Proceedings of the 2007 IEEE/RSJ International
Conference on Intelligent Robots and Systems
San Diego, CA, USA, Oct 29 - Nov 2, 2007

Real-Time Multitarget Visual Tracking with an Active Camera

Cheng-Ming Huang, Chuan-Wen Lai, and Li-Chen Fu, *Fellow, IEEE*

Abstract—**This paper presents a real-time surveillance system to track multiple moving objects by controlling a pan-tilt camera platform. In order to describe the relationship between the targets and camera in this surveillance system, the input/output hidden Markov model (HMM) is applied here in the well-defined spherical camera coordinate. Since the targets are hard to be distinguished with one single camera when they are close to each other, we extend the particle filter for multitarget tracking with depth level estimate to track interacting targets. The targets overlapping each other still can be tracked in the images captured by single camera. Furthermore, an optimal camera action selection strategy is proposed to track multitarget within its limited field of view. The maximization of mutual information for the action design is formalized and implemented by the Monte Carlo method. The overall performance has been validated in the experiments of real-time tracking.**

I. Introduction

VISUAL surveillance in dynamic environment has drawn much attention nowadays. It has a wide spectrum of researches, including access control, human or vehicle detection and identification, multi-target visual tracking, detection of anomalous behaviors, crowd statistics or congestion analysis, etc. In order to construct a wide-area surveillance system economically, we utilize the limited field of view of one single camera to track multiple targets with might and main. This is the key concept and contribution in this paper. Detection and tracking of multitarget at the same time is an important issue in a wide variety of fields recently. Our goal is to develop a multitarget tracking system with real-time performance. Otherwise, the overall system, including the visual perception and camera action, is easily unstable and loses tracking [1].

For the multitarget tracking or high-dimensional estimate, the comprehensively search in the state space is computational expensive, thus making the system incapable of being real-time. The Monte Carlo method is one solution to this obstacle. By approximating the probability density function in state space with discrete samples, we can obtain the estimate from the sample set. Particle filter or sequential Monte Carlo (SMC), which is based on the Bayesian filtering framework, has been presented to estimate non-Gaussian and non-linear dynamic processes [17]. The sampling importance sampling (SIS) particle filter is also applied to visual tracking and cooperates with auxiliary information knowledge, which is well-known as ICONDENSATION algorithm [4]. Markov chain Monte Carlo (MCMC) [2] has been proven to perform excellently for drawing the particles. Single [3, 4] or multiple [2, 18] particle filters have their nature to efficiently represent multi-model distributions, but they are still insufficient to track multitarget when they overlap with each other.

There are a lot of researches about the camera tracking and active vision. In general, single camera is commanded to move the pan and tilt motors for tracking single target after the target detection and localization [6, 7]. In the visual servoing works [8], the camera mounted on a moving platform, such as the robot arm or vehicle, is controlled to follow the target through the feedback of the target information in the images. Multi-camera surveillance systems have been developed recently to increase the surveillance area for continuously tracking multiple targets of interest [10] or observe the same targets from different viewpoints [9], but the cameras in these systems are pre-located and static. Matsuyama *et al.* [1] have proposed the cooperative distributed vision system to track multitarget with multiple active cameras. However, one camera is still controlled to follow one target each time. From the literature, we can find that single camera is utilized conservatively to track single target, even when the targets crowd.

In order to design the camera moving strategy for tracking multitarget, we utilize the mutual information for evaluation of the camera action selection. In information theory, the mutual information is a quantity commonly used to measure the mutual dependence of two variables. It has been widely applied in computer vision, such as image registration [11], feature selection in machine learning [12], and viewpoint selection in recognition [13], etc. In addition, the mutual information is used to manage the sensors for tracking targets in the collaborative sensor networks [14, 15].

This paper is organized as follows. In Section II, we present the input/output HMM to model the overall surveillance system, and the corresponding SMC inference for the multitarget visual tracking is described. The sampling importance resampling (SIR) particle filter for multitarget tracking is extended with depth level estimate in Section III to track multiple interacting targets with overlap between each other. Then in Section IV, the camera action is designed by exploiting the estimate of the targets states to maximize the mutual information. Section V presents the real-time experimental results of the developed algorithms. Finally, conclusion is given in Section VI.

This research is sponsored by NSC95-2725-E-002-007-PAE and 95-EC-17-A-04-S1-054.

C. M. Huang and C. W. Lai are with the Department of Electrical Engineering, National Taiwan University, Taiwan, ROC.

L. C. Fu is with the Department of Electrical Engineering and Computer Science and Information Engineering, National Taiwan University, Taiwan, ROC (e-mail: lichen@ntu.edu.tw).

