Multi-target Tracking via Max-Entropy Target Selection and Heterogeneous Camera 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. Nowadays, dual-camera systems, which consist of a static camera and a pan-tilt-zoom (PTZ) camera, have become popular in video surveillance, since they can offer wide area coverage and highly detailed images of the interesting starget simultaneously. Different from most previous multi-target tracking methods without information fusion, we propose a multi-target tracking framework based on information fusion of the heterogeneous cameras. Specifically, a conservative online multi-target tracking method is introduced to generate reliable tracklets in both cameras in real time. A max-entropy target selection strategy is proposed to determine which target should be observed by the PTZ camera at a higher resolution to reduce the ambiguity of multi-target tracking. Finally, the information from the static camera and the PTZ camera is fused into a tracking-by-detection framework for more robust multi-target tracking. The proposed method is tested in an outdoor scene, and the experimental results show that our method significantly improves the multi-target tracking performance.

1 Introduction

Multi-target tracking is important for activity analysis and anomaly detection in video surveillance systems. Surveillance cameras used in public areas (such as squares, parking lots, railway stations, airports, etc.) usually cover a large area. Therefore, the target size observed from these cameras is small, and the appearance information of the targets is not discriminative enough to distinguish different targets due to the low resolution. If the scene is crowded and long-time occlusion occurs frequently, the number of ID switches would increase significantly. This greatly hampers the performance of multi-target tracking. Increasing the number of cameras can solve this problem, but the costs would also increase. Recently, hybrid camera systems which consist of static cameras and PTZ cameras (which are also referred to as active cameras in this paper) have been widely used in video surveillance [1–10], since they can offer a wide range monitoring and close-up view simultaneously. The static camera can cover a large area providing the motion information of all observed targets, and the active camera is able to zoom in and focus on individual targets to obtain their discriminative appearance information.

Y.-S. Ho et al. (Eds.): PCM 2015, Part I, LNCS 9314, pp. 149–159, 2015.
DOI: 10.1007/978-3-319-24075-6_15

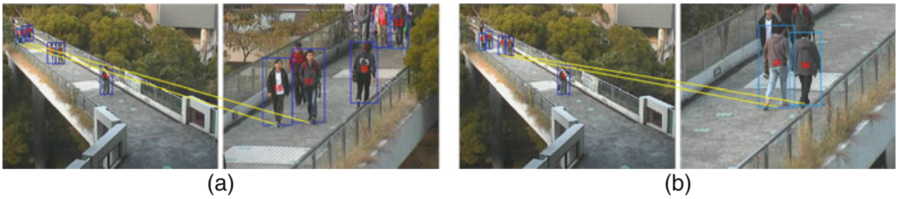

Fig. 1. Examples of tracking results in our dual-camera system. The left images in (a) and (b) are from the static camera, and the right images in (a) and (b) are from the active camera. Same persons moving in static and active cameras are linked with yellow solid lines (Colour figure online).

Previous tracking methods in dual-camera systems mainly focus on the spatial mapping between static cameras and active cameras [1–6], active camera control [7] and scheduling strategies [3]. Most of them use the static camera to detect and track all the targets that appear in the scene, and the active camera to track individual targets or simply capture the images of them at a higher resolution. During tracking, no information fusion from static cameras and active cameras is used to enhance the multi-target tracking performance. Although some of them like [8,9] fuse information from two cameras to improve the tracking accuracy, they focus on improving the single-target tracking accuracy. To the best of our knowledge, there is no research work that focuses on improving the overall multi-target tracking performance in such a dual-camera system.

In multi-target tracking, tracking-by-detection methods, which build long trajectories of targets by associating detection responses or tracklets, have become popular in recent years. Many state-of-the-art trackers [11–14] follow the tracking-by-detection framework. However, they suppose the cameras are homogeneous, and suffer from the low resolution of targets when applied to large-area monitoring systems.

