2009 Ninth International Conference on Intelligent Systems Design and Applications

Intelligent Cooperative Tracking in Multi-Camera Systems

Yan Lu, and Shahram Payandeh,

***Abstract*— In this paper, an approach for intelligent integration of indoor visual tracking system for event detection and movement is proposed. This surveillance system is composed of a stationary camera and a pan tilt zoom (PTZ) camera, where the two cameras have been intrinsically and extrinsically calibrated. The stationary camera detects events such as fall and wandering using motion-based visual tracking. In this initial study, the PTZ camera tracks and follows the person who triggered the event using intelligent color-based particle filtering which is defined based on the expected dynamics of the scene. The purpose of tracking in view of the PTZ camera is to continuously keep the person in the full view of the camera which can further be processed for identifying details of the person. Preliminary experimental results for camera calibration, event detection, and human tracking are presented to demonstrate the performance of the proposed cooperative hybrid visual tracking system.**

I. Introduction

Configurations of camera systems for people surveillance on how the cameras are deployed to track targets have been topics of various research and development which involve expanding the capacity of cameras and increasing the number of cameras. For example, some initial studies stationary cameras were mounted on mobile platforms in order to enhance the coverage and the flexibility of robot vision [1] [2]. Later on, stationary cameras were substituted by active cameras of PTZ capabilities in order to facilitate the collection of scenes. Lalonde et al. [3] has proposed a surveillance system to automatically track humans and vehicles using a PTZ camera. The flexible perspective and resolution of live video recorded by a single PTZ camera reduces the number of stationary cameras that would otherwise be used to build up an equivalent surveillance system.

Collins et al. [4] and Costello et al. [5] have built up visual systems of multiple PTZ cameras in order for multiple targets can be tracked simultaneously. In [5], a distributed scheduling algorithm was proposed for identifying each person in the scene by a network of PTZ cameras. However, tracking algorithms run for every PTZ camera, which can result in high computational cost. Such a system is for highly secured and specialized environments; for example, in safe-deposit rooms or consulates, which may not be applicable for general settings in public areas. Typically, identifying people are of no interest until they trigger events. As a result, using multiple PTZ cameras may lead to redundancy. In [6], Zhou proposed a configuration of a camera system featured as "master-slave". The system was used to detect a moving human at a distance, with the master camera taking wide images of the person and the slave camera zooming in to obtain close images of the person. However, the slave camera was kept active, and it was not event triggered. This set-up may result in system redundancy for the same reason as that for visual systems composed of multiple PTZ cameras.

In this paper, an intelligent cooperative hybrid visual tracking system composed of a stationary camera and a pan tilt zoom (PTZ) camera is proposed. The stationary camera has a wide field of view, and it is attentive about the scene for event detection. The PTZ camera is activated if an event has been detected in view of the stationary camera. Unlike the method of [11], the method of this paper uses the calibrated camera setting and it then pans and tilts to center the target in its view, and zooms in to obtain identifying details of the target that may not be clear in view of the stationary camera. Such a system can be used for attentive surveillance in various locations. Compared with the single PTZ camera system proposed in [3], our hybrid camera system can still observe the overall pictures of the scene when close images of a person are being obtained. Compared with the configurations of multiple cameras proposed in [4], [5], and [6], our camera system is more efficient in surveillance because algorithms of visual tracking and event detection only run for the stationary camera if no event is detected.

II. Event Detection in View of the Stationary Camera

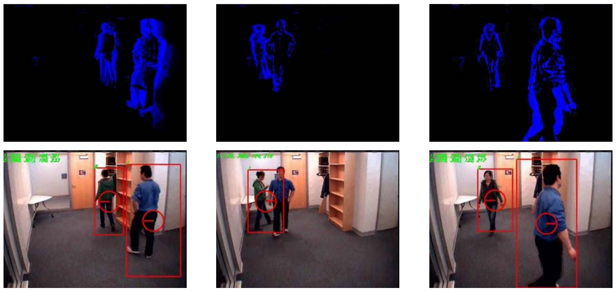
The stationary camera is responsible for detecting events that are triggered by people in the camera view. In this paper, two events, fall and wandering, are used as examples to demonstrate the functionality of the cooperative tracking system. In the camera view, motion of people is detected by comparing the difference between the current video frame and a reference video frame. In this paper, motion history image (MHI) [8] is utilized used to motion detection and representation. A MHI successively layers N frame differences over the last N time steps.

By using the MHI, for example, moving people are represented by bounding boxes, and their positions and sizes are determined. Fig. 1 shows examples of integration of MHI.