II. Problem Formulation and Bayesian Inference

In general surveillance system, the pan-tilt camera is set up at a fixed location. The camera can expand its field of view by commanding the pan and tilt motors. The target is tracked in the image plane, i.e., we observe the target motion with respect to the camera platform coordinate. In order to control the pan-tilt camera to dynamically track moving objects, we construct a well defined spherical camera platform coordinate (r,θ,ϕ) as Fig. 1. The original image coordinate is transformed into the spherical camera platform coordinate and located on the surface with $r=f$ (f : camera constant), and the center of image plane is at $(f,0,0)$. The four corners of the image are at (f,H_θ,H_ϕ), $(f,-H_\theta,H_\phi)$, $(f,-H_\theta,-H_\phi)$, and $(f,H_\theta,-H_\phi)$, respectively. Furthermore, the images, targets states, and camera action in this paper are all defined on the spherical camera platform coordinate system.

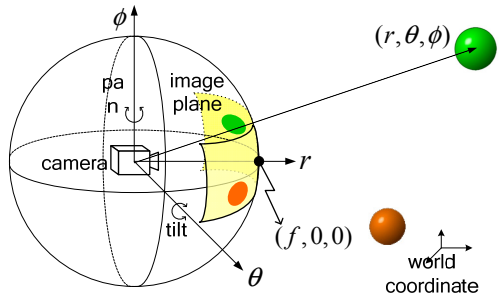


Fig. 1. Spherical camera platform coordinate.

Let the set $\mathcal{X}_t=\{\mathbf{x}_{1,t},...,\mathbf{x}_{M_t,t}\}$ combines the state vectors of current M_t targets in the spherical camera platform coordinate. The number of targets M_t, which is assumed to be known, may change over time due to the limited field of view of the camera. Each target state may contain its position and other parameters which we are interested in. Given the image observation $\mathbf{z}_t$ and the camera action $\mathbf{u}_t$ at current time t, the problem of tracking multitarget on a moving camera can be formalized by the probabilistic form as

$$p(\mathcal{X}_t|\mathbf{u}_0,\mathbf{z}_0,\mathbf{u}_1,\mathbf{z}_1,\ldots,\mathbf{u}_t,\mathbf{z}_t)=p(\mathcal{X}_t|\mathbf{u}_{0:t},\mathbf{z}_{0:t}), \quad (1)$$

where $\mathbf{u}_{0:t}$ and $\mathbf{z}_{0:t}$ are the camera action and image observations from the beginning to current time, respectively. Our goal is to estimate the states of the targets $\mathcal{X}_t$ through the evaluation of the conditional distribution.

In the general state estimation or visual tracking problem, it may be sufficient to describe the observation and state transition without control input by using the hidden Markov model (HMM). Since the targets states $\mathcal{X}_t$ are estimated from the camera's viewpoint and the designed camera control input $\mathbf{u}_t$ will also influence the targets states in the spherical camera platform coordinate, we apply the input/output HMM [16] to represent the overall surveillance system. Figure 2 shows the graphical model of the input/output HMM, which can be utilized to visualize the dependencies between the targets and active camera in the spherical camera platform coordinate.

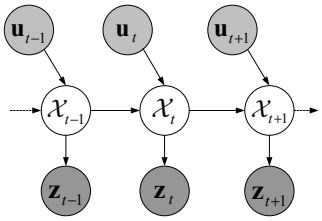


Fig. 2. Graphical model of the input/output HMM for the mutitarget visual tracking on an active camera.

Utilizing the recursive Bayesian filtering [17] and assuming that the targets states transition relative to the camera is Markov, the posterior probability (1) can be expressed as

$$p(\mathcal{X}_t|\mathbf{u}_{0:t},\mathbf{z}_{0:t}) \cong \alpha p(\mathbf{z}_t|\mathcal{X}_t) \cdot\int p(\mathcal{X}_t|\mathbf{u}_t,\mathcal{X}_{t-1})p(\mathcal{X}_{t-1}|\mathbf{u}_{0:t-1},\mathbf{z}_{0:t-1})d\mathcal{X}_{t-1}, \quad (2)$$

where α is a normalization constant, the joint likelihood $p(\mathbf{z}_t|\mathcal{X}_t)$ describes the observation measurement in the current image frame given the set of the targets, and $p(\mathcal{X}_t|\mathbf{u}_t,\mathcal{X}_{t-1})$ predicts the targets' states from the previous states with their motion models. The prediction and update from the posterior $p(\mathcal{X}_{t-1}|\mathbf{u}_{0:t-1},\mathbf{z}_{0:t-1})$ at time $t-1$ in (2) are highly related and joint between targets.

In general, the multitarget tracking problem has to solve the interaction between neighboring targets, especially when the targets with similar appearances overlap. Since the posteriors of some targets are independent, we can separate them to reduce the computational complexity. When the distance between two targets is larger than the sum of their size or a threshold, they are considered as one dependent pair. Then, the dependent pairs having the same targets will be collected as one interacting group.

