptively to deal with appearance changes of the objects. The SHM model is generally applicable to various dynamic processes with multiple alternative measurement methods.

1. Introduction

Visual tracking is important in application areas including human-computer interaction, surveillance, and visual reconstruction. Tracking could be difficult due to the potential variability such as partial or full occlusions of objects, appearance changes caused by variation of object poses or illumination conditions, as well as distractions from background clutter.

The variability in visual environments usually results in a multimodal state space probability distribution. Thus, one principle challenge for visual tracking is to develop an accurate and effective model representation. The Kalman filter [3] [17], a classical choice employed in tracking work, is restricted to representing unimodal probability distributions. Joint probabilistic data association (JPDA) [2] and multiple hypothesis tracking (MHT) [5] techniques are able to represent multimodal distributions by constructing data association hypotheses. A measurement may either belong to a target or be a false alarm. The multiple hypotheses arise when there are more than one target and many measurements in the scene. Dynamic Bayesian networks (DBN) [8], especially switching linear dynamic systems (SLDS) [18] [19] and their equivalents [14] [22] have been used to track dynamic processes. Intuitively, a complex dynamic system is represented with a set of linear models controlled by a switching variable. Moreover, Monte

Carlo methods such as the Condensation algorithm [12] [15] 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 [4].

On the other hand, 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. Thus, the measurement process is another essential issue in visual tracking to deal with the potential variability. Parametric models can be used to describe appearance changes of target regions [10]. In the work of [6] and [7], adaptive or virtual snakes are used to resolve the occlusion. A joint measurement process for tracking multiple objects is described in [20]. Moreover, layered approach [13] [24] is an efficient way to represent multiple moving objects during visual tracking. A moving object is characterized by a coherent motion model over its support region.

The idea of hypothesized measurements, which results in a switching hypothesized measurements (SHM) model that differs from the above mentioned state space models, is proposed in this paper. The ability to support multimodality makes the model suitable for handling the potential variability in visual tracking. At each time instant, the approach acquires a set of hypothesized measurements for different occlusion hypotheses rather than uses a uniform measurement process. A computationally efficient filtering algorithm is derived for tracking multiple objects jointly. Both occlusion relationships and states of the objects are estimated from the history of hypothesized measurements. The proposed method helps prevent distractions from background clutter. When there is a high confidence in nonocclusion, the reference regions can be adaptively updated to deal with object appearance changes.

Previously, Ghahramani and Hinton introduced a DBN framework for learning and inference in switching state space models [9]. Pavlovic et al. proposed a SLDS approach for human motion analysis [18]. A switching model framework for the Condensation algorithm is also proposed by Isard and Blake [11]. In their work, the switching variable determines which dynamic model is in effect at each time instant. Rather than switches among a set of models, the SHM approach switches among a set of known hypothesized measurements. The JPDA algorithm [2] can be cast in the framework of SLDS as well.

Moreover, in our model each measurement component corresponds to one and only one given target region (see section 3).

Rasmussen and Hager describe a joint measurement process enumerating all possible occlusion relationships [20]. The measurement with respect to the most possible occlusion relationship is determined using the information from the current frame. The corresponding measurement is then plugged into a Kalman tracker. In our approach, the estimation is based on the history of all the (hypothesized) measurements. In the work of Galvin et al. [7], two virtual snakes, a background and a foreground snake for each object, are generated to resolve the occlusion when two objects intersect. Their manner parallels to the case of acquiring measurements under a set of two hypotheses in our method.

The remainder of the paper is arranged as follows: Section 2 presents the formulation of the SHM model. Section 3 proposes the measurement process for joint region tracking. Section 4 derives the filtering algorithm. Section 5 describes the implementation details. Section 6 discusses the experimental results. At the end, our technique is concluded in section 7.

2. Model

2.1. Generative SHM model

To model a dynamic system with state space representation, consider the evolution of a hidden state sequence $\{\mathbf{z}_k\}$ ($k \in \mathbf{N}$), given by

$$\mathbf{z}_{k+1} = \mathbf{f}_k(\mathbf{z}_k, \mathbf{n}_k), \quad (1)$$

where $\mathbf{f}_k: \mathbf{R}^{n_z} \times \mathbf{R}^{n_n} \rightarrow \mathbf{R}^{n_z}$ is a state transition function, and $\{\mathbf{n}_k\}$ is a process noise sequence. The objective of online tracking is to recursively estimate \mathbf{z}_k from a measurement sequence. In a complex system with dynamic mode control, there exists a mode or switching state sequence $\{s_k\}$, with $s_k \in \{1, 2, \dots, L\}$ ($L \in \mathbf{N}$). The switching state s_k determines which mode is in effect at time k . Usually the sequence $\{s_k\}$ is modeled as the outcome of an unobserved discrete first order Markov process.

Specifically, the mode switching is correlated with the measurement process in our work. The notion of a uniform measurement is extended to a set of L hypothesized measurements $\mathbf{y}_k = (\mathbf{y}_{k,1}, \mathbf{y}_{k,2}, \dots, \mathbf{y}_{k,L})$. Each $\mathbf{y}_{k,j}$ ($1 \leq j \leq L$) is called a hypothesized measurement since it is obtained by assuming that the switching state s_k is j at time k . For the measurement under the j th hypothesis,

$$\mathbf{y}_{k,j} = \mathbf{h}_{k,j}(s_k, \mathbf{z}_k, \mathbf{v}_{k,j}), \quad (2)$$

where $\mathbf{h}_{k,j}: \mathbf{N} \times \mathbf{R}^{n_z} \times \mathbf{R}^{n_v} \rightarrow \mathbf{R}^{n_y}$ is the measurement function, and $\mathbf{v}_{k,j}$ is the measurement noise under the j th hypothesis. To make the model computationally efficient, we assume that the hypothesized measurements are

conditionally independent on each other when both the hidden state \mathbf{z}_k and the switching state s_k are given. This switching hypothesized measurements (SHM) model can be represented by a dynamic Bayesian network shown in Figure 1.