For example, for detecting sudden change in the vertical motion patterns of a person and in the view of a stationary camera a fall happens when the height of a bound box in the camera view at time t, h_t, significantly decreases, and the vertical location of the bounding box, v_t, simultaneously decreases. Or when $h'_t = \frac{h_t - h_{t-1}}{h_{t-1}} < \Delta_h, \;\; v'_t = \frac{v_t - v_{t-1}}{h_{t-1}} <$

Y. Lu and S. Payandeh are with Experimental Robotics Lab (ERL), School of Engineering Science, Simon Fraser University, 8888 University Drive, Burnaby, B.C., V5A 1S6, Canada luyanl@sfu.ca, shahram@cs.sfu.ca

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.12



(a) F_{182} (b) F_{226} (c) F_{256}

Fig. 1. Three snapshots of the tracking results together with their MHI ($N = 5$) when motion of two people is being tracked in view of the stationary camera. The position and the size of each bounding box displayed at the upper-left of the video indicate the position and the size of the moving person(s), and the clock of each bounding box indicates the direction of motion of the person(s).

Δ_v. Here, h'_t and v'_t are the instantaneous changes of the height and the vertical position of the bounding box with respect to its height at the previous time $t-1$. Δ_h and Δ_v are the thresholds used to define the occurrence of a fall. The value of Δ_h should be negative, and the value of Δ_v should also be negative if the origin of the image is defined at the bottom-left of the image.

Another example can be the case of crowd monitoring and the case of wandering person. The event of wandering is defined when a person is separated from a crowd. In this case, the total number of bounding boxes increases by one, and the distance between the separated bounding box that represents the wanderer and the bounding box that represents the crowd is between a pre-defined thresholds L and L'. Or, we can write: $L \leq l_t = \frac{\sqrt{(u_{c_t}-u_{w_t})^2+(v_{c_t}-v_{w_t})^2}}{h_{c_t}} \leq L'$, where h_{c_t} is the height of the crowd at time t, and (u_{c_t}, v_{c_t}) and (u_{w_t}, v_{w_t}) are the centers of the crowd and the wanderer at time t, respectively. The threshold L is set to judge the separation, and the threshold L' is set to judge whether the small bounding box is separated from the big one.

III. Centering Initialization and Fitting in View of the PTZ Camera

For our intelligent cooperative tracking system, two cameras should be geometrically related so that the PTZ camera "knows" where to pan and tilt in order to center and fit in its view the person who triggered an event. The calibration matrices of the two cameras, K_1 and K_2, and the rotation and translation matrices between two camera frames, R_C and t_C, are known.

Given the center and the height of the person in view of the stationary camera, $\mathbf{x}_1 = (u_1, v_1, 1)$ and h_1, and the camera parameters the center of the person in view of the PTZ camera, $\mathbf{x}_2 = (u_2, v_2, 1)$ can be obtained.

Given $\mathbf{x}_1 = (u_1, v_1, 1)$ and based on the pinhole camera model [7], $\mathbf{X}_{C_1} = (x_{c_1}, y_{c_1}, z_{c_1}, 1)^T$ is recovered from

$$\mathbf{x}_1 = K_1 I [I|\mathbf{0}] \mathbf{X}_{C_1}, \tag{1}$$

which is written explicitly as

$$z_{c_1} \begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = \begin{bmatrix} f_{u_1} & s_1 & u_{o_1} & 0 \\ 0 & f_{v_1} & v_{o_1} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{c_1} \\ y_{c_1} \\ z_{c_1} \\ 1 \end{bmatrix}. \tag{2}$$

From equation (2), we have two sub-equations:$x_{c_1} = \frac{u_1 z_{c_1} - s_1 y_{c_1} - u_{o_1} z_{c_1}}{f_{u_1}}$, and $y_{c_1} = \frac{v_1 z_{c_1} - v_{o_1} z_{c_1}}{f_{v_1}}$. In these equations , z_{c_1} is the depth of the person in view of the stationary camera;

$$\frac{z_{c_1}}{f_1} = \frac{H^m}{h_1^m} \Rightarrow \frac{z_{c_1} m_{v_1}}{f_{v_1}} = \frac{H^m m_{v_1}}{h_1} \Rightarrow z_{c_1} = \frac{H^m f_{v_1}}{h_1}, \tag{3}$$

where f_1 is the focal length of the stationary camera, h_1^m is the height of the person in meters on the image plane of the camera, and H^m is the actual height of the person in meters. Therefore, the center of the person in the frame of the stationary camera $\mathbf{X}_{C_1} = (x_{c_1}, y_{c_1}, z_{c_1}, 1)^T$ is obtained.

Given $\mathbf{X}_{C_1}$, R_C, and t_C, the center of the person in the frame of the PTZ camera, $\mathbf{X}_{C_2} = (x_{c_2}, y_{c_2}, z_{c_2}, 1)^T$, is computed from

$$\mathbf{X}_{C_2} = \begin{bmatrix} R_C & t_C \\ 0 & 1 \end{bmatrix} \mathbf{X}_{C_1}. \tag{4}$$

The center of the person in view of the PTZ camera, $\mathbf{x}_2 = (u_2, v_2, 1)^T$, is obtained by projecting $\mathbf{X}_{C_2}$ onto the image plane of the PTZ camera:

$$z_{c_2} \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix} = \begin{bmatrix} f_{u_2} & s_2 & u_{o_2} & 0 \\ 0 & f_{v_2} & v_{o_2} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_{c_2} \\ y_{c_2} \\ z_{c_2} \\ 1 \end{bmatrix}. \tag{5}$$

From equation (5), the expressions for u_2 and v_2 are: $u_2 = \frac{f_{u_2} x_{c_2} + s_2 y_{c_2} + u_{o_2} z_{c_2}}{z_{c_2}}$, and $v_2 = \frac{f_{v_2} y_{c_2} + v_{o_2} z_{c_2}}{z_{c_2}}$. Therefore, $\mathbf{x}_2 = (u_2, v_2, 1)^T$ is obtained.