After investigating the dependence of the previous estimates $\mathcal{X}_{t-1}$, we assemble the interacting targets' states as joint states $\mathbf{X}_{g,t}$, $g=1,\ldots,G$, which are the sets of targets' states in G interacting groups. Hence, the posterior in (2) is equivalent to

$$p(\mathcal{X}_t|\mathbf{u}_{0:t},\mathbf{z}_{0:t}) \cong \alpha\prod_{g=1}^{G} p(\mathbf{X}_{g,t}|\mathbf{u}_{0:t},\mathbf{z}_{0:t}). \quad (3)$$

Let $M_{g,t}$ be the numbers of the dependent targets in each interacting group, and then $M_t=\sum_{g=1}^{G}M_{g,t}$. Note that $M_{g,t}=1$ represents an independent target, and the posterior of one independent target in (3) can be efficiently evaluated by one sequential importance sampling (SIS) particle filter [4]. On the other hand, the importance functions of other overlapped targets may not be distinguishable. For the dependent terms ($M_{g,t}\neq 1$) in (3), the sampling importance resampling (SIR) particle filter [3, 17], which is resampled from the joint posterior of overlapped targets at previous time step, is employed and described in the next section.

III. MULTITARGET TRACKING WITH DEPTH ESTIMATE

In this section, we focus on the multitarget visual tracking in a certain interacting group g. When tracking multiple interacting targets in an image, one challenge is that targets may overlap with each other. To efficiently estimate each target's state, we extend the SIR particle filter with depth estimate. The depth level is the sequence which represents the relative distance from every target to the camera, as shown in Fig. 3. The target with lower depth level is located closer to the camera than the target with higher depth level. We define the depth level set of joint targets in one interacting group as $D_{g,t}$. Obviously there might be $(M_{g,t}!)$ depth level hypotheses for inferring the overlapping situation of the joint state $\mathbf{X}_{g,t}$. For example, in the case of Fig. 4, there are six depth level hypothesis events for three overlapping targets: {1,2,3}, {1,3,2}, {2,1,3}, {2,3,1}, {3,1,2}, {3,2,1}.

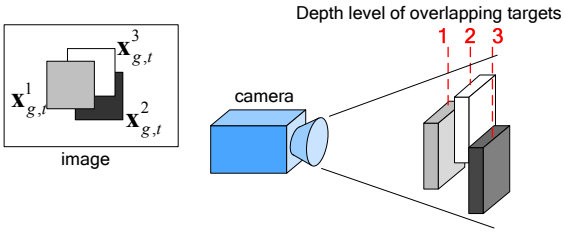


Fig. 3. The depth level definition for occlusion handling. The depth level set of this joint state $\mathbf{X}_{g,t}$ is $D_{g,t}=\{1,3,2\}$.

With this augmentation of the depth level hypotheses, the posterior of the dependent targets $p(\mathbf{X}_{g,t}\,|\,\mathbf{u}_{0:t},\mathbf{z}_{0:t})$ in (3) can be redefined as $p(D_{g,t},\mathbf{X}_{g,t}\,|\,\mathbf{u}_{0:t},\mathbf{z}_{0:t})$, and (2) can be rederived as

$$p(D_{g,t},\mathbf{X}_{g,t}\,|\,\mathbf{u}_{0:t},\mathbf{z}_{0:t}) \cong \beta p(\mathbf{z}_t\,|\,\mathbf{X}_{g,t},D_{g,t}) \cdot\sum\nolimits_{D_{g,t-1}}\int p(D_{g,t},\mathbf{X}_{g,t}\,|\,D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t) \cdot p(D_{g,t-1},\mathbf{X}_{g,t-1}\,|\,\mathbf{u}_{0:t-1},\mathbf{z}_{0:t-1})d\mathbf{X}_{g,t-1}, \quad (4)$$

where β is a normalization constant, $p(\mathbf{z}_t\,|\,\mathbf{X}_{g,t},D_{g,t})$ is the joint image likelihood [5] with the depth level concerned, and the mixed transition model $p(D_{g,t},\mathbf{X}_{g,t}\,|\,D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t)$ denotes the overlapping targets' states transition and depth level variation in the two dimensional image plane after the camera action $\mathbf{u}_t$. Using the chain rule, we can divide the mixed transition probability for the filtering update into the following two terms:

$$p(D_{g,t},\mathbf{X}_{g,t}\,|\,D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t) = p(\mathbf{X}_{g,t}\,|\,D_{g,t},D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t)p(D_{g,t}\,|\,D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t). \quad (5)$$

The first factor $p(\mathbf{X}_{g,t}\,|\,D_{g,t},D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t)$ in (5) is the joint state transition model of dependent targets relative to the camera action $\mathbf{u}_t$. It is the conditional probability given the current depth level hypothesis $D_{g,t}$ and the joint states with depth level estimates at previous time t–1. The second factor in (5) $p(D_{g,t}\,|\,D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t)$ is the depth level transition probability, and it models how the depth levels change over time based on the targets estimates at previous time t–1.

