Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2008, Article ID 253039, 13 pages
doi:10.1155/2008/253039

*Research Article*

# Multiview-Based Cooperative Tracking of Multiple Human Objects

**Kuo-Chin Lien[1] and Chung-Lin Huang[1, 2]**

[1] *Institute of Electrical Engineering, National Tsing Hua University (NTHU), Hsin-Chu 30013, Taiwan*
[2] *Department of Informatics, Fo-Guang University, I-Lan 26247, Taiwan*

Correspondence should be addressed to Chung-Lin Huang, clhuang@ee.nthu.edu.tw

Received 31 January 2007; Revised 28 July 2007; Accepted 3 December 2007

Recommended by Nikos Nikolaidis

Human tracking is a popular research topic in computer vision. However, occlusion problem often complicates the tracking process. This paper presents the so-called multiview-based cooperative tracking of multiple human objects based on the homographic relation between different views. This cooperative tracking applies two hidden Markov processes (tracking and occlusion processes) for each target in each view. The tracking process locates the moving target in each view, whereas the occlusion process represents the possible visibility of the specific target in that designated view. Based on the occlusion process, the cooperative tracking process may reallocate tracking resources for different trackers in different views. Experimental results show the efficiency of the proposed method.



## 1. INTRODUCTION

Currently, multiple-view multiple-object tracking has become an essential technology for many applications such as video surveillance system. Modern video-based surveillance systems [1] employ real-time image analysis techniques for the target tracking and identification. The major issue of target tracking is to identify the multiple moving objects. However, occlusions among the objects will complicate the tracking process and make it difficult for the system to identify the object after occlusion. This paper proposes a novel method for multiple human tracking in multiple views.

Recently, researchers have shown a great interest in using particle filters for visual tracking [2–5]. For analyzing the occlusion between targets, Wu et al. [6] propose modeling occlusion relations as an extra hidden process in a dynamic Bayesian network. A hidden variable was used to indicate the three possible relations between two moving objects. The transition process is described with a three-state finite state machine. Hu et al. [7] extend the framework to human tracking. Analyzing the depth order around occlusions is indeed helpful to maintain tracking. However, as the number of targets increases, occlusion relations among targets get more complicated. Another problem in object tracking is that the appearance of object changes quite often. Zhou et al. [8] present an approach incorporating appearance-adaptive models into a particle filter to realize robust visual tracking.

Kang et al. [9] use time weighted color information (or temporal color) for multiple people tracking. Recently, graph-based multiple hypothesis tracking (MHT) algorithms have been proposed [10–12] to track multiple targets. However, the hypothesis tree grows exponentially as more measurements are received. Khan et al. [13] replace the traditional sampling step in the particle filter with a novel Markov chain Monte Carlo (MCMC) sampling step to obtain an efficient multitarget tracking. These methods try to limit the growth via a series of clustering and pruning operations. However, the major limiting factor is that these algorithms lag noticeably in a crowded scene because of high computation complexity.

Because the single viewpoint loses the depth information, there are too many hidden regions when targets present complicated depth order. Matshuyama and Ukaita [14] present a cooperative tracking system which consists of a group of active vision agents. Each agent may include more than one camera and is dynamically established to handle one single

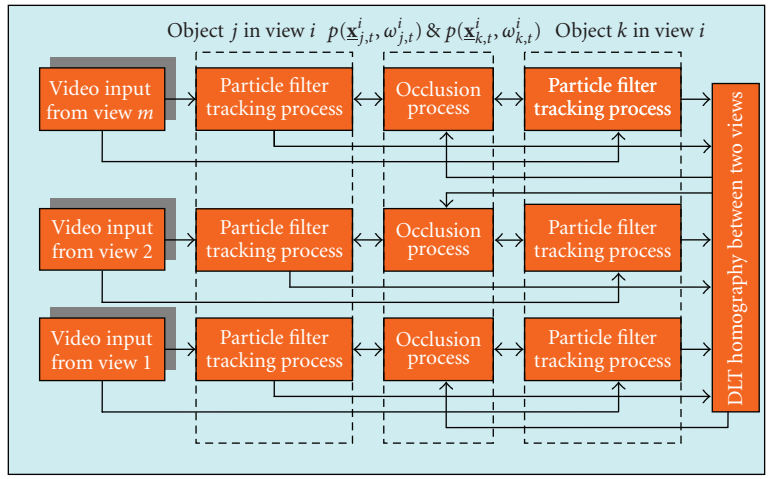



Figure 1: System flow diagram.

target. Utsumi et al. [15] propose a multiview tracking algorithm using synchronous cameras to locate the objects. To detect human rotation angles and body sides, they project 3D object model onto every 2D image plane to choose a proper view.

Chang and Gong [16] use Bayesian modality function to track people by using multiple cameras to overcome the occlusion problem. Otsuka and Mukawa [17] introduce a dual-loop particle filter operating in multicamera environments. For eliminating the phantom cells produced in back-projection, a hypothesis support ratio is added to estimate the most possible global structure. Canton-Ferrer et al. [18] present a Bayesian approach to find the correspondence of moving objects in multicamera environments. A simple point-based feature of each foreground region is extracted which is used to establish the correspondence. Khan and Shah [19] propose a method to find the limits of field of view (FOV) of each camera as visible in the other camera. The FOV can be used to recover the homograph between the views. Snidaro and Foresti [20] propose an approach to automatically evaluate and select the sensors of a multisensor system by measuring their efficiency in detecting the target. Lopez et al. [21] propose a 3D tracking method using multiple particle filters and model the interaction among them through a 3D blocking scheme.

This paper presents the so-called *multiview-based cooperative tracking* system by using particle filter. Different from other multiview-based works on multiple targets tracking [14–21], this paper aims at solving the occlusion problem by combining the multiple camera inputs. This approach is based on the concepts of data sharing and resource sharing of the tracking processes for all the targets. Since view-to-view mapping is preconstructed, the precise trajectory of each target can be obtained based on the reliable observations in which the designated target is more visible.

This paper applies the particle filtering to track each object. In the multiview cooperative tracking system, the tracking status of each target in each view is described by a hidden Markov process, the occlusion status of each target is also modeled by another hidden Markov process, and these two processes are related. Different from single view tracking [5–7], this approach solves the complex occlusion problem by allowing different trackers to share their common computation resources and sensor data. For each tracking process, the total allocated computation resources are always the same; that is, the number of sampled particles used for tracking is fixed. The cooperative tracking allocates fewer resources for the tracking process in unreliable views, and distributes more resources to track the object in the reliable view. In comparison with the other multiview noncooperative tracking methods [14–21], our method can track the objects more effectively.

## 2. SYSTEM OVERVIEW

In surveillance videos, the background scenery is assumed to be stationary and homogeneous. The observations of the same target are supposed to be similar. Moreover, every person can be identified by some visual features different from the background. However, when multiple objects approach each other, the tracking process of one object will interfere with the process of the other one. For particle-filter-based tracking algorithms, the impact is that some of the particles will be located on the other target and the biased expectation of the hypothesis will be generated. Besides, the interference among different tracking processes may also occur because of the occlusion or shadow. In the crowded scene, it becomes significant. To overcome these difficulties, we introduce another hidden Markov process to model the occlusion status among different tracking processes.

The overall system is demonstrated in the flow diagram shown in Figure 1. For each object in each view, there is a corresponding tracking process. Between two tracking processes in the same view, there is an occlusion process that provides the occlusion status between these two moving objects. For each object appearing in different views, we apply the direct linear transformation (DLT) homograph to determine

whether the previously occluded object has reappeared and will be allocated with more tracking resources.

### 2.1. Overview of particle filter

The basic idea of particle filtering is to estimate how a probability density propagates. The object to be tracked can be modeled according to its visual feature. Based on this probabilistic model, the presence of the object in a scene can be described as a probability density function. In time series, the propagation of density function is a stochastic process $\underline{\mathbf{X}}_t$. Estimating the state propagation of $\underline{\mathbf{X}}_t$ means locating the object frame by frame.

In particle filtering, we assume that $\{\underline{\mathbf{X}}_1, \underline{\mathbf{X}}_2, \ldots, \underline{\mathbf{X}}_T\}$ is a sequence of hidden Markov processes; the present state $\underline{\mathbf{X}}_t$ is only related to the preceding state $\underline{\mathbf{X}}_{t-1}$. Mutually independent observations $\{Z_1, \ldots, Z_T\}$ are introduced, and conditional state density $p(\underline{\mathbf{X}}_t \mid Z_t)$ can be calculated as

$$p(\underline{\mathbf{X}}_t \mid Z_t) = p(Z_t \mid \underline{\mathbf{X}}_t) \int p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1})\, p(\underline{\mathbf{X}}_{t-1} \mid Z_{t-1})\, d\underline{\mathbf{X}}_{t-1}, \tag{1}$$

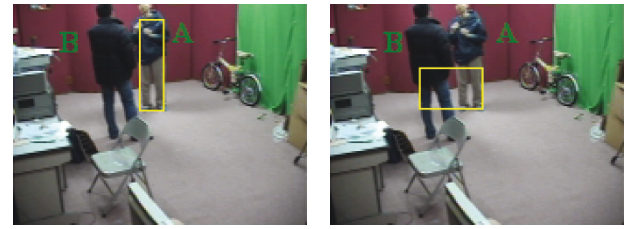
where $p(\underline{\mathbf{X}}_{t-1} \mid Z_{t-1})$ is the posterior from the previous time step, and $p(Z_t \mid \underline{\mathbf{X}}_t)$ is the observation likelihood. The dynamics $p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1})$ can be implemented as a first-order process $\underline{\mathbf{X}}_t = A\underline{\mathbf{X}}_{t-1} + \mathbf{w}_t$, where $\mathbf{w}_t$ is a Gaussian noise. In our experiments, $A$ is defined as a constant value that represents targets moving at a constant velocity.