Given $\mathbf{x}_2$ and the camera parameters, the angles of pan and tilt and the zoom-in amount, which are used for person centering and fitting in the PTZ camera view, are computed here. The desirable center of the person in view of the PTZ camera after person centering, $\mathbf{x}'_2$, is the image center, $(u_{o_2}, v_{o_2}, 1)^T$. From above relationships equations, the 3D coordinates of $\mathbf{x}'_2$, $\mathbf{X}_{C'_2} = (x_{c'_2}, y_{c'_2}, z_{c'_2}, 1)^T$, satisfies $x_{c'_2} = y_{c'_2} = 0$. The original PTZ camera frame, (X_{C_2}-Y_{C_2}-Z_{C_2}), should be rotated, so that the original coordinates of $\mathbf{X}_{C_2}$ turns to be $\mathbf{X}_{C'_2}$, which satisfies the conditions $x_{c'_2} = y_{c'_2} = 0$ in the rotated camera frame, ($X_{C'_2}$-$Y_{C'_2}$-$Z_{C'_2}$). Here, α defines rotating the camera frame by α about the Y_{C_2} axis. Tilting the camera by an angle β means rotating the camera frame by β about the $X_{C'_2}$ axis. Therefore, the relationship between $\mathbf{X}_{C_2}$ and $\mathbf{X}_{C'_2}$ is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & c\beta & -s\beta \\ 0 & s\beta & c\beta \end{bmatrix} \begin{bmatrix} c\alpha & 0 & s\alpha \\ 0 & 1 & 0 \\ -s\alpha & 0 & c\alpha \end{bmatrix} \begin{bmatrix} x_{c_2} \\ y_{c_2} \\ z_{c_2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ z_{c'_2} \end{bmatrix}, \tag{6}$$

where $s(\cdot)$ denotes $\sin(\cdot)$ and $c(\cdot)$ denotes $\cos(\cdot)$. From equation (6), α, β, and $z_{c'_2}$ are: $\alpha = -\arctan \frac{x_{c_2}}{z_{c_2}}$,,

$\beta = \arctan \frac{y_{c_2}}{z_{c_2}\cos\alpha - x_{c_2}\sin\alpha}$,, $z_{c'_2} = -x_{c_2}\sin\alpha\cos\beta + y_{c_2}\sin\beta + z_{c_2}\cos\alpha\cos\beta$.

Once the person has been centered in view of the PTZ camera after pan and tilt, the camera zooms in to obtain close images of the person with the zoom-in amount $\delta_f = f'_2 - f_2$, where f_2 is the original focal length and f'_2 is the focal length after zoom-in. The value of f'_2 is dependent on the desirable height of the person in view of the PTZ camera after zoom-in, h'_2; $h'_2 = a_r H$, where H is the height of the image and a_r ($a_r \leq 1$) is the ratio of the desirable height of the person to H. Therefore,

$$f'_2 = \frac{f'_{v_2}}{m_{v_2}} = \frac{z_{c'_2} h'_2}{H^m m_{v_2}} = \frac{z_{c'_2} a_r H}{H^m m_{v_2}}. \tag{7}$$

In equations (3) and (7), H^m is the height of the person in meters in the world.

IV. Tracking in View of the PTZ Camera

Once the person is centered and fitted in view of the PTZ camera the color-based particle filtering (CPF) is applied to track the person. One of the main challanges of integrating the CPF is a suitable selection of the color histogram of the person in relationships the background and also development of predictive dynamical model of tracking. This paper, a color histogram for the person is then used as a reference to weigh samples propagate by our novel implementation of *particle filter*. In parallel, based on the current state of the person from the CPF, the camera zooms in to obtain clear images of the person, and keeps the person in the camera view by panning and tilting.

A color histogram for a person is used as a reference to weigh samples in the CPF, and the histogram is established based on the color distribution within the rectangular area of the person. In this paper, a 2D hue and saturation histogram of the person with m_h and m_s bins, respectively, is used. Such a histogram is created by counting the number of pixels for each bin that has the respective values of hue and saturation. Divided by the total number of pixels occupied by the person, the color histogram is normalized, and hence represents the probability of each hue and saturation value that the person has.