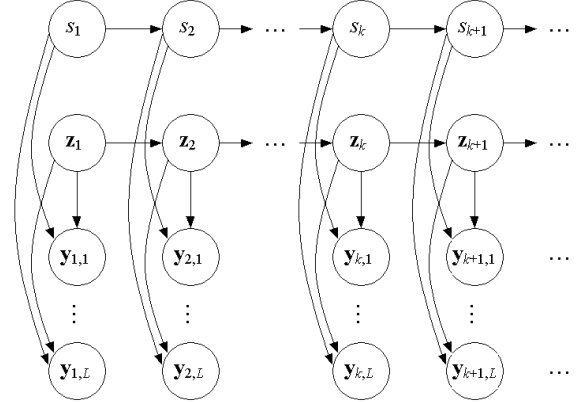


Figure 1. Bayesian network representation of the SHM model.

2.2. Example of hypothesized measurements

To illustrate the idea of hypothesized measurements in the SHM model, a simple example of the measurement process for jointly tracking a rectangle and a circle is studied in this section. The two objects translationally move in an image sequence $\{g_k\}$.

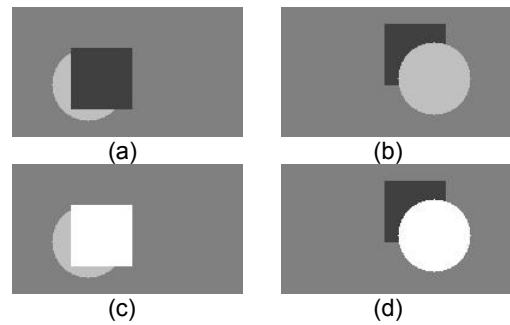


Figure 2. (a) (b) Two frames of the “rectangle and circle” sequence under different occlusion relationships. (c) Masked image under the first occlusion hypothesis. (d) Masked image under the second occlusion hypothesis.

When measuring the centroids of these two objects from the k th frame g_k , two occlusion relationship hypotheses (hypotheses corresponding to the rectangle being in front of the circle and the circle being in front of the rectangle, see Figure 2a and 2b) should be considered. The switching state s_k is introduced to describe the depth ordering at time k . 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 $\mathbf{y}_{k,j}$ ($1 \leq j \leq 2$) is written as $(\mathbf{y}_{k,j}^{(1)}, \mathbf{y}_{k,j}^{(2)})^T$, where $\mathbf{y}_{k,j}^{(1)}$ is the measurement of the

rectangle centroid, and $\mathbf{y}_{k,j}^{(2)}$ is the measurement of the circle centroid under the j th 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 (see Figure 2c). The occluded area of the circle is ignored and only the visible region is matched normally to get $\mathbf{y}_{k,1}^{(2)}$. Thus, the occlusion will not affect the measurement result. 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 (see Figure 2d) is used to measure $\mathbf{y}_{k,2}^{(1)}$.

Given the occlusion relationship s_k at time k , the hypothesized measurement $\mathbf{y}_{k,j}$ for $j \neq s_k$ may bias the true value since the measurement is obtained under a false hypothesis. Unfortunately, whether the rectangle occludes the circle or the circle occludes the rectangle is not given before hand. So it is not known whether $\mathbf{y}_{k,1}$ or $\mathbf{y}_{k,2}$ is the proper measurement for frame g_k . To handle this uncertainty, the occlusion relationship could be estimated from the history of all the hypothesized measurements.

Moreover, it is obvious that both hypothesized measurements support the condition of nonocclusion since different depth orderings of nonoverlapping objects are visually equivalent. The values of $p(s_k = 1)$ and $p(s_k = 2)$ should be equal in the case of nonocclusion.

2.3. 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 (or depth ordering) at time k , $s_k \in \{1, \dots, L\}$. The number of all occlusion relationship hypotheses is $L = M!$. The switching state transition probability is given as

$$p(s_{k+1} = i | s_k = j) = \alpha_{i,j}, \text{ with } \sum_i \alpha_{i,j} = 1. \quad (3)$$

The hidden state \mathbf{z}_k is denoted as $(\mathbf{z}_k^{(1)}, \mathbf{z}_k^{(2)}, \dots, \mathbf{z}_k^{(M)})^T$, with $\mathbf{z}_k^{(m)}$ ($1 \leq m \leq M$) being the state of the m th object (e.g. position and velocity) at time k . For a linear process with Gaussian noise, the hidden state transition function is

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{F}\mathbf{z}_k + \mathbf{n}, \\ p(\mathbf{z}_{k+1} | \mathbf{z}_k) &= N(\mathbf{z}_{k+1}; \mathbf{F}\mathbf{z}_k, \mathbf{Q}), \end{aligned} \quad (4)$$