The distribution $p(\underline{\mathbf{X}}_t)$ is represented by a weighted sample set $\{(\mathbf{s}_1, \pi_1), \ldots, (\mathbf{s}_n, \pi_n)\}$ which represents the so-called "particles". The weighting factor $\pi_n$ is proportional to $p(Z_t \mid \underline{\mathbf{X}}_t = \mathbf{s}_n)$, and the estimated state of object can be determined from the expectation of the sample set $\{\mathbf{s}_n\}$ as $\hat{\mathbf{x}}_t = \sum_{n=1}^{N} \pi_n \mathbf{s}_n$. The particle filter iteratively resamples $\underline{\mathbf{X}}_t$ and reweights the samples.

To construct the observation model, feature selection is a crucial issue. A proper feature should have high discriminating power to maintain the identities of objects and low complexity in extracting the feature information. For vision-based human tracking, color histogram analysis has been widely used [3, 4]. Here, we adopt a hue-saturation (HS) color histogram model and uniformly quantize $H$ and $S$ into 10 levels. So a color region in an image can be given a statistical description by $10 \times 10$ bins on the HS plane. Besides, for those points with $R = G = B$, an additional bin is used to count them, because we cannot evaluate their hue information. Thus, the resulting complete histogram is composed of $10 \times 10 + 1 = 101$ bins.

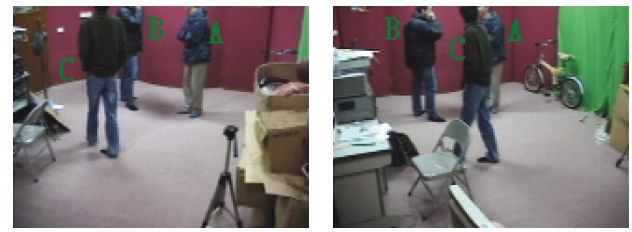
In single object tracking, the observation of every particle sample is in a rectangular area. Each rectangle can be seen as a candidate region of the human target, and each sample $\mathbf{s}$ represents a four-dimension entity, that is, $\mathbf{s} = \{(x, y), (h, r)\}$, where $(x, y)$ represents the bottom center of the rectangle, defining the location of a person, and $(h, r)$ represents the height and the aspect ratio. Evaluating the similarities between these hypothesis boxes and the target model indicates the weights of these particles.

The measure between two color distributions $p(u)$ and $q(u)$ is the Bhattacharyya coefficient which is defined as




(a) (b)

Figure 2: Both of the two candidate regions (yellow rectangles) highly match the target model.




(a) (b)

Figure 3: Different views provide different observations of the same object.

$\rho[p, q] = \sum_u \sqrt{p^{(u)} q^{(u)}}$. The larger $\rho$ indicates higher similarity of the two color histograms and this hypothetical region is more probable to be the target. Then, we define the distance between two distributions as $d = \sqrt{1 - \rho[p, q]}$. Dissimilar distributions result in a larger $d$. The sample weight is proportional to the observation likelihood which is written as

$$\pi = \frac{e^{-d^2/2\sigma^2}}{\sqrt{2\pi}\sigma}. \tag{2}$$

The set of particles used for human tracking represents the enclosing blocks of the human target with various positions and sizes. Since more particles provide more measurements, constructing a very large particle set will support all hypotheses. To reduce the number of particles (less computation complexity) and increase the tracking accuracy, we have two assumptions. First, we assume the motion smoothness constraint. For human object tracking, the states of targets are not supposed to be changing rapidly. Under this assumption, the sampling set only has to support a small region around the previous state so that a reduced number of particle samples are sufficient for tracking. Second, we impose a specific range constraint on the aspect ratio of the tracking object block to avoid some ambiguities. For instance, there is an ambiguous match between the target color model and either one of the two hypotheses, that is, the two rectangular boxes shown in Figures 2(a) and 2(b). Here, we apply a range constraint on the aspect ratio of the box to avoid this problem. This constraint on aspect ratio is useful for multiple human objects tracking, because they may dress in similar colors.

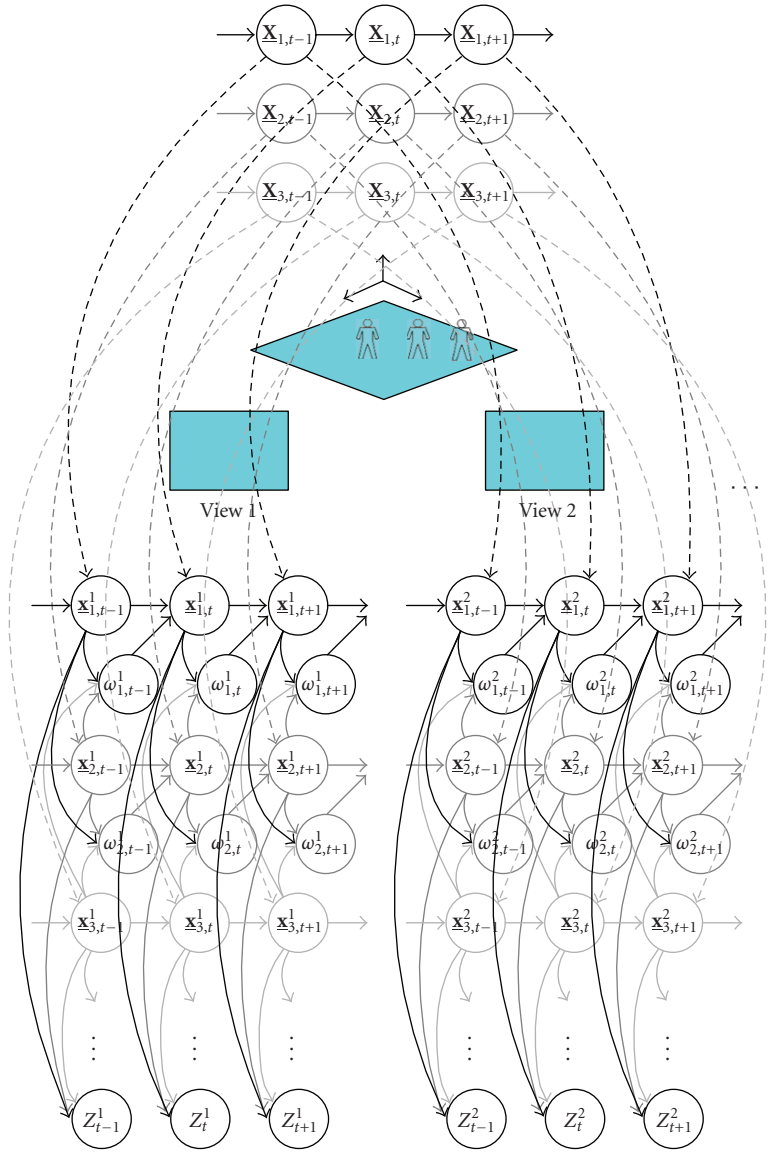



Figure 4: The state transition model.

In single view object tracking, if the color models of the two human objects are very similar, then the two tracking processes may fail when the two objects approach each other and make occlusion. However, in multiple view tracking, we do not need this restriction since the occluded object disappears in one view but appears in the other view. The tracking process in the other view may provide adequate tracking results.

Here, we allow a sampling set with variable size. Fox [22] proposed a sampling technique based on Kullback-Leibler distance (KLD) to choose a large number of samples if the uncertainty is high. Otherwise, it chooses a small number of samples. Instead, we do not insist on choosing more samples for the tracker in the occluded view, but on allowing more samples for the tracker in the visible view. With more reliable observation, the tracker will generate more accurate estimation of the target. So, the number of particles for the process in the occluded view diminishes, whereas the number of particles for the process in the visible view increases.

### 2.2. Homographic relation between two views

The purpose of applying homography for tracking is based on the assumption that each human object has its footprint on the ground. The two corresponding footprints (i.e., the bottom center of the rectangles) of the same object in two different views can be used to locate the object. To develop the homography between different views, we introduce a mechanism to correlate one image plane with the others by using the direct linear transformation (DLT) algorithm [23]. We assume that a set of points $\{\mathbf{x}_i\}$ and a corresponding set of points $\{\mathbf{x}_i'\}$ are located on two different images. The DLT method uses a set of four pairs of 2D corresponding points to estimate a $3 \times 3$ matrix $\mathbf{H}$ such that $\mathbf{H}\mathbf{x}_i = \mathbf{x}_i'$.

