Abstract—Occlusion reasoning is one of the most challenging issues in visual surveillance. In this letter, we propose a new approach for reasoning about occlusions between multiple people. In our approach, occlusion relationships between people are explicitly defined and deduction of the occlusion relationships is integrated into the whole tracking framework. The prior knowledge is supplied by a set of models which include a 2-D elliptical shape model, a spatial-color mixture of Gaussians appearance model, and a motion model with constant velocity. An observation likelihood function is constructed based on the similarity between the observations and the object appearance models with given states. The occlusion relationships are deduced from the current states of the objects and the current observations, using the observation likelihood function. The previous occlusion relationships are not required for deducing the current occlusion relationships. The problem of tracking and occlusion reasoning for more than two people is formulated mathematically, and a solution is proposed based on particle filtering. Experimental results on several real video sequences from indoor and outdoor scenes show the effectiveness of our approach.

Index Terms—Occlusion reasoning, particle filtering, tracking multiple people, visual surveillance.

I. INTRODUCTION

Occlusion reasoning for object tracking is one of the most challenging problems in visual surveillance. This is because during occlusion, only portions of each occluded object are visible and the correspondences between objects and their features become ambiguous [22]. The aim of occlusion reasoning for tracking is to explicitly determine the occlusion relationships between objects and then accurately localize the objects. As people activities are of key interest in monitored scenes, we focus on reasoning about occlusions between multiple people.

A. Related Work

There exist approaches for tracking multiple objects, in which the occlusion relationships are not explicitly defined and determined. For example, Bar-Shalon and Fortmann [23] use the joint probabilistic data association filter (JPDAF) to track multiple objects, by finding the exclusive correspondence between different objects. Rasmussen and Hager [11] propose a framework that describes data association to improve the performance of tracking multiple objects. The accurate localization of objects present to occlusions needs a very long running time. Khan et al. [16], [17] use a Markov random field to model motion interaction, and the motion model is naturally incorporated into a MCMC (Markov Chain Monte Carlo)-based particle filter for effective multiple object tracking. However, their model is used to describe the interactions between different objects rather than reason about occlusions.

Some approaches track multiple people through occlusions using 3-D calibration information. For instance, Isard and MacCormick [21], and Zhao and Nevatia [18], [19] investigate tracking in calibrated images, where calibration information is utilized to reason about occlusions. Mittal and Davis [20] reason about occlusions using multiple calibrated cameras. As camera calibration is costly, we only consider un-calibrated occlusion reasoning in this letter.

The 2-D un-calibrated people tracking approaches base occlusion reasoning on people appearance models. Non-parametric people appearance models including appearance template-based [1], [7], [12], color histogram-based [2], [13]–[15], and kernel density-based [3], [4], and parametric appearance models [5], [6] are used to track multiple people through occlusions. For instance, Senior et al. [1], [12] use an appearance template to record the RGB color of each pixel of each person. Each appearance template is associated with a set of probability masks, each of which represents the probability of the corresponding pixel belonging to that template. Before the occurrence of an occlusion, the template is continuously updated on each new image as it arrives. During occlusions, foreground pixels are segmented into different regions corresponding to different templates according to the learned templates and the predicted people positions. People are then tracked according to the segmented regions. Wu et al. [7] apply the Bayesian network to track two faces through occlusions using multi-cameras, with the participation of the appearance templates. They put the occlusion relationships between objects into the object states, and thereby combine prior expectations of occlusion with the likelihood function. They assume that the occlusion relationships between objects are independent of the current states of the objects, and dependent on the previous occlusion relationships. Qu et al. [15] propose a distributed Bayesian formulation using multiple interactive trackers that require low complexity for real time tracking applications. They extend the conventional Bayesian tracking framework by modeling object interaction in terms of potential functions. They combine the color histogram model and the PCA-based mode to efficiently model object appearance.

B. Our Work

Considering that most current methods [1]–[6] cast occlusion reasoning as a classical segmentation problem, i.e., classify foreground pixels into several sub-regions according to prior knowledge and then track people based on the segmented...
regions, in this letter we omit the process of segmenting the motion region into sub-regions corresponding to different people, and propose a new approach for occlusion reasoning between multiple people. In the approach occlusion relationships between people are explicitly defined and deduced, the models for the different people are matched against the image in the order consistent with the occlusion relationships, and then the localizations of the people are optimized. The main contributions of our approach are:

- Tracking and occlusion reasoning for more than two people are formulated mathematically. The occlusion relationships between multiple people are incorporated into a tracking framework in which the main focus is on finding the positions of the people.
- The occlusion relationships are deduced from the current states of the people and the current observations, with the participation of the observation likelihood function. This is in contrast with previous work such as [7] in which it is assumed that the occlusion relationships between objects are independent of the current states of the objects, and dependent only on the previous occlusion relationships.

To implement these contributions, we apply the ellipse shape model, the spatial-color mixture of Gaussians appearance model [24], and the motion model with constant velocity. An observation likelihood function is constructed based on the appearance model used. A standard particle filter is adopted to optimize object states and occlusion relationships.

II. OVERVIEW OF OUR APPROACH

In our approach, pixels are classified as foreground/background for each frame using background subtraction, without segmenting the foreground pixels into objects. Motion regions are extracted from foreground pixels. The extracted motion regions in the current frame are compared with the motion regions in the previous frame according to their positions and appearances. If the current frame contains a motion region which does not correspond to any motion region in the previous frame, then a new object is detected as a tracked object and the appearance model for the object is initialized according to the new motion region. If a motion region gradually becomes smaller to the point where it can be ignored, then object disappearance occurs and the tracking of the object terminates. If two or more objects come into the scene with one occluding the other, then they are tracked as a single object. If it is detected that a motion region tracked as an object in the previous frame has divided into some smaller motion regions in the current frame, then the objects are tracked separately: pixels in these smaller motion regions are used to initialize the models of these objects. The appearance model for each person is updated in each of the frames in which the person is absent from any occlusion. If it is detected that a person is present to occlusion, the appearance model of the person is kept constant. The motion model is used to predict the states of the people at time \( t \) according to the tracking results at time \( t - 1 \). The occlusion relationships between the people are deduced according to the predicted states, the observations, and their shape and appearance models. The human shape model and the appearance model with predicted states and deduced occlusion relationships are projected to the image plane, and the observation likelihood is used to further update the states of the objects and their occlusion relationships for maximizing the posterior probability of the states and the occlusion relationships given the observations.

We consider occlusions between \( n \) objects. We denote the state of object \( k \) at time \( t \) by \( S^k_t \) (\( k = 1, 2, \ldots, n \)). The observation image (contained in the foreground pixels) at time \( t \) is denoted by \( O_t \), with history \( O_{t-1} \cdots O_0 O_1 \). The occlusion relationship \( \pi_t(i,j) \) between objects \( i \) and \( j \) at time \( t \) is represented as

\[
\pi_t(i,j) = \begin{cases} 
-1, & \text{if } i \text{ occludes } j \\
0, & \text{if } \text{no occlusion} \\
1, & \text{if } j \text{ occludes } i. 
\end{cases}
\]

So, the object state vector is \( (S^1_t, \ldots, S^n_t) \), and the occlusion relationship vector is \( (\pi_t(1,2), \ldots, \pi_t(n-1,n)) \) (\( 1 \leq i < j \leq n \)) which is simplified with \( (\pi_t(1,2), \ldots, \pi_t(n-1,n)) \). The state \( S^k_t \) is predicted using probability \( p(S^k_{t+1} | S^k_t, O_{t}, \pi_t(1,2), \ldots, \pi_t(n-1,n)) \) and the occlusion relationships \( (\pi_t(1,2), \ldots, \pi_t(n-1,n)) \) are deduced according to the predicted states \( (S_{t+1}^1, \ldots, S_{t+1}^n) \), with the participation of the shape and appearance models of the people. Then, the observation likelihood is used to further update the state vector \( (S^1_t, \ldots, S^n_t) \) and the occlusion relationship vector \( (\pi_t(1,2), \ldots, \pi_t(n-1,n)) \) for maximizing the posterior probability \( p(S^1_t, \ldots, S^n_t, \pi_t(1,2), \ldots, \pi_t(n-1,n) | O_{t-1}, O_t) \). To estimate \( p(S^1_t, \ldots, S^n_t, \pi_t(1,2), \ldots, \pi_t(n-1,n)) \), we should construct proper human body models and a proper observation function.