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

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 true

value, while every other $\mathbf{y}_{k,j}$ for $j \neq s_k$ 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 function is simplified as

$$\begin{aligned} \mathbf{y}_{k,j} &= \begin{cases} \mathbf{H}\mathbf{z}_k + \mathbf{v}_{k,j}, & \text{if } j = s_k, \\ \mathbf{w}, & \text{otherwise,} \end{cases} \\ p(\mathbf{y}_{k,j} | s_k, \mathbf{z}_k) &= \begin{cases} N(\mathbf{y}_{k,j}; \mathbf{H}\mathbf{z}_k, \mathbf{R}_{k,j}), & \text{if } j = s_k, \\ \text{a constant,} & \text{otherwise,} \end{cases} \end{aligned} \quad (5)$$

where \mathbf{H} is the measurement matrix and $\mathbf{v}_{k,j}$ is a zero-mean Gaussian noise with covariance matrix $\mathbf{R}_{k,j}$. \mathbf{w} is a uniformly distributed noise, whose density is a small positive constant. For the measurement of M objects (e.g. translation), $\mathbf{y}_{k,j}$ is denoted as $(\mathbf{y}_{k,j}^{(1)}, \mathbf{y}_{k,j}^{(2)}, \dots, \mathbf{y}_{k,j}^{(M)})^T$, and $\mathbf{v}_{k,j}$ is written as $(\mathbf{v}_{k,j}^{(1)}, \mathbf{v}_{k,j}^{(2)}, \dots, \mathbf{v}_{k,j}^{(M)})^T$.

Combining with the conditional independence among the hypothesized measurements, we know that

$$\begin{aligned} p(\mathbf{y}_k | s_k = j, \mathbf{z}_k) &= p(\mathbf{y}_{k,1}, \mathbf{y}_{k,2}, \dots, \mathbf{y}_{k,L} | s_k = j, \mathbf{z}_k) \\ &= \prod_l p(\mathbf{y}_{k,l} | s_k = j, \mathbf{z}_k) \\ &= p(\mathbf{y}_{k,j} | s_k = j, \mathbf{z}_k) \prod_{l \neq j} p(\mathbf{y}_{k,l} | s_k = j, \mathbf{z}_k) \\ &\propto N(\mathbf{y}_{k,j}; \mathbf{H}\mathbf{z}_k, \mathbf{R}_{k,j}). \end{aligned} \quad (6)$$

3. Measurement

Multiple, occluding objects are modeled using layer representation. Layers are indexed by $m = 1, 2, \dots, M$, 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. Each depth ordering permutation is tagged with a index j ($1 \leq j \leq L$). For the example in section 2.2, it is known that $M = 2$ and $L = 2$.

Under each occlusion relationship 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. Thus occluded points are not matched when measuring the objects. Measurement results of nonoverlapping objects should be equivalent for different depth ordering permutations. During the measurement process, the motion of a point \mathbf{x} within the target region is described by a parametric model $\mathbf{d}(\boldsymbol{\theta}, \mathbf{x})$, with $\mathbf{d}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{x}$. $\boldsymbol{\theta} = (\theta^1, \theta^2, \dots, \theta^{n_\theta})$ is a set of motion parameters. The dimension of the motion vector $\boldsymbol{\theta}$, i.e. n_θ , changes under different motion models (two for the translational model, six for the affine model, and nine for the perspective model, for example). Under

the j th hypothesis, the measurement for the m th object $\mathbf{y}_{k,j}^{(m)}$ is denoted as $(y_{k,j}^{(m,1)}, y_{k,j}^{(m,2)}, \dots, y_{k,j}^{(m,n_\theta)})$. 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. The m th object is located at area D_m in the reference image. For each measured $\mathbf{y}_{k,j}^{(m)}$, $e_{k,j}^{(m)}$ is the corresponding minimum squared difference mean.

The measurement noise for the m th object $\mathbf{v}_{k,j}^{(m)}$ is denoted as $(v_{k,j}^{(m,1)}, v_{k,j}^{(m,2)}, \dots, v_{k,j}^{(m,n_\theta)})$ under the j th hypothesis. From appendix A, it can be known that

$$E[(v_{k,j}^{(m,i)})^2] \approx \frac{|D_m|}{\sum_{\mathbf{x} \in D_m} [g_r(\mathbf{d}(\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x})]^2} e_{k,j}^{(m)}, \quad (7)$$

where \mathbf{e}_i is the unit vector of dimension n_θ with a non-zero element in the i th position. To reduce the computation, it is assumed that the components of the measurement noise are uncorrelated to each other. Thus the diagonal matrix $\mathbf{R}_{k,j}$ can be easily computed from (7). Moreover, it should be noted that other measurement approaches (e.g. the snake methods in [6] and [7]) are also applicable for the SHM model.

4. Filtering

From a Bayesian perspective, the online tracking problem is to recursively calculate the posterior state space distribution. Given the measurement data $\mathbf{y}_{1:k} = \{\mathbf{y}_i\}_{1 \leq i \leq k}$ up to time k , the probability density function (pdf) $p(s_k, \mathbf{z}_k | \mathbf{y}_{1:k})$ is expressed as

$$\begin{aligned} p(s_k = j, \mathbf{z}_k | \mathbf{y}_{1:k}) &= p(s_k = j | \mathbf{y}_{1:k}) p(\mathbf{z}_k | s_k = j, \mathbf{y}_{1:k}) \\ &= \beta_{k,j} N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j}), \end{aligned} \quad (8)$$

where $p(s_k = j | \mathbf{y}_{1:k})$ is denoted as $\beta_{k,j}$, with $\sum_j \beta_{k,j} = 1$,

and the pdf $p(\mathbf{z}_k | s_k = j, \mathbf{y}_{1:k})$ is modeled as a normal distribution $N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j})$ under each switching state hypothesis. Hence $p(\mathbf{z}_k | \mathbf{y}_{1:k})$ 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 $\{\beta_{k,j}, \mathbf{m}_{k,j}, \mathbf{P}_{k,j}\}_{1 \leq j \leq L}$ to $\{\beta_{k+1,i}, \mathbf{m}_{k+1,i}, \mathbf{P}_{k+1,i}\}_{1 \leq i \leq L}$. From appendix B, the filtering algorithm is