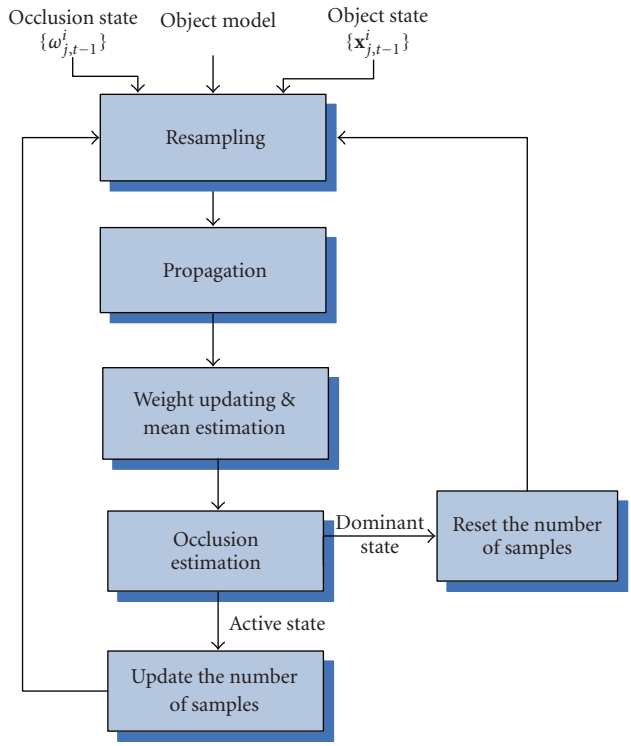



FIGURE 5: The flow diagram of tracking a target $j$ in view $i$.

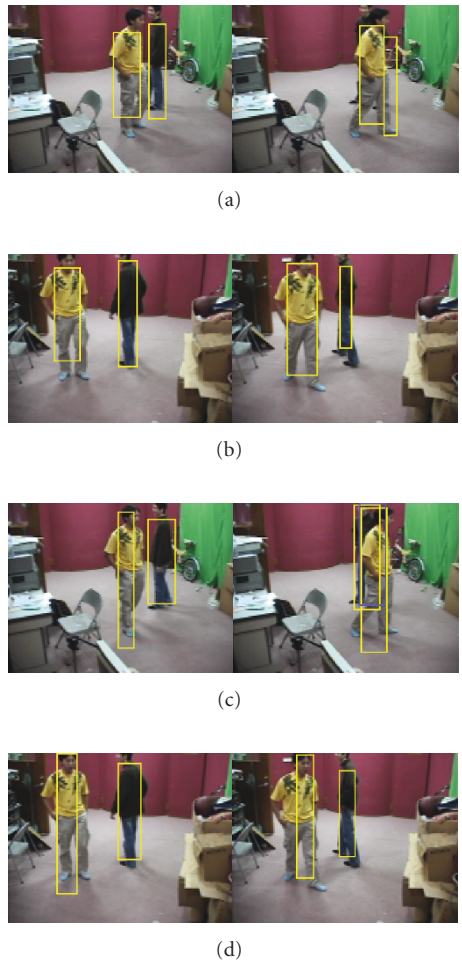



FIGURE 6: Sequence no. 1 shows the tracking of 2 persons in 2 views with 40 particles per tracker. (a) and (b) show the tracking results (yellow blocks) in two individual views by using regular PF. (c) and (d) show the results (yellow blocks) of our method. The blue block indicates the estimated $\hat{\underline{\mathbf{x}}}^i_{j,t}$, $i = 1$ or 2, which is obtained from the other view, that is, $\hat{\underline{\mathbf{x}}}^i_{j,t} = \mathbf{H}(k,i)\cdot\underline{\mathbf{x}}^k_{j,t}$, $k, i = 1$ or 2.

This basic DLT algorithm is not invariant to different choices of the image coordinates [23]. More than four correspondences do not guarantee an acceptable precision. Here, we normalize the data before performing the basic DLT algorithm, and then we obtain the homography $\mathbf{H}$ after denormalization. The normalization undoes the effect of coordinate changes and improves the accuracy of the basic DLT.

First, each point $\mathbf{x}_i$ is normalized to ensure unit-scaled homogeneous coordinate. The origin of the coordinates is then moved to the centroid of the set of points which are then normalized. All of the above procedures can be carried out by a transformation $\mathbf{T}$, written as $\tilde{\mathbf{x}} = \mathbf{T}\mathbf{x}$. We also compute a transformation $\mathbf{T}'$ for the points in the second image which map points $\mathbf{x}'$ to $\tilde{\mathbf{x}}'$. Then, we apply DLT algorithm to the two sets of points $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ to determine homography $\tilde{\mathbf{H}}$ and then obtain $\underline{\mathbf{X}}_{j,t} = \bigcup_i \underline{\mathbf{x}}^i_{j,t}$. In our experiments, we find that the mapping generated by the normalized DLT illustrates better results.

## 3. COOPERATIVE TRACKING OF MULTIPLE OBJECTS IN MULTIPLE VIEWS

The benefit of multiple view tracking is that different observations of the same target can be found in different views. In cooperative tracking, we apply a homography transformation $\mathbf{H}$ between different processes in different views to reveal the occlusion information (occluded or not) of the nearby targets. The target occluded in one view may be visible in the other view. Therefore, the tracking process for the target in the occluded view may be assisted by another process in the visible view.

Suppose we want to monitor $n$ targets at the same time, and the state of the $j$th target at time $t$ is defined as $\underline{\mathbf{X}}_{j,t} = \{\mathbf{X}_{j,t}, (H,R)_{j,t}\}$, where $\mathbf{X}_{j,t}$ indicates the position of the sample of $\underline{\mathbf{X}}_{j,t}$ and $(H,R)_{j,t}$ denotes the dimension of the sample. Obviously, our goal is to estimate the states of all targets, $\underline{\mathbf{X}}_t = \{\underline{\mathbf{X}}_{1,t}, \underline{\mathbf{X}}_{2,t}, \ldots, \underline{\mathbf{X}}_{n,t}\}$. The tracking process is treated as the density propagation from $p(\underline{\mathbf{X}}_{t-1} \mid \mathbf{Z}_{t-1})$ to $p(\underline{\mathbf{X}}_t \mid \mathbf{Z}_t)$ governed by the dynamic model $p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1})$ and the observation model $p(\mathbf{Z}_t \mid \underline{\mathbf{X}}_t)$ as

$$p(\underline{\mathbf{X}}_t \mid \mathbf{Z}_t) = p(\mathbf{Z}_t \mid \underline{\mathbf{X}}_t) \int p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1})\, p(\underline{\mathbf{X}}_{t-1} \mid \mathbf{Z}_{t-1})\, d\underline{\mathbf{X}}_{t-1}. \tag{3}$$

The motion activities of the $n$ targets are independent, and there are $n$ independent processes, that is, $\underline{\mathbf{X}}_t = \bigcup_j \underline{\mathbf{X}}_{j,t}$, $j = 1, \ldots, n$, developed for tracking. For target $j$, we may

also decompose the process of $\underline{\mathbf{X}}_{j,t}$ into $m$ view-dependent processes distributed in $m$ views (i.e., $\underline{\mathbf{X}}_{j,t} = \bigcup_i \underline{\mathbf{X}}^i_{j,t}$). The tracking can be modified as the density propagation from $p(\underline{\mathbf{x}}^i_{j,t-1} \mid Z^i_{t-1})$ to $p(\underline{\mathbf{x}}^i_{j,t} \mid Z^i_t)$, which is governed by the dynamic model $p(\underline{\mathbf{x}}^i_{j,t} \mid \underline{\mathbf{x}}^i_{j,t}\underline{\mathbf{x}}^i_{j,t-1})$ and the observation likelihood function $p(Z^i_t \mid \underline{\mathbf{x}}^i_{j,t})$. We define $\underline{\mathbf{x}}^i_{j,t} = \{\mathbf{x}^i_{j,t}, (h,r)^i_{j,t}\}$, where $\mathbf{x}^i_{j,t}$ indicates the location of the random sample, and $(h,r)^i_{j,t}$ denotes the dimension of the sample. If there is no occlusion between objects in view $i$, then the tracking processes applied for each target are independent, that is, $p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1}) = \prod_j p(\underline{\mathbf{X}}_{j,t} \mid \underline{\mathbf{X}}_{j,t-1})$. In multiview environment, we apply the particle filtering to calculate posterior density $p(\underline{\mathbf{X}}_t \mid \mathbf{Z}_t)$ based on the observations $\mathbf{Z}_t = \{Z^1_t, Z^2_t, \ldots, Z^m_t\}$ in $m$ views. For each different observation of the same object, we have different processes. The processes of the same object in different views (i.e., $\underline{\mathbf{x}}^i_{j,t}$ and $\underline{\mathbf{x}}^v_{j,t}$) are related. Thus, we may manipulate these processes and observations for robust object tracking.