III. HUMAN BODY MODELS

Human body models include the shape model, the appearance model and the motion model.

A. Shape Model

There are many different shape models for the human body. The more parameters there are in a shape model, the more accurate the tracking results, but the more expensive the computation. The aim in this letter is to reason about the occlusions between multiple people. We do not consider tracking the articulated motions of a human body and do not estimate the pose of each part of a human body. Therefore, we select a simple elliptical model involving only a few parameters, to simplify the computation. As shown in Fig. 1, a human body is modeled as a 2-D vertical ellipse with a parameter vector \((x, y, h, \psi)\), where \((x, y)\) is the center of the ellipse, \(h\) is the length of the major axis of the ellipse, and \(\psi\) is the eccentricity of the ellipse. For the tracking, the state of the ellipse is taken to be \(S = (x, y, h, \psi)\) which is an example of a state introduced in Section II.

B. Appearance Model

In this letter, we use the spatial-color mixture of Gaussians appearance model in [24] to model the spatial layout of the colors in a person in an image. Let \(\mu_j^a\) and \(\Sigma_j^a\) be the mean and
the covariance of the color feature in the $j$th mode of the mixture of Gaussians, and $\mu_j^c$ and $\Sigma_j^c$ be the mean and the covariance of the spatial feature. Let $\eta(x^c, \mu_j^c, \Sigma_j^c)$ be the probability density function value of the color $x^c$ of a pixel $x$ conditional on the Gaussian $N(\mu_j^c, \Sigma_j^c)$, and $\eta(x^s, \mu_j^s, \Sigma_j^s)$ be the probability density function value of the position $x^s$ of the pixel $x$ conditional on the Gaussian $N(\mu_j^s, \Sigma_j^s)$. Then, the probability density function value $P_x$ of the pixel $x$ under the spatial-color mixture of Gaussians appearance model is:

$$P_x = \sum_{j=1}^{\gamma} \omega_j \eta(x^c, \mu_j^c, \Sigma_j^c) \eta(x^s, \mu_j^s, \Sigma_j^s) \sum_j$$

where $\gamma$ is the number of Gaussians and $\omega_j$ is the normalized weight of the $j$th mode of the mixture of Gaussians. The K-means clustering algorithm is used to initialize the parameters of the mixture of Gaussians and the Expectation-Maximization (EM) algorithm is used to refine these parameters. Please see [24] for detail.

The spatial-color mixture of Gaussians appearance model improves the popular color histogram-based appearance model because it considers not only the colors in a region but also the spatial layout of the colors.

### C. Motion Model

Application of motion models, in the form of prior knowledge, to object tracking can reduce the search space and make the tracking results more accurate. As the changes in motion velocity and object sizes between successive frames are very small, for simplicity we assume a constant velocity and a constant height and eccentricity of the ellipse shape model. The motion model is described by the function below:

$$\begin{cases} x_{t+1} = x_{t-1} + (x_{t-1} - x_{t-2}) \\ y_{t+1} = y_{t-1} + (y_{t-1} - y_{t-2}) \\ h_{t+1} = h_{t-1} \\ \rho_{t+1} = \psi_{t-1} \end{cases}$$

where $x_t$ and $y_t$ are the $x$ and $y$ coordinates of the position of a person, and $h_t$ and $\psi_t$ are the height and eccentricity of the ellipse corresponding to the person in the image at time $t$.

### IV. Observation Likelihood

The observation likelihood function $p\left(O_t|S_t^1 \ldots S_t^n, \pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)}\right)$ represents the probability of an observed image $O_t$ (contained in the foreground pixels), given states $\{S_t^k, 1 \leq k \leq n\}$ ($S_t^k = (x_t^k, y_t^k, h_t^k, \rho_t^k)$) of objects and their occlusion relationships $\{\pi_t^{(i,j)}, 1 \leq i < j \leq n\}$. The observation likelihood depends on the similarity [8] between the observations and the appearance reflected by the appearance models of objects, the shape models with given states, and the occlusion relationships. The similarity is measured according to the comparison between the appearance model of each person and the observation specified by the shape model of the person with given states and occlusion relationships.

We first determine the corresponding observation of an object $k$ under its given states and occlusion relationships. The given occlusion relationships $\{\pi_t^{(1,k)}, \ldots, \pi_t^{(k-1,k)}, \pi_t^{(k,k+1)}, \ldots, \pi_t^{(n,n)}\}$ involving object $k$ specify all objects occluding object $k$. Let $R^k_t$ be the region of foreground pixels within the ellipse of object $i$ with the given state at time $t$. If object $k$ is not occluded, its corresponding observation $V^k_t$ is $R^k_t$; and if object $k$ is occluded by $c$ objects which are indexed by $I_1, I_2, \ldots, I_c$, its corresponding observation is:

$$V^k_t = R^k_t - \left(\bigcup_{i=1}^{c} R^i_t\right).$$

Based on the determined corresponding observation $V^k_t$ of object $k$, we estimate the likelihood of the pixels in $V^k_t$ given the spatial-color mixture of Gaussians appearance model $M_k$ of object $k$. Let $x$ represent a pixel in $V^k_t$, let $|V^k_t|$ be the number of pixels in $V^k_t$, and let $p(x|M_k)$ be the probability density function value of pixel $x$ given $M_k$. The likelihood $P \left(V^k_t|M_k\right)$ of the pixels in $V^k_t$ given the model $M_k$ is defined as:

$$P \left(V^k_t|M_k\right) = \exp \left\{ \frac{1}{|V^k_t|} \sum_{x \in V^k_t} \log p(x|M_k) \right\}.$$  

The larger the $P \left(V^k_t|M_k\right)$, the more the compatibility between the pixels in $V^k_t$ and the model $M_k$. The log function and the exp function in (5) are used to transform the product of probability density function values to sum of the logarithmic values of the probability density function to avoid the underflow produced by the product of probability density function values.

As the motions of objects are independent, the similarity between the appearance models of all people and the observations is measured by the sum of the similarities between the appearance model of each person and the corresponding observation of the person. Then, the observation likelihood function is defined as:

$$p\left(O_t|S_t^1 \ldots S_t^n, \pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)}\right) \propto \sum_{k=1}^{n} P \left(V^k_t|M_k\right).$$

where $n$ is the number of people in the scene at time $t$. 

