Cooperative Hybrid Multi-camera Tracking for People Surveillance

Yan Lu, IEEE Student Member, and Shahram Payandeh, IEEE Member

Abstract—Developing and improving people surveillance in public areas has been a topic of interest over past several years. A novel hybrid visual tracking system for event detection and people tracking is proposed in this paper. Our surveillance system consists of a calibrated stationary and a pan, tilt, and zoom (PTZ) camera. The stationary camera has a wide field of view, and it detects global events. The PTZ camera is triggered once an event is detected, and it pans and tilts to center the target based on the camera geometry. Then, the camera zooms in to obtain levels of details from the target which may not be clear in the wide view of the stationary camera. By applying color-based particle filtering (PF), the active camera is able to continuously centers and tracks the target in high resolution images. Experimental results are carried to further demonstrate the practicality of our proposed event monitoring system.

I. INTRODUCTION

Efforts have been made by researchers over past several years for developing and improving attentive surveillance systems. Visual tracking techniques are proposed for automatic target tracking and event detection. For example, [1] uses a coarse ellipse model and a particle filter for fall detection in a supportive home environment. [2] monitors urban environments and detects people merging and splitting by crowd analysis. In the majority of previous papers, little work has been done for after event detection. For example, it is desirable to obtain levels of details of the person who triggered an event in people surveillance.

In this paper, the stationary camera has a wide field of view, and it is attentive about the scene to detect events. The PTZ camera is triggered active once an event is detected in the other camera. Then, it pans and tilts to center the target in its view, and zooms to obtain details of the target which may not be clear in the stationary camera. The flexible perspective and resolution of the PTZ camera facilitate scene collection and enhance the scope and depth of our surveillance system.

The advantage of our camera configuration for events detection over a single PTZ camera system [3] is that the stationary camera still keeps the overall visual information of the scene, while the monitoring control system can decide to obtain various levels of details using single or multiple active cameras. Compared with the multiple active camera system in [4], our hybrid configuration reduces computational cost, because algorithms of visual tracking and event detection are run on the stationary camera only if no event is detected. Collins group proposes a similar configuration of cameras, which is called “master-slave” in [5]. Their system is for detecting a moving human at a distance, where the master camera takes wide images and the slave camera zooms in onto the person to obtain his/her close images. However, the slave camera is not event triggered, and it is kept active. This may lead to system redundancy, because the person’s identification is not of interests until he/she triggers an event, for example, falling.

In order to build up our cooperative tracking system, two cameras should be geometrically related before hand, so that the active camera “knows” where to pan and tilt in order to center the target in its view when an event is triggered. Building up geometric relations between cameras, we intrinsically and extrinsically calibrate cameras using the method proposed in [6]. Intrinsic parameters can also be known directly from camera specifications. Thus, we know their projection matrices $K_1$ and $K_2$, and the rotation $R_C$ and the translation $t_C$ between their camera coordinates.

The organization of the rest of the paper is as follows. In section II, we discuss motion-based people tracking and event detection in the stationary camera. Section III focuses on target fitting and tracking in the active camera. We present the method to calculate pan and tilt angles, and zoom in amount. Color-based particle filtering for target tracking is also presented in this section. Experimental results for people surveillance using our proposed system are shown in section IV. Finally, we conclude in section V.

II. EVENT DETECTION IN THE STATIONARY CAMERA

The stationary camera is responsible for monitoring a wide area and reporting to the active camera any events detected. Typically, events, such as falling, happen when motion of people change in extraordinary ways. Therefore, we use the motion-based visual tracking technique to track and analyze moving people. The criteria to judge extraordinary motions are based on positions and speed of people represented in rectangles.