$$\begin{aligned} \beta_{k+1,i} &= p(s_{k+1} = i | \mathbf{y}_{1:k+1}) \\ &= \frac{\sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{Hm}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}{\sum_i \sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{Hm}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}, \end{aligned} \quad (9)$$

$$p(\mathbf{z}_{k+1} | s_{k+1} = i, \mathbf{y}_{1:k+1}) \approx N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i}, \mathbf{P}_{k+1,i}), \quad (10)$$

where

$$\begin{aligned} \mathbf{m}_{k+1|k,j} &= \mathbf{Fm}_{k,j}, \\ \mathbf{P}_{k+1|k,j} &= \mathbf{FP}_{k,j} \mathbf{F}^T + \mathbf{Q}, \end{aligned}$$

$$\begin{aligned} \mathbf{S}_{k+1,i|j} &= \mathbf{HP}_{k+1|k,j} \mathbf{H}^T + \mathbf{R}_{k+1,i}, \\ \mathbf{K}_{k+1,i|j} &= \mathbf{P}_{k+1|k,j} \mathbf{H}^T \mathbf{S}_{k+1,i|j}^{-1}, \\ \mathbf{m}_{k+1,i|j} &= \mathbf{m}_{k+1|k,j} + \mathbf{K}_{k+1,i|j} (\mathbf{y}_{k+1,i} - \mathbf{Hm}_{k+1|k,j}), \\ \mathbf{P}_{k+1,i|j} &= \mathbf{P}_{k+1|k,j} - \mathbf{K}_{k+1,i|j} \mathbf{HP}_{k+1|k,j}, \\ \beta_{k+1,i|j} &= \frac{\alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{Hm}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}{\sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{Hm}_{k+1|k,j}, \mathbf{S}_{k+1,i|j})}, \\ \mathbf{m}_{k+1,i} &= \sum_j \beta_{k+1,i|j} \mathbf{m}_{k+1,i|j}, \\ \mathbf{P}_{k+1,i} &= \sum_j \beta_{k+1,i|j} [\mathbf{P}_{k+1,i|j} + \\ &\quad (\mathbf{m}_{k+1,i|j} - \mathbf{m}_{k+1,i})(\mathbf{m}_{k+1,i|j} - \mathbf{m}_{k+1,i})^T]. \end{aligned} \quad (11)$$

The state at time $k+1$ is estimated as

$$\begin{aligned} \hat{s}_{k+1} &= \arg \max_i p(s_{k+1} = i | \mathbf{y}_{1:k+1}) = \arg \max_i \beta_{k+1,i}, \\ \hat{\mathbf{z}}_{k+1} &= \arg \max_{\mathbf{z}_{k+1}} p(\mathbf{z}_{k+1} | s_{k+1} = \hat{s}_{k+1}, s_k = \hat{s}_k, \mathbf{y}_{1:k+1}) \\ &= \mathbf{m}_{k+1, \hat{s}_{k+1} | \hat{s}_k}. \end{aligned} \quad (12)$$

It can be seen that the computation of the SHM filter is slightly more complex than the computation of multiple Kalman filters (or Gaussian sum filters [1]).

5. Implementation

When an object is totally (or mostly) occluded by the other objects at time k , no (or few) points of the target region will be matched. The corresponding squared difference mean is computed as $\lambda_1 e_{k-1,j}^{(m)}$ for the m th object under the j th hypothesis, where λ_1 ($\lambda_1 > 1$) is a penalty term. The estimation is based on the result of time $k-1$ when no visible region of the object is expected at time k . The penalty λ_1 helps prevent interpreting an object as being completely occluded when there is image evidence for its visibility.

Due to the variation of the object poses and illumination conditions, the reference image should be updated throughout the tracking process to deal with the object appearance changes. Frame g_k can be used as the reference image when the following is satisfied.

$$\min_j \beta_{k,j} > \frac{\lambda_2}{L}, \quad \max_j \beta_{k,j} < \frac{1}{L\lambda_2}. \quad (13)$$

The value of λ_2 is a little bit smaller than one. From (13), it is known that the update is with a high confidence in nonocclusion.

The switching state transition probability is set as

$$\alpha_{i,j} = \begin{cases} 1 - \lambda_3, & \text{if } i = j, \\ \frac{\lambda_3}{L-1}, & \text{otherwise,} \end{cases} \quad (14)$$

where λ_3 is a small positive value so that two successive switching states are more likely to be of the same label. The transition matrix \mathbf{F} , covariance matrix \mathbf{Q} , and measurement matrix \mathbf{H} are defined in the same way as in a classical Kalman tracker with second order model [21]. The objects are assumed to be separated from each other in the initial image g_0 . At the beginning, the reference image is set as $g_r = g_0$. The target regions are detected from the initial image using an adaptive foreground detection method [23]. The initial β_{0j} , \mathbf{m}_{0j} , and \mathbf{P}_{0j} should be equal for different j because of nonocclusion. $\beta_{0j} = p(s_0 = j) = \frac{1}{L}$. According to the definition of the motion model \mathbf{d} , the initial mean \mathbf{m}_{0j} is set as a zero vector. The initial covariance matrix \mathbf{P}_{0j} is set as diagonal with small variances since the initialization is assumed to be accurate.