---

**Fig. 1. Elliptical model.**
V. OCCLUSION RELATIONSHIP DEDUCTION

We found that, given the states of the objects, their occlusion relationships are fixed. So the occlusion relationships between objects are dependent on the current states of the objects, and independent of their previous occlusion relationships. The transition of occlusion relationships [7] from the previous step to the current step is not needed. It follows that we use the likelihood function \( p\left(O_t|\pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)}, S_t^1 \ldots S_t^n\right) \) whose definition involves the current observations and the shape and appearance models of the people, to estimate the occlusion relationships. The optimal occlusion relationships maximize the likelihood function given the states of the people

\[
\pi_t^k = \arg\max_{\pi_t^k} \max\left(p\left(O_t|\pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)}, S_t^1 \ldots S_t^n\right)\right)
\]  

(7)
i.e., the optimal type of occlusion relationships is selected from the combinations of occlusion relationships between the people, to maximize the value of the likelihood function given the states of the people.

In implementing the deduction of occlusion relationships between people, it is checked whether the ellipses specified by the states of the people intersect or overlap. Only for the people whose ellipses intersect or overlap, the observation likelihood function is used to determine “who occludes whom”. So when there are a large number of people in the scene, but most of them are absent from occlusions, only the combinations of occlusion relationships between the people present to occlusions are mainly considered. Then, the computational complexity is decreased.

VI. TRACKING WITH OCCLUSION REASONING

The process of tracking with occlusion reasoning is to maximize the posterior probability \( p\left(S_t^1 \ldots S_t^n, \pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)}|O_t \ldots O_t\right) \) according to the motion model \( p\left(S_t^k|S_{t-1}^k\right) \) (\( k = 1, \ldots, n \)) and the observation likelihood \( p\left(O_t|S_t^1 \ldots S_t^n, \pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)}\right) \).

Particle filtering is a simple and effective method for computing posterior probabilities from non-Gaussian, non-linear, and high-dimensional observation data. In this letter, a standard particle filtering approach [7], [9] is applied to find the maximum of the posterior probability density function. The posterior density \( p\left(S_t^1 \ldots S_t^n, \pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)}|O_t \ldots O_t\right) \) is represented by \( M \) particles \( \left\{S_t^{1(m)} \ldots S_t^{n(m)} \right\} \) (\( m = 1, \ldots, M \)) where \( S_t^{i(m)} \) corresponds to the state of object \( i \), deduced occlusion relationships \( \left\{\pi_t^{(1,2)(m)} \ldots \pi_t^{(n-1,n)(m)}\right\} \), and the weights \( \bar{W}_t^{(m)} \) of the particles. The propagation of the filter is outlined as follows.

1) Initialization: \( t = 0 \). Particles \( \left\{S_0^{1(m)} \ldots S_0^{n(m)}\right\} \) are sampled according to the prior probability distribution \( p\left(S_0^{1(m)} \ldots S_0^{n(m)}\right) \) which is assumed to be the uniform distribution. The ranges for \( x, y, h, \) and \( \psi \) in each object are \([x-\Delta x, x+\Delta x], [y-\Delta y, y+\Delta y], [h-\Delta h, h+\Delta h], \) and \([\psi-\Delta \psi, \psi+\Delta \psi]\), respectively. The occlusion relationships \( \left\{\pi_0^{(1,2)}(m) \ldots \pi_0^{(n-1,n)}(m)\right\} \) are deduced. The weights \( \left\{\bar{W}_0^{(m)}\right\} \) are computed by the observation likelihood:

\[
\bar{W}_t^{(m)} = p\left(O_0|S_0^{1(m)} \ldots S_0^{n(m)}, \pi_0^{(1,2)(m)} \ldots \pi_0^{(n-1,n)(m)}\right).
\]

2) \( t = 1, 2, \ldots \)

a) Re-sampling: In order to avoid degeneration of particles, particles \( \left\{S_t^{1(m)} \ldots S_t^{n(m)}\right\} (m = 1, \ldots, M) \) at time \( t-1 \) are re-sampled to produce a new particle set \( \left\{S_t^{1(m)} \ldots S_t^{n(m)}\right\} \) according to \( \left\{W_t^{(m)}\right\} \). In this process, particles with large weights are sampled several times, and particles with small weights are ignored. The total number of particles is kept constant.

b) Prediction: Particles \( \left\{S_t^{1(m)} \ldots S_t^{n(m)}\right\} \) at time \( t \) are generated according to the prediction model \( p\left(S_t^k|S_{t-1}^k\right) \) (\( k = 1, 2, \ldots, n \)) which is described by (3).

c) Deduction: The corresponding occlusion relationships \( \left\{\pi_t^{(1,2)(m)} \ldots \pi_t^{(n-1,n)(m)}\right\} \), given the states \( \left\{S_t^{1(m)} \ldots S_t^{n(m)}\right\} \), are deduced according to \( p\left(O_t|S_t^{1(m)} \ldots S_t^{n(m)}, \pi_t^{(1,2)(m)} \ldots \pi_t^{(n-1,n)(m)}\right) \), as described in Section V.

d) Measurement: Each particle is re-weighted by the observation likelihood

\[
\bar{W}_t^{(m)} = p\left(O_t|S_t^{1(m)} \ldots S_t^{n(m)}, \pi_t^{(1,1)(m)} \ldots \pi_t^{(n-1,n)(m)}\right).
\]

Then, the weights are normalized

\[
\bar{W}_t^{(m)} = \frac{\bar{W}_t^{(m)}}{\sum_{l=1}^{M} \bar{W}_l^{(l)}}
\]

(9)
e) Estimation: The states of objects at time \( t \) are estimated as the summation of the weighted particles:

\[
S_t^k = \sum_{m=1}^{M} \bar{S}_t^k \bar{W}_t^{(m)}, \quad k = 1, 2, \ldots, n.
\]

(10)

Then the occlusion relationships at time \( t \) are deduced according to the estimated object states.

In our approach, the particles are defined in the joint state space of all objects. If a new object enters the scene, then the state components of the object are added to the joint state space and the particles are updated. The particle components of the object in all the particles are sampled initially according to the prior probability distribution which is assumed to be the uniform distribution. For example, if an object indexed by \( n+1 \) enters the scene, particle components \( \left\{S_t^{n+1(m)}\right\} \) of object \( n+1 \) are sampled and then added to the previous particles \( \left\{S_t^{1(m)} \ldots S_t^{n(m)}\right\} \) (\( m = 1, \ldots, M \))—the particles become \( \left\{S_t^{1(m)} \ldots S_t^{n(m)}, S_t^{n+1(m)}\right\} \). If an object leaves the scene, the state components of the object are
removed from the joint state space (the dimension of the state space is reduced), and the particle components of the object are removed from all the particles.

For example, if object \( i \) leaves the scene, particles \( \{S_i^1(m), S_i^2(m), S_i^3(m), \ldots, S_i^n(m)\} \) for \( m = 1, \ldots, M \) are changed to be \( \{S_i^1(m), S_i^2(m), S_i^3(m), \ldots, S_i^m(m)\} \).

To adapt different parameters and increase the robustness and the effectiveness of the inference component, a hierarchical strategy is used to estimate the states of objects. Firstly, the size of the ellipse of each person is set equal to be the size of the ellipse of the person in the previous frame. The particle filtering process is used to estimate the optimal positions of the objects. Then, the positions are fixed and the particle filtering process is used again to estimate the optimal sizes of the objects.

### VII. Experimental Results

The above algorithm is implemented using MatLab on the Windows XP platform. Computation time is measured on a P4-2.8G computer with 512M RAM. A simple background subtraction algorithm [10] is applied to extract the foreground pixels. The success of tracking and occlusion reasoning is demonstrated using the recovered occlusion relationship diagram whose \( x \)-coordinate is the frame number, and \( y \)-coordinate is the occlusion relationship. In the following, we demonstrate four examples of tracking with occlusion reasoning, and then illustrate some comparison results.