Particle Filtering (PF) [14] implements a recursive Bayesian filter using Monte Carlo simulations. The key idea of the PF is to represent the required posterior probability density function (PDF) of the system by a set of random samples, $S = \{(\mathbf{s}_t^{(i)}, \pi_t^{(i)}) | i = 1, \cdots, N\}$, with associated weights $\pi_t^{(i)}$. The estimate of the current state $\mathbf{M}_t$, $E(\mathbf{M}_t)$, is computed based on these samples and weights. As the number of samples grows, PF can recover the true posterior PDF. The PF is good at dealing with visual tracking in cluttered environments because it can recover bimodal, multi-modal, and heavily skewed PDFs. The system dynamic model is defined as

$$\mathbf{M}_t = \mathbf{F}_t(\mathbf{M}_{t-1}, \mathbf{n}_{t-1}), \tag{8}$$

where $\mathbf{M}_t$ is the current state, $\mathbf{F}_t$ is a possibly non-linear function of the previous state $\mathbf{M}_{t-1}$, and $\mathbf{n}_{t-1}$ is the sequence of process. The set of samples is propagated according to the system dynamic model:

$$\mathbf{s}_t^{(i)} = \mathbf{F}_t(\mathbf{s}_{t-1}^{(i)}, \mathbf{n}_{t-1}). \tag{9}$$

Each sample in the set is weighted by the normalized probability:

$$\pi_t^{(i)} = p(\mathbf{z}_t | \mathbf{M}_t = \mathbf{s}_t^{(i)}), \qquad (\sum_{i=1}^{N} \pi_t^{(i)} = 1), \tag{10}$$

where $\mathbf{z}_t$ is the observation at time t. Therefore, the estimated state vector at time t is

$$E(\mathbf{M}_t) = \sum_{i=1}^{N} \pi_t^{(i)} \mathbf{s}_t^{(i)}. \tag{11}$$

A person in the camera view is represented by a bounding box, whose state vector, $\mathbf{M}_t$, is an 8-tuple vector

$$\mathbf{M}_t = \{u_t, v_t, u'_t, v'_t, w_t, h_t, w'_t, h'_t\}. \tag{12}$$

In the above definition, (u_t, v_t) is the center of the bounding box at time t, and u'_t and v'_t are the velocities of the box moving in the directions of the axes u and v, respectively, at time t. w_t and h_t are the width and height of the box at time t, and w'_t and h'_t are the instantaneous changes of the width and height at time t. The system dynamic model is a first-order, auto-regressive dynamic model:

$$\mathbf{M}_t = \mathbf{A}\mathbf{M}_{t-1} + \mathbf{n}_{t-1}, \tag{13}$$

whose $\mathbf{F}_t(\mathbf{M}_{t-1}, \mathbf{n}_{t-1})$ in equation (8) is $\mathbf{F}_t = \mathbf{A}\mathbf{M}_{t-1} + \mathbf{n}_{t-1}$ here. In equation (13), $\mathbf{A}$ is the deterministic component of the state model, and $\mathbf{n}_{t-1}$ is the stochastic component of the model. $\mathbf{n}_{t-1}$ is an 8-tuple vector whose ith element satisfies the distribution $\mathbf{n}_{t-1}(i, 1) \sim \tau_i N(\mu_i, \sigma_i^2)$ $(1 \leq i \leq 8)$, which is normally distributed with mean μ_i, variance σ_i^2, and amplifier τ_i. Both $\mathbf{A}$ and $\mathbf{n}_{t-1}$ can be determined based on the knowledge of the scene and the target being tracked. The set of samples is propagated based on the model in equation (13):

$$\mathbf{s}_t^{(i)} = \mathbf{A}\mathbf{s}_{t-1}^{(i)} + \mathbf{n}_{t-1}. \tag{14}$$

For the CPF, each sample $\mathbf{s}_t^{(i)}$ is weighted by the Bhattacharyya distance [15] between the color histogram for the sample, $hist_t^{(i)}$, and the color histogram for the target, $hist_r$. The Bhattacharyya distance, d_B, measures the similarity of two discrete probability distributions. Given two color histograms, $hist_t^{(i)}$ and $hist_r$, d_B is computed as

$$d_B = \sqrt{1 - \sum_{j=1}^{m} \sqrt{hist_t^{(i)}(j) hist_r(j)}}, \tag{15}$$

where m is the total number of bins. For a 2D color histogram, $m = m_h \times m_s$, where m_h and m_s are the numbers of bins of hue and saturation, respectively. The closer two color histograms (distributions) are, the smaller the value of d_B is. The exponential of the squared d_B [13] is chosen to be the weight function

$$\pi_t^{(i)} = p(\mathbf{z}_t | \mathbf{M}_t = \mathbf{s}_t^{(i)}) = e^{-\lambda d_B^2}, \tag{16}$$