In this paper, we propose a novel framework to improve the multi-target tracking performance in the surveillance systems where static cameras cover large areas and the target size is usually small (some examples are shown in Fig. 1). In our framework, the static camera covers a large area, detects and tracks targets online. The active camera observes one target at each time according to the tracking results in the static camera. The static camera provides the motion information of all observed targets, as the static camera can observe and track all targets simultaneous. The active camera is able to provide discriminative appearance information since it can observe targets at a higher resolution. We first introduce a conservative multi-target tracking method to generate reliable tracklets online in both cameras. Then, a max-entropy target selection method is proposed to choose one target to be observed by the active camera at each time, since there may be more than one target tracked by the static camera. The key idea is that the target with the most ambiguity may lead to the most tracking errors, and should be observed with the highest priority. According to the position of the selected target, the active camera adjusts its parameters to obtain close-up views of the target. The tracking results from both cameras

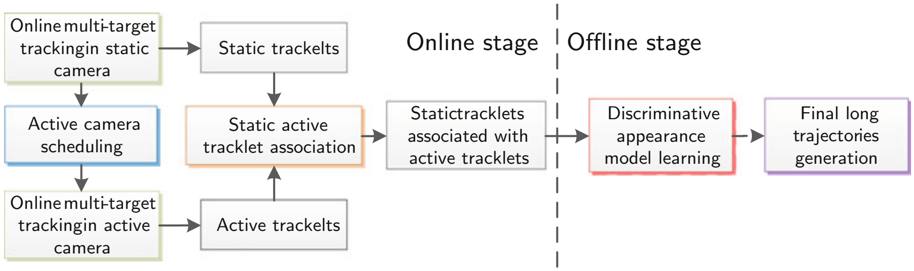
Online multi-target trackingin static camera
Static trackelts
Online stage
Offline stage
Active camera scheduling
Static active tracklet association
Statictracklets associated with active tracklets
Discriminative appearance model learning
Final long trajectories generation
Online multi-target trackingin active camera
Active trackelts

Fig. 2. The overview of our proposed system.

are finally fused into a tracking-by-detection framework to improve the tracking performance. The proposed multi-target tracking system is tested in a challenging outdoor environment, and compared with the state-of-the-art multi-target trackers. The experimental results show that with the cooperation of the active camera, the multi-target tracking performance in the static camera is improved significantly.

2 Our Method

The framework of our proposed method is shown in Fig. 2. It consists of two main stages: the online stage, and the offline stage. During the online stage, an efficient online multi-target tracking algorithm is used to generate tracklets in the static camera which are called *static tracklets*. The tracking results are used to guide the action of the active camera according to the proposed max-entropy target selection strategy. Then, the active camera focuses on the selected target for a short period of time. During this time, the same online multi-target tracking algorithm is used to generate reliable tracklets in the active camera, which are called *active tracklets*. Meanwhile, the active tracklets are associated with static tracklets. During the offline stage, static tracklets which may have associated active tracklets are associated to generate final long trajectories according to the affinity between them. To improve the association accuracy, a discriminative appearance model is learned for each static tracklet using the target image sequences from the static tracklets, which is called *static appearance model*. For each static tracklet which has associated active tracklets, an additional discriminative appearance model, which is called *active appearance model*, is learned using the target image sequences from the associated active tracklets. The appearance affinity between different static tracklets is computed based on the learned static and active appearance models.

2.1 Online Multi-target Tracking

During the online stage, reliable tracklets should be generated efficiently to guide the action of the active camera. We introduce an efficient and reliable online multi-target tracking method. At each frame, pairwise association is performed to associate detection responses with tracklets. The affinity measure to determine how well a detection and a tracklet (or two tracklets) are matched is defined the same as [13]:

$$\Lambda = \Lambda^A(X,Y)\Lambda^S(X,Y)\Lambda^M(X,Y) \tag{1}$$

where X and Y can be tracklets or detections. The affinity is the product of affinities of appearance, shape and motion models, which are computed as follows:

$$\begin{aligned} \Lambda^A(X,Y) &= \sum_{u=1}^{m} \sqrt{h_u(X)h_u(Y)} \\ \Lambda^S(X,Y) &= \exp\left(-\left\{\left|\frac{h_X - h_Y}{h_X + h_Y}\right| + \left|\frac{w_X - w_Y}{w_X + w_Y}\right|\right\}\right) \\ \Lambda^M(X,Y) &= \mathcal{N}(P_X + v_X \varDelta t; P_Y, \Sigma) \end{aligned} \tag{2}$$

The appearance affinity $\Lambda^A(X,Y)$ is the Bhattacharyya coefficient between the color histogram $h_u(X)$ of X and $h_u(Y)$ of Y. The bin number of the histogram is m. For each tracklet, a Kalman Filter is applied to refine the positions and sizes of its detection responses and predict its location. The shape affinity $\Lambda^S(X,Y)$ is computed with the height h and width w of targets. $\Lambda^M(X,Y)$ is the motion affinity between the position P_X of X and the position P_Y of Y with the frame gap $\varDelta t$ and velocity v_X estimated by the Kalman Filter. $\mathcal{N}$ is a Gaussian distribution with covariance matrix Σ. Given the affinity matrix, a conservative strategy is used to link detections with tracklets. A detection response is linked to a tracklet if and only if their affinity is higher than a threshold θ_1, and exceeds their conflicting pairs' affinities by a threshold θ_2. This strategy can avoid ID switches effectively when performing online multi-target tracking. New tracklets are generated from detection responses which are not associated to any tracklet. Tracklets which are not associated with any detection response for a certain period are terminated.