#### A. Example 1

The test video for this example is captured in an outdoor environment. Its resolution is 320×240. Fig. 2 contains some key frames illustrating the tracking of two people through occlusion. The recovered occlusion relationship is shown in Fig. 3. Before Frame 741, there was no occlusion between the two people, and thus the occlusion relationship equals to 0. From Frame 741 to Frame 760, the front person (Person A) occluded the back person (Person B), and thus the occlusion relationship equals to −1. The two people separated after Frame 760, so the occlusion relationship then equals to 0.

#### B. Example 2

The test video for this example is captured in an outdoor environment. Its resolution is 320×240. Fig. 4 contains some frames illustrating the results of tracking from Frame 390 to Frame 458. In this portion of the video sequence, People A entered the scene from the right and walked to left; Person B entered the scene from the left and walked to right; and Person C entered the scene from the right and walked behind Person A. From Frame 411, Person B occluded Person A. From Frame 421, Person B occluded Person C. As Person C walked more quickly than Person A, Person C caught up with Person A from Frame 426, and then Person C occluded Person A. From Frames 426 to Frame 429, there is mutual occlusion between the three people, i.e., Person B occluded people A and C, and Person C occluded Person B and C. From Frame 429 to Frame 445, there is mutual occlusion between the three people, i.e., Person B occluded people A and C, and Person C occluded Person A. From Frames 445 to Frame 458, there is mutual occlusion between the three people, i.e., Person B occluded people A and C, and Person C occluded Person A. Fig. 5(a)–(c) show respectively the recovered occlusion relationships between People A and B, People A and C, and People B and C. The results show that all the three people are successfully tracked, and their occlusion relationships are deduced correctly. In this example, the three people occlude each other for a long time. The example shows that if there are no large...
changes in people’s appearances, then our approach can handle long-term multiple occlusions.

C. Example 3

This is an example in which a person appeared within the image. The test video for this example is in the PETS 2004 database which is an open database for the research on visual surveillance. The video is available on http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ShopAssistant2Cor.avi. Its resolution is 384×288. It shows a shopping center. Fig. 6 shows some frames of the results of tracking people in the video. Before Frame 201, three people are tracked. Two of them occluded each other in the midsection of the scene and one of them moved at the right-bottom corner of the scene. From Frame 201, a person entered the scene by emerging from a door placed away from the edge of the image. Tracking of the person is successfully initialized. From Frame 314, the person who moved at the right-bottom corner of the scene left the scene. For the video, all the four people are accurately tracked and their occlusion relationships are correctly determined.

D. Example 4

This video is taken from the same scene as Example 3. It is available on http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. In the video, there are many people walking or standing and talking. Fig. 7 illustrates some frames of tracking results when large mutual occlusions between five people occur from Frame 200 to Frame 400. During these occlusions, these people are still tracked successfully and their occlusions are deduced correctly using our approach.

We contrast the tracking results of our approach with the ground truth which is labeled by the suppliers of the PETS 2004 database: the pixel position of the center of the ellipse of each person is contrasted with the center of the manually labeled image bounding box of the person. In the following, the contrast results for three people occluded extensively are shown. Fig. 8 shows the trajectory of Person A who is the first person counting from left to right in Frame 200, where the dashed line is the ground truth trajectory, and the solid line is the trajectory obtained by our approach. From Frame 200 to Frame 400, for Person A the average warp between the trajectory obtained by our approach and the ground truth is 9.2 pixels. This warp is comparatively large. The reason is that the lower part of this person is occluded in most of the time that the person appeared in the scene, and therefore the height parameter of the ellipse of the person cannot be updated adaptively. Fig. 9 shows the trajectory of Person D who is the fourth person counting from left to right in Frame 200. From Frame 200 to Frame 400, for Person D the average warp between the trajectory obtained by our approach and the ground truth is 5.0 pixels. Fig. 10 shows the trajectory of Person E who is the fifth person counting from left to right in Frame 200. From Frame 200 to Frame 400, for Person E the average warp between the trajectory obtained by our approach and the ground truth is 4.9 pixels.
To evaluate the performance of our approach, we compare our approach with the one in [1], [12]. The appearance template-based occlusion reasoning approach in [1], [12] is a typical segmentation-based one, which has high reliability in segmenting foreground pixels, but needs a large area of memory to store the template.

Example 1 is used to compare the performance of our approach with the performance of the segmentation-based approach. Fig. 11 illustrates the tracking results obtained by the approach in [1], [12], corresponding to Example 1. From Frame 751 to Frame 754, the occluded person is not correctly tracked. From Frame 755 to Frame 757 the occluded person is lost in tracking.

Fig. 12 compares the ellipse centers for our approach with those for the segmentation-based approach. The horizontal and vertical coordinates in the image are respectively the x-coordinate and y-coordinate of the center of each ellipse. The trajectory curves of the un-occluded person (Person A) are shown in (a) and the trajectory curves of the occluded person (Person B) are shown in (b). From Fig. 12, we can see that the curve produced by the segmentation-based approach is farther from the ground truth than the one produced by our approach. This is more noticeable for the occluded person as shown in (b).

Table I shows the quantitative comparison between our approach and the segmentation-based approach for Example 1 with respect to the following quantitative criteria:

- the number of frames in which tracking was successful, corresponding to the tracking success rate;
- the average position error—the average pixel distance between the estimated ellipse center and the center of the corresponding ground truth bounding box;
- the average size error—the average warp between the estimated ellipse size and the size of the corresponding ground truth bounding box;
- runtime per frame.

From this table, it can be seen that position errors and size errors of tracking the two people using our approach are less than those using the segmentation-based approach. However, the runtime of our approach is more than that of the segmentation-based approach. This is due to the time taken to estimate the parameters of the spatial-color mixture of Gaussians appearance model.

### Table I

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Segmentation-based</th>
<th>Our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of frames in which tracking was successful</td>
<td>Person A</td>
<td>147/147</td>
</tr>
<tr>
<td></td>
<td>Person B</td>
<td>140/147</td>
</tr>
<tr>
<td>Position error (pixels)</td>
<td>Person A</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td>Person B</td>
<td>7.7</td>
</tr>
<tr>
<td>Size error (pixels)</td>
<td>Person A</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>Person B</td>
<td>5.4</td>
</tr>
<tr>
<td>Runtime (seconds per frame)</td>
<td></td>
<td>26.3s</td>
</tr>
</tbody>
</table>

### VIII. CONCLUSION

In this letter, we have presented a new approach for occlusion reasoning for tracking multiple people. In our approach, occlusion reasoning is integrated into the whole process of tracking, in order to obtain better tracking accuracy during occlusions. Prior knowledge is represented by a set of models: a 2-D elliptical shape model, a spatial-color mixture of Gaussians appearance model, and a motion model with constant velocity. We have constructed an observation likelihood based on the similarity between the appearance models of the people and the observations. The occlusion relationships are deduced from the current states of the objects and the current observations. The previous occlusion relationships are not required for deducing
the current occlusion relationships. Tracking and occlusion reasoning for more than two people are formulated in the particle filtering framework, and a solution is proposed. Experimental results on several real video sequences from indoor and outdoor scenes have demonstrated the effectiveness of our approach.

ACKNOWLEDGMENT

The authors thank X. Zhang and M. Liu for their help for implementing some experiments in this letter.