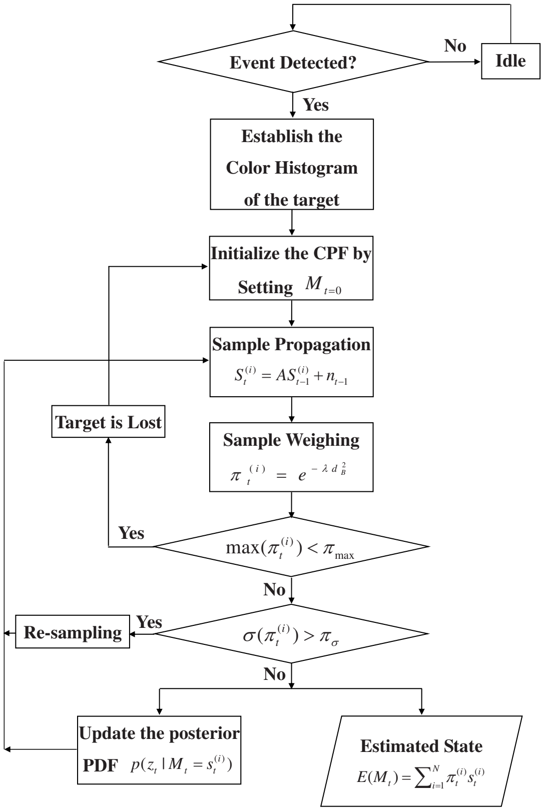

Event Detected?
No
Idle
Yes
Establish the Color Histogram of the target
Initialize the CPF by Setting $M_{t=0}$
Sample Propagation
$S_t^{(i)} = AS_{t-1}^{(i)} + n_{t-1}$
Target is Lost
Sample Weighing
$\pi_t^{(i)} = e^{-\lambda d_B^2}$
Yes
$\max(\pi_t^{(i)}) < \pi_{\max}$
No
Re-sampling
Yes
$\sigma(\pi_t^{(i)}) > \pi_\sigma$
No
Update the posterior PDF $p(z_t \mid M_t = s_t^{(i)})$
Estimated State
$E(M_t) = \sum_{i=1}^{N} \pi_t^{(i)} s_t^{(i)}$

Fig. 2. Flowchart for the CPF used to track people in the view of the PTZ camera after an event was detected in view of the stationary camera.

where λ is a scale determined by experiments. The weight, $\pi_t^{(i)}$, is normalized to satisfy $\sum_{i=1}^{N} \pi_t^{(i)} = 1$. Fig. 2 shows the flowchart for the CPF used to track people in view of the PTZ camera after an event was detected in view of the stationary camera.

In order to initialize the CPF for visual tracking, the state of the bounding box at $t = 0$, $\mathbf{M}_0$, should be set. The values set for the center, the height, and the width of the box at $t = 0$ are associated with the initial centering and fitting of the person in the PTZ camera view. The center of the box, (u_0, v_0), is set as the center of the camera view because the PTZ camera initially pans, tilts, and zooms in to center the person in its view. The height of the box, h_0, is set as $a_r H$ (H is the height of the image) and the width of the box, w_0, is set as $b_r h_0$, where a_r and b_r $(0 < a_r, b_r < 1)$ are scalars. In order to initialize the CPF for visual tracking, the state of the bounding box at $t = 0$, $\mathbf{M}_0$, should be set. Based on the center of the box, (u_0, v_0), is set as the center of the camera view because the PTZ camera initially pans, tilts, and zooms in to center the person in its view. The height of the box, h_0, is set as $a_r H$, and the width of the box, w_0, is set as $b_r h_0$, where a_r and b_r $(0 < a_r, b_r < 1)$ are scalars. The values for u_0', v_0', w_0', and h_0' are determined based on specific event.

V. Experimental Results

Once the PTZ camera is activated, based on camera parameters and the geometric relationship between the two cameras, the PTZ camera initially centers and fits the person who triggered an event. A board was used as an example to demonstrate target centering and fitting in view of the PTZ camera. The purpose of the experiment is to obtain clear images of the board in view of the PTZ camera by pan, tilt, and zoom, given the center and the height of the board in view of the stationary camera. Fig. 3-(a) shows that the board, whose height is $0.6m$ in the world, was placed in view of the stationary camera in four different locations $(L_1, \cdots, L_4)$. The centers and the heights of the board, $\mathbf{x}_1$ and h_1, are indicated by crosses and double-arrowed lines in Fig. 3-(a), and are listed in the second and the third columns of table I. The mapping between the two camera frames is

$$R_C = \begin{bmatrix} 0.98 & 0 & -0.17 \\ -0.08 & 0.90 & -0.43 \\ 0.16 & 0.43 & 0.89 \end{bmatrix}, t_C = \begin{bmatrix} 0.73 \\ 1.67 \\ 0.96 \end{bmatrix}. \quad (17)$$

The computed $\mathbf{X}_{C_1}$, $\mathbf{X}_{C_2}$, and $\mathbf{x}_2$ of the centers of the board are listed in table I, II and $\mathbf{x}_2$ for each board are indicated by crosses in Fig. 3-(b). For comparison, the ground truth (GT) of $\mathbf{x}_2$ for each location is listed in the last column of the table.

TABLE I

Parameters for the board shown in Fig. 3-(a) and Fig. 3-(b).