The view-dependent occlusion of the targets complicates the tracking processes. Since occlusion between objects may occur, we develop a sequence of hidden Markov processes to model the appearance of each object tracked in each view based on the occlusion status of object in the previous state. During the multiview tracking process, these hidden processes help the system relocate the tracking resources to trace the same object from unreliable views to reliable views. For the same object in all views, the outcomes of the corresponding tracking processes are combined to locate the individual object more precisely.

By considering the occlusion variable $\mathbf{\Omega}_t$, the tracking process can be modified as the density propagation from $p(\underline{\mathbf{X}}_{t-1}, \mathbf{\Omega}_{t-1} \mid \mathbf{Z}_{t-1})$ to $p(\underline{\mathbf{X}}_t, \mathbf{\Omega}_t \mid \mathbf{Z}_t)$ governed by the dynamic model $p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1})$ and the observation model $p(\mathbf{Z}_t \mid \underline{\mathbf{X}}_t, \mathbf{\Omega}_t)$ as

$$p(\underline{\mathbf{X}}_t, \mathbf{\Omega}_t \mid \mathbf{Z}_t) = p(\mathbf{Z}_t \mid \underline{\mathbf{X}}_t, \mathbf{\Omega}_t) \int p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1}) \times p(\mathbf{\Omega}_t \mid \mathbf{\Omega}_{t-1}) p(\underline{\mathbf{X}}_{t-1}, \mathbf{\Omega}_{t-1} \mid \mathbf{Z}_{t-1}) d\underline{\mathbf{X}}_{t-1}, \tag{4}$$

where $\underline{\mathbf{X}}_t = \bigcup_j \underline{\mathbf{X}}_{j,t}$, $\underline{\mathbf{X}}_{j,t} = \bigcup_i \underline{\mathbf{x}}^i_{j,t}$, and $p(\underline{\mathbf{X}}_t \mid \underline{\mathbf{X}}_{t-1}) = \prod_j p(\underline{\mathbf{X}}_{j,t} \mid \underline{\mathbf{X}}_{j,t-1})$. For each view $i$, there is a hidden occlusion variable $\mathbf{\Omega}^i_t$, and $\mathbf{\Omega}_t = \bigcup_i \mathbf{\Omega}^i_t$, where $i = 1, \ldots, m$. Occlusion occurs due to the unpredictable motion activities of the nearby objects so that the occlusion processes for different targets are independent, that is, $p(\mathbf{\Omega}^i_t \mid \mathbf{\Omega}^i_{t-1}) = \prod_j p(\omega^i_{j,t} \mid \omega^i_{j,t-1})$. The expectation of the hidden occlusion variable $E(\omega^i_{j,t})$ indicates the visibility of target $j$ in view $i$ at time $t$. As shown in Figure 3, the more reliable view for observing person A is view $a$; however, for person B, view $b$ is better.

For target $j$ in specific view $i$, the tracking process can be described as

$$p\left(\underline{\mathbf{x}}^i_{j,t}, \omega^i_{j,t} \mid Z^i_t\right) = p\left(Z^i_t \mid \underline{\mathbf{x}}^i_{j,t}, \omega^i_{j,t}\right) \int p\left(\underline{\mathbf{x}}^i_{j,t} \mid \underline{\mathbf{x}}^i_{j,t-1}\right) \times p\left(\omega^i_{j,t} \mid \omega^i_{j,t-1}\right) p\left(\underline{\mathbf{x}}^i_{j,t-1}, \omega^i_{j,t-1} \mid Z^i_{t-1}\right) d\underline{\mathbf{x}}^i_{j,t-1}. \tag{5}$$

The overall tracking process model is shown in Figure 4. Tracking object $j$ in view $i$ is modeled as a sequence of hidden process $\{p(\underline{\mathbf{x}}^i_{j,t} \mid \underline{\mathbf{x}}^i_{j,t-1})\}$, another sequence of hidden process $p(\omega^i_{j,t} \mid \omega^i_{j,t-1})$, and a sequence of observation likelihood $p(Z^i_t \mid \underline{\mathbf{x}}^i_{j,t}, \omega^i_{j,t})$. The hidden random variable $\omega^i_{j,t}$ containing the occlusion information of object $j$ is determined by $\underline{\mathbf{x}}^i_{j,t}$ and $\underline{\mathbf{x}}^i_{k,t|k\neq i}$. For each variable $\underline{\mathbf{x}}^i_{j,t}$, there are random samples $\{\mathbf{s}_n \mid n = 1, \ldots, \mathrm{NS}^i_{j,t}\}$, and the weight of each sample is denoted as $\pi^i_{j,t}(n)$, which is proportional to the observation likelihood defined in (2), that is, $\pi^i_{j,t}(n) \cong p(Z^i_t \mid \underline{\mathbf{x}}^i_{j,t}, \omega^i_{j,t})$.

Here, we assume that the hidden processes of $\underline{\mathbf{x}}^i_{j,t}$ and $\omega^i_{j,t}$ are related. If the samples of $\underline{\mathbf{x}}^i_{j,t}$ and $\underline{\mathbf{x}}^i_{k,t}$ are very similar, then there is a great probability of occlusion. For two objects $j$ and $k$ in view $i$, if $[\sum_n \pi^i_{j,t}(n)/\mathrm{NS}^i_j] < [\sum_n \pi^i_{k,t}(n)/\mathrm{NS}^i_k]$ and $|\mathbf{x}^i_{j,t} - \mathbf{x}^i_{k,t}| < \theta_{\mathrm{dis}}$ ($\theta_{\mathrm{dis}}$ is determined by $h$ and $r$ as $\theta_{\mathrm{dis}} \propto f(h,r)$; here we let $\theta_{\mathrm{dis}} = 0.5\,\mathrm{hr}$), then object $j$ is occluded by object $k$, and the likelihood of $\omega^i_{j,t}$ will decrease. For every two samples $\mathbf{s}_{x,j}$ and $\mathbf{s}_{x,k}$ of $\underline{\mathbf{x}}^i_{j,t}$ and $\underline{\mathbf{x}}^i_{k,t}$, we may find the normalized overlapped area of rectangles $(h,r)^i_{j,t}$ and $(h,r)^i_{k,t}$ and determine the likelihood of $\omega^i_{j,t}$ as

$$p\left(\omega^i_{j,t}\right) = \frac{\int p\left(\omega^i_{j,t} \mid \underline{\mathbf{x}}^i_{j,t}, \underline{\mathbf{x}}^i_{k,t}\right) p\left(\underline{\mathbf{x}}^i_{j,t}, \underline{\mathbf{x}}^i_{k,t}\right) d\underline{\mathbf{x}}^i_{j,t} d\underline{\mathbf{x}}^i_{k,t}}{\sum_i p\left(\omega^i_{j,t}\right)}, \tag{6}$$

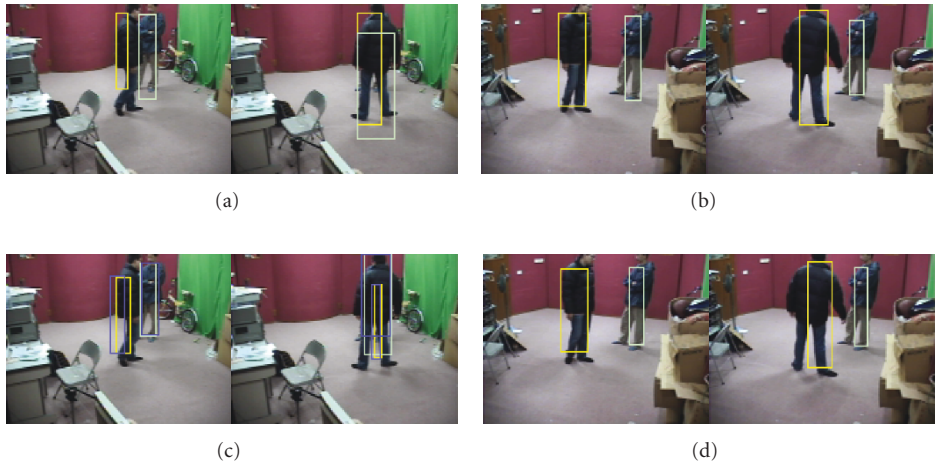
where $\underline{\mathbf{x}}^i_{k,t}$ indicates the state of the closest neighbor $k$, the $n$th sample of $p(\omega^i_{j,t} \mid \underline{\mathbf{x}}^i_{j,t}, \underline{\mathbf{x}}^i_{k,t})$ is defined as $|\pi^i_{j,t}(n) - \pi^i_{k,t}(n)| / \max(\pi^i_{j,t}(n), \pi^i_{k,t}(n))$, and the $n$th sample of $p(\underline{\mathbf{x}}^i_{j,t}, \underline{\mathbf{x}}^i_{k,t})$ is determined by the normalized overlapped area of rectangles $(h,r)^i_{j,t}$ and $(h,r)^i_{k,t}$.