REFERENCES

Occlusion Reasoning for Tracking Multiple People

Weiming Hu, Xue Zhou, Min Hu, and Steve Maybank, Senior Member, IEEE

Abstract—Occlusion reasoning is one of the most challenging issues in visual surveillance. In this letter, we propose a new approach for reasoning about occlusions between multiple people. In our approach, occlusion relationships between people are explicitly defined and deduction of the occlusion relationships is integrated into the whole tracking framework. The prior knowledge is supplied by a set of models which include a 2-D elliptical shape model, a spatial-color mixture of Gausians appearance model, and a motion model with constant velocity. An observation likelihood function is constructed based on the similarity between the observations and the object appearance models with given states. The occlusion relationships are deduced from the current states of the objects and the current observations, using the observation likelihood function. The previous occlusion relationships are not required for deducing the current occlusion relationships. The problem of tracking and occlusion reasoning for more than two people is formulated mathematically, and a solution is proposed based on particle filtering. Experimental results on several real video sequences from indoor and outdoor scenes show the effectiveness of our approach.

Index Terms—Occlusion reasoning, particle filtering, tracking multiple people, visual surveillance.

I. INTRODUCTION

Occlusion reasoning for object tracking is one of the most challenging problems in visual surveillance. This is because during occlusion, only portions of each occluded object are visible and the correspondences between objects and their features become ambiguous [22]. The aim of occlusion reasoning for tracking is to explicitly determine the occlusion relationships between objects and then accurately localize the objects. As people activities are of key interest in monitored scenes, we focus on reasoning about occlusions between multiple people.

A. Related Work

There exist approaches for tracking multiple objects, in which the occlusion relationships are not explicitly defined and determined. For example, Bar-Shalon and Fortmann [23] use the joint probabilistic data association filter (JPDAF) to track multiple objects, by finding the exclusive correspondence between different objects. Rasmussen and Hager [11] propose a framework that describes data association to improve the performance of tracking multiple objects. The accurate localization of objects present to occlusions needs a very long running time. Khan et al. [16], [17] use a Markov random field to model motion interaction, and the motion model is naturally incorporated into a MCMC (Markov Chain Monte Carlo)-based particle filter for effective multiple object tracking. However, their model is used to describe the interactions between different objects rather than reason about occlusions.

Some approaches track multiple people through occlusions using 3-D calibration information. For instance, Isard and MacCormick [21], and Zhao and Nevatia [18], [19] investigate tracking in calibrated images, where calibration information is utilized to reason about occlusions. Mittal and Davis [20] reason about occlusions using multiple calibrated cameras. As camera calibration is costly, we only consider un-calibrated occlusion reasoning in this letter.

The 2-D un-calibrated people tracking approaches base occlusion reasoning on people appearance models. Non-parametric people appearance models including appearance template-based [1], [7], [12], color histogram-based [2], [13]–[15], and kernel density-based [3], [4], and parametric appearance models [5], [6] are used to track multiple people through occlusions. For instance, Senior et al. [1], [12] use an appearance template to record the RGB color of each pixel of each person. Each appearance template is associated with a set of probability masks, each of which represents the probability of the corresponding pixel belonging to that template. Before the occurrence of an occlusion, the template is continuously updated on each new image as it arrives. During occlusions, foreground pixels are segmented into different regions corresponding to different templates according to the learned templates and the predicted people positions. People are then tracked according to the segmented regions. Wu et al. [7] apply the Bayesian network to track two faces through occlusions using multi-cameras, with the participation of the appearance templates. They put the occlusion relationships between objects into the object states, and thereby combine prior expectations of occlusion with the likelihood function. They assume that the occlusion relationships between objects are independent of the current states of the objects, and dependent on the previous occlusion relationships. Qu et al. [15] propose a distributed Bayesian formulation using multiple interactive trackers that require low complexity for real time tracking applications. They extend the conventional Bayesian tracking framework by modeling object interaction in terms of potential functions. They combine the color histogram model and the PCA-based mode to efficiently model object appearance.

B. Our Work

Considering that most current methods [1]–[6] cast occlusion reasoning as a classical segmentation problem, i.e., classify foreground pixels into several sub-regions according to prior knowledge and then track people based on the segmented
regions, in this letter we omit the process of segmenting the motion region into sub-regions corresponding to different people, and propose a new approach for occlusion reasoning between multiple people. In the approach occlusion relationships between multiple people are explicitly defined and deduced, the models for the different people are matched against the image in the order consistent with the occlusion relationships, and then the localizations of the people are optimized. The main contributions of our approach are:

- Tracking and occlusion reasoning for more than two people are formulated mathematically. The occlusion relationships between multiple people are incorporated into a tracking framework in which the main focus is on finding the positions of the people.
- The occlusion relationships are deduced from the current states of the people and the current observations, with the participation of the observation likelihood function. This is in contrast with previous work such as [7] in which it is assumed that the occlusion relationships between objects are independent of the current states of the objects, and dependent only on the previous occlusion relationships.

To implement these contributions, we apply the ellipse shape model, the spatial-color mixture of Gaussians appearance model [24], and the motion model with constant velocity. An observation likelihood function is constructed based on the appearance model used. A standard particle filter is adopted to optimize object states and occlusion relationships.

II. OVERVIEW OF OUR APPROACH

In our approach, pixels are classified as foreground/background for each frame using background subtraction, without segmenting the foreground pixels into objects. Motion regions are extracted from foreground pixels. The extracted motion regions in the current frame are compared with the motion regions in the previous frame according to their positions and appearances. If the current frame contains a motion region which does not correspond to any motion region in the previous frame, then a new object is detected as a tracked object and the appearance model for the object is initialized according to the new motion region. If a motion region gradually becomes smaller to the point where it can be ignored, then object disappearance occurs and the tracking of the object terminates. If two or more objects come into the scene with one occluding the other, then they are tracked as a single object. If it is detected that a motion region tracked as an object in the previous frame has divided into some smaller motion regions in the current frame, then the objects are tracked separately: pixels in these smaller motion regions are used to initialize the models of these objects. The appearance model for each person is updated in each of the frames in which the person is absent from any occlusion. If it is detected that a person is present to occlusion, the appearance model of the person is kept constant. The motion model is used to predict the states of the people at time $t$ according to the tracking results at time $t-1$. The occlusion relationships between the people are deduced according to the predicted states, the observations, and their shape and appearance models. The human shape model and the appearance model with predicted states and deduced occlusion relationships are projected to the image plane, and the observation likelihood is used to further update the states of the objects and their occlusion relationships for maximizing the posterior probability of the states and the occlusion relationships given the observations.

We consider occlusions between $n$ objects. We denote the state of object $k$ at time $t$ by $S_{k,t}^l$ ($k = 1, 2, \ldots, n$). The observation image (contained in the foreground pixels) at time $t$ is denoted by $O_t$, with history $(O_{t-1} \ldots O_{0} O_t)$. The occlusion relationship $\pi_t^{(i,j)}$ between objects $i$ and $j$ at time $t$ is represented as

$$\pi_t^{(i,j)} = \begin{cases} -1, & \text{if } i \text{ occludes } j \\ 0, & \text{if no occlusion} \\ 1, & \text{if } j \text{ occludes } i. \end{cases} \quad (1)$$

So, the object state vector is $(S_{1,t} \ldots S_{k,t} \ldots S_{n,t})$, which is simplified with $(S_{1,t} \ldots S_{k,t} \ldots S_{n-1,t})$ if $i < j < n$, which is simplified with $(S_{i,t} \ldots S_{n,t})$. The state $S_{k,t}$ is predicted using probability $p(S_{k,t} | S_{k-1,t} \ldots S_{i,t})$ and the occlusion relationships $(\pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)})$ are deduced according to the predicted states $\{S_{k,t}^l, 1 \leq k \leq n\}$, with the participation of the shape and appearance models of the people. Then, the observation likelihood is used to further update the state vector $(S_{1,t} \ldots S_{n,t})$ and the occlusion relationship vector $(\pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)})$ for maximizing the posterior probability $p(S_{1,t} \ldots S_{n,t} | S_{1,t}^{l}, \ldots, S_{n,t}^{l}, O_t \ldots O_{0} O_t)$. To estimate $p(S_{k,t}^{l} | S_{k-1,t} \ldots S_{i,t}^{l})$ and $p(O_t | S_{1,t}^{l} \ldots S_{n,t}^{l}, \pi_t^{(1,2)} \ldots \pi_t^{(n-1,n)})$, we should construct proper human body models and a proper observation function.