	Stationary Camera		
	$\mathbf{x}_1^T$	h_1	$\mathbf{X}_{C_1}^T$
L_1	$(119, 143)$	70	$(1.72, 0.83, 4.65)$
L_2	$(115, 360)$	130	$(0.95, -0.55, 2.51)$
L_3	$(353, 244)$	82	$(-0.24, -0.03, 3.97)$
L_4	$(418, 191)$	86	$(-0.68, 0.34, 3.79)$

TABLE II

Parameters for the board shown in Fig. 3-(b).

	PTZ Camera		
	$\mathbf{X}_{C_2}^T$	$\mathbf{x}_2^T$	$\mathbf{x}_2^T$ (GT)
L_1	$(1.74, -0.21, 5.52)$	$(149, 260)$	$(167, 210)$
L_2	$(1.29, -0.19, 2.88)$	$(76, 276)$	$(78, 305)$
L_3	$(-0.10, -0.42, 4.24)$	$(333, 294)$	$(323, 290)$
L_4	$(-0.51, 0.02, 4.23)$	$(385, 237)$	$(375, 239)$

Using the values for $\mathbf{X}_{C_2}$ and $\mathbf{x}_2$ listed in table I and II, the pan and tilt angles, α and β, and the zoom-in amount, δ_f, were computed.

For the case of detecting an abnormal behavior pattern of a person such as a class of fall, both values of h_t' and v_t' are negatively large. Based on the experimental study, the values for Δ_h and Δ_v were set as $\Delta_h = -0.40$ and $\Delta_v = -0.20$ in order to detect a fall and leave the normal activities, such as sitting down, undetected.

Fig. 4 shows screen shots of the tracking results for fall detection in view of the stationary camera. The height and

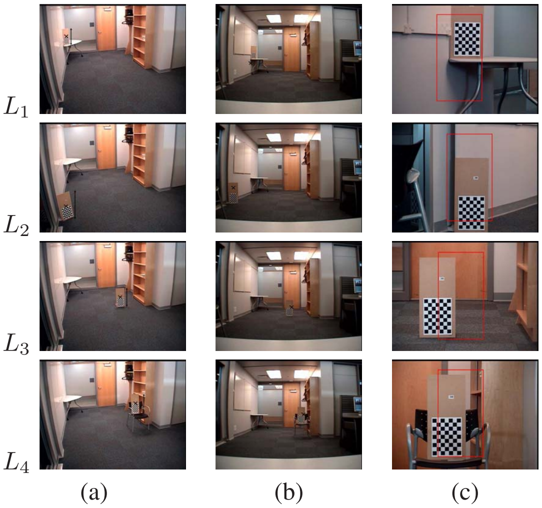


Fig. 3. (a) The board was placed in view of the stationary camera in four different locations ($L_1, \cdots, L_4$), with their centers and heights indicated by crosses and double-arrowed lines; (b) Centers of the board in view of the PTZ camera indicated by crosses; (c) Close images of the board obtained by pan, tilt, and zoom of the PTZ camera.

the vertical location of the falling person decrease quickly and simultaneously from video frame F_{262} to F_{271}. Based on the criteria $h'_t \leq -0.4$ and $v'_t \leq -0.2$, a fall was detected at F_{271}. The center of the person at time $t = 271$, $\mathbf{x}_{271}$, is $(254, 259)$, and the width w_{271} and the height h_{271} of the person at the time $t = 271$ are 90 and 132, respectively.

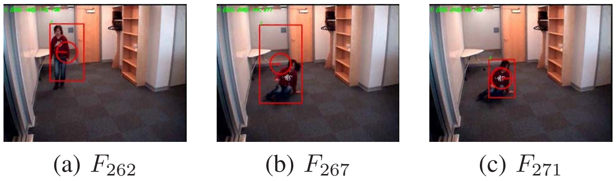

(a) F_{262} (b) F_{267} (c) F_{271}

Fig. 4. Snapshots of the tracking results for fall detection in the view of the stationary camera. A fall was detected at video frame F_{271}.

Fig. 5-(b) shows the zoomed-in image of the person with her position and size indicated by the bounding box. Fig. 5-(c) is the 2D hue and saturation histogram for the fallen person, which is used as a reference to weigh samples propagated by the PF.

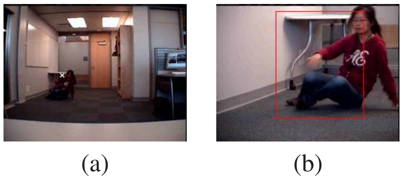

(a) (b)

Fig. 5. (a) Computed center of the fallen person in view of the PTZ camera, which is indicated by a cross; (b) A close image of the fallen person obtained by initial pan, tilt, and zoom of the PTZ camera.

