Cooperative Target Tracking in Dual-Camera System with Bidirectional Information Fusion

Jingjing Wang(✉) and Nenghai Yu

CAS Key Laboratory of Electromagnetic Space Information,
University of Science and Technology of China, Hefei, China
kkwang@mail.ustc.edu.cn, ynh@ustc.edu.cn

Abstract. The Dual-Camera system which consists of a static camera and a pan-tilt-zoom (PTZ) camera, plays an importance role in public area monitoring. The superiority of this system lies in that it can offer wide area coverage and highly detailed images of the interesting target simultaneously. Most existing works in Dual-Camera systems only consider simplistic scenarios, which are not robust in real situations, and no quantitative comparison between different tracking algorithms is provided. In this paper, we propose a cooperative target tracking algorithm with bidirectional information fusion which is robust even in moderately crowded scenes. Moreover, we propose a method to compare the algorithms quantitatively by generating a virtue PTZ camera. The experimental results on realistic simulations and the implementation on a real surveillance system validate the effectiveness of the proposed algorithm.

Keywords: Cooperative tracking · PTZ camera · Information fusion

1 Introduction

With rapidly growing demands of security in public area monitoring, multiple-camera surveillance system has become a hot subject in the field of computer vision. Among them, one popular example is the Dual-Camera system, which consists of a static camera and a PTZ camera. A static camera can cover a large public area. However, it cannot provide high resolution images of the interesting target which are useful for abnormal behaviour detection, gesture recognition, face identification, etc. This is where PTZ cameras compensate for the deficiencies of static cameras. A PTZ camera can pan and tilt to center the target in its view and zoom in to obtain desirable high-resolution images. The Dual-Camera system which combines these two types of cameras can monitor the large surveillance area and obtain close-up observation of the interesting target simultaneously. Figure 1 shows an example of images obtained from the static camera and the PTZ camera.

Dual-camera systems have been widely studied in surveillance [1–8]. References [2–5] use tracking results from the static camera to guide the movement of the PTZ camera. They use background subtraction algorithms to detect targets

Y.-J. Zhang (Ed.): ICIG 2015, Part II, LNCS 9218, pp. 60–74, 2015.
DOI: 10.1007/978-3-319-21963-9_6

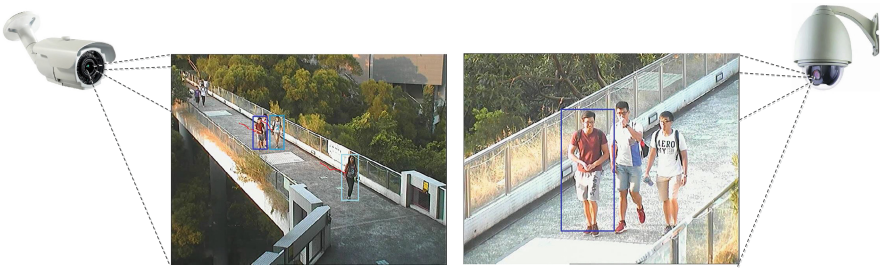

Fig. 1. The left is the image obtained from the static camera, and heights of targets in the image are about 50 pixels. The right is the image obtained from the PTZ camera, and heights of tragets in the image are about 250 pixels.

and track targets by associating the detection responses with the corresponding targets. They pay more attention to the calibration between the static and PTZ cameras. However, it requires a level of pointing accuracy to keep a highly zoomed camera pointing at a moving target, that is not achievable from calibration alone [6]. Instead, [6–8] use control signals from the static camera only initially, to make the target within the view of the PTZ camera, and then perform real-time tracking in the PTZ camera to keep the camera centered on the target with desirable resolution. Considering the requirement of real time processing, most existing tracking systems which use PTZ cameras adopt simple and efficient algorithms to perform tracking in PTZ cameras. Reference [6] uses Mean-Shift algorithm [13] to track the target. References [7,8] use color-based particle filter algorithm to keep following the target. In [14], Mean-Shift tracker and KLT [15] tracker are combined for target tracking. Although these methods can guarantee real-time performance, they are not robust enough in practice. Mean-shift and color-based particle filter trackers may fail to differentiate between the interesting target and the background with similar color, by using color histogram. KLT tracker is not robust to background clutters and occlusion. Moreover, all these methods consider situations where only a few of targets appear without frequent occlusion. However, in real pubic surveillance areas, occlusion may frequently occur especially in crowded scenes. No information is fused between the cameras in these methods, which is very helpful for resolving occlusion.

In video tracking using PTZ cameras, comparing different methods directly is very difficult. It is not possible to work offline with recorded videos, since each frame in the PTZ camera depends on the pan, tilt and zoom parameters, and such parameters are differently set by the different tracking algorithms. To deal with this problem, [11] proposes an experimental framework which allows to compare different algorithms in repeatable scenarios. The key idea consists in projecting a video containing the target on a screen in front of the camera. However, it is difficult to use this framework in a cooperative tracking setting, since this framework cannot generate different image sequences for different cameras which have different view points. In [16], a synthetic camera network is placed in a virtual scene which is created through computer graphics. However, modelling realistic human behavior within a virtual environment is difficult. It is still different from real situations.