III. HUMAN BODY MODELS

Human body models include the shape model, the appearance model and the motion model.

A. Shape Model

There are many different shape models for the human body. The more parameters there are in a shape model, the more accurate the tracking results, but the more expensive the computation. The aim in this letter is to reason about the occlusions between multiple people. We do not consider tracking the articulated motions of a human body and do not estimate the pose of each part of a human body. Therefore, we select a simple elliptical model involving only a few parameters, to simplify the computation. As shown in Fig. 1, a human body is modeled as a 2-D vertical ellipse with a parameter vector $(x,y,h,\psi)$, where $(x,y)$ is the center of the ellipse, $h$ is the length of the major axis of the ellipse, and $\psi$ is the eccentricity of the ellipse. For the tracking, the state of the ellipse is taken to be $S = (x,y,h,\psi)$ which is an example of a state introduced in Section II.

B. Appearance Model

In this letter, we use the spatial-color mixture of Gaussians appearance model in [24] to model the spatial layout of the colors in a person in an image. Let $\mu_j^l$ and $\Sigma_j^l$ be the mean and
the covariance of the color feature in the \(^j\)th mode of the mixture of Gaussians, and \(\mu_j^c\) and \(\Sigma_j^c\) be the mean and the covariance of the spatial feature. Let \(\eta(x_c^j, f_j^c, \Sigma_j^c)\) be the probability density function value of the color \(x^c\) of a pixel \(x\) conditional on the Gaussian \(N(\mu_j^c, \Sigma_j^c)\), and \(\eta(x^s_j, \mu_j^s, \Sigma_j^s)\) be the probability density function value of the position \(x^s\) of the pixel \(x\) conditional on the Gaussian \(N(\mu_j^s, \Sigma_j^s)\). Then, the probability density function value \(P_x\) of the pixel \(x\) under the spatial-color mixture of Gaussians appearance model is:

\[
P_x = \sum_{j=1}^{\gamma} \omega_j \eta(x^c, \mu_j^c, \Sigma_j^c, x^s, \mu_j^s, \Sigma_j^s)\]  \(\text{(2)}\)

where \(\gamma\) is the number of Gaussians and \(\omega_j\) is the normalized weight of the \(j\)th mode of the mixture of Gaussians. The K-means clustering algorithm is used to initialize the parameters of the mixture of Gaussians and the Expectation-Maximization (EM) algorithm is used to refine these parameters. Please see [24] for detail.

The spatial-color mixture of Gaussians appearance model improves the popular color histogram-based appearance model because it considers not only the colors in a region but also the spatial layout of the colors.

C. Motion Model

Application of motion models, in the form of prior knowledge, to object tracking can reduce the search space and make the tracking results more accurate. As the changes in motion velocity and object sizes between successive frames are very small, for simplicity we assume a constant velocity and a constant height and eccentricity of the ellipse shape model. The motion model is described by the function below

\[
\begin{align*}
x_t &= x_{t-1} + (x_{t-1} - x_{t-2}) \\
y_t &= y_{t-1} + (y_{t-1} - y_{t-2}) \\
h_t &= h_{t-1} \\
\psi_t &= \psi_{t-1}
\end{align*}
\]  \(\text{(3)}\)

where \(x_t\) and \(y_t\) are the \(x\) and \(y\) coordinates of the position of a person, and \(h_t\) and \(\psi_t\) are the height and eccentricity of the ellipse corresponding to the person in the image at time \(t\).

IV. Observation Likelihood

The observation likelihood function \(p(O_t|\pi_t^1 \ldots \pi_t^m, \pi_t^1 \ldots \pi_t^{(n-1,m)})\) represents the probability of an observed image \(O_t\) (contained in the foreground pixels), given states \(\{\pi_t^1, \ldots, \pi_t^n\}\) of objects and their occlusion relationships. The observation likelihood depends on the similarity [8] between the observations and the appearance reflected by the appearance models of objects, the shape models with given states, and the occlusion relationships. The similarity is measured according to the comparison between the appearance model of each person and the observation specified by the shape model of the person with given states and occlusion relationships.

We first determine the corresponding observation of an object \(k\) under its given states and occlusion relationships. The given occlusion relationships \(\pi_t^{(1,k)}, \ldots, \pi_t^{(k-1,k)}, \pi_t^{(k,k+1)}, \ldots, \pi_t^{(n,m)}\) involving object \(k\) specify all objects occluding object \(k\). Let \(R_k\) be the region of foreground pixels within the ellipse of object \(i\) with the given state at time \(t\). If object \(k\) is not occluded, its corresponding observation \(V_t^k\) is \(R_k\); and if object \(k\) is occluded by \(c\) objects which are indexed by \(I_1, I_2, \ldots, I_c\), its corresponding observation is

\[
V_t^k = R_t - \left( \bigcup_{i=1}^c R_t^i \right) .
\]  \(\text{(4)}\)

Based on the determined corresponding observation \(V_t^k\) of object \(k\), we estimate the likelihood of the pixels in \(V_t^k\) given the spatial-color mixture of Gaussians appearance model \(M_k\) of object \(k\). Let \(x\) represent a pixel in \(V_t^k\), let \(|V_t^k|\) be the number of pixels in \(V_t^k\), and let \(p(x|M_k)\) be the probability density function value of pixel \(x\) given \(M_k\). The likelihood \(P(V_t^k|M_k)\) of the pixels in \(V_t^k\) given the model \(M_k\) is defined as

\[
P(V_t^k|M_k) = \exp \left\{ \frac{1}{|V_t^k|} \sum_{x \in V_t^k} \log p(x|M_k) \right\}.
\]  \(\text{(5)}\)

The larger the \(P(V_t^k|M_k)\), the more the compatibility between the pixels in \(V_t^k\) and the model \(M_k\). The log function and the exp function in (5) are used to transform the product of probability density function values to sum of the logarithmic values of the probability density function to avoid the underflow produced by the product of probability density function values.