Assume that the targets are solid without any hole in the projected image frame. Observing the tracking history of dependent targets in one interacting group, we can find that the depth level of each target is not expected to have sudden change during the occlusion. Even if the depth level of each target changes, it seldom changes more than one level. Furthermore, a phenomenon can be observed that once the targets are severely overlapped, the depth level is almost impossible to change. Moreover, as shown in Fig. 4, if the targets slightly overlap in the image, then the possibility of the depth level variation of each target is comparatively high. Hence, we model the depth level transition $p(D_{g,t}\,|\,D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t)$ as the normal distribution with adaptive covariance:

$$p(D_{g,t}|D_{g,t-1},\mathbf{X}_{g,t-1},\mathbf{u}_t) \propto \prod_v \mathcal{N}\left(D_{g,t}(v);D_{g,t-1}(v),\sigma_D^2(\mathbf{X}_{g,t-1})\right), \quad (6)$$

where $D_{g,t}(v)$ is the vth element of the depth level $D_{g,t}$, i.e., the depth of the target v, $\mathcal{N}\left(D_{g,t}(v);D_{g,t-1}(v),\sigma_D^2(\mathbf{X}_{g,t-1})\right)$ is the normal distribution of the variable $D_{g,t}(v)$ with mean as the previous depth level $D_{g,t-1}(v)$, and the adaptive covariance $\sigma_D^2(\mathbf{X}_{g,t-1})$ defined as

$$\sigma_D(\mathbf{X}_{g,t-1})=\frac{\sigma_g}{M_{g,t-1}-1}\sum_{s=1,s\neq v}^{M_{g,t-1}}\left\|\mathbf{x}_{g,t-1}^v-\mathbf{x}_{g,t-1}^s\right\|, \quad (7)$$

in which σ_g is a predefined constant of covariance, $\mathbf{x}_{g,t-1}^v$ is the state of one target v at previous time t–1 in the interacting group g. If the targets in one interacting group are crowded, the covariance becomes small; otherwise the covariance is large.

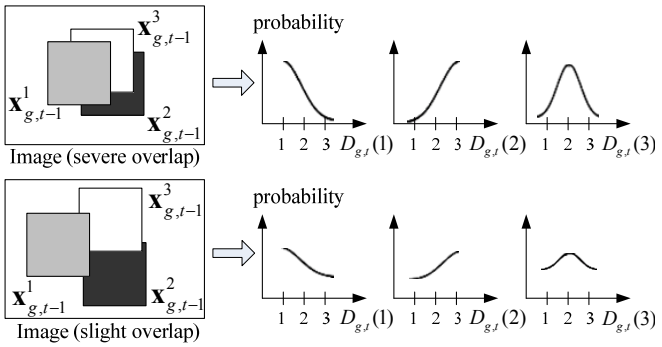


Fig. 4. The depth level transition probability is modeled as a normal distribution with adaptive covariance. Previous depth $D_{g,t-1}=\{1,3,2\}$.

Finally, the Monte Carlo approximation of (4) can then be obtained through the SIR particle filter with a set of weighted samples $\{D_{g,t}^i,\mathbf{X}_{g,t}^i,w_{g,t}^i\}_{i=1}^{N_g}$:

$$p(D_{g,t},\mathbf{X}_{g,t}\,|\,\mathbf{u}_{0:t},\mathbf{z}_{0:t}) \approx \beta\sum_{i=1}^{N_g} w_{g,t}^i\,\delta\left(\begin{bmatrix}D_{g,t}\\ \mathbf{X}_{g,t}\end{bmatrix}-\begin{bmatrix}D_{g,t}^i\\ \mathbf{X}_{g,t}^i\end{bmatrix}\right), \quad (8)$$

where the weights $w_{g,t}^i = p\left(\mathbf{z}_t \mid \mathbf{X}_{g,t}^i, D_{g,t}^i\right)$ measure the joint image likelihood [5]. Note that the multidimensional sample $[D_{g,t}^i, \mathbf{X}_{g,t}^i]^T$ is drawn as follows:

$$[D_{g,t}^i, \mathbf{X}_{g,t}^i]^T \sim \sum_{i=1}^{N_g} w_{g,t-1}^i p(D_{g,t}, \mathbf{X}_{\mathrm{g},t} \mid D_{g,t-1}^i, \mathbf{X}_{g,t-1}^i, \mathbf{u}_t), \quad (9)$$

where $\{D_{g,t-1}^i, \mathbf{X}_{g,t-1}^i, w_{g,t-1}^i\}_{i=1}^{N_g}$ are the weighted samples yielded from approximation of the posterior probability $p(D_{g,t-1}, \mathbf{X}_{g,t-1} \mid \mathbf{u}_{0:t-1}, \mathbf{z}_{0:t-1})$ at previous time t–1.