In this paper, we use the idea of frame difference, and create a motion history image (MHI) that contains $N$ differences between recent $N+1$ frames. Greyscale values for frame differences decrease according to the time sequence. The difference between frames $F_{t-N}$ and $F_{t-N-1}$ has the lowest greyscale value (but greater than zero) among $N$ buffers. The MHI is updated every time step by adding newly detected frame difference, reducing the greyscale value for the rest frame difference by one level, and zeroing the oldest frame difference among $N$ buffers. We indicate a moving object by a geometrically defined region of interest. Having a MHI at time $t$, the center $I(t(u_x,v_y))$ is calculated by a weighted sum of $N$ centers $I_t(u_i,v_i)$ ($i = 1 \ldots N$) of $N$
frame differences. Recent differences (small $i$) are assigned higher weights than old differences (large $i$). Fig. 1 shows an example of the MHI to represent a walking person. The green solid region is the weighted sum of $N$ red dashed regions. Higher weighted red regions are drawn in thicker lines than lower weighted ones.

A. Fall Detection

Fall happens with a high probability at the time when our geometrically defined region of interest has significant decreases in its height $h_t$ and vertical location $v_t$ with respect to the previous height $h_{t-1}$. Mathematically, fall happens when

$$
\begin{align*}
\delta_{ht} &= \frac{h_t - h_{t-1}}{h_{t-1}} < \Delta_h, \\
\delta_{vt} &= \frac{v_t - v_{t-1}}{h_{t-1}} < \Delta_v,
\end{align*}
$$

(1)

where $\delta_{ht}$ and $\delta_{vt}$ are instantaneous change rate of the rectangle height and the vertical position with respect to the previous height. $\Delta_h$ and $\Delta_v$ are thresholds. $\Delta_h$ is negative, and $\Delta_v$ is also negative if the image origin is at the bottom left. Two relations in the equation (1) should be satisfied at the same time to claim an event of fall, or false detection will happen. Thresholds $\Delta_h$ and $\Delta_v$ are determined in experiments.

B. Wanderer Detection

In wanderer detection, the stationary camera should have an overview of the scene such that no person is hidden from other people. The event happens when a person is separated from the crowd, which means that

$$
\Delta_t \leq l = \sqrt{(u_{it} - u_{jt})^2 + (v_{it} - v_{jt})^2} / h_{jt} \leq \Delta_t',
$$

(2)

$h_{jt}$ is the height of wanderer in pixel at time $t$. $(u_{it}, v_{it})$ and $(u_{jt}, v_{jt})$ are centers of the crowd and wanderer at time $t$, respectively. $\Delta_t$ is a criteria set for separation judgement, and $\Delta_t'$ is set to judge whether wanderer is separated from the crowd. Same as $\Delta_h$ and $\Delta_v$, $\Delta_t$ and $\Delta_t'$ are determined in experiments as well.

III. Active Camera Process

The active camera is triggered after an event is detected. The stationary camera passes target’s central pixel position $I_1(u_t, v_t)$ on its image plane to the active camera. Initially, the active camera centers and zooms in onto the target based camera geometry. After that, the active camera attentively tracks the target using color-based PF [9]. Due to the reason that pan and tilt motion of the camera will easily be confused with the target motion, we choose color instead of motion as the cue to track a target at this stage. Moreover, color is a rotation and scale invariant feature of the target, so it is suitable for tracking the target under different zooming levels.

A. Target Fitting in the Active Camera

$I_1$ and $K_1$ (the projection matrix of the stationary camera) are used to recover the person’s 3-D position $P_1 = (x_1, y_1, z_1)^T$ in the stationary camera coordinates (refer to the Appendix for details). The position of the person in the active camera coordinates $P_2 = (x_2, y_2, z_2)^T$ is calculated by $P_2 = R_cP_1 + t_c$. If the target is in the center of the active camera view, $I_2(u_t, v_t) = (W_{ij}, H_{ij})$, which is that $x_2 = y_2 = 0$ based on the equation (12). Thus, the active camera should pan and tilt by angles $\alpha$ and $\beta$, respectively. Panning the camera by an angle $\alpha$ means to rotating the original camera coordinates by $\alpha$ about the $y$ axis. Tilting by $\beta$ means to rotating the coordinates by $\beta$ about the $x$ axis. The relation between $P_2 = (x_2, y_2, z_2)^T$ and $P_1 = (x_1, y_1, z_1)^T$ is that

$$
R_c R_y \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ z_2 \end{bmatrix},
$$