To overcome the drawbacks of previous methods, we propose a cooperative target tracking algorithm with bidirectional information fusion. Specifically, an efficient multi-target tracking algorithm is introduced for online tracking in the static camera, a robust single-target tracking algorithm is proposed in the PTZ camera and a bidirectional information fusion strategy is proposed to enhance the algorithm. The single-target tracking algorithm combines a state-of-the-art category detector and an online trained classifier. The category detector is offline trained and robust against the challenges in PTZ cameras, such as background clutters, abrupt motion and motion blur. The online trained classifier can differentiate the interesting target from the other detected targets and adapt to the appearance changes of the interesting target through online updating. The bidirectional information fusion method makes the algorithm robust even in moderately crowded scenes with frequent interactions and occlusions. Moreover, unlike the existing works in Dual-Camera systems which provide no quantitative comparison of different tracking algorithms, we propose a method to quantitatively compare different algorithms by generating a virtue PTZ camera. The experimental results on realistic simulations and the implementation on a real surveillance system show the effectiveness of our proposed algorithm.

2 System Overview

In our Dual-Camera system, the static camera detects and tracks multiple targets in a wide scene. When an interesting target is detected by an anomaly detection algorithm or specified by the user, the PTZ camera is directed to gaze at the target according to camera-to-camera calibration. Then a cooperative tracking algorithm is used to track the target. The camera control module adjusts parameters of the PTZ camera according to the tracking results in the PTZ camera to follow the target at high resolution. The purpose of the Dual-Camera system is to continuously keep the interesting target in the PTZ camera view to obtain high resolution images of the target.

The rest of the paper is organized as follows. The camera calibration and control strategy are introduced in Sect. 3. Section 4 describes the proposed cooperate tracking algorithm. Section 5 presents the experimental results. Section 6 summarises this paper.

3 Camera Calibration and Control

In this paper, we focus on the cooperative tracking algorithm, but camera calibration and camera control strategy are indispensable parts. So we first briefly introduce the camera calibration and control strategy. We denote the parameters of the PTZ camera at time t as (P_a^t, T_a^t, Z_a^t), where P_a^t and T_a^t represent the pan-tilt angles, and Z_a^t means the optical zoom of the PTZ camera. (P_a^t, T_a^t, Z_a^t) can be read from the interface of the PTZ camera, or estimated using the method proposed by [18]. In order to perform cooperative tracking, calibration between the static camera and the PTZ camera is needed. We use the method proposed by [5]

to calibrate the two cameras for its simpleness. In [5], ground plane homography is exploited to realise camera collaboration, assuming that the two cameras share a common ground plane which is reasonable in typical surveillance scenes. The ground plane homography H_{gd} between the static camera and the PTZ camera at parameter (P_a^0, T_a^0, Z_a^0) is estimated offline using people correspondence. In online stage, the bottom center of the bounding box which contains the tracked target in the static camera image is mapped to the corresponding point in the image plane of the PTZ camera at parameter (P_a^0, T_a^0, Z_a^0). The pan and tilt angles which are needed to bring the interesting target to the center of the PTZ camera image can be computed. Please refer to [5] for more details. Once the same target is found in the PTZ camera, the cooperative tracking module is activated to track the target. During tracking, the projective transformation $H_{s\to a}^t$ which maps the bottom center of the target in the static camera image to the position in the PTZ camera image at frame t is computed as:

$$H_{s\to a}^t = K^t R^t (R^0)^{-1} (K^0)^{-1} H_{gd} \tag{1}$$

where R^t is the rotation matrix corresponding to the pan and tilt angles (P_a^t, T_a^t) at frame t, and K^t is the intrinsic matrix corresponding to the zoom Z_a^t. The intrinsic matrix K^t is estimated using the method proposed by [17]. For the camera control strategy during tracking, the current and previous distances between the position of the target and the center of the PTZ camera image are used to deal with the camera speed. If the distance becomes larger, we give a higher speed, and vice versa. This strategy can give a smoother tracking than absolute positioning.

4 Proposed Cooperative Tracking Algorithm

The overview of the proposed cooperative tracking algorithm is shown in Fig. 2. It consists of multi-target tracking in the static camera, single-target tracking in the PTZ camera, and information fusing through transferred positions from the previous tracking result. Details are described in subsections.

4.1 Multi-target Tracking in Static Camera

The multi-target tracking algorithm in the static camera follows the tracking-by-association framework [21]. At each frame, pairwise association is performed to associate detection responses with tracklets. In Dual-Camera systems, most existing methods use background subtraction algorithms to detect objects. However, they can only detect moving objects and it is difficult to model the background in PTZ cameras. Hence, we use a fast state-of-the-art object detector [19] to detect objects. For each tracklet, a Kalman Filter is applied to refine the positions and sizes of its detection responses and estimate its velocity. The affinity measure $A_{i,j}$ to determine how well a detection D_i and a tracklet T_j are matched is defined as:

$$A_{i,j} = A_{appr}(D_i|T_j) A_{pos}(D_i|T_j) A_{pos}(D_i|T_j) \tag{2}$$

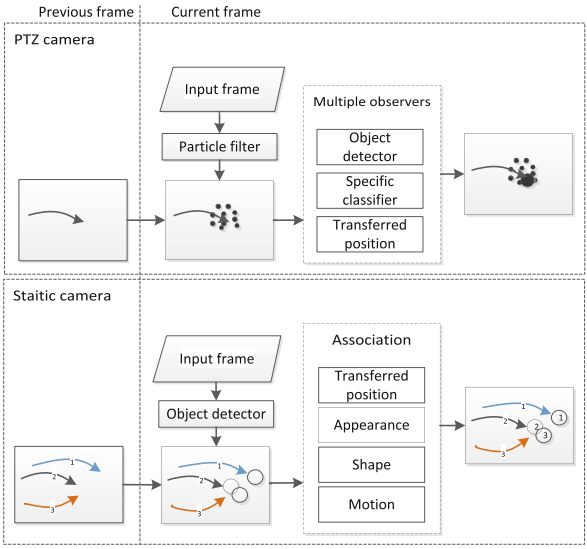

Fig. 2. The overview of the proposed cooperative target tracking algorithm.