6. Results and discussion

The proposed approach is tested on both synthetic data and realistic data. The parameter values are set as $\lambda_1 = 1.1$, $\lambda_2 = 0.98$, and $\lambda_3 = 0.1$. Using a Pentium 4 1.4GHz PC, our C program can process 5 frames per second in the experiments.

Figure 3 shows quantitatively the results of jointly tracking a rectangle, a diamond, and a circle under noisy background in a synthetic image sequence of 200 frames. The state of the tracker is the position, diameter, orientation, and the velocities of these parameters. Each measurement is a translation, scaling, and rotation. Figure 3a shows the 10th, 70th, 90th, and 130th frame of the sequence. It could be seen that the circle is totally occluded in Figure 3a.3. Figure 3b shows the true horizontal trajectories of the three objects. Figure 3c and 3d demonstrate the tracking results of the SHM filter and the Kalman filter. Comparing with the Kalman filter, the tracking performance is greatly improved by our algorithm when heavy occlusions take place among the

three objects. The objects are correctly tracked even when total occlusion occurs.

Figure 4 shows the tracking of two hands as they cross twelve times in a realistic image sequence of 800 frames. The state of the tracker is the position and orientation, and the velocities of these parameters. Each measurement is a translation and rotation. Figure 4a shows the 30th, 65th, 165th, and 230th frame of the sequence. Appearance variation due to hand pose changes is obvious (see Figure 4a.4). Figure 4b and 4c demonstrate the tracking efficacy of the SHM filter versus the Kalman filter. The SHM filter successfully tracks both hands under different occlusion relationships (the left hand being in the front or the right hand being in the front). In Figure 4b, 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 4c.4, the distraction from background clutter causes the Kalman tracker to fail. The posterior distributions for the vertical position of the occluded hand in Figure 4a.3 and 4a.4 are shown in Figure 4d and 4e. When the occlusion is not severe, measurements under the two hypotheses are similar, and the distribution is unimodal (see Figure 4d). Under heavy occlusions, the distribution becomes multimodal (see Figure 4e) because the two hypothesized measurements turn to be different. The measurement under true hypothesis matches the hand correctly, while the measurement under false hypothesis is distracted by background clutter. Figure 4f shows the probabilities of the first occlusion hypothesis (the left hand being in the front) over the first 300 frames. The probabilities for the four frames shown in Figure 4a are circled in Figure 4f. The probabilities of the two hypotheses are equal in the nonoverlapping cases, while the probability of the true hypothesis becomes dominant under occlusions. As a byproduct of the SHM filter, the quantitative information helps update reference regions correctly to deal with the object appearance changes.

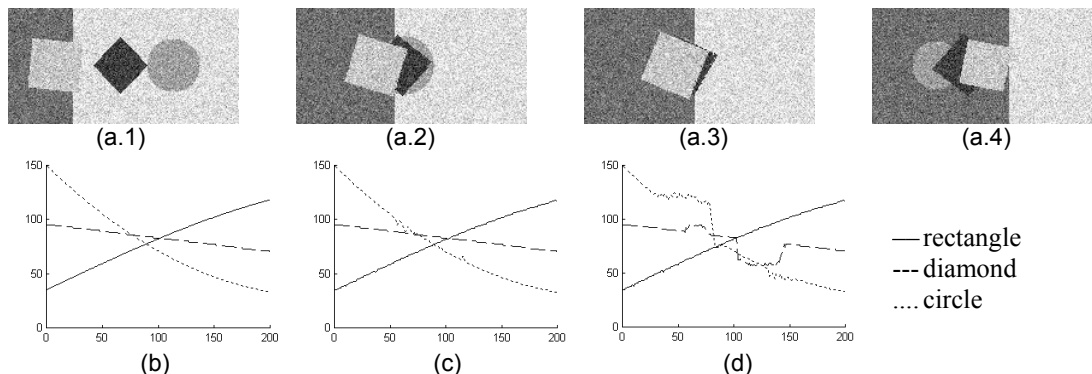


Figure 3. (a) Four frames of the “three objects” sequence. (b) True horizontal trajectories of the objects. (c) Tracking result of the SHM filter. (d) Tracking result of the Kalman filter.