IV. Camera Action Selection

Due to the limited filed of view of camera, if we want to keep on tracking the moving targets, the pan and tilt motors of the camera platform must be commanded to move. The ideal scenario is that the camera follows the moving targets and then the targets can be continuously observed from the captured images. However, poorly designed camera action will lead to lose tracking easily. The mutual information is employed to evaluate the dependence between the future targets' states and the image observation through the camera action at time t+1, and it can be formalized as [13]:

$$I(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1}) = \int_{\mathcal{X}_{t+1}} \int_{\hat{\mathbf{z}}_{t+1}} p(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1}) \cdot \log \frac{p(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})}{p(\mathcal{X}_{t+1} \mid \mathbf{u}_{t+1}) p(\hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})} d\hat{\mathbf{z}}_{t+1} d\mathcal{X}_{t+1}, \quad (10)$$

where $\hat{\mathbf{z}}_{t+1}$ is defined to express the future image observation which is not captured yet. Since we desire to maintain the relationship between the targets and images captured by camera, the optimal camera action $\mathbf{u}_{t+1}^*$ will be selected to maximize the mutual information:

$$\mathbf{u}_{t+1}^* = \arg\max_{\mathbf{u}_{t+1} \in \mathbf{U}} I(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1}), \quad (11)$$

where $\mathbf{U}$ is the set of feasible moving steps of the camera pan and tilt motors.

From the definition of the expected value in probability theory and using the chain rule, (10) can be rewritten as

$$\begin{aligned} & I(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1}) \\ &= E_{p(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})} \left[\log \frac{p(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})}{p(\mathcal{X}_{t+1} \mid \mathbf{u}_{t+1}) p(\hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})} \right] \\ &= E_{p(\hat{\mathbf{z}}_{t+1} \mid \mathcal{X}_{t+1}, \mathbf{u}_{t+1}) p(\mathcal{X}_{t+1} \mid \mathbf{u}_{t+1})} \left[\log \frac{p(\hat{\mathbf{z}}_{t+1} \mid \mathcal{X}_{t+1}, \mathbf{u}_{t+1})}{p(\hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})} \right], \end{aligned} \quad (12)$$

where we call the term $p(\hat{\mathbf{z}}_{t+1} \mid \mathcal{X}_{t+1}, \mathbf{u}_{t+1})$ as the "camera likelihood" and the term $p(\hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})$ can be expanding by the law of total probability as

$$p(\hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1}) = \int p(\hat{\mathbf{z}}_{t+1} \mid \mathcal{X}_{t+1}, \mathbf{u}_{t+1}) p(\mathcal{X}_{t+1} \mid \mathbf{u}_{t+1}) d\mathcal{X}_{t+1}. \quad (13)$$

Since we want to utilize one camera for tracking multitarget simultaneously, the camera has to capture as most targets as possible in the limited field of view. Furthermore, the best way to continuously monitor these targets is to maintain the targets around the center of the image. Let $\mathcal{X}_{\hat{\mathbf{z}},t+1}$ be the set of targets' states within image plane $\hat{\mathbf{z}}_{t+1}$, $M_{\hat{\mathbf{z}}_{t+1}}$ is the number of these targets, and $(\mathcal{P}_{\hat{\mathbf{z}}}^{\theta}(\mathcal{X}_{\hat{\mathbf{z}},t+1}), \mathcal{P}_{\hat{\mathbf{z}}}^{\phi}(\mathcal{X}_{\hat{\mathbf{z}},t+1}))$ are the positions of these targets in the image plane. Then, the camera likelihood function is modeled as

$$p(\hat{\mathbf{z}}_{t+1} \mid \mathcal{X}_{t+1}, \mathbf{u}_{t+1}) \propto \exp\left[-c_0 \prod_{v \in \mathcal{X}_{\hat{\mathbf{z}},t+1}} \lambda_v - c_1 \left(1 - \frac{M_{\hat{\mathbf{z}}_{t+1}}}{M_t}\right) - c_2 \cdot \frac{\bar{\mathcal{P}}_{\hat{\mathbf{z}}}^{\theta}}{H_\theta} \cdot \frac{\bar{\mathcal{P}}_{\hat{\mathbf{z}}}^{\phi}}{H_\phi} \right], \quad (14)$$

where c_0, c_1 and $c_2 \in [0, 1]$ are the relative weightings, $\lambda_v \in [0, 1]$ is the user defined priority factor for each tracked target, $\mathcal{X}_{\hat{\mathbf{z}},t+1}$ is the set of the targets within the region $\hat{\mathbf{z}}$, and

$$\begin{aligned} \bar{\mathcal{P}}_{\hat{\mathbf{z}}}^{\theta} &= \left| \max\left(\mathcal{P}_{\hat{\mathbf{z}}}^{\theta}(\mathcal{X}_{\hat{\mathbf{z}},t+1})\right) + \min\left(\mathcal{P}_{\hat{\mathbf{z}}}^{\theta}(\mathcal{X}_{\hat{\mathbf{z}},t+1})\right) \right| / 2, \\ \bar{\mathcal{P}}_{\hat{\mathbf{z}}}^{\theta} &= \left| \max\left(\mathcal{P}_{\hat{\mathbf{z}}}^{\theta}(\mathcal{X}_{\hat{\mathbf{z}},t+1})\right) + \min\left(\mathcal{P}_{\hat{\mathbf{z}}}^{\theta}(\mathcal{X}_{\hat{\mathbf{z}},t+1})\right) \right| / 2. \end{aligned} \quad (15)$$