The affinity is the product of affinities of appearance, shape and motion models, which are computed as follows:

$$\begin{aligned} A_{appr}(D_i|T_j) &= \sum_{u=1}^{m} \sqrt{h_u(D_i)h_u(T_j)} \\ A_{size}(D_i|T_j) &= \exp\left(-\left\{\left|\frac{h_{D_i}-h_{T_j}}{h_{D_i}+h_{T_j}}\right| + \left|\frac{w_{D_i}-w_{T_j}}{w_{D_i}+w_{T_j}}\right|\right\}\right) \\ A_{pos}(D_i|T_j) &= \mathcal{N}(P_{T_j} + v_{T_j}\Delta t; P_{D_i}, \Sigma_s) \end{aligned} \tag{3}$$

Color histograms of the detection responses of T_j are computed and averaged as the appearance model of T_j. The appearance affinity $A_{appr}(D_i|T_j)$ is the Bhattacharyya coefficient between the color histogram $h_u(D_i)$ of D_i and $h_u(T_j)$ of T_j. The bin number of the histogram is m. The shape affinity $A_{size}(D_i|T_j)$ is computed with the height h and width w of targets. $A_{pos}(D_i|T_j)$ is the motion affinity between P_{T_j} the last refined position of T_j and the position P_{D_i} of D_i with the frame gap Δt and velocity v_{T_j}. v_{T_j} is estimated by the Kalman Filter. The difference between the predicted position and the observed position is assumed to follow a Gaussian distribution $\mathcal{N}$.

Once the affinity matrix is computed, the optimal association pairs, which maximize the global association affinity in A, are determined using the Hungarian algorithm [20]. Detection responses which are not associated with any tracklet are used to generate new tracklets. To avoid false alarms, new tracklets are generated from detection responses with overlap bigger than 90 % in five consecutive frames.

4.2 Single-Target Tracking in PTZ Camera

For single-target tracking in the PTZ camera, we propose a robust algorithm within the particle filter framework. The particle filter approach is popularly

used for visual tracking. It can approximate multi-modal probability density function, so it is suitable for tracking in clutter [9]. Let Z_t and Y_t denote the latent state and observation, respectively, at time t. A particle filter approximates the true posterior state distribution $p(Z_t|Y_{1:t})$ by a set of samples $\{Z_t^i\}_{i=1}^N$ with corresponding weights $\{w_t^i\}_{i=1}^N$ which sum to 1. This can be done by the well-known two-step recursion. One is the Prediction, which uses a motion model to propagate the particles; the other is the Update, which uses observation models to compute the weight of each particle. In our implementation, we define the state at time t as $Z_t = \{x_t, y_t, u_t, v_t, s_t\}$, where (x_t, y_t) is the 2D image position, (u_t, v_t) is the velocity, and s_t is the scale of the bounding box.

Motion Model. We use a constant velocity motion model, which is defined as:

$$\begin{aligned}(x_t, y_t) &= (x_{t-1}, y_{t-1}) + (u_{t-1}, v_{t-1}) + \varepsilon_1 \\ (u_t, v_t, s_t) &= (u_{t-1}, v_{t-1}, s_{t-1}) + \varepsilon_2\end{aligned} \tag{4}$$

where ε_1 and ε_2 are drawn from zero-mean Gaussian distributions.

Observation Model. The observation model is a critical issue in the particle filter framework. The observation model should be robust against the main challenges in the PTZ camera tracking: abrupt motion, background clutters and appearance changes which are caused by occlusion, illumination variations and motion blur. To overcome these challenges, we propose a robust observation model. Firstly we use the category detector [19] to detect all the targets which belong to the same category. The category detector is trained offline, so it is more robust against background clutters, illumination variations and motion blur. Then an online classifier is trained to distinguish the interesting target from the other detections. If a detection is near the particles and is classified as positive by the classifier, we call this detection a strong observation. For each detection D_i in the PTZ camera, the matching score between a detection D_i and the interesting target T_a is measured as follows:

$$S(D_i, T_a) = c(D_i)(\frac{1}{N}\sum_{k=1}^{N}\mathcal{N}(P_{D_i}; P_k, \Sigma_a)) \tag{5}$$

where $c(D_i)$ is the classifier score; P_{D_i} and P_k are the positions of the detection D_i and the particle p_k; N is the number of particles. The detection with the maximum matching score among all the detections and the score of which is above a threshold θ, is considered as a strong observation D_*. The weight w_k for each particle p_k is computed as:

$$w_k = c(P_k) + \alpha \cdot \mathcal{I}(D_*) \cdot \mathcal{N}(P_k; P_{D_*}, \Sigma_a)) \tag{6}$$

where $c(P_k)$ is the classifier score of p_k, α is the parameter which controls the weight of the distance between the the particle p_k and the strong observation D_*, and $\mathcal{I}(D_*)$ is an indicator function that returns 1 if the strong observation is observed and 0 otherwise. If the strong observation is observed, it robustly guides the particles.