As the motions of objects are independent, the similarity between the appearance models of all people and the observations is measured by the sum of the similarities between the appearance model of each person and the corresponding observation of the person. Then, the observation likelihood function is defined as

\[
p(O_t|\pi_t^1 \ldots \pi_t^n, \pi_t^1 \ldots \pi_t^{(n-1,m)}) \propto \sum_{k=1}^n P(V_t^k|M_k)
\]  \(\text{(6)}\)

where \(n\) is the number of people in the scene at time \(t\).
V. OCCLUSION RELATIONSHIP DEDUCTION

We found that, given the states of the objects, their occlusion relationships are fixed. So the occlusion relationships between objects are dependent on the current states of the objects, and independent of their previous occlusion relationships. The transition of occlusion relationships [7] from the previous step to the current step is not needed. It follows that we use the likelihood function \( p \left( O_t | \pi^{(1,2)}_t, \ldots, \pi^{(n-1,n)}_t, S^1_t, \ldots, S^n_t \right) \) which involves the current observations and the shape and appearance of the objects, to estimate the occlusion relationships. The optimal occlusion relationships maximize the likelihood function given the states of the people

\[
\pi^*_t = \arg \max_{\pi_t} \left( p \left( O_t | \pi^{(1,2)}_t, \ldots, \pi^{(n-1,n)}_t, S^1_t, \ldots, S^n_t \right) \right)
\]

i.e., the optimal type of occlusion relationships is selected from the combinations of occlusion relationships between the people, to maximize the value of the likelihood function given the states of the people.

In implementing the deduction of occlusion relationships between people, it is checked whether the ellipses specified by the states of the people intersect or overlap. Only for the people whose ellipses intersect or overlap, the observation likelihood function is used to determine “who occludes whom”. So when there are a large number of people in the scene, but most of them are absent from occlusions, only the combinations of occlusion relationships between the people present to occlusions are mainly considered. Then, the computational complexity is decreased.

VI. TRACKING WITH OCCLUSION REASONING

The process of tracking with occlusion reasoning is to maximize the posterior probability

\[
p \left( S^1_t, \ldots, S^n_t | \pi^{(1,2)}_t, \ldots, \pi^{(n-1,n)}_t | O_t, \ldots, O_l \right)
\]

according to the motion model \( p \left( S^k_t | S^k_{t-1} \right) \) and the observation likelihood \( p \left( O_t | S^1_t, \ldots, S^n_t, \pi^{(1,2)}_t, \ldots, \pi^{(n-1,n)}_t \right) \).

Particle filtering is a simple and effective method for computing posterior probabilities from non-Gaussian, non-linear, and high-dimensional observation data. In this letter, a standard particle filtering approach [7], [9] is applied to find the maximum of the posterior probability density function. The posterior density \( p \left( S^1_t, \ldots, S^n_t, \pi^{(1,2)}_t, \ldots, \pi^{(n-1,n)}_t | O_t, \ldots, O_l \right) \) is represented by \( M \) particles \( \{ S^1_t^{(m)}, \ldots, S^n_t^{(m)} \} \) \( (m = 1, \ldots, M) \) where \( S^i_t^{(m)} \) corresponds to the state of object \( i \), deduced occlusion relationships \( \{ \pi^{(1,2)}_t^{(m)}, \ldots, \pi^{(n-1,n)}_t^{(m)} \} \), and the weights \( W^t_t^{(m)} \) of the particles. The propagation of the filter is outlined as follows.

1) Initialization: \( t = 0 \). Particles \( \{ S^1_0^{(m)}, \ldots, S^n_0^{(m)} \} \) are sampled according to the prior probability distribution \( p \left( S^1_0, \ldots, S^n_0 \right) \) which is assumed to be the uniform distribution. The ranges for \( x, y, h, \) and \( \psi \) in each object are \( [x-\Delta x, x+\Delta x], [y-\Delta y, y+\Delta y], [h-\Delta h, h+\Delta h], \) and \( [\psi-\lambda \psi, \psi+\lambda \psi] \), respectively. The occlusion relationships \( \{ \pi^{(1,2)}_0^{(m)}, \ldots, \pi^{(n-1,n)}_0^{(m)} \} \) are deduced. The weights \( \{ W^0_0^{(m)} \} \) are computed by the observation likelihood:

\[
W^0_0^{(m)} = p \left( O^0_0 | S^1_0^{(m)}, \ldots, S^n_0^{(m)}, \pi^{(1,2)}_0^{(m)}, \ldots, \pi^{(n-1,n)}_0^{(m)} \right).
\]

2) \( t = 1, 2, \ldots \)

a) Re-sampling: In order to avoid degeneration of particles, particles \( \{ S^1_t^{(m)}, \ldots, S^n_t^{(m)} \} \) \( (m = 1, \ldots, M) \) at time \( t-1 \) are re-sampled to produce a new particle set \( \{ S^1_t^{(m)}, \ldots, S^n_t^{(m)} \} \) according to \( \{ W^t_0^{(m)} \} \). In this process, particles with large weights are sampled several times, and particles with small weights are ignored. The total number of particles is kept constant.

b) Prediction: Particles \( \{ S^1_t^{(m)}, \ldots, S^n_t^{(m)} \} \) at time \( t \) are generated according to the prediction model \( p \left( S^k_t | S^k_{t-1} \right) \) \( (k = 1, 2, \ldots, n) \) which is described by (3).

c) Deduction: The corresponding occlusion relationships \( \{ \pi^{(1,2)}_t^{(m)}, \ldots, \pi^{(n-1,n)}_t^{(m)} \} \), given the states \( \{ S^1_t^{(m)}, \ldots, S^n_t^{(m)} \} \), are deduced according to \( p \left( O_t | S^1_t, \ldots, S^n_t, \pi^{(1,2)}_t, \ldots, \pi^{(n-1,n)}_t \right) \), as described in Section V.

d) Measurement: Each particle is re-weighted by the observation likelihood:

\[
W^t_t^{(m)} = p \left( O_t | S^1_t^{(m)}, \ldots, S^n_t^{(m)}, \pi^{(1,2)}_t^{(m)}, \ldots, \pi^{(n-1,n)}_t^{(m)} \right).
\]

Then, the weights are normalized

\[
\tilde{W}^t_t^{(m)} = \frac{W^t_t^{(m)}}{\sum_{l=1}^{M} W^t_l^{(l)}}.
\]

e) Estimation: The states of objects at time \( t \) are estimated as the summation of the weighted particles:

\[
S^i_t = \sum_{m=1}^{M} \tilde{W}^t_t^{(m)} S^i_t^{(m)}, \quad k = 1, 2, \ldots, n.
\]

Then the occlusion relationships at time \( t \) are deduced according to the estimated object states.

In our approach, the particles are defined in the joint state space of all objects. If a new object enters the scene, then the state components of the object are added to the joint state space and the particles are updated. The particle components of the object in all the particles are sampled initially according to the prior probability distribution which is assumed to be the uniform distribution. For example, if an object indexed by \( n+1 \) enters the scene, the particles components \( \{ S_t^{n+1,m} \} \) of object \( n+1 \) are sampled and then added to the previous particles \( \{ S_t^{1,m}, \ldots, S_t^{n,m} \} \) \( (m = 1, \ldots, M) \)—the particles become \( \{ S_t^{1,m}, \ldots, S_t^{n,m}, S_t^{n+1,m} \} \). If an object leaves the scene, the state components of the object are
removed from the joint state space (the dimension of the state space is reduced), and the particle components of the object are removed from all the particles. For example, if object $i$ leaves the scene, particles $\left\{ S_{i}(m), \ldots, S_{i}^{-1}(m), S_{i}^{-1+1}(m), \ldots, S_{i}^{-n}(m) \right\}$ ($m = 1, \ldots, M$) are changed to be $\left\{ S_{i}^{-1}(m), S_{i}^{-1+1}(m), \ldots, S_{i}^{-n}(m) \right\}$.

To adapt different parameters and increase the robustness and the effectiveness of the inference component, a hierarchical strategy is used to estimate the states of objects. Firstly, the size of the ellipse of each person is set equal to be the size of the ellipse of the person in the previous frame. The particle filtering process is used to estimate the optimal positions of the objects. Then, the positions are fixed and the particle filtering process is used again to estimate the optimal sizes of the objects.