2.2 Active Camera Scheduling

In our dual-camera system, multiple targets are tracked in the static camera. It is necessary for the active camera to observe these targets according to their priority levels at each time. We observe that when performing final tracklet association, a tracklet which can be associated with more than one traklet with similar affinities, may change its ID with a higher probability. The priorities of targets are computed based on the ambiguity of the targets and the earliest deadline first policy. During online multi-target tracking, for each alive static tracklet t_i, we compute its affinity vector A with a tracklet set $T^o = \{t_k^o | k = 1 : m\}$, where t_k^o is the k-th tracklet in T^o. T^o is collected from previous tracklets whose end time is before the start time of t_i with a frame gap less than a certain length. $A = \{a_k | k = 1 : m\}$, where a_k is the affinity between t_i and t_k^o, and a_k is computed using Eq. (1). A is normalized such that a_k $(k = 1 : m)$ sum up to one. The ambiguity of target i is defined as $h_i = -\sum_{k=1}^{m} a_k \ln(a_k)$ following the definition of entropy. The priority of the target i is defined as:

$$w_i = \exp(-(t_e)/\sigma) \cdot h_i \cdot \delta(i) \tag{3}$$

where t_e is the predicted time for the target i to exit from the scene or be occluded, and σ controls the importance of t_e. $\delta(i)$ is an indicator function, it equals 0 if the target i has been observed by the active camera, otherwise it equals 1. We predict the future position of the target using the Kalman Filter. If the overlap ratio between the bounding box of the target and the ones of other targets exceeds 0.5, the target is considered in occlusion. The target with the highest priority would be observed by the active camera at each time.

2.3 Static and Active Camera Tracklet Association

Once the target with the highest priority is determined, the active camera is directed to observe the target for a short period of time at a higher resolution. It might be possible to detect multiple targets in the active camera, so we need to associate the targets in the active camera with the tracklets in the static camera. Associating the detection responses in the active camera with the static tracklets would be sensitive to the detection noise. Instead, we suggest to associate the active tracklets with the static tracklets. The score ψ_{ij} between static tracklet i and active tracklet j is defined as:

$$\psi_{ij} = \begin{cases} \exp(-\frac{\sum_{k=k_1}^{k_2} d(x_s^k, H(x_a^k))}{(k_2-k_1+1)*\gamma}) & \text{if } k_2 - k_1 >= 3 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where k_1 and k_2 are the start and end frame indexes of the time overlap between the two tracklets. x_s^k and x_a^k are the foot positions of the detection responses in the corresponding tracklets in frame k. d is the Euclidean distance. γ is a parameter controlling the smoothness of the score function. $H(\cdot)$ is the function mapping the position in the active camera to the position in the static camera using a homegraphy, which can be estimated by the calibration method [15].

To obtain the optimal assignment, we use the Hungarian algorithm [17] to assign the active tracklets to the static tracklets. Some association results are shown in Fig. 2.

2.4 Final Trajectory Generation

After the online stage, we obtain the reliable static tracklets from the static camera which may have associated active tracklets. These static tracklets are used to generate final long trajectories based on the affinity between them. To distinguish visually similar targets, discriminative appearance models should be learned for the tracklets. For each static tracklet, we train a static appearance model using the static camera image sequences and an active appearance model using the active camera image sequences if the static tracklet has an associated active tracklet.

Collecting Training Samples. We propose a method to collect positive and negative training samples to train the static appearance model and the active appearance model for each static tracklet t_i.

For the static appearance model, we randomly choose responses in the static tracklet t_i as positive samples. It is intuitive that one target can not appear at different locations at one time and targets in the static camera can not change its positions drastically. Tracklets, that have time overlap with t_i, or are far enough from t_i which makes the targets reach t_i within their time gaps impossibly, represent different targets. Negative samples are collected from these static tracklets.