Classifier. In our implementation, we use online linear SVM as our classifier. The classifier score is measured as:

$$c(P) = \frac{1}{1 + \exp(-W \cdot F(I_P))} \tag{7}$$

where I_P is the image patch at the particle or detection location P with the corresponding size; $F(I_P)$ is the feature vector and W is the weights on features learned by linear SVM.

For efficiency, the same features (normalized gradient magnitude, histogram of oriented gradients, and LUV color) as the detector [19] are used by the classifier. When the same target is detected in the PTZ camera, positive samples are sampled near the detection, and negative samples are sampled from the other detections and background. In order to make the classifier robust against the detection noise, six image patches are sampled around the detection as positive samples with translations $\pm 0.05w$ in horizontal, translations $\pm 0.05h$ in vertical and scale changes $\pm 1/2^{\frac{1}{8}}$. w and h are the width and height of the detection bounding box. A linear SVM is trained using these samples. During tracking, to adapt to the appearance changes of the target, W is online updated using a passive-aggressive algorithm [10]. We only update the classifier when the strong observation is observed to prevent the tracking noise from being introduced into the classifier. The same sampling strategy is adopted to generate positive and negative samples for online updating.

4.3 Information Fusion

In situations with moderate crowd, in order to make the tracker robust against occlusions and target interactions, we use information fusion to improve the tracking accuracy.

We first define the tracking confidences in two cameras which are useful for information fusion. We denote the interesting target in the static camera as T_s, and in the PTZ camera as T_a. The multi-target tracking confidence for T_s at frame t is defined as:

$$conf_s^t = \frac{1}{M} \sum_{k=t-M+1}^{t} A_s(k) \tag{8}$$

where M is the length of the time window to compute the confidence, and $A_s(k)$ is the affinity score between T_s and the associated detection at frame k. $A_s(k)$ equals 0, if no detection is associated with T_s at that time. $conf_s^t$ is the average affinity score within the time window. The tracking confidence in the PTZ camera at frame t is defined as:

$$conf_a^t = \begin{cases} c^t(P_{D_*}) & \text{if strong observation is observed} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $c^t(P_{D_*})$ is the classifier score of the strong observation D_* at frame t.

In the association stage of multi-target tracking in the static camera, the affinity $A^f_{i,s}$ between T_s and the detection D_i with information fusion at frame t is measured as:

$$A^f_{i,s} = A^f_{pos}(D_i|T_s)A_{size}(D_i|T_s)A_{appr}(D_i|T_s) \tag{10}$$

The only difference between Eqs. (2) and (10) is the motion affinity $A^f_{pos}(D_i|T_s)$ for the interesting target T_s. For other targets, the affinities are computed by Eq. (2). Let P_{T_s} and P_{T_a} represent the bottom centers of the target bounding boxes of T_s and T_a in corresponding camera images respectively, and P_{D_i} denotes the bottom center of the detection D_i. $A^f_{pos}(D_i|T_s)$ is calculated as:

$$A^f_{pos}(D_i|T_s) = \mathcal{N}(P_{T_s} + V_{T_s}\Delta t; P_{D_i}, \Sigma_s) + \beta \cdot conf_a^{t-1} \cdot \mathcal{N}(P_{a\to s}; P_{D_i}, \Sigma_s) \tag{11}$$

where $P_{a\to s} = (H^{t-1}_{s\to a})^{-1}P^{t-1}_{T_a}$ is the target position in the static camera image which is projected from the position of the target in the PTZ camera image at frame $t-1$, using the homography $H^{t-1}_{s\to a}$ computed by Eq. (1), and β controls the importance of the transferred position $P_{a\to s}$. Compared with the motion affinity defined in Eq. (3), $A^f_{pos}(D_i|T_s)$ incorporates the target position in the PTZ camera image. After long occlusion, the predicted position $P_{T_s} + V_{T_s}\Delta t$ is unreliable, and may cause the association to fail. At this time, the tracking result in the PTZ camera can help the target find the correct association.

In the PTZ camera, incorporating the tracking result in the static camera as an extra cue, the matching score between a detection D_i and the target T_a is measured as:

$$S^f(D_i, T_a) = c(D_i)\left(\frac{1}{N}\sum_{k=1}^{N}\mathcal{N}(P_{D_i}; P_k, \Sigma_a) + \gamma \cdot conf_s^{t-1} \cdot \mathcal{N}(P_{D_i}; P_{s\to a}, \Sigma_a)\right) \tag{12}$$

where P_k is the bottom center of particle p_k. Since the PTZ camera has changed its parameters at frame t, we use $H^t_{s\to a}$ rather than $H^{t-1}_{s\to a}$ to map the target position $P^{t-1}_{T_s}$ in the static camera image to the position $P_{s\to a}$ in the PTZ camera image, i.e. $P_{s\to a} = H^t_{s\to a}P^{t-1}_{T_s}$. The weight of the transferred position $P_{s\to a}$ is controlled by γ. The difference between Eqs. (5) and (12) is that the transferred position $P_{s\to a}$ is added into Eq. (12) as a motion cue. Since the particles may deviate from the true target position after long occlusion, the transferred position can help the target to be matched with the correct detection.