For any two samples $\mathbf{s}_{xj}(n)$ and $\mathbf{s}_{xk}(n)$ of $\underline{\mathbf{x}}^i_{j,t}$ and $\underline{\mathbf{x}}^i_{k,t}$, we have the corresponding weight factors $\pi^i_{j,t}(n)$ and $\pi^i_{k,t}(n)$, as well as a corresponding sample $s_{\omega j}(n)$ for $\omega^i_{j,t}$ defined as $s_{\omega j}(n) \propto f(\pi^i_{j,t}(n), \pi^i_{k,t}(n))$. After estimating all the samples in view $i$, we have a density function of $p(\omega^i_{j,t})$ for each target $j$. By registering all density functions, $p(\omega^i_{j,t})$, we normalize the density function and discretize the variable $\omega^i_{j,t}$ into $k$ different levels. The mean value $E(\omega^i_{j,t}) = \omega^i_{j,t} p(\omega^i_{j,t})$ indicates the overall visibility of object $j$ in view $i$. For smaller $E(\omega^i_{j,t})$, the number of particle samples required for the tracking process $p(\underline{\mathbf{x}}^i_{j,t} \mid \underline{\mathbf{x}}^i_{j,t-1})$ decreases.

For each object $j$, we have two hidden variables $\mathbf{x}^i_{j,t}$ and $\omega^i_{j,t}$ in different views $i$. $\mathbf{H}(v, i)$ is the homography matrix relating the location of $\underline{\mathbf{x}}^v_{j,t}$ ($\mathbf{x}^v_{j,t}$ in view $v$) to the location of $\underline{\mathbf{x}}^i_{j,t}$ ($\mathbf{x}^i_{j,t}$ in view $i$). Based on different $\omega^i_{j,t}$, we may integrate different $\mathbf{x}^i_{j,t}$ in multiple views into the canonical (overhead) view $v$ as

$$\hat{\underline{\mathbf{x}}}^v_{j,t} = \frac{\sum_{i=1}^m E\left(\omega^i_{j,t}\right) \cdot \mathbf{H}(v,i) \cdot \underline{\mathbf{x}}^i_{j,t}}{\sum_{i=1}^m E\left(\omega^i_{j,t}\right)}, \tag{7}$$

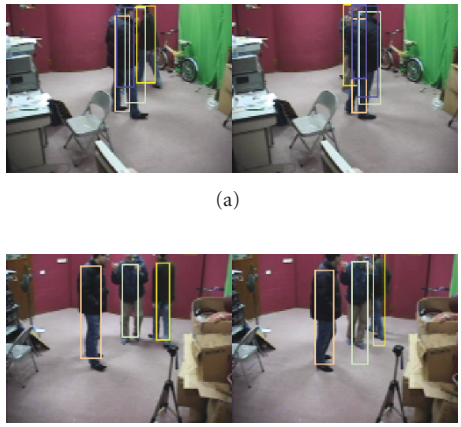
where $E(\omega^i_{j,t})$ indicates the overall average visibility of target object $j$ in view $i$. Since the tracking process of object $j$ in



(a) (b)

(c) (d)

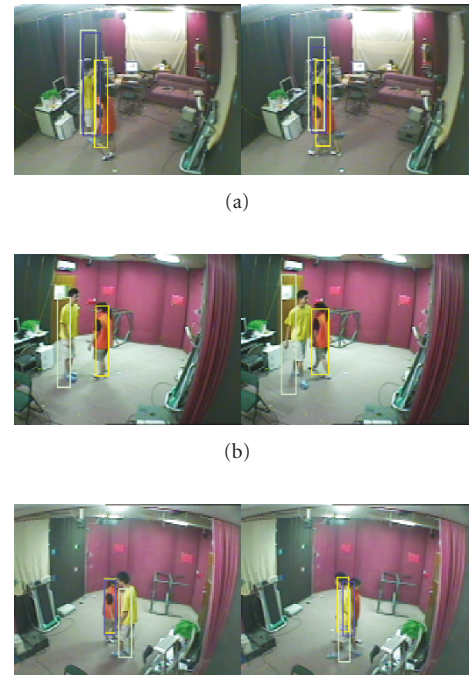
Figure 7: Tracking 2 similar persons in 2 views with 100 particles per tracker. (a) and (b) show the results (yellow blocks) of tracking in two views using regular PF. (c) and (d) show the results (yellow blocks) of our method. The blue block indicates the estimated $\hat{\underline{\mathbf{x}}}^i_{j,t}$, $i = 1$ or 2, which is obtained from the other view, that is, $\hat{\underline{\mathbf{x}}}^i_{j,t} = \mathbf{H}(k,i)\cdot\underline{\mathbf{x}}^k_{j,t}$, $k, i = 1$ or 2.



(a)

(b)

Figure 8: Tracking 3 persons dressed in similar color in two views with 100 particles per tracker. One person occludes the other two at the same time. The blue block indicates the estimated $\hat{\underline{\mathbf{x}}}^i_{j,t}$, $i = 1$ or 2, which is obtained from the other view, that is, $\hat{\underline{\mathbf{x}}}^i_{j,t} = \mathbf{H}(k,i)\cdot\underline{\mathbf{x}}^k_{j,t}$, $k, i = 1$ or 2.



(a)

(b)

(c)

Figure 9: Three-view cooperative tracking in video sequence no. 5. The blue block indicates the estimated $\hat{\underline{\mathbf{x}}}^i_{j,t}$, $i = 1, 2,$ or 3, which is obtained from the mapping of the most visible view, that is, $\hat{\underline{\mathbf{x}}}^i_{j,t} = \mathbf{H}(k,i)\cdot\underline{\mathbf{x}}^k_{j,t}$, $k, i = 1, 2,$ or 3.

view $i$ is not reliable, the number of particles in the propagation decreases. For each target, the total number of samples needed for different tracking processes in various views is fixed. The target in more reliable view will be tracked with more particle samples, whereas the target in less reliable view will be tracked with fewer samples.

In the unreliable view, as the target is partially occluded, the tracker is allocated with fewer particle samples. Once the target is completely occluded, the corresponding tracker fails and loses tracking. At any time instance, the tracking process can be in either an *active state* or a *dormant state*. In the active state, the tracking process is successful. Once the object is completely occluded, that is, $E(\omega^i_{j,t}) < \theta_{\text{occluded}}$ (i.e., $\theta_{\text{occluded}} = 0.5$), the tracking process enters the dormant state. The dormant tracking process will have fewer computing resources to do the tracking. The resampling process of $\underline{\mathbf{x}}^i_{j,t}$ is

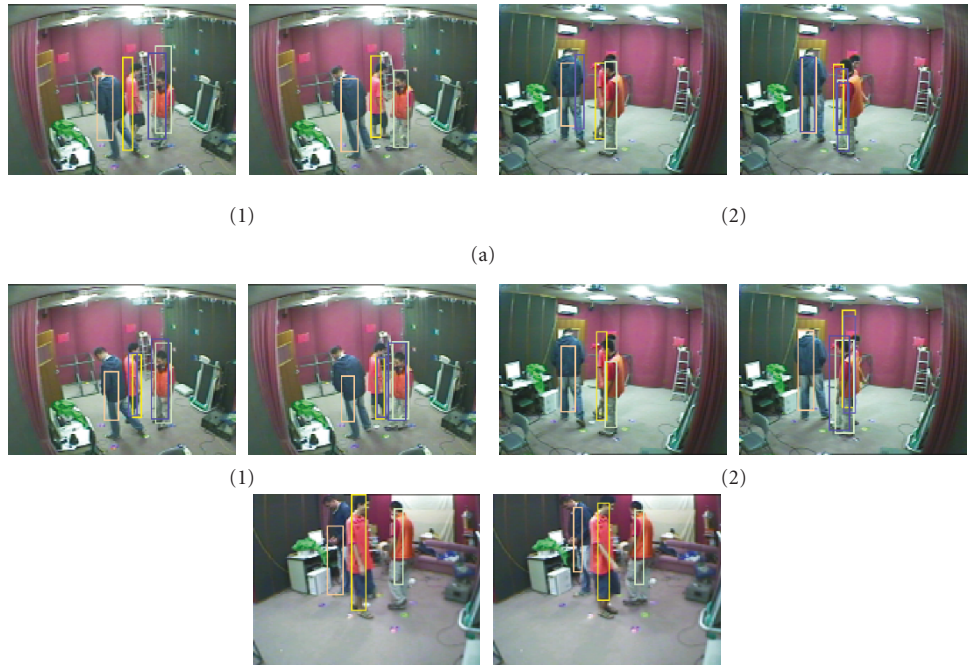



Figure 10: (a) Tracking 3 objects with two-view cooperative tracking in video sequence no. 6. The blue block indicates the estimated $\hat{\underline{\mathbf{x}}}^i_{j,t}$, $i = 1$ or 2, which is obtained from the other view, that is, $\hat{\underline{\mathbf{x}}}^i_{j,t} = \mathbf{H}(k,i)\cdot\underline{\mathbf{x}}^k_{j,t}$, $k, i = 1$ or 2. (b) Tracking 3 objects with three cameras in video sequence no. 6. The blue block indicates the estimated $\hat{\underline{\mathbf{x}}}^i_{j,t}$, $i = 1, 2,$ or 3, which is obtained from the mapping of the most visible view, that is, $\hat{\underline{\mathbf{x}}}^i_{j,t} = \mathbf{H}(k,i)\cdot\underline{\mathbf{x}}^k_{j,t}$, $k, i = 1, 2,$ or 3.