For the active appearance model, if t_i has associated active tracklet t^a, we randomly choose responses in t^a as positive samples. Any active tracklet which has time overlap with the tracklet t^a can be matched with t^a impossibly and represents a different target. Active tracklets associated with the impossibly matched static tracklets of t_i can also unlikely represent the same target as t_i. Negative samples are collected from these active tracklets.

Appearance Model Learning. The goal of appearance model learning is to learn a model which determines the affinity score $S_{i,j}$ between two tracklets t_i and t_j. In each detection bounding box from tracklets, the color histogram and HOG histogram at different local patches are computed as features. Given a pair of detection responses d_1 and d_2 from two tracklets t_i and t_j, a linear combination of the similarity scores between local patches is learned, which takes the form:

$$S_{i,j}(d_1, d_2) = 1/2(\sum_{k=1}^{n} \alpha_k^i s_k(d_1, d_2) + \sum_{k=1}^{n} \alpha_k^j s_k(d_1, d_2)) \tag{5}$$

where s_k is the similarity computed at k-th local image region, and we use Bhattachayya distance to measure it. α_k^i and α_k^j are the target specific coefficients which are learned using the Adaboost algorithm [18] similar as [12]. The largest affinity score $S_{i,j}(d_1, d_2)$ between randomly sampled detection responses from t_i and t_j is chosen as the appearance affinity $S_{i,j}$ between t_i and t_j.

For a static tracklet, a static appearance model is learned using training samples from the static camera by Adaboost. If the static tracklet has an associated active tracklet, an active appearance model is learned in the similar way using training samples from the active camera.

Finally, the tracklet affinity is computed using Eq. (1), where $\Lambda^A(t_i, t_j) = S_{i,j}$. For tracklets which have active appearance models, the affinity computed from the active appearance model is chosen as $S_{i,j}$, since it is more discriminative than the static appearance model. Given the affinity matrix of different tracklets, the final trajectories are generated by applying the Hungarian algorithm [17].

It should be noted that, different from [12], we learn active appearance models for static tracklets which have associated active tracklets to enhance the multi-target tacking performance.

3 Experiments

3.1 Experiment Setting

To evaluate the effectiveness of our proposed algorithm, we implement it on a real dual-camera system and test the system in a challenging outdoor environment.

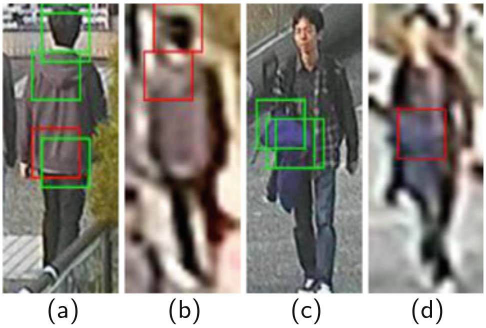

Fig. 3. Examples of the most discriminative features selected by static appearance models and active appearance models for two targets. (a) and (c) is the most discriminative features selected by active appearance models for target a and target b respectively. (b) and (d) are the corresponding discriminative features selected by static appearance models. The color and HOG descriptors are indicated by red and green bounding boxes, respectively.

We use two off-the-shelf AXIS PTZ Network Cameras Q6032-E in our experiments. One is fixed to serve as the static camera to monitor a wide area, and the other is used as the active camera. The typical height of the target is about 50 pixels in the static camera. Due to the zoom ability of the active camera, the height of the target is about 250 $\sim$ 300 pixels in the active camera. A detector [19] is trained to detect targets in the scene for its efficiency. The online multi-target tracking algorithm is run in both cameras in real time. The thresholds θ_1 and θ_2 of the online multi-target tracking algorithm are set to 0.4 and 0.1 repectively. We collected 10 videos at different time to evaluate our system.

Standard metrics for multi-target tracking are used to evaluate the proposed method: the multiple object tracking precision (MOTP), the multiple object tracking accuracy (MOTA), the number of fragments (FM) and identity switches (IDS). MOTP and MOTA are the higher the better. FM and IDS are the lower the better. We compare our method with two state-of-the-art multi-target trackers [16] and [12]. Both of them only use the images from the static camera. The difference between them is that [12] trains a discriminative appearance model to compute the appearance similarity, while [16] uses the color histogram to compute the appearance similarity. Different from [16] and [12], our method fuses information from both cameras through learning static discriminative appearance models and active discriminative appearance models.