The state vector of the fallen person, $\mathbf{M}_t$, is an 8-tuple vector defined in equation (12). The set of samples was propagated based on the first-order, auto-regressive dynamic model expressed in equation (13). For the experiment of tracking a fallen person, the deterministic component of the state model in equation (13), $\mathbf{A}$, is an 8×8 matrix:

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{A}_2 \\ \hline \mathbf{A}_2 & \mathbf{A}_1 \end{array} \right], \tag{18}$$

where

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{19}$$

and $\mathbf{A}_2$ is a 4×4 zero matrix. The stochastic component of the model, $\mathbf{n}_{t-1}$, is an 8-tuple vector. Each element in the vector, $\mathbf{n}_{t-1}(i, 1)$, is normally distributed as $\mathbf{n}_{t-1}(i, 1) \sim \tau_i N(0, 1)$ $(1 \leq i \leq 8)$. For the experimental studies of tracking a fallen person we have selected, $\tau_1 = \tau_2 = 5$, $\tau_3 = \tau_4 = 2$, $\tau_5 = \tau_6 = 5$, and $\tau_7 = \tau_8 = 2$.

To initialize the CPF, the center of the bounding box at $t = 0$, (u_0, v_0), was set as the center of the camera view, $(320, 240)$. The height of the box at $t = 0$, h_0, was set as $a_r = 4/5$ of the image height, and the width of the box at $t = 0$, w_0, was set as $b_r = 4/5$ of h_0. The velocities of the box moving in the directions of the axes u and v at $t = 0$, u'_0 and v'_0, were set as zero because a fallen person usually remains on the floor for a while after he/she falls. For the same reason, the rates of change for the width and the height of the bounding box at $t = 0$, w'_0 and h'_0, were set as zero. The scale in the weight function (16), λ, was set as $\lambda = 1$. Fig. 6 shows the results for tracking the fallen person using the CPF, when the number of samples N was set as $N = 100$. From Fig. 6, the PTZ camera was able to obtain clear identifying details of the fallen person, which was not clear in view of the stationary camera as shown in Fig. 4-(c).

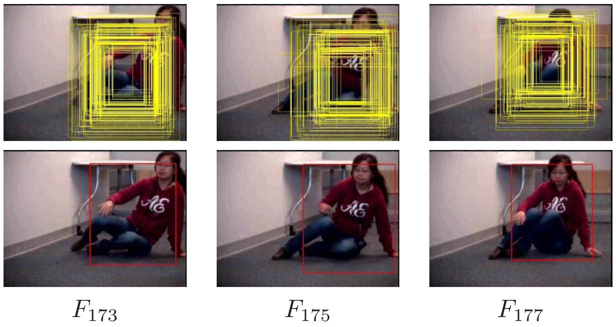
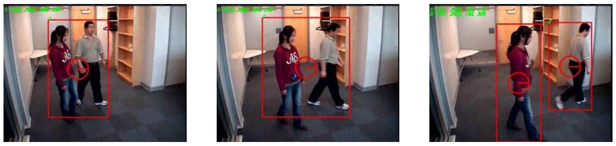

F_{173} F_{175} F_{177}

Fig. 6. Results of tracking a fallen person using the CPF; first row: samples propagated ($N = 100$); second row: weighted sum of the samples.

For demonstration of the second example, we have two people walked in a group initially, and one of them wandered away from the other at the end. Similar to determining the thresholds in fall detection, the values for L and L', which are the thresholds in wandering detection, should be determined experimentally. the camera view. Wandering occurs when the separation of the

Fig. 7 shows screen shots of the tracking results for wandering detection in view of the stationary camera. Based

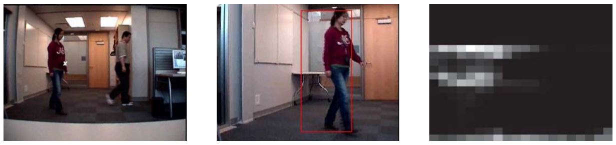


(a) F_{152} (b) F_{156} (c) F_{160}

Fig. 7. Snapshots of the tracking results for wandering detection in view of the stationary camera. Wandering was detected at video frame F_{160}.

on the criteria $0.55 \leq l_t \leq 0.70$, wandering was detected at video frame F_{160}. The center of the wanderer at time $t = 160$, $\mathbf{x}_{160}$, is $(310, 274)$, and the width w_{160} and the height h_{160} of the person at time $t = 160$ are 157 and 410, respectively.

Fig. 8-(b) shows the zoomed-in image of the person with her position and size indicated by the bounding box. Fig. 8-(c) is the 2D hue and saturation histogram for the person.



(a) (b) (c)

Fig. 8. (a) Computed center of the wanderer in view of the PTZ camera, which is indicated by a cross; (b) A close image of the wanderer obtained by initial pan, tilt, and zoom of the PTZ camera; (c) The 2D hue and saturation histogram for the wanderer.