based on the samples of the other processes by using the following equation:

$$\underline{\mathbf{x}}^i_{j,t} = \frac{\sum_{k=1,k\neq i}^{m} E\left(\omega^k_{j,t}\right)\cdot\mathbf{H}(v,i)\cdot\underline{\mathbf{x}}^k_{j,t}}{\sum_{k=1,k\neq i}^{m} E\left(\omega^k_{j,t}\right)}. \tag{8}$$

After resampling, similar to the active process, the dormant process continues doing the density propagation $p(\mathbf{x}^i_{j,t} \mid \mathbf{x}^i_{j,t-1})$ and calculating $E(\omega^i_{j,t})$. The dormant tracking process has very limited tracking resource, that is, $\mathrm{NS}^i_{j,t} = N_{\mathrm{dorman}}$, to do the prediction, where $N_{\mathrm{dorman}}$ is defined as $N_{\mathrm{dorman}} = (N_{\mathrm{total}}/m)\cdot E(\omega^i_{j,t})$. The rest of the tracking resources, $\mathrm{NS}^k_{j,t} = (N_{\mathrm{total}} - N_{\mathrm{dorman}})/(m-1)$, will be assigned to the other active processes of the object $k$ $(k \neq i)$. With the information provided by the other tracking processes, it can be woken up later and enters the active state once the target becomes partially visible, that is, $E(\omega^i_{j,t}) > \theta_{\mathrm{occluded}}$. Usually, the lost target will reappear at a very different place from where it disappeared. Compared with a regular noncooperative tracking, our method will continue tracking the object after occlusion successfully.

## 4. IMPLEMENTATION

The state transition model is shown in Figure 4 in which the number of particle samples is determined by the likelihood of the hidden state $\omega^i_{j,t}$. The tracking processes monitoring the same target in different views share a fixed number of particle samples. The number of samples in each view is defined as

$$\mathrm{NS}^i_{j,t} \propto p\left(\omega^i_{j,t-1}\right), \quad \sum_i p\left(\omega^i_{j,t-1}\right) = 1, \quad \sum_i \mathrm{NS}^i_{j,t} = N, \tag{9}$$

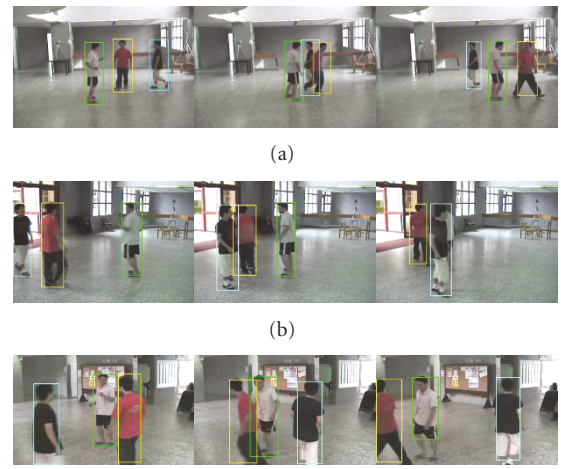
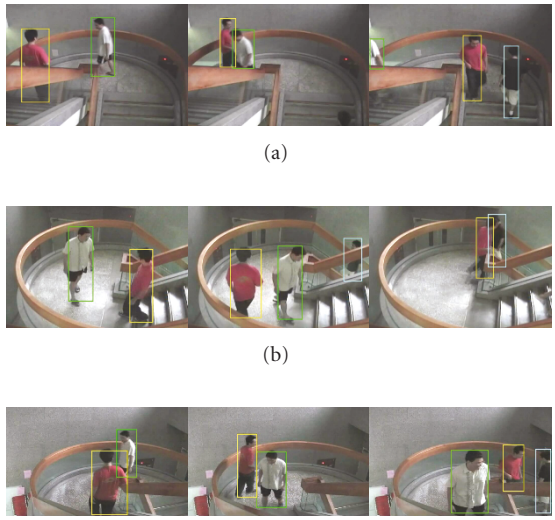



Figure 11: Tracking 3 objects with three cameras in video sequence no. 7. The different color block indicates different tracked object.



(a)

(b)

(c)

FIGURE 12: Tracking 3 objects with three cameras in video sequence no. 8. The different color block indicates the different tracked object.

where $N$ is the total number of random samples assigned to each target, and $p(\omega_{j,t}^i)$ is determined by the interaction of object $j$ and its nearest neighbors.

The cumulated sample weight also represents the confidence of tracking result. In each tracker, we define the normalized cumulated weight as $\Pi_{j,t}^i = \Sigma_{s=1}^N \pi_{j,t}^i(s)/N$, where $N$ is the total number of samples. Then (9) is rewritten as

$$\mathrm{NS}_{j,t}^i \propto \alpha \frac{\Pi_{j,t-1}^i}{\sum_i \Pi_{j,t-1}^i} + \beta E\left(\omega_{j,t-1}^i\right), \quad \alpha + \beta = 1. \tag{10}$$

Initially, we define $\alpha = \beta = 0.5$, then $\beta$ is adjusted based on the variance of $\{\omega_j^i\}$. The target with low visibility will be tracked with less number of samples. Once the occlusion occurs, we increase the weight $\beta$ for moreinfluence of $\{\omega_j^i\}$ on the number of samples, that is, $\mathrm{NS}_{j,t}^i$.

The flow diagram of the cooperative target tracking algorithm is shown in Figure 5 and summarized as follows.

### *Cooperative target tracking algorithm*

*Definitions*: (a) $\{\underline{\mathbf{x}}_{j,t}^i\}$ and $\{\omega_{j,t}^i\}$ are two sequences of hidden variables modeling the object tracking and the occlusion state of object $j$ in view $i$. (b) $N$ is the number of samples for each target. (c) $\mathrm{NS}_{j,t}^i$ is the number of current samples for tracking target $j$ in view $i$. (d) $\{\mathbf{s}_{j,t}^i\} = \{\mathbf{x}_{j,t}^i, (h,r)_{j,t}^i\}$ is the sample of the set of $\underline{\mathbf{x}}_{j,t}^i$. (e) $p(u)$ is the object model. (f) $q(u)$ is the observation model of the sample $\mathbf{s}_{j,t}^i$. (g) $\pi_{j,t}^i$ is the sample weights defined in (2).

*Inputs*: the object model $p_j(u)$ for each target $j$, the sample set $\{\mathbf{s}_{j,t-1}^i(n) \mid n = 1, \ldots, \mathrm{NS}_{j,t-1}^i\}$ of $\underline{\mathbf{x}}_{j,t-1}^i$, and the probability $\pi_{j,t-1}^i(n)$ of each sample.

*Output*: generate $\{\underline{\mathbf{x}}_{j,t}^i, \pi_{j,t}^i, \omega_{j,t}^i\}$ from $\{\underline{\mathbf{x}}_{j,t-1}^i, \pi_{j,t-1}^i, \omega_{j,t-1}^i\}$.

(1) *Resample* $\{\mathbf{s}_{j,t-1}^i\}$ with $\mathrm{NS}_{j,t-1}^i$ samples and probability $\{\pi_{j,t-1}^i\}$.

(a) Calculate the normalized cumulative probabilities as

$$c_{j,t-1}^i(n) = c_{j,t-1}^i(n-1) + \pi_{j,t-1}^i(n),$$
$$c_{j,t-1}^{\prime i}(n) = \frac{c_{j,t-1}^i(n)}{c_{j,t-1}^i(\mathrm{NS}_{j,t-1}^i)}. \tag{11}$$

(b) Selectively resample $\{\mathbf{s}_{j,t-1}^i\}$ by randomly drawing $m$ samples of which the cumulative probabilities $\{c_{j,t-1}^{\prime i}(n),\ n = 1, \ldots, m\}$ are larger than $\{c_{j,t-1}^{\prime i}(n),\ n = 1, \ldots, \mathrm{NS}_{j,t-1}^i - m\}$, and extend the $m$ samples to $\mathrm{NS}_{j,t-1}^i$ samples as $\{\mathbf{s}_{j,t-1}^{\prime i}\}$.

(2) *Propagate* the new sample set $\{\mathbf{s}_{j,t}^{\prime i}\}$ based on the density propagation $p(\mathbf{x}_{j,t}^i \mid \mathbf{x}_{j,t-1}^i)$ governed by the random walk as $\mathbf{s}_{j,t}^i(n) = \mathbf{s}_{j,t-1}^{\prime i}(n) + \mathbf{w}_t(n)$, $n = 1, \ldots, \mathrm{NS}_{j,t-1}^i$, where $\mathbf{w}_t(n)$ is a multivariate Gaussian random variable.

(3) *Correction*: calculate the observation of each sample as $q(u)$ and update the weight of each particle of the sample set $\{\mathbf{s}_{j,t}^i\}$ as $\pi_{j,t}^i(n) = p(Z_t^i \mid \mathbf{x}_{j,t}^i = \mathbf{s}_{j,t}^i(n))$ using (2).