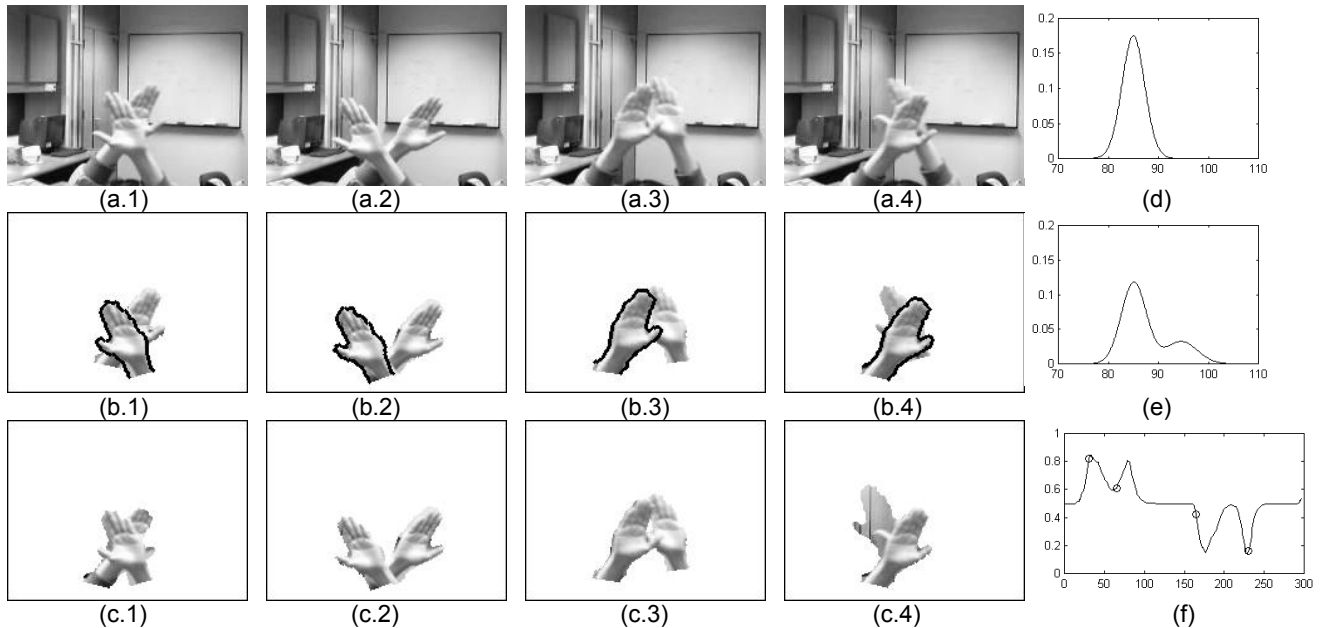


Figure 4. (a) Four frames of the “crossing hands” sequence. (b) Tracking results of the SHM filter. (c) Tracking results of the Kalman Filter. (d) (e) Posterior distributions of the left hand’s vertical position in (a.3) and (a.4). (f) Probabilities of the left hand being in the front over time.

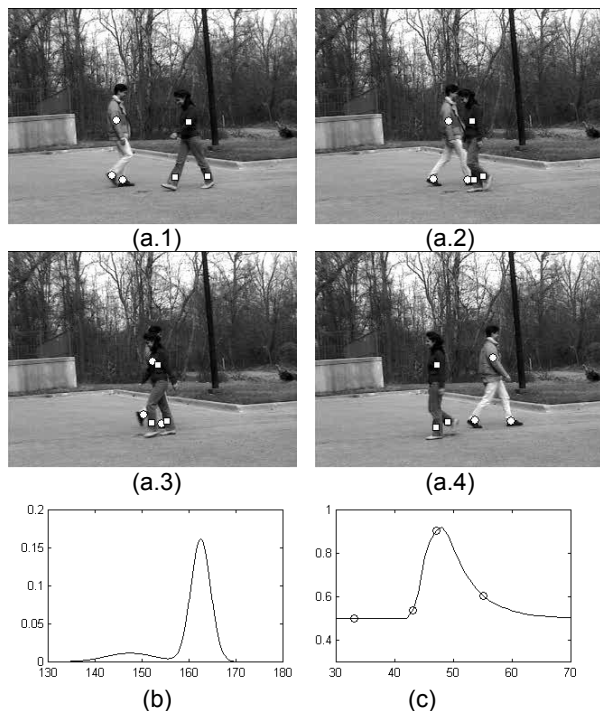


Figure 5. (a) Results of tracking the four shanks of two persons. (b) Posterior distribution of the occluded body’s horizontal position in (a.3). (c) Probabilities of the woman’s body being in the front over time.

Figure 5 shows the results of jointly tracking the four shanks of a man and a woman as they cross in a sequence of 80 frames. There should be totally $4! = 24$ hypotheses if we directly apply the SHM filter. 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 5a shows the tracking results for the 32nd, 42nd, 46th, and 54th frame of the sequence (circles are marked on the man’s body and shanks, and rectangles are marked on the woman). The man’s right shank has been totally occluded when they cross. Figure 5b shows the posterior distribution for the horizontal position of the occluded body in Figure 5a.3. Figure 5c shows the probabilities of the woman’s body being in the front. The probabilities for the four frames in Figure 5a are circled. The number of occlusion relationship hypotheses grows nonlinearly with the increase of objects. To reduce the computation, less plausible hypotheses should be (progressively) pruned when the number of the objects for joint tracking is large.

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 a smaller variance. Thus, the true information (the switching state and the hidden state) could be enhanced through the propagation. In addition, comparing with a uniform measurement process, the acquirement of multiple hypothesized measurements

helps decrease the information loss (e.g. caused by background clutter) in complex visual environments before filtering.

7. Conclusion

This paper makes two main contributions. First, we propose a switching hypothesized measurements model for multimodal state space representation of dynamic systems. Second, we describe a measurement process and derive an efficient filtering algorithm for joint region tracking in image sequences.

Our approach reasons about the occlusion relationships explicitly. The occlusion relationships are quantitatively estimated 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 [9] [11] [19]. Effective combining of these two ideas may result in a more powerful framework for visual tracking. Furthermore, from section 2.1 it can be known that the SHM model is generally applicable to describe various dynamic processes in which there are multiple alternative measurement methods.

Acknowledgement

The authors acknowledge Dr. Ismail Haritaoglu et al. for providing the test data on the website.