(3)

where $R_x$ is the rotation matrix containing $\beta$ and $R_y$ is the matrix of $\alpha$. Solving equation (3), we obtain

$$
\alpha = -\arctan \frac{x_2}{z_2}, \beta = \arctan \frac{y_2}{\cos \alpha z_2 - \sin \alpha x_2},
$$

(4)

$$
z_2 = \sin \beta y_2 + \cos \beta \cos \alpha z_2 - \cos \beta \sin \alpha x_2.
$$

(5)

Once the target is centered, the camera zooms in by the amount $\delta_{f_2} = f_2' - f_2$. $f_2'$, the focal length after zoom in, is calculated as follows

$$
f_2' = \frac{h_{CCD_2}}{H} z_2 = \frac{a H_{CCD_2}}{H} z_2',
$$

(6)

$a$ is the ratio of desired zoomed in target height in pixel $h_{CCD_2}'$ to $H_{CCD_2}$, and $a$ satisfies $\frac{h_{CCD_2}'}{H_{CCD_2}} \leq a \leq 1$. For general applications, the exact value of $H$ may not be known, but it can be approximated. In wanderer detection, we assume $H$ is the average human height, for example, $H = 1.72 m$ for North Americans. In fall detection, $H$ is selected to be a half of the average human height.

B. Color-based Particle Filtering

PF is a technique to build the tracking system within a stochastic frame. Compared with traditional color-based trackers, such as Meanshift and CAMSHIFT, color-based PF is able to deal with clutters, such as partial occlusion and background distraction, which occur in a high probability for wanderer tracking. We use 2-D hue-saturation histogram $r$ as the color model to represent a target. The
state $X_t$ for our visual tracking case is an 8-tuple vector, and $X_t = \{u_t, v_t, u'_t, v'_t, w_t, h_t, w'_t, h'_t\}$. $(u_t, v_t)$ is the center of the region, and $u'_t$ and $v'_t$ are velocities in corresponding directions. $w_t$ and $h_t$ are width and height of the region, and $w'_t$ and $h'_t$ are change rates of the corresponding lengths.

The sample set is propagated according to a first-order autoregressive dynamic model, which is

$$X_t = AX_{t-1} + \omega_{t-1}. \tag{7}$$

$A$ is the deterministic component. $\omega_{t-1}$, the stochastic component, is an 8-tuple vector such that $\omega_{t-1}(i, 1) \sim N(0, \sigma_i^2)$ $(1 \leq i \leq 8)$.

Each sample $q^{(n)}_t$ is weighed by the Bhattacharyya distance $d$ between the sample histogram $q^{(n)}_t$ and $r$ [9]. Samples are weighted by the exponential of $d^2$

$$\pi^{(n)}_t \propto e^{-\lambda d^2}. \tag{8}$$

$p^{(n)}_t$ is normalized such that $\sum_{i=1}^{N} p^{(n)}_t = 1$. $\lambda$ is determined based on experiments. To start tracking, $X_{t=0}$ should be known for initialization. Positions of samples are manually spread about the center of the view, because, for zoomed images, the target is most likely to appear around the center.

IV. EXPERIMENTAL RESULTS

We use two Sony EVI-D100 PTZ color video cameras for our experiments, and the PTZ capacity is utilized on one of them. For fall detection, $R_C = 1$ and $t_C = [-0.71, 0, 0]^T$ $m$. For wanderer detection,

$$R_C = \begin{bmatrix} 0.866 & 0 & 0.5 \\ 0 & 1 & 0 \\ -0.5 & 0 & 0.866 \end{bmatrix}, \quad t_C = \begin{bmatrix} -2.00 \\ 0 \\ -0.23 \end{bmatrix}. $$

We try two different settings to test the practicality of our hybrid tracking system.

A. Experiments for Fall Detection

We put two cameras facing an area where two persons are walking. Thresholds $\Delta_h = -0.4$ and $\Delta_v = -0.2$ are determined experimentally. Fig. 2 (a)-(c) show motion-based tracking results in the stationary camera. When the event happens, $I_1 = (360, 130)$, $h_1 = 90$, and $w_1 = 71$. Table I are geometric parameters of the target and pan, tilt, and zoom amount of the active camera.