(4) *Estimate* the mean state of the sample set $\{\mathbf{s}_{j,t}^i\}$ as $E(\mathbf{s}_{j,t}^i) = \sum_n \pi_{j,t}^i(n)\mathbf{s}_{j,t}^i(n)$.

(5) *Canonical view estimation*: calculate $\hat{\underline{\mathbf{x}}}_{j,t}^v$ by using (7).

(6) *Occlusion estimation*: calculate $p(\omega_{j,t}^i)$ using (6) and find $E(\omega_{j,t}^i)$.

(7) *Dormant state*: if $E(\omega_{j,t}^i) < \theta_{\text{occluded}}$, enter the dormant state with $\mathrm{NS}_{j,t}^i = N_{\text{dormant}}$, and go to step (1).

(8) *Active state*: if $E(\omega_{j,t}^i) > \theta_{\text{occluded}}$, enter the active state and update the number of samples $\mathrm{NS}_{j,t}^i$ by using (10), then go to step (1).

In the estimation and updating process, we need to integrate intercamera information and use homography $\mathbf{H}(g, i)$ to transform the variables $\omega_{j,t}^i$ and $\mathbf{x}_{j,t}^i$ onto a reference plane $g$ that represents the birds-eye view of the ground. This ground plane is intuitionally used to show the trajectories of moving objects without occlusion in the birds-eye view. We rewrite (7) as

$$\mathbf{x}_{j,t}^g = \frac{\sum_{i=1}^m p\left(\omega_{j,t}^i\right) \cdot \mathbf{H}(g,i) \cdot \mathbf{x}_{j,t}^i}{\sum_{i=1}^m p\left(\omega_{j,t}^i\right)}, \tag{12}$$

where $g$ indicates the ground plane and $\mathbf{H}(g,i)$ transforms the target position from view $i$ to the birds-eye view $g$. Besides, as the number of views increases, it is more convenient to calculate $2m$ homographic matrices rather than $m \times (m-1) = m^2 - m$ homographic matrices.

Table 1: Comparison of the regular PF and cooperative PF.

| | | Particle no. | MOTP | $\overline{m}$ | $\overline{\text{fp}}$ | $\overline{\text{mme}}$ | MOTA |
|---|---|---|---|---|---|---|---|
| Sequence no. 1 (Figure 6) (2 objects) | Cooperative PF | 80 | 240 mm | 6.1% | 5.2% | 0% | 88.7% |
| | Regular PF | 600 | 235 mm | 7.1% | 5.4% | 0% | 87.5% |
| | Regular PF | 300 | 267 mm | 8.3% | 7.1% | 0% | 84.6% |
| | Regular PF | 100 | 295 mm | 55.6% | 48.4% | 12.2% | −15.8% |
| | Regular PF | 80 | 363 mm | 62.2% | 54.2% | 15.1% | −31.5% |
| Sequence no. 2 (Figure 7) (2 objects) | Cooperative PF | 100 | 220 mm | 4.6% | 4.2% | 0% | 91.2% |
| | Regular PF | 600 | 268 mm | 11.2% | 10.1% | 0% | 78.7% |
| | Regular PF | 400 | 330 mm | 58.1% | 45.5% | 11.6% | −15.2% |
| | Regular PF | 300 | 340 mm | 59.2% | 57.3% | 9.2% | −25.7% |
| | Regular PF | 100 | 358 mm | 76.4% | 69.6% | 13.3% | −39.3% |
| Sequence no. 3 (Figure 8) (3 objects) | Cooperative PF | 100 | 189 mm | 8.1% | 6.3% | 0% | 85.4% |
| | Regular PF | 200 | 224 mm | 7.5% | 8.3% | 0% | 84.2% |
| | Regular PF | 160 | 365 mm | 59.2% | 46.6% | 10.6% | −16.4% |
| | Regular PF | 100 | 389 mm | 58.6% | 56.3% | 8.2% | −23.1% |
| Sequence no. 4 (3 objects) | Cooperative PF | 200 | 194 mm | 9.5% | 8.4% | 0% | 82.3% |
| | Regular PF | 600 | 237 mm | 60.9% | 56.6% | 12.8% | −28.3% |
| | Regular PF | 400 | 287 mm | 63.9% | 56.3% | 10.2% | −30.4% |
| | Regular PF | 300 | 367 mm | 68.9% | 65.4% | 13.3% | −44.6% |
| | Regular PF | 200 | 389 mm | 72.9% | 68.4% | 14.4% | −52.7% |

## 5. EXPERIMENTAL RESULTS

In the experiments, we use a 4-channel Winnov Videum video card to capture the synchronized video sequences by using three CCD cameras fixed on the ceiling. We put some markers on the ground for camera calibration by using the DLT. The color image frame resolution is $160 \times 120$ and the frame rate is 30 Hz. In the experiments, the indoor scene of multiple human objects walking in different paths is captured in different views. With eight different video sequences, we test our method and compare it with the noncooperative regular tracking.

As shown in Figure 6 (video sequence no. 1 with 3600 frames), once a long occlusion occurs, our method can continue tracking, whereas the regular noncooperative tracking may lose the target. Figure 6(a) shows that when the target is occluded in view 1, the tracking process fails. The samples located around the previous mean position are used for estimating the occluded target. They may not match the real target. In Figure 6(b), the target may be lost when occlusion occurs in view 2. Once the two targets approach each other, the number of particles for tracking the partially occluded object (human object dressed in black) in view 1 decreases, whereas the number of particles for tracking the same target in view 2 increases. The tracking process for the more reliable view is granted more computation resources to trace the target. Once the view becomes unreliable, the weighted tracking results from the other views will be used as a new position to reinitialize the tracker.

The comparisons of tracking three image sequences using regular particle filtering (PF) algorithm and ours are shown in Table 1. We apply a different number of particle samples per target to trace the target after occlusions. When target is partially occluded, the tracking process still can track the object until it is completely occluded. When the target is completely occluded, the corresponding tracking process enters the dormant state, and it still continues trying to locate the object based on the information provided from the other active trackers.

The tracking results are evaluated frame by frame based on the distance between the real target and the hypothesis. The frame-based evaluation counts the number of frames of successful tracking in which the targets may be completely visible, partially occluded, or completely occluded. To evaluate the performance for each testing video sequence, we adopt the metric proposed in [24]. Two metrics employed are the multiple object tracking precision (MOTP) and the multiple object tracking accuracy (MOTA). The former (i.e., MOTP) is the total position error for matched object hypothesis pairs over all frames, averaged by the total number of matches, which is defined as follows:

$$\text{MOTP} = \frac{\Sigma_{i,t} d_{i,t}}{\Sigma_t c_t}, \tag{13}$$

where $c_t$ is the number of matches found in time $t$. The match at time $t$ is found when the distance between the object $o_i$ and its corresponding hypothesis $d_{i,t}$ is less than certain threshold, $d_{i,t} < T$. If $d_{i,t} > T$, then it is a mismatch. We may estimate the position of human object on the ground using (12); $d_{i,t}$ is measured in *minimeter* (*mm*) and the threshold $T = 500$ mm. MOTP shows the ability of the tracker to estimate precise object positions, independent of its skill at recognizing object configurations, keeping

Table 2: Tracking the objects with 100 particles per tracker.

| | | MOTP | $\overline{m}$ (%) | $\overline{\text{fp}}$ (%) | $\overline{\text{mme}}$ (%) | MOTA |
|---|---|---|---|---|---|---|
| Sequence no. 5 (Figure 9) (2 objects) | PF in single view | 194 mm | 7.3% | 6.1% | 0% | 86.6% |
| | Cooperative PF (2 views) | 210 mm | 7.3% | 5.5% | 0% | 87.2% |
| | Cooperative PF (3 views) | 220 mm | 6.3% | 5.3% | 0% | 88.4% |
| Sequence no. 6 (Figure 10) (3 objects) | PF in single view | 367 mm | 27.3% | 28.1% | 0% | 44.5% |
| | Cooperative PF (2 views) | 224 mm | 9.8% | 9.1% | 0% | 81.1% |
| | Cooperative PF (3 views) | 189 mm | 6.3% | 5.2% | 0% | 88.5% |

Table 3: Tracking the objects with 80 particles per tracker.

| | | MOTP | $\overline{m}$ (%) | $\overline{\text{fp}}$ (%) | $\overline{\text{mme}}$ (%) | MOTA |
|---|---|---|---|---|---|---|
| Sequence no. 7 (Figure 11) (3 objects) | PF in single view | 232 mm | 20.3% | 18.1% | 0% | 61.6% |
| | Cooperative PF (3 views) | 192 mm | 7.4% | 6.2% | 0% | 86.4% |
| Sequence no. 8 (Figure 12) (3 objects) | PF in single view | 325 mm | 12.3% | 11.2% | 0% | 76.5% |
| | Cooperative PF (3 views) | 204 mm | 9.7% | 7.8% | 0% | 82.5% |

consistent trajectories, and so forth. The multiple object tracking accuracy (MOTA) is defined as

$$\text{MOTA} = 1 - \left(\frac{\Sigma_t m_t + \text{fp}_t + \text{mme}_t}{\Sigma_t g_t}\right), \tag{14}$$

where $m_t, \text{fp}_t$ and $\text{mme}_t$ are the number of misses of false positives and mismatches, respectively, for time $t$. $g_t$ denotes the number of objects at time $t$. The MOTA is composed of 3 error ratios in the sequence: (1) the ratio of misses, $\overline{m} = \Sigma_t m_t / \Sigma_t g_t$, (2) the ratio of false positives, $\overline{\text{fp}} = \Sigma_t \text{fp}_t / \Sigma_t g_t$, and (3) the ratio of mismatches, $\overline{\text{mme}} = \Sigma_t \text{mme}_t / \Sigma_t g_t$, computed over the total number of objects $\Sigma_t g_t$ presented in all frames. The MOTA accounts for all object configuration errors made by the tracker over all frames.