Appendix A

Using the first order Taylor expansion and ignoring the high order terms, we have that

$$|g_k(\mathbf{d}(\boldsymbol{\theta} + v\mathbf{e}_i, \mathbf{x})) - g_k(\mathbf{d}(\boldsymbol{\theta}, \mathbf{x}))| \propto |v|, \quad (15)$$

where v is a small random disturbance in the i th component of the motion vector $\boldsymbol{\theta}$. For the points within the m th object,

$$E[(g_k(\mathbf{d}(\boldsymbol{\theta} + v\mathbf{e}_i, \mathbf{x})) - g_k(\mathbf{d}(\boldsymbol{\theta}, \mathbf{x})))^2] = c^{(m,i)} E[v^2], \quad (16)$$

where $\mathbf{x} \in D_m$, and $c^{(m,i)}$ is the proportional factor. $c^{(m,i)}$ can be learned from the reference frame by substituting r for k , $\mathbf{0}$ for $\boldsymbol{\theta}$, and fixing the variable v as 1 in (16). Since $\mathbf{d}(\boldsymbol{\theta}, \mathbf{x}) = \mathbf{x}$,

$$\begin{aligned} c^{(m,i)} &= E[(g_r(\mathbf{d}(\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x}))^2] \\ &\approx \frac{1}{|D_m|} \sum_{\mathbf{x} \in D_m} [g_r(\mathbf{d}(\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x})]^2. \end{aligned} \quad (17)$$

From (17) we know that $c^{(m,i)}$ is computed as the mean of the squared intensity differences in the reference region.

If the hidden state \mathbf{z}_k is given, the true value of the motion parameters can be considered as $\mathbf{H}\mathbf{z}_k$ in our model.

Denote $(\mathbf{H}\mathbf{z}_k)^{(m)}$ as the true motion vector for the m th

object. Assume that the intensity distribution remains constant along a motion trajectory, $g_k(\mathbf{d}((\mathbf{H}\mathbf{z}_k)^{(m)}, \mathbf{x}))$ should equal $g_r(\mathbf{x})$ for a visible point of the m th object. Hence, variances of the measurement noise components can be estimated by substituting $(\mathbf{H}\mathbf{z}_k)^{(m)}$ for $\boldsymbol{\theta}$, and $v_{k,j}^{(m,i)}$ for v in (16). Combing with (5) under the j th hypothesis,

$$\begin{aligned} &E[(v_{k,j}^{(m,i)})^2] \\ &= \frac{1}{c^{(m,i)}} E[(g_k(\mathbf{d}((\mathbf{H}\mathbf{z}_k)^{(m)} + v_{k,j}^{(m,i)}\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x}))^2] \\ &\approx \frac{1}{c^{(m,i)}} E[(g_k(\mathbf{d}((\mathbf{H}\mathbf{z}_k)^{(m)} + \mathbf{v}_{k,j}^{(m)}, \mathbf{x})) - g_r(\mathbf{x}))^2] \\ &= \frac{1}{c^{(m,i)}} E[(g_k(\mathbf{d}(\mathbf{y}_{k,j}^{(m)}, \mathbf{x})) - g_r(\mathbf{x}))^2] \\ &\approx \frac{1}{c^{(m,i)}} e_{k,j}^{(m)} = \frac{|D_m|}{\sum_{\mathbf{x} \in D_m} [g_r(\mathbf{d}(\mathbf{e}_i, \mathbf{x})) - g_r(\mathbf{x})]^2} e_{k,j}^{(m)}. \end{aligned} \quad (18)$$

Appendix B

Using Bayes' rule, we know that

$$\begin{aligned} &p(s_{k+1}, \mathbf{z}_{k+1} | \mathbf{y}_{1:k+1}) \\ &= \frac{1}{p(\mathbf{y}_{k+1} | \mathbf{y}_{1:k})} p(\mathbf{y}_{k+1} | s_{k+1}, \mathbf{z}_{k+1}) p(s_{k+1}, \mathbf{z}_{k+1} | \mathbf{y}_{1:k}) \\ &\propto p(\mathbf{y}_{k+1} | s_{k+1}, \mathbf{z}_{k+1}) p(s_{k+1}, \mathbf{z}_{k+1} | \mathbf{y}_{1:k}). \end{aligned} \quad (19)$$

In principle, the filtering process has three stages: prediction, update, and collapsing.

With the transition probabilities in (3) and (4), the predictive distribution for time $k+1$ is computed as

$$\begin{aligned} &p(s_{k+1} = i, \mathbf{z}_{k+1} | \mathbf{y}_{1:k}) \\ &= \sum_j \int p(s_{k+1} = i, \mathbf{z}_{k+1} | s_k = j, \mathbf{z}_k) \cdot \\ &\quad p(s_k = j, \mathbf{z}_k | \mathbf{y}_{1:k}) d\mathbf{z}_k \\ &= \sum_j p(s_{k+1} = i | s_k = j) p(s_k = j | \mathbf{y}_{1:k}) \cdot \\ &\quad \int p(\mathbf{z}_{k+1} | \mathbf{z}_k) p(\mathbf{z}_k | s_k = j, \mathbf{y}_{1:k}) d\mathbf{z}_k \\ &= \sum_j \alpha_{i,j} \beta_{k,j} \int N(\mathbf{z}_{k+1}; \mathbf{F}\mathbf{z}_k, \mathbf{Q}) N(\mathbf{z}_k; \mathbf{m}_{k,j}, \mathbf{P}_{k,j}) d\mathbf{z}_k \\ &= \sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1|k,j}, \mathbf{P}_{k+1|k,j}). \end{aligned} \quad (20)$$

After receiving the measurement set \mathbf{y}_{k+1} at time $k+1$, the posterior density is updated as follows,

$$\begin{aligned} &p(s_{k+1} = i, \mathbf{z}_{k+1} | \mathbf{y}_{1:k+1}) \\ &\propto p(\mathbf{y}_{k+1} | s_{k+1} = i, \mathbf{z}_{k+1}) p(s_{k+1} = i, \mathbf{z}_{k+1} | \mathbf{y}_{1:k}) \\ &\propto \sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{z}_{k+1}, \mathbf{R}_{k+1,i}) \cdot \\ &\quad N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1|k,j}, \mathbf{P}_{k+1|k,j}). \end{aligned} \quad (21)$$