The weight of each particle k with information fusion is computed as:

$$w_k = c(P_k) + \eta \cdot \mathcal{I}(D_*) \cdot \mathcal{N}(P_k; P_{D_*}, \Sigma_a) + \zeta \cdot conf_s^{t-1} \cdot \mathcal{N}(P_k; P_{s\to a}, \Sigma_a) \tag{13}$$

where the parameters η and ζ balance the weights of the corresponding items. Compared with Eq. (6), the transferred position $P_{s\to a}$ is fused into Eq. (13) which can help to guide the particles and make them surround the true position when occlusion occurs.

5 Experiments

Due to the difficulty in repeating the experiments with PTZ cameras, so far now, there is no unique real video benchmark which allows a genuine global testing. Most existing works only provide qualitative experiments in real scenarios. In the experiments, we firstly compare our algorithm with some popular tracking algorithms which are popularly used in the PTZ camera tracking, on a realistic data set quantitatively. Then, our algorithm is implemented on a real Dual-Camera system to show its effectiveness qualitatively, as in the other literatures.

5.1 Parameter Setting

All the parameters are set experimentally and kept fixed for all experiments. The covariance matrix Σ_s in Eqs. (3) and (11) is set to $\text{diag}[10^2, 20^2]$. The variances of the ε_1 in Eq. (4) are set proportionally to the width w of the tracking target, i.e. $(0.01w^2, 0.01w^2)$. The variances of the ε_2 in Eq. (4) are set to $(2^2, 4^2, 0.01^2)$. The covariance Σ_a in Eqs. (5, 6, 12 and 13) is set to $\text{diag}[0.04w^2, 0.04w^2]$. The parameter α in Eq. (6), β in Eq. (11), γ in Eq. (12), η and ζ in Eq. (13) are set to 10, 2, 0.5, 10, 1 respectively. The number of particles N in Eq. (5) is set to 100. The threshold θ which is used to determinate the strong observation is set to 0.4. The length of the time window M in Eq. (8) is set to 10. For all the algorithms which use color histogram, the histogram is calculated in the HSV color space using 10×10×5 bins.

5.2 Realistic Experiments

Data Set. We captured two synchronized videos at a shopping plaza using two cameras from different view points. The resolution of both videos is 1920×1080. One is served as the static camera video, and the other is served as the PTZ camera panorama video. For the static camera video, the resolution is reduced to 480×270 to simulate the static camera which monitors a large area. The static camera video is further cropped to 480×221 to ensure that all the pedestrians can be seen in the PTZ camera panorama video. For the PTZ camera panorama video, we generate the virtual PTZ camera view according to pan, tilt and zoom parameters using the method similar to [12]. The virtual view resolution of PTZ camera is set to 640×480. Some frames are shown in Fig. 3. The tracking results in the PTZ camera panorama video are labeled by us as ground truth (GT).

Evaluation Metrics. GT is projected to the PTZ camera view according to the camera pose and compared with the tracking result of the tracker in the PTZ camera image at the same time. The metrics proposed by [11] are used to compare different tracking algorithms:

– r_A^T: the mean overlap ratio between the bounding boxes of tracking results B_{tck} and the GT B_{GT}. The overlap ratio is computed as: $r_o = \frac{\|B_{GT} \cap B_{trk}\|}{\|B_{GT} \cup B_{trk}\|}$, where $\|\cdot\|$ is the area of the bounding box.

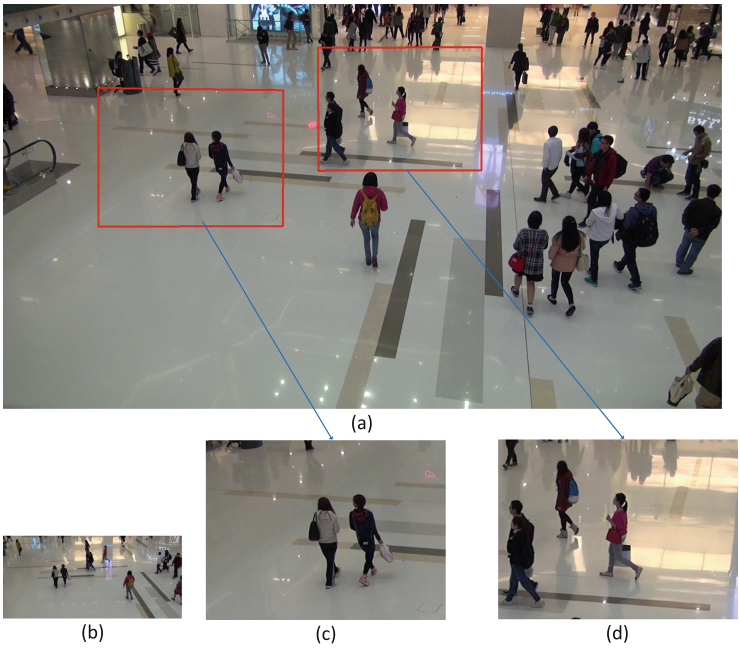

Fig. 3. (a) is the frame from the PTZ camera panorama view; (b) is the frame from the static camera view; (c) is the view generated from (a) with the virtual PTZ camera at parameters (pan:7.62, tilt:−2.17, focal length:4500); (d) is the view generated from (a) with the virtual PTZ camera at parameters (pan:−1.68, tilt:−4.37, focal length:5000)

- d_{ct}: the mean distance between the centers of B_{trk} and B_{GT}
- r_c: the percentage of correctly tracked frames (if $r_o \geq 0.5$)
- r_A^i: the mean overlap ratio between B_{GT} and the image bounding box B_i
- d_{ci}: the mean distance between the GT and the image center.