In video sequence no. 1, there are two persons making five occlusions. Each time after a long occlusion, the occluded person changes moving direction. After the occlusion, the tracking process may either resume tracking or lose the target. If it has lost tracking the target, we will consider it a failure. It was not until we had assigned at least 300 particle samples for the tracking process that the regular PF could maintain tracking the targets after all occlusions.

In Figure 7 (video sequence no. 2 with 3600 frames), after a long occlusion, the regular PF lost one of the separating objects, because they are dressed in similar color, as shown in Figure 7(a). Because the number of particles is not sufficient (less than 400), the tracking process is trapped in a local minimum. Comparing Figures 7(a) and 7(c), our method can track the two separating objects after occlusion effectively by using a fewer number of particles (see Table 1).

In Figure 8 (video sequence no. 3 with 3600 frames), three human objects follow a regular moving pattern. Two persons circle around the third one and make these three human objects in the same projection line of sight in one of the two views. At that moment, it looks like there is only one person appearing in that specific view. As shown in Table 1, the regular PF tracker requires twice the number of particles to maintain successful tracking in this scenario.

In video sequence no. 4 (with 4200 frames), there are three persons dressed in similar color with 3 short occlusions and 1 long occlusion. The regular PF lost tracking the occluded target after it reappeared. In the experiments, we find that allowing more particles for the regular PF does not guarantee better tracking results. The tracking process often fails after a long occlusion. The regular PF cannot find the new location of the reappearing object that is quite different from where it disappeared. Expanding the size of particle set does not work since it only increases the possibility of retracing the target only if it is still in the neighborhood. Our method will continue tracking the target which is completely occluded and which will then appear in a totally different place.

Figure 9 (video sequence no. 5 with 4500 frames) shows the tracking results of three cooperative views with many occurrences of occlusions. In the beginning, occlusion occurs in two views. The cooperative tracking can still locate the objects occluded simultaneously in view 1 and view 3. The two trackers will continue tracking after the targets reappear. The third column shows that these two trackers help the third tracker to locate the object occluded in the 2nd view. Finally, all of the three views become clear and all trackers continue tracking the object precisely.

Figure 10 (video sequence no. 6 with 4500 frames) shows the comparison of the two-view-based and three-view-based cooperative tracking. In Figure 10(a), the two-view-based cooperative tracking cannot track the object when it is occluded in both views at the same time. In Figure 10(b), we add the 3rd view to provide more reliable information for tracking the object. Hence, the tracker in the 3rd view contributes much more reliable tracking information for the other two views to make the tracking continue after the occlusion. The precisions of the tracking results are shown in Table 2.

Figure 11 (video sequence no. 7 with 8000 frames) shows three-view-based cooperative tracking of the human objects

moving in the lobby of EECS building. The number of particles assigned to each tracker is 80. Figure 12 (video sequence no. 8 with 6000 frames) shows another three-view-based cooperative tracking in another scene of the EECS building. In Figures 11 and 12, we show three-view cooperative tracking which provides more reliable information for tracking the object. Hence, the tracker in the 3rd view contributes much more reliable tracking information for the other two views to make the tracking continue after the occlusion. The precisions of the tracking results are listed in Table 3.

In the above experiments, by properly distributing the tracking resource, we may avoid the risk of losing the object after occlusion. We find that using more cameras provides more chances of obtaining clear observation, and the system will have more flexibility in dealing with the occlusions. However, there are still some problems for multiple view tracking. For example, the color distribution of the same object may be different in different views because of the changes in lighting of the environments and the quality of CCD sensors. This can be solved by developing adaptive multiple target models for tracking the same object in different views.

## 6. CONCLUSIONS

We have developed a cooperative tracking model by integrating tracking results across views and applying a sequence of hidden processes containing human interaction information. This hidden information reveals the instability of the trackers. As a result, we then allocate computations among multiple views efficiently. In the experiments, we have proved that our cooperative tracking system is more effective than the regular noncooperative tracking system.

## ACKNOWLEDGMENT

This work is finically supported by National Science Foundation under Project NSC 95-2221-E-007-053-MY3.

## REFERENCES

[1] G. L. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, "Active video-based surveillance systems: the low-level image and video processing techniques needed for implementation," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 25–37, 2005.

[2] M. Isard and A. Blake, "Condensation—conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[3] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proceedings of the 7th European Conference on Computer Vision (ECCV '02)*, vol. 2350 of *Lecture Notes in Computer Science*, pp. 661–675, Copenhagen, Denmark, May 2002.

[4] K. Nummiaro, E. Koller-Meier, and L. Van Gool, "An adaptive color-based particle filter," *Image and Vision Computing*, vol. 21, no. 1, pp. 99–110, 2003.

[5] J. MacCormick and A. Blake, "Probabilistic exclusion principle for tracking multiple objects," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.

[6] Y. Wu, T. Yu, and G. Hua, "Tracking appearances with occlusions," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 1, pp. 789–795, Madison, Wis, USA, June 2003.

[7] M. Hu, W. Hu, and T. Tan, "Tracking people through occlusions," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 2, pp. 724–727, Cambridge, UK, August 2004.

[8] S. K. Zhou, R. Chellappa, and B. Moghaddam, "Visual tracking and recognition using appearance-adaptive models in particle filters," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1491–1506, 2004.

[9] S. Kang, B.-W. Hwang, and S.-W. Lee, "Multiple people tracking based on temporal color feature," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 17, no. 6, pp. 931–949, 2003.

[10] Y. B. Shalom, *Multitarget-Multisensor Tracking: Advanced Applications*, Artech House, Norwood, Mass, USA, 1990.

[11] M. Han, W. Xu, H. Tao, and Y. Gong, "An algorithm for multiple object trajectory tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 1, pp. 864–871, Washington, DC, USA, June 2004.

[12] A. Y. S. Chia, W. Huang, and Y. Li, "Multiple objects tracking with multiple hypotheses graph representation," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 1, pp. 638–641, Hong Kong, August 2006.

[13] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1805–1819, 2005.

[14] T. Matshuyama and N. Ukaita, "Real-time multitaget tracking by a cooperative distributed vision system," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1136–1150, 2002.

[15] A. Utsumi, H. Mori, J. Ohya, and M. Yachida, "Multiple-view-based tracking of multiple humans," in *Proceedings of the 14th International Conference on Pattern Recognition (ICPR '98)*, vol. 1, pp. 597–601, Brisbane, Qld., Australia, August 1998.

[16] T.-H. Chang and S. Gong, "Tracking multiple people with a multi-camera system," in *Proceedings of IEEE Workshop on Multi-Object Tracking*, pp. 19–26, Vancouver, BC, Canada, July 2001.

[17] K. Otsuka and N. Mukawa, "A particle filter for tracking densely populated objects based on explicit multiview occlusion analysis," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 4, pp. 745–750, Cambridge, UK, August 2004.

[18] C. Canton-Ferrer, J. R. Casas, and M. Pardàs, "Towards a Bayesian approach to robust finding correspondences in multiple view geometry environments," in *Proceedings of the 5th International Conference on Computational Science (ICCS '05)*, vol. 3515 of *Lecture Notes in Computer Science*, pp. 281–289, Atlanta, Ga, USA, May 2005.

[19] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1355–1360, 2003.

[20] L. Snidaro and G. L. Foresti, "Sensor quality evaluation in a multi-camera system," in *Proceedings of the 8th International Conference on Information Fusion (FUSION '05)*, vol. 1, pp. 387–393, Philadelphia, Pa, USA, July 2005.

[21] A. Lopez, C. Canton-Ferrer, and J. R. Casas, "Multi-person 3D tracking with particle filters on voxels," in *Proceedings of the*

*32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, vol. 1, pp. 913–916, Honolulu, Hawaii, USA, April 2007.

[22] D. Fox, "KLD-sampling: adaptive particle filters," in *Advances in Neural Information Processing Systems (NIPS '01)*, Vancouver, BC, Canada, December 2001.

[23] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, Cambridge, UK, 2003.

[24] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in smart room environment," in *Proceedings of the 6th IEEE International Workshop on Visual Surveillance (VS '06)*, pp. 53–68, Graz, Austria, May 2006.