### VII. EXPERIMENTAL RESULTS

The above algorithm is implemented using MatLab on the Windows XP platform. Computation time is measured on a P4-2.8G computer with 512M RAM. A simple background subtraction algorithm [10] is applied to extract the foreground pixels. The success of tracking and occlusion reasoning is checked by our own judgment. The results of occlusion reasoning are illustrated using the recovered occlusion relationship diagram whose $x$-coordinate is the frame number, and $y$-coordinate is the occlusion relationship. In the following, we demonstrate four examples of tracking with occlusion reasoning, and then illustrate some comparison results.

#### A. Example 1

The test video for this example is captured in an outdoor environment. Its resolution is 320 x 240. Fig. 2 contains some key frames illustrating the tracking of two people through occlusion. The recovered occlusion relationship is shown in Fig. 3. Before Frame 741, there was no occlusion between the two people, and thus the occlusion relationship equals to 0. From Frame 741 to Frame 760, the front person (Person A) occluded the back person (Person B), and thus the occlusion relationship equals to $-1$. The two people separated after Frame 760, so the occlusion relationship then equals to 0.

#### B. Example 2

The test video for this example is captured in an outdoor environment. Its resolution is 320 x 240. Fig. 4 contains some frames illustrating the results of tracking from Frame 390 to Frame 458. In this portion of the video sequence, People A entered the scene from the right and walked to left; Person B entered the scene from the left and walked to right; and Person C entered the scene from the right and walked behind Person A. From Frame 411, Person B occluded Person A. From Frame 421, Person B occluded Person C. As Person C walked more quickly than Person A, Person C caught up with Person A from Frame 426, and then Person C occluded Person A. From Frames 426 to Frame 429, there is mutual occlusion between the three people, i.e., Person B occluded people A and C, and Person C occluded Person A. From Frames 429 to Frame 458, there is mutual occlusion between the three people, i.e., Person B occluded people A and C, and Person C occluded Person A. Fig. 5(a)–(c) show respectively the recovered occlusion relationships between People A and B, People A and C, and People B and C. The results show that all the three people are successfully tracked, and their occlusion relationships are deduced correctly. In this example, the three people occlude each other for a long time. The example shows that if there are no large
changes in people’s appearances, then our approach can handle long-term multiple occlusions.

C. Example 3

This is an example in which a person appeared within the image. The test video for this example is in the PETS 2004 database which is an open database for the research on visual surveillance. The video is available on http://home-pages.inf.ed.ac.uk/rbf/CAVIAR/ShopAssistant2Cor.avi. Its resolution is 384 × 288. It shows a shopping center. Fig. 6 shows some frames of the results of tracking people in the video. Before Frame 201, three people are tracked. Two of them occluded each other in the midsection of the scene and one of them moved at the right-bottom corner of the scene. From Frame 201, a person entered the scene by emerging from a door placed away from the edge of the image. Tracking of the person is successfully initialized. From Frame 314, the person who moved at the right-bottom corner of the scene left the scene. For the video, all the four people are accurately tracked and their occlusion relationships are correctly determined.

D. Example 4

This video is taken from the same scene as Example 3. It is available on http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. In the video, there are many people walking or standing and talking. Fig. 7 illustrates some frames of tracking results when large mutual occlusions between five people occur from Frame 200 to Frame 400. During these occlusions, these people are still tracked successfully and their occlusions are deduced correctly using our approach.

We contrast the tracking results of our approach with the ground truth which is labeled by the suppliers of the PETS 2004 database: the pixel position of the center of the ellipse of each person is contrasted with the center of the manually labeled image bounding box of the person. In the following, the contrast results for three people occluded extensively are shown. Fig. 8 shows the trajectory of Person A who is the first person counting from left to right in Frame 200, where the dashed line is the ground truth trajectory, and the solid line is the trajectory obtained by our approach. From Frame 200 to Frame 400, for Person A the average warp between the trajectory obtained by our approach and the ground truth is 9.2 pixels. This warp is comparatively large. The reason is that the lower part of this person is occluded in most of the time that the person appeared in the scene, and therefore the height parameter of the ellipse of the person cannot be updated adaptively. Fig. 9 shows the trajectory of Person D who is the fourth person counting from left to right in Frame 200. From Frame 200 to Frame 400, for Person D the average warp between the trajectory obtained by our approach and the ground truth is 5.0 pixels. Fig. 10 shows the trajectory of Person E who is the fifth person counting from left to right in Frame 200. From Frame 200 to Frame 400, for Person E the average warp between the trajectory obtained by our approach and the ground truth is 4.9 pixels.
Fig. 10. Trajectory contrast for Person E.

Fig. 11. Tracking two people with occlusion reasoning using the segmentation-based method.

E. Comparison

To evaluate the performance of our approach, we compare our approach with the one in [1], [12]. The appearance template-based occlusion reasoning approach in [1], [12] is a typical segmentation-based one, which has high reliability in segmenting foreground pixels, but needs a large area of memory to store the template.

Example 1 is used to compare the performance of our approach with the performance of the segmentation-based approach. Fig. 11 illustrates the tracking results obtained by the approach in [1], [12], corresponding to Example 1. From Frame 751 to Frame 754, the occluded person is not correctly tracked. From Frame 755 to Frame 757 the occluded person is lost in tracking.

Fig. 12 compares the ellipse centers for our approach with those for the segmentation-based approach. The horizontal and vertical coordinates in the image are respectively the \(x\)-coordinate and \(y\)-coordinate of the center of each ellipse. The trajectory curves of the un-occluded person (Person A) are shown in (a) and the trajectory curves of the occluded person (Person B) are shown in (b). From Fig. 12, we can see that the curve produced by the segmentation-based approach is farther from the ground truth than the one produced by our approach. This is more noticeable for the occluded person as shown in (b).

Table I shows the quantitative comparison between our approach and the segmentation-based approach for Example 1 with respect to the following quantitative criteria:

- the number of frames in which tracking was successful, corresponding to the tracking success rate;
- the average position error—the average pixel distance between the estimated ellipse center and the center of the corresponding ground truth bounding box;
- the average size error—the average warp between the estimated ellipse size and the size of the corresponding ground truth bounding box;
- runtime per frame.

From this table, it can be seen that position errors and size errors of tracking the two people using our approach are less than those using the segmentation-based approach. However, the runtime of our approach is more than that of the segmentation-based approach. This is due to the time taken to estimate the parameters of the spatial-color mixture of Gaussians appearance model.

VIII. CONCLUSION

In this letter, we have presented a new approach for occlusion reasoning for tracking multiple people. In our approach, occlusion reasoning is integrated into the whole process of tracking, in order to obtain better tracking accuracy during occlusions. Prior knowledge is represented by a set of models: a 2-D elliptical shape model, a spatial-color mixture of Gaussians appearance model, and a motion model with constant velocity. We have constructed an observation likelihood based on the similarity between the appearance models of the people and the observations. The occlusion relationships are deduced from the current states of the objects and the current observations. The previous occlusion relationships are not required for deducing
the current occlusion relationships. Tracking and occlusion reasoning for more than two people are formulated in the particle filtering framework, and a solution is proposed. Experimental results on several real video sequences from indoor and outdoor scenes have demonstrated the effectiveness of our approach.

ACKNOWLEDGMENT

The authors thank X. Zhang and M. Liu for their help for implementing some experiments in this letter.

REFERENCES