The first three criteria evaluate the accuracy of the algorithms, while the last two evaluate the ability of the system to keep the target in the center of the field of view and at the desirable resolution.

Compared Algorithms. According to which camera's tracking results are used to guide the PTZ camera, the tracking algorithms in Dual-Camera systems can be divided into two classes, static camera guided (SG) and PTZ camera guided (AG). Based on the information fusion strategy, the tracking algorithms can be categorized into four categories: with bidirectional fusion (BF), only with fusion from the static camera to the PTZ camera (AF), only with fusion from the PTZ camera to the static camera (SF), and without fusion (NF). Four popular tracking algorithms are compared: color-based particle filter(PF), KLT, Mean-shift (MS) and TLD [22] which is a real time and effective tracker shown by [23]. In addition to these algorithms, we also compare our cooperative tracking algorithm (TD) with different fusion and camera guiding strategies: TD-NF-SG, TD-NF-AG, TD-SF-SG, TD-AF-AG and TD-BF-AG, where NF, SF, AF and

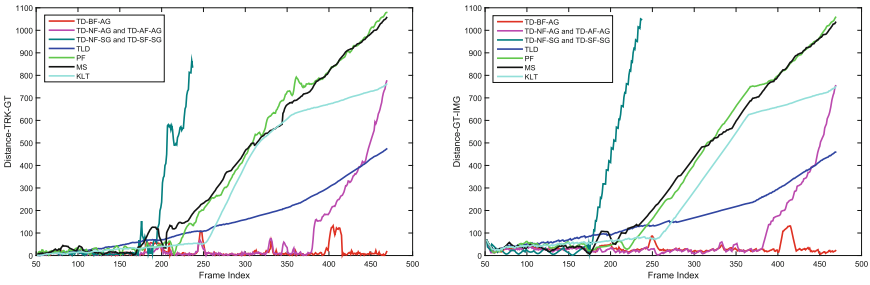

Fig. 4. (a) and (b) are the frame-by-frame tracking results in terms of the distance (in pixels) between the tracking result and the GT, and the distance (in pixels) between the GT and the image center, respectively. **Best viewed in color** (Color figure online).

Table 1. Comparison of different trackers. Bold font indicates best performance.

Methods	r_A^T	d_{ct}	r_c	r_A^i	d_{ci}
KLT	0.215	264.412	0.191	0.023	254.231
MS	0.163	337.003	0.170	0.021	326.09
PF	0.184	335.261	0.189	0.022	323.581
TLD	0.172	139.861	0.191	0.023	151.084
TD-SF-SG	0.415	104.334	0.299	0.035	152.855
TD-NF-SG	0.415	104.334	0.299	0.035	152.855
TD-AF-AG	0.508	76.725	0.756	0.041	80.662
TD-NF-AG	0.508	76.725	0.756	0.041	80.662
TD-BF-AG	**0.618**	**14.243**	**0.883**	**0.051**	**27.882**

BF indicate the information fusion strategies; SG and AG represent the camera guiding strategies. For example, TD-BF-AG represent our cooperative tracking algorithm with bidirectional information fusion, and using the tracking results in the PTZ camera to guide the movement of the PTZ camera. If not otherwise specified, in the tracking algorithms, the tracking result in the PTZ camera is used to guide the movement of the PTZ camera, due to its accuracy, and no information fusion is used as in the other literatures.

Results. Figure 4 illustrates the tracking results frame by frame in terms of the distance between the tracking result and the GT, and the distance between the GT and the image center. Table 1 shows the quantitative comparison results. From the tracking results, we can see that: due to the robust tracking algorithm in the PTZ camera, without information fusion, our method TD-NF-AG outperforms the other tracking algorithms; the tracking performance of the algorithm with one-directional information fusion is the same as the algorithm without information fusion; with bidirectional information fusion, our method TD-BF-AG achieves the best performance. Since the tracking results of TD-NF-SG and