A. Discussions

When tracking the fallen person and the wanderer, the first-order, auto-regressive dynamic model was used to propagate samples in the CPF because, at time $t+1$, the person usually appeared near the area where she appeared at time t. Therefore, the deterministic component of the model, $\mathbf{A}$, was the same for the two tracking. However, the difference between the two tracking is that the fallen person sit on the floor after a fall was detected while the wanderer was still walking after wandering was detected. Therefore, different stochastic components of the system dynamic model and different initial states were set for the CPF to track the fallen person and the wanderer. In the stochastic component, setting τ_i as zero means that the target is totally static. The larger the value of τ_i is, the more mobile the target is. When tracking the fallen person, the amplifiers $(\tau_1, \cdots, \tau_8)$ were set small compared with those set for tracking the wanderer. The value for specific τ_i was determined based on the characteristic of the parameter it associates with (u with τ_1, v with τ_2, u' with τ_3, v' with τ_4, w with τ_5, h with τ_6, w' with τ_7, and h' with τ_8). For example, in fallen person tracking, τ_1 and τ_2 for the magnitudes of noises presented in u_t and v_t were both set as 5 because the person was sitting on the floor. However, in wanderer tracking, τ_1 and τ_2 were set as 20 and 10, respectively, because the variance of the position of wanderer in horizontal direction was generally larger than the variance of the position of the wanderer in vertical direction.

VI. Conclusions

The intelligent cooperative hybrid tracking system proposed in this paper is composed of two cameras. The stationary camera has a fixed location and focal length, and it is used to monitor a wide area to detect events. The PTZ camera can be controlled to change perspective and levels of zoom, so that different levels of detail of the target who has triggered an event can be obtained. Two cameras were geometrically related by camera calibration, so that the PTZ camera "knows" how to pan, tilt, and zoom in order to fit the target in its view. For the stationary camera, motion-based visual tracking is used to monitor moving people and detect events. Two events, fall and wandering, are used as examples to demonstrate the functionality of the cooperative tracking system. Motion of people is represented by the motion history image, and people are represented by bounding boxes. An event is detected by analyzing the states of bounding boxes based on predefined criteria. For the PTZ camera, the CPF is used to track the person who has triggered an event. The color histogram for the person is used as a reference to weigh samples propagate by the particle filter. In the meanwhile, the camera zooms in to obtain clear images of the person, and keeps the person in the camera view by panning and tilting. The camera is assigned to the person until, for example, enough identifying details of the person have been collected, or another person triggers a new event.

References

[1] N.Papanikolopoulos, P.Khosla, T.Kanade, "Visual tracking of a moving target by a camera mounted on a robot: A combination of control and vision", *IEEE Transactions on Robotics and Automation*, 9(1), 1993.

[2] D.Murray, A.Basu, "Motion tracking with an active camera", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):449-459, May 1994.

[3] M.Lalonde, S.Foucher, L.Gagnon, E.Pronovost, M.Derenne, A.Janelle, "A system to automatically track humans and vehicles with a PTZ camera", *Visual Information Processing XVI, Proceedings of the SPIE*, Volume 6575, pp. 657502, Orlando, 2007.

[4] R.Collins, O.Amidi, T.Kanade, ”An Active Camera System for Acquiring Multi-View Video“, *International Conference on Image Processing (ICIP)*, pp.517-520, Rochester, NY, September 2002.

[5] C.J.Costello, I.J.Wang, "Surveillance Camera Coordination Through Distributed Scheduling", *44th IEEE Conference on Decision and Control, and the European Control Conference*, pages 1485- 1490, 2005.

[6] X.Zhou, R.Collins, T.Kanade, P.Metes, "A Master-slave System to Acquire Biometric Imagery of Humans at Distance", *ACM International Workshop on Video Surveillance*, Berkeley, USA, November 2003.

[7] R.Hartley, A.Zisserman, "Multiple View Geometry in Computer Vision", Cambridge University Press, 2000.

[8] J.W.Davis, "Hierarchical motion history images for recognizing human motion", *Proc. IEEE Workshop on Detection and Recognition of Events in Video*, 2001.

[9] P.Kilambi, O.Masoud, N.Papanikolopoulos, "Crowd Analysis at Mass Transit Sites", *IEEE International Conference on Intelligent Transportation Systems*, pp. 753-758, Seattle, WA, Sep. 2006.

[10] C.Rougier, J.Meunier, A.St-Arnaud, J.Rousseau, "Fall Detection from Human Shape and Motion History Using Video Surveillance", *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW)*, pp. 875-880, 2007.

[11] A.W.Senior, A.Hampapur, M. Lu, ”Aquiring Multi-Scale Imahes by Pan-Tilt-Zoom Control and Automatic Multi-Camera Calibration", *Proceedings of the Seventh IEEE Workshop on Application of Computer Vision*, pp. 1-6, 2005

[12] Y. Lu, S. Payandeh, ”Dumbbell Calibration for a Multi-Camera Tracking System", *P*roceedings of Canadian Conference on Electrical and Computer Engineering, pages 1472-1475, 2007

[13] K.Nummiaro, E.Koller-Meier, L. V. Gool, "An Adaptive Color-Based Particle Filter", *Image and Vision Computing*, 2002.

[14] M.S.Arulampalam, S.Maskell, N.Gordon, T.Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian Bayesian tracking", *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174-188, 2002.

[15] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions", *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99C109, 1943.