If the covariances in $\mathbf{P}_{k+1|k,j}$ are small [1], the product in (21) can be approximated by

$$\begin{aligned} & N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{z}_{k+1}, \mathbf{R}_{k+1,i}) N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1|k,j}, \mathbf{P}_{k+1|k,j}) \\ & \approx N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j}) N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i|j}, \mathbf{P}_{k+1,i|j}). \end{aligned} \quad (22)$$

The conditional probability of the switching state is updated as

$$\begin{aligned} \beta_{k+1,i} &= p(s_{k+1} = i | \mathbf{y}_{1:k+1}) \\ &= \int p(s_{k+1} = i, \mathbf{z}_{k+1} | \mathbf{y}_{1:k+1}) d\mathbf{z}_{k+1} \\ &\propto \sum_j \alpha_{i,j} \beta_{k,j} N(\mathbf{y}_{k+1,i}; \mathbf{H}\mathbf{m}_{k+1|k,j}, \mathbf{S}_{k+1,i|j}). \end{aligned} \quad (23)$$

Since $\sum_i \beta_{k+1,i} = 1$, (9) can be obtained by normalizing.

From (21) – (23), the pdf $p(\mathbf{z}_{k+1} | s_{k+1} = i, \mathbf{y}_{1:k+1})$ becomes a mixture of L Gaussians.

$$\begin{aligned} & p(\mathbf{z}_{k+1} | s_{k+1} = i, \mathbf{y}_{1:k+1}) \\ &= \sum_j \beta_{k+1,i|j} N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i|j}, \mathbf{P}_{k+1,i|j}). \end{aligned} \quad (24)$$

It could be derived that

$$\begin{aligned} & p(\mathbf{z}_{k+1} | s_{k+1} = i, s_k = j, \mathbf{y}_{1:k+1}) \\ &= N(\mathbf{z}_{k+1}; \mathbf{m}_{k+1,i|j}, \mathbf{P}_{k+1,i|j}). \end{aligned} \quad (25)$$

At time k , the distribution $p(\mathbf{z}_k | \mathbf{y}_{1:k})$ is represented as a mixture of L Gaussians, one for each hypothesis of s_k . Then each Gaussian is propagated through state transition, so that $p(\mathbf{z}_{k+1} | \mathbf{y}_{1:k+1})$ will be a mixture of L^2 Gaussians. The number of Gaussians grows exponentially with time. To deal with this problem, the mixture of Gaussians in (24) is collapsed to a single Gaussian in (10) using moment matching [16]. Collapsing is processed under each hypothesis of s_{k+1} . Therefore, the possibility of each hypothesis will not be cast throughout the propagation.

References

- [1] B. D. O. Anderson and J. B. Moore, *Optimal filtering*, Prentice-Hall, 1979.
- [2] Y. Bar-Shalom and T. E. Fortmann, *Tracking and data association*, Academic Press, 1988.
- [3] R. G. Brown, *Introduction to random signal analysis and Kalman filtering*, John Wiley & Sons, 1983.
- [4] T.-J. Cham and J. M. Rehg, "A multiple hypothesis approach to figure tracking," *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 239–245, 1999.
- [5] 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.
- [6] Y. Fu, A. T. Erdem, and A. M. Tekalp, "Tracking visible boundary of objects using occlusion adaptive motion snake," *IEEE Trans. Image Processing*, vol. 9, pp. 2051–2060, 2000.
- [7] B. Galvin, B. McCane, and K. Novins, "Virtual snakes for occlusion analysis," *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 294–299, 1999.
- [8] Z. Ghahramani, "Learning dynamic Bayesian networks," in *Adaptive processing of temporal information* (C. L. Giles and M. Gori, eds.), Lecture notes in artificial intelligence, pp. 168–197, Springer-Verlag, 1998.
- [9] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, pp. 963–996, 1998.
- [10] 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.
- [11] M. Isard and A. Blake, "A mixed-state Condensation tracker with automatic model-switching," *Proc. International Conf. Computer Vision*, pp. 107–112, 1998.
- [12] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," *Proc. European Conf. Computer Vision*, pp. 343–356, 1996.
- [13] N. Jojic and B. J. Frey, "Learning flexible sprites in video layers," *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 199–206, 2001.
- [14] C.-J. Kim, "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, pp. 1–22, 1994.
- [15] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *Proc. International Conf. Computer Vision*, vol. 1, pp. 572–578, 1999.
- [16] K. P. Murphy, "Learning switching Kalman filter models," Technical Report 98-10, Compaq Cambridge Research Lab, 1998.
- [17] H. T. Nguyen, M. Worrington, and R. van den Boomgaard, "Occlusion robust adaptive template tracking," *Proc. International Conf. Computer Vision*, vol. 1, pp. 678–683, 2001.
- [18] V. Pavlovic and J. M. Rehg, "Impact of dynamic model learning on classification of human motion," *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 788–795, 2000.
- [19] V. Pavlovic, J. M. Rehg, T.-J. Cham, and K. P. Murphy, "A dynamic Bayesian network approach to figure tracking using learned dynamic models," *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 94–101, 1999.
- [20] 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.
- [21] 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.
- [22] R. H. Shumway and D. S. Stoffer, "Dynamic linear models with switching," *Journal of the American Statistical Association*, vol. 86, pp. 763–769, 1991.
- [23] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 22, pp. 747–757, 2000.
- [24] 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.