Note that if we consider one target is important, λ_v of that target can be assigned relatively smaller than others.

In order to evaluate (12) efficiently, the Monte Carlo method [15] and the weighted samples from Bayesian filtering in Section III are exploited here. First, we resample N weighted samples $\{\mathcal{X}_t^i, W_t^i\}_{i=1}^N$ from the posterior distribution $p(\mathcal{X}_t \mid \mathbf{u}_{0:t}, \mathbf{z}_{0:t})$ approximated in Section III. By the law of total probability and the assumption of input/output HMM as shown in Fig. 2, the probability distribution $p(\mathcal{X}_{t+1} \mid \mathbf{u}_{t+1})$ can then be approximated by the samples drawn from:

$$\mathcal{X}_{t+1}^i \sim \sum_{i=1}^{N} W_t^i p(\mathcal{X}_{t+1} \mid \mathbf{u}_{t+1}, \mathcal{X}_t^i), \quad (16)$$

and the corresponding weight $W_{t+1}^i = W_t^i$, where $p(\mathcal{X}_{t+1} \mid \mathbf{u}_{t+1}, \mathcal{X}_t^i)$ is the states transition model used to predict the future targets states relative to a camera action $\mathbf{u}_{t+1}$. Second, we can yield the samples $\hat{\mathbf{z}}_{t+1}^i \sim p(\hat{\mathbf{z}}_{t+1} \mid \mathcal{X}_{t+1}^i, \mathbf{u}_{t+1})$ from the camera likelihood function. Through the two sampling procedures, the set of weighted samples $\{\mathcal{X}_{t+1}^i, \hat{\mathbf{z}}_{t+1}^i, W_{t+1}^i\}_{i=1}^N$ appropriately distributed over $p(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1})$ in (12). Finally, the expected value of mutual information in (12) can be approximated by

$$I(\mathcal{X}_{t+1}, \hat{\mathbf{z}}_{t+1} \mid \mathbf{u}_{t+1}) \approx \sum_{i=1}^{N} W_{t+1}^i \cdot \log \frac{p(\hat{\mathbf{z}}_{t+1}^i \mid \mathcal{X}_{t+1}^i, \mathbf{u}_{t+1})}{p(\hat{\mathbf{z}}_{t+1}^i \mid \mathbf{u}_{t+1})}, \quad (17)$$

where the term $p(\hat{\mathbf{z}}_{t+1}^i \mid \mathcal{X}_{t+1}^i, \mathbf{u}_{t+1})$ can be measured by (14) and the term $p(\hat{\mathbf{z}}_{t+1}^i \mid \mathbf{u}_{t+1})$ can be evaluated by the law of total probability and approximated as:

$$p(\hat{\mathbf{z}}_{t+1}^i \mid \mathbf{u}_{t+1}) \approx \sum_{k=1}^{N} W_{t+1}^k p(\hat{\mathbf{z}}_{t+1}^i \mid \mathcal{X}_{t+1}^k, \mathbf{u}_{t+1}). \quad (18)$$

The set of feasible camera action $\mathbf{U}$ is decided by the precision of the pan and tilt motors and their moving range in the processing period between two consecutive image frames. Instead of exhaustive selection in the feasible camera action set $\mathbf{U}$, we could employ the coarse search followed by the fine search to reduce the computational cost. The optimal

selection $\mathbf{u}_{t+1}^{*}$ will also be taken into consideration for multitarget tracking at time t+1.

V. Experimental Results

In this section, we will test our proposed multitarget visual tracking system with the image sequences captured in real-time. The image size of all frames is 352 × 240, and the pan-tilt camera platform is the AXIS 2130 camera. The computer processor is Intel Pentium 4 2.8G Hz.

In the following experiments, the states of each target are defined as $\mathbf{x}_{m,t} = [\theta,\ \phi,\ s]^T$, where $(\theta,\ \phi)$ is the position in the spherical camera platform coordinate and s is its size. For the state transition model, we assume the predicted state is a normal distribution around the original one [2]:

$$\mathbf{x}_{m,t+1} = \mathbf{x}_{m,t} - \mathbf{u}_t + [\Delta\theta,\ \Delta\phi,\ \Delta s]^T, \tag{19}$$

where $[\Delta\theta,\ \Delta\phi,\ \Delta s]^T \sim [\mathcal{N}(0,\sigma_\theta^2),\ \mathcal{N}(0,\sigma_\phi^2),\ \mathcal{N}(0,\sigma_s^2)]^T$. The image likelihood evaluation for updating the weights of samples considers the combination of contour and template model [7]. When the targets are independent, each target is tracked by one particle filter with 100 particles. On the other hand, 1000 particles are used to estimate joint states and depth levels when the targets are interacting in one group.