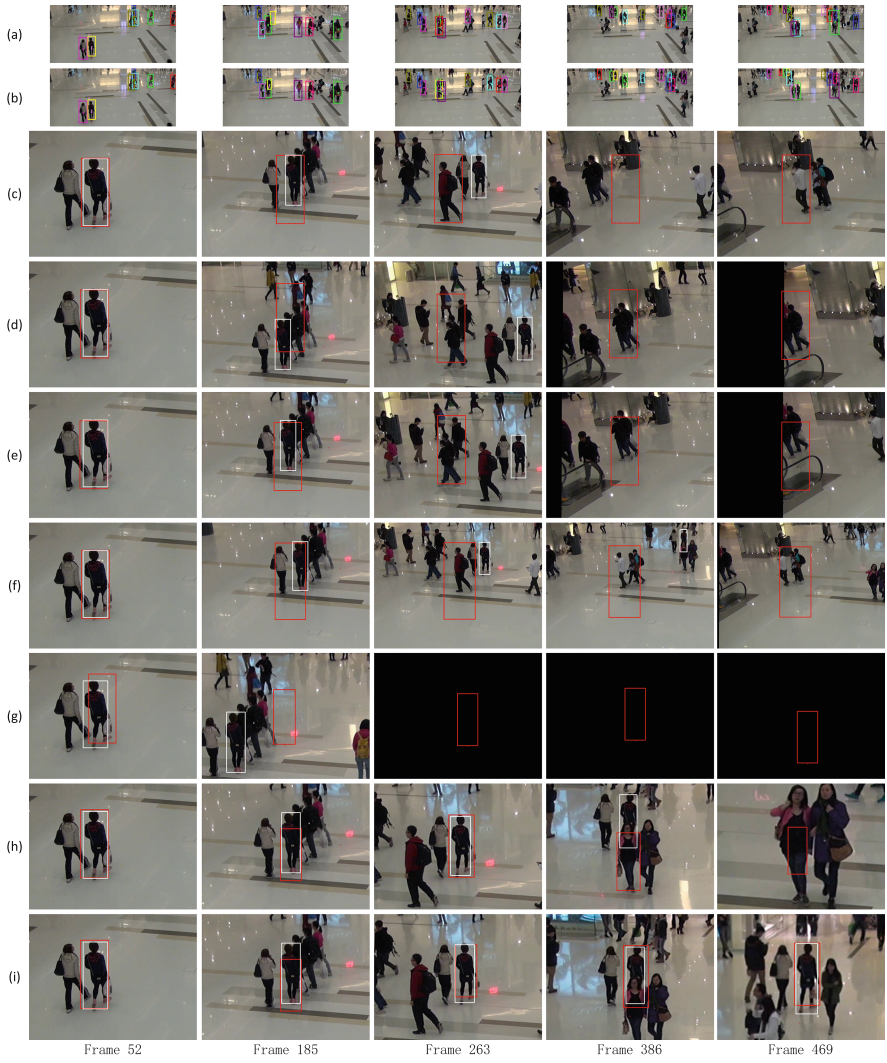

Fig. 5. (a) shows the tracking results in the static camera using methods excluding TD-BF-AG. (b) shows the tracking results in the static camera using TD-BF-AG. The interesting target in the static camera is indicated by the yellow bounding box. (c)∼(i) show tracking results in the PTZ camera, where (c)∼(f) use KLT, MS, PF and TLD trackers respectively; (g) shows the tracking results of TD-SF-SG and TD-NF-SG; (h) shows the results of TD-AF-AG and TD-NF-AG; (i) shows results using TD-BF-AG. Tracking results of the interesting target are represented by the red bounding boxes in the PTZ camera, and the ground truth is shown by white bounding boxes.

TD-SF-SG exceed the view of the PTZ camera panorama video after frame 263, their tracking results in Table 1 only consider frames from 52 to 263. This shows that static camera guided tracking is less accurate than PTZ camera guided

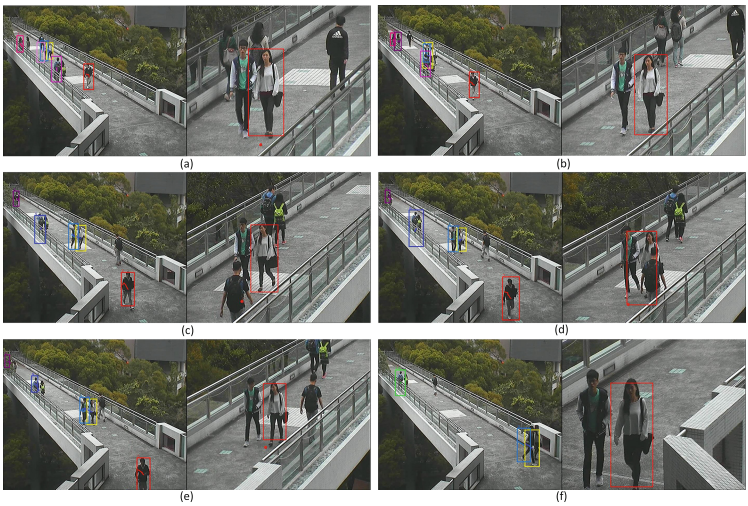

Fig. 6. An illustration of our cooperative tracking system. The left images in (a)∼(f) are the images in the static camera, and the right images in (a)∼(f) are the corresponding images in the PTZ camera. The tracking results in the static camera and the PTZ camera are indicated by the yellow bounding box and red bounding box respectively. (a) shows the interesting target is tracked by our algorithm. In (b), due to occlusion, the tracker fails in the static camera. The information fused from the PTZ camera helps the static camera find the correct association as shown in (c). In (d), due to occlusion in the PTZ camera, the tracker deviates from the true position of the target. The information fused from the static camera helps the tracker track the target correctly again when the target reappears, as shown in (e). Our cooperative tracking system keeps following the target at high resolution until the target exits from the view of the static camera as shown in (f).

tracking. The reason is that tracking in the static camera is difficult since the targets are small in the static camera images.