After zoom in, color-based PF is run on the active camera. In our experiment, $A$ in equation (7) for the system dynamic model is an 8 $\times$ 8 matrix.

$$A = \begin{bmatrix} M_1 & M_2 \\ M_2 & M_1 \end{bmatrix}, \quad M_1 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and $M_2$ is a 4 $\times$ 4 zero matrix. Here, we have selected $\lambda = 1$ based on experiments. Fig. 2 (d)-(e) are screen shots in the active camera, where yellow rectangles in (d) are samples generated, and the red rectangle in (e) is the target location and sized, which is the weighted sum of samples.

B. Experiments for Wanderer Detection

We put two cameras facing a hallway. Three persons walk in a group at the beginning, and one of them walk away from the majority later on. Total number of rectangles in the scene is monitored, and thresholds $\Delta_l = 0.5$ and $\Delta_v = 0.55$ are determined during several sample runs under the same experimental setting. Experimental results for wanderer detection are shown in Fig. 3. $I_1 = (414, 158)$, $h_1 = 193$, and $w_1 = 60$. Table II presents geometric parameters of the wanderer and pan, tilt, and zoom amount of the active camera.

The system dynamic model and $\lambda$ for wandering detection are the same with those in fall detection. An extra challenge for wanderer detection, which might not be a problem for fall detection, is that the target is still moving after the event happens. It is possible that the wanderer is partially occluded by settings in the background. PF can deal with...
such difficulties. However, PF loses the target when he is totally blocked by a bookshelf. In such a case, the active camera zooms out and searches in the target in a wider view. After finding the target, the camera pans, tilts, and zooms again to follow the target in its view.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a hybrid visual tracking system for people surveillance to detect events and obtain target details. The system has a calibrated stationary and a PTZ camera. The stationary camera is responsible for event detection, and the active camera for target following and identification acquisition. Once an event is detected, the active camera is triggered. Then, it pans, tilts, and zooms to center and track the target in high resolution images.

As to tracking strategy, we choose different methods according to different tracking situations. In the stationary camera, motion-based tracking is utilized for events detection, because events happen of high probabilities when people are moving. However, when it comes to follow the target in the active camera, color is picked up as the cue for tracking, because color of the camera will easily be confused with motion of the target.

One of the future work of our topic is role changing between the cameras, which means that the fixed camera becomes active and the active camera becomes fixed. Role switching facilitates event detection and collection of people’s identifications if the target is blocked in one of the camera views. Another future work is to expand our dual camera setting to a triple camera setting that consists two stationary and one active camera. Such a configuration is necessary if one stationary camera is not able to cover the area under surveillance. The active camera is shared between two fixed cameras.

APPENDIX

POINTS GEOMETRY UNDER PINHOLE CAMERA MODEL

Fig. 4(a)-(c) show how a point $I(u, v)$ on the image plane is related to a point $P = (x, y, z)^T$ in the camera coordinates according to the pinhole camera geometry. The camera is calibrated, and therefore, the depth or $z$ of the target is recovered by

$$ z = \frac{f}{h_{CCD}} H = \frac{W_{CCD} f_x}{h_{CCD} W_I} H = \frac{H_{CCD} f_y}{h_{CCD} H_I} H, $$(9)

<table>
<thead>
<tr>
<th>$P_1$ (m)</th>
<th>$P_2$ (m)</th>
<th>$P_3$ (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.836, -0.731, 4.84)</td>
<td>(1.14, -0.731, 3.54)</td>
<td>(0, 0, 3.79)</td>
</tr>
</tbody>
</table>

TABLE II

PARAMETERS FOR WANDERER DETECTION

<table>
<thead>
<tr>
<th>$a$</th>
<th>$x$ (degree)</th>
<th>$\beta$ (degree)</th>
<th>$f_1$ (mm)</th>
<th>$\delta_f$ (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>17.9</td>
<td>-11.1</td>
<td>5.14</td>
<td>2.04</td>
</tr>
</tbody>
</table>

REFERENCES