A. Multitarget tracking by a static camera

First, we track two coins by a static camera to examine the performance of the multitarget tracker, i.e. $\mathbf{u}_t = 0$ in (19). The targets with homogeneous appearance are difficult to be distinguished when they get close. The color of the coin is used to detect the entering target and act as the importance function of SIS particle filter for tracking independent target.

As shown in Fig. 5, when the distance between two coins is smaller than a threshold, they are collected into one interacting group. Our proposed tracker will distinguish the targets by estimating their depth levels. Unlike the other multitarget tracking filters [2, 18], they employ the motion model that may reject the occurrence of overlap and only measure the individual likelihood. Here, the samples of SIR particle filter are predicted by the mixed transition probability of the state and depth variation. The joint image likelihood [5] is evaluated according to the image pattern of each sample generated through the prediction step. Hence, our multitarget visual tracking system can overcome the occlusion problem between dependent targets.

Note that in the frame #1100 of Fig. 5, the green and red targets overlap seriously, and the red target is almost fully occluded by the green one. The depth level transition probability with adaptive covariance will become much smaller when the targets are getting closer. Through such mechanism, we can still successfully accomplish the tracking mission and distinguish these two homogeneous targets under such circumstance.

B. Multitarget tracking with an active camera

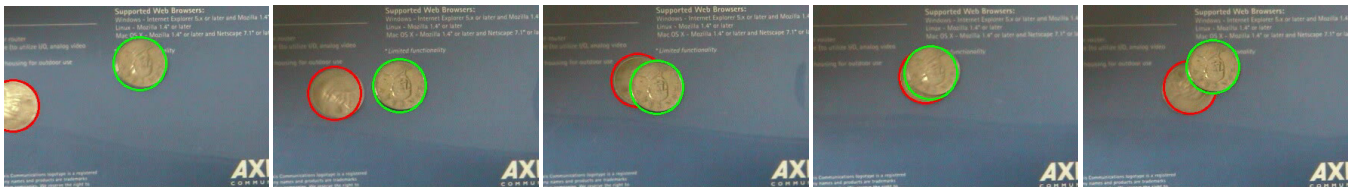
In the following experiments, we utilize the active camera platform to implement the overall proposed surveillance system and verify its performance. We define the human head as the target here. The skin color is utilized to detect the person and draw the samples for SIS particle filter. During evaluation of the mutual information, 50 particles for each target are resampled from the posterior estimation of the multitarget tracking. The red number written in the bottom left of images in Fig. 6 and 7 denotes the processing period with unit as millisecond between two consecutive frames.

Figure 6 shows the tracking results of two people. The relative weightings of the camera likelihood (14) are assigned as: $c_0 = 0$, $c_1 = 1$ and $c_2 = 0.5$. The priority factor λ_v of each target is equivalent to 1. It implies that none of the targets is especially important to be continuously monitored. These two people are successfully tracked when they cross. We can see that the camera always attempts to capture the two targets at the same time through the action design with maximizing the mutual information. From frame #115 to frame #295, these two people move apart, and the camera gradually can not observe them with its limited field view. At that moment, the system must decide to track target 1 or 2. Since the two choices have the same camera likelihood in (14), the optimal camera action will be determined by the weight W_{t+1}^i in (17). The target with more reliable posterior estimate would have higher image likelihood weight than that of the other. Therefore, the mutual information will be maximized by the camera action that can observe the more credible target, which is labeled as 2. After the camera action is selected, the system decreases the total number of tracked targets M_{t+1}.

We track more people in the experiment shown in Fig. 7. In this scenario, we appoint the person labeled 4 as an important target which needs to be continuously monitored. The relative weightings in (14) are assigned as: $c_0 = 1$, $c_1 = 1$ and $c_2 = 0.3$, and the priority factor of target 4 is set to 0. Although the target 4 is always the leftmost one in the images, it is still covered in the camera's field of view after target 2 and 3 have walked away. The experimental results demonstrate that the active camera successfully utilizes its limited field of view to track multiple moving targets at the same time. On the other hand, we can find that the center of the image may not always locate at desired position, because the feasible camera moving range is finite and we approximate the evaluations by samples to achieve real-time performance.