We try to give a qualitative comparison by showing in Fig. 5 some key frames. Due to targets with similar color, MS and PF trackers drift at about frame 263. The KLT tracker also drifts at about frame 263 due to occlusion. The TLD tracker fails at about frame 263, because the scale change is not estimated by the TLD tracker accurately, and makes the TLD detector fail. TD-SF-SG and TD-NF-SG fail at about frame 185, due to the target interactions. Although TD-SF-SG fuses information from the PTZ camera, it uses the results in the static camera to guide the PTZ camera which makes the target exit from the view of the PTZ camera before the information fusion helps it find the correct association. TD-AF-AG and TD-NF-AG fail at about frame 386 due to long occlusion. Although TD-AF-AG fuses information from the static camera, but the tracker has already failed in the static camera at about frame 185. The TD-BF-AG tracker can follow the target correctly through the whole sequence. The tracker fails in the static camera at about frame 185, but due to bidirectional

information fusion, the tracker correctly tracks the target again. At frame 386, due to long occlusion, the online classifier fails in the PTZ camera, but the particles are guided by the transferred position from the static camera. The target is correctly tracked again when the target reappears.

5.3 Real Trials

We apply the proposed robust cooperative tracking algorithm with bidirectional information fusion in a Dual-Camera system. Our Dual-Camera system is consist of two off-the-shelf AXIS PTZ Network Cameras Q6032-E. One is fixed to serve as the static camera to monitor a wide area, and the other is used as the PTZ camera. Our algorithm is implemented in C++ using OpenCV on an Intel Core I7-4700MQ 2.40 GHz with 8 GB RAM. Duo to the multithreaded implementation, our algorithm can run in real time at 20 fps. Some key frames are shown in Fig. 6 with a detailed description.

6 Conclusion

In this paper, we have proposed a robust tracking algorithm within the particle filter framework in the PTZ camera, which combines a category detector and an online classifier to make the algorithm robust against background clutters. Furthermore a bidirectional information fusion method is proposed to enhance the performance of cooperative tracking in crowded scenes. Finally, we compare different tracking algorithms which are frequently used by other researchers in realistic experiments, and also show the efficiency of our method in real trials. To the best of our knowledge, this is the first time different tracking algorithms in Dual-Camera systems are evaluated, and the results show our method outperform the others.

Acknowledgement. This work is supported by National Natural Science Foundation of China (NO. 61371192).

References

1. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recogn. Lett. **34**(1), 3–19 (2013)
2. Scotti, G., Marcenaro, L., Coelho, C., Selvaggi, F., Regazzoni, C.: Dual camera intelligent sensor for high definition 360 degrees surveillance. IEE Proc.-Vis. Image Sig. Process. **152**(2), 250–257 (2005)
3. Chen, C., Yao, Y., Page, D., Abidi, B., Koschan, A., Abidi, M.: Heterogeneous fusion of omnidirectional and PTZ cameras for multiple object tracking. IEEE Trans. Circ. Syst. Video Technol. **18**(8), 1052–1063 (2008)
4. Ghidoni, S., Pretto, A., Menegatti, E.: Cooperative tracking of moving objects and face detection with a dual camera sensor. In: 2010 IEEE International Conference on Robotics and Automation, pp. 2568–2573 (2010)

5. Cui, Z., Li, A., Feng, G., Jiang, K.: Cooperative object tracking using dual-pan-tilt-zoom cameras based on planar ground assumption. In: IET Computer Vision (2014)
6. Zhou, X., Collins, R.T., Kanade, T., Metes, P.: A master-slave system to acquire biometric imagery of humans at distance. In: First ACM SIGMM International Workshop on Video Surveillance, pp. 113–120 (2003)
7. Lu, Y., Payandeh, S.: Cooperative hybrid multi-camera tracking for people surveillance. Can. J. Elect. Comput. Eng. **33**(3/4), 145–152 (2008)
8. Fahn, C., Lo, C.: A high-definition human face tracking system using the fusion of omni-directional and PTZ cameras mounted on a mobile robot. In: The 5th IEEE Conference on Industrial Electronics and Applications, pp. 6–11 (2010)
9. Nummiaroa, K., Koller-Meierb, E., Van Gool, L.: An adaptive color-based particle filter. Image Vis. Comput. **21**(1), 99–110 (2003)
10. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. J. Mach. Learn. Res. **7**, 551–585 (2006)
11. Salvagnini, P., Cristani, M., Del Bue, A., Murino, V.: An experimental framework for evaluating PTZ tracking algorithms. In: Crowley, J.L., Draper, B.A., Thonnat, M. (eds.) ICVS 2011. LNCS, vol. 6962, pp. 81–90. Springer, Heidelberg (2011)
12. Possegger, H., Rüther, M., Sternig, S., Mauthner, T., Klopschitz, M., Roth, P.M., Bischof, H.: Unsupervised calibration of camera networks and virtual PTZ cameras. In: 17th Computer Vision Winter Workshop (2012)
13. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 564–577 (2003)
14. Bernardin, K., Van De Camp, F., Stiefelhagen, R.: Automatic person detection and tracking using fuzzy controlled active cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
15. Shi, J., Tomasi, C.: Good features to track. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)
16. Qureshi, F.Z., Terzopoulos, D.: Surveillance in virtual reality: system design and multi-camera control. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
17. Wu, Z., Radke, R.J.: Keeping a pan-tilt-zoom camera calibrated. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1994–2007 (2013)
18. Del Bimbo, A., Lisanti, G., Masi, I., Pernici, F.: Continuous recovery for real time pan tilt zoom localization and mapping. In: 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance, pp. 160–165 (2011)
19. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. **36**(8), 1532–1545 (2014)
20. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows (1988)
21. Bae, S., Yoon, K.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1218–1225 (2014)
22. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1409–1422 (2012)
23. Wu, Y., Lim, J., Yang, M.: Online object tracking: a benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2411–2418 (2013)