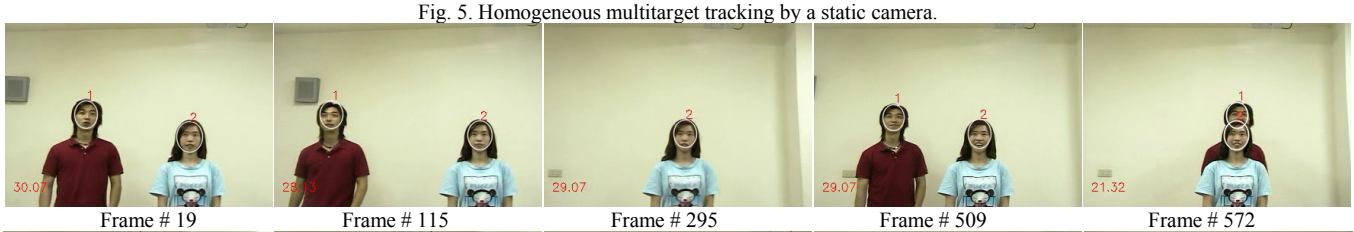
VI. Conclusion

In this paper, we proposed a real-time multitarget visual tracking system with an active camera. The camera is not only assigned to track single target, and its limited field of view is fully utilized to track a crowd of targets simultaneously. The input/output HMM is employed to model the overall surveillance system in the spherical camera platform coordinate. For tracking multiple dependent targets with overlapping between each other, the SIR particle filter is extended with depth estimate to resolve this obstacle.



Frame # 850 Frame # 910 Frame # 1070 Frame # 1100 Frame # 1120

Fig. 5. Homogeneous multitarget tracking by a static camera.



Frame # 19 Frame # 115 Frame # 295 Frame # 509 Frame # 572

Frame # 611 Frame # 646 Frame # 741 Frame # 761 Frame # 806

Fig. 6. Two people tracking by an active camera.

Frame # 1 Frame # 16 Frame # 31 Frame # 68 Frame # 146

Fig. 7. Multiple people tracking by an active camera. The person labeled 4 is set to be continuously monitored.

Furthermore, exploiting the results of multitarget tracking, the pan-tilt camera moving strategy is developed. The mutual information between the future prediction and observation is maximized to select the optimal camera action, and it is implemented with the Monte Carlo method.

In the future, we will consider a more complicated joint state transition model of dependent targets and apply MCMC to draw samples of the particle filter more efficiently. The cooperation of multiple active cameras will also be constructed to achieve multitarget seamless tracking.

REFERENCES

[1] T. Matsuyama and N. Ukita, "Real-time multitarget tracking by a cooperative distributed vision system," *Proc. of IEEE*, vol. 90, pp. 1136-1150, 2002.

[2] K. Zia, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1805-1819, 2005.

[3] M. Isard and A. Blake, "CONDENSATION - Conditional density propagation for visual tracking," *Int. J. Computer Vision*, vol. 29, pp. 5-28, 1998.

[4] M. Isard and A. Blake, "ICONDENSATION: Unifying low level and high level tracking in a stochastic framework," *Proc. 5th European Conf. Computer Vision*, Freiburg, Germany, 1998.

[5] C. Rasmussen and G. D. Hager, "Probabilistic data association methods for tracking complex visual objects," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 560-576, 2001.

[6] D. Murray and A. Basu, "Motion tracking with an active camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, pp. 449-459, 1994.

[7] P. Y. Chen, C. M. Huang and L. C. Fu, "A robust visual servo system for tracking an arbitrary-shaped object by a new active contour method," *American Control Conf.*, vol. 2, pp. 1516-1521, 2004.

[8] F. Chaumette and S. Hutchinson, "Visual servo control. I. Basic approaches," *IEEE Robotics and Automation Magazine*, vol. 13, pp. 82-90, 2006.

[9] Q. Cai and J. K. Aggarwal, "Tracking human motion in structured environments using a distributed-camera system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, pp. 1241-1247, 1999.

[10] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, pp. 1355-1360, 2003.

[11] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients," *IEEE Trans. Medical Image*, vol. 23, pp. 1430-1444, 2004.

[12] H. C. Peng, F. H. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 1226-1238, 2005.

[13] J. Denzler and C. M. Brown, "Information theoretic sensor data selection for active object recognition and state estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 145-157, 2002.

[14] T. Vercauteren, D. Guo and X. Wang, "Joint multiple target tracking and classification in collaborative sensor networks," *IEEE J. Selected Areas in Communications*, vol. 23, pp. 714-723, 2005.

[15] A. Doucet, B. Vo, C. Andrieu, and M. Davy, "Particle filtering for multitarget tracking and sensor management," *Int. Conf. Information Fusion*, vol. 1, pp. 474-481, 2002.

[16] K. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," PhD. thesis, Dept. Computer Science, UC Berkeley, 2002.

[17] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Processing*, vol. 50, pp. 174-188, 2002.

[18] T. Yu and Y. Wu, "Collaborative tracking of multiple targets," *IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. I-834-I-841, 2004.