3.2 Results

The most discriminative features selected by static appearance models and active appearance models for two targets are shown in Fig. 3. The HOG features are selected more by active appearance models, and the color features are selected more by static appearance models. This is reasonable, since the resolution of the target images in the static camera is rather low, and there is little useful shape

Table 1. Multi-target tracking results. Comparison of our method with the state-of-the-art methods.

Methods	MOTP[%]	MOTA[%]	FM	IDS
(a) Huang et al. [16]	76.15	74.33	23	20
(b) Kuo et al. [12]	82.23	81.75	18	15
(c) Our method (hist)	83.89	81.44	16	15
(d) Our method	92.88	93.13	5	4

Table 2. Multi-target tracking results. Comparison of different methods using different features.

Methods	MOTP[%]	MOTA[%]	FM	IDS
Kuo et al. [12] (color)	80.23	78.48	19	17
Kuo et al. [12] (hog)	77.66	75.78	21	19
Kuo et al. [12] (both)	82.23	81.75	18	15
Our method (color)	85.93	84.19	11	12
Our method (hog)	89.37	91.55	8	6
Our method (both)	92.88	93.13	5	4

details on the cloth of people. Compared with color information, the shape details are more discriminative when the color of people's cloth is similar. Therefore, when the resolution is high, more shape features are selected.

Table 1 records the multi-target tracking results. In Table 1, (c) is the result of our method without discriminative model learning, using the similarity of color histogram instead. (d) is the result of our method with discriminative model learning. The results shows that discriminative appearance models can improve the tracking performance, and active discriminative models learned from higher resolution images of the active camera can make the multi-target tracking performance even better, reduce the ID switch errors significantly.

Table 2 records the tracking results using different features when learning discriminative appearance models, which shows that hog features are more discriminative when using active discriminative models, and color features are more discriminative when using static discriminative models. This is consistent with the most features selected by them.

Figure 4 shows some sample results of multi-target tracking. The top row, middle row and bottom row show the results of [12,16] and our method respectively. Person #1 is tracked by [12,16] and our system in frame #441 (left column). Then the person are occluded for a long time, as shown in Frame #469 (middle column). The motion similarity is unreliable when long time occlusion happens. If the appearance information is not discriminative enough, ID switches may happen. When the person #1 reappears, ID switch occurs in results of both [16] and [12]. While our method associates it with the correct person. The discriminative models of person #1 are shown in Fig. 3 (a)-(b). The discriminative

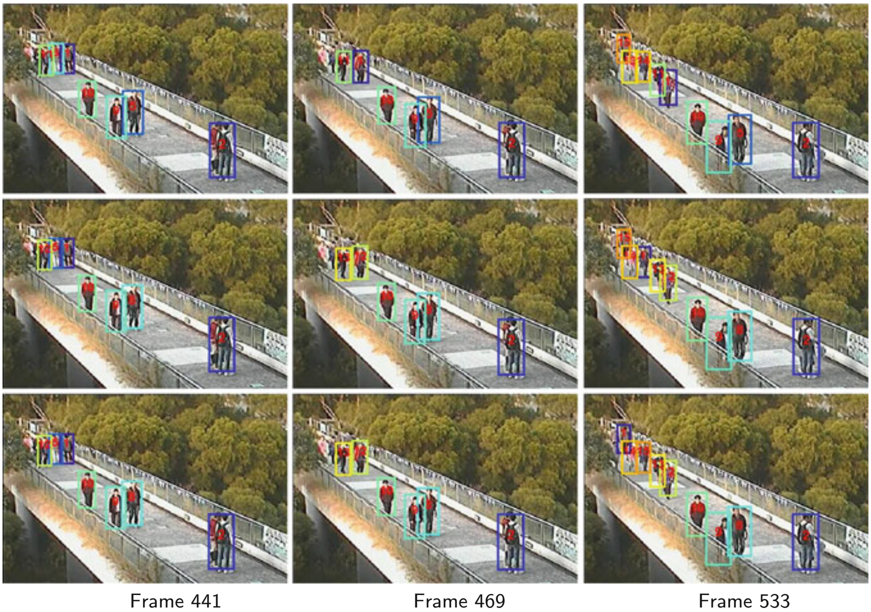

Fig. 4. Some tracking results on our collected video. The top row shows the results of [16], and the middle row shows the results of [12]. The results of our method are shown in the bottom row.

models of the person which the ID of person#1 changed to using method [12], are shown in Fig. 3 (c)-(d). The appearance similarity of the two persons computed using the active appearance model is 0.18, while the score is 0.63 using the static appearance model.

4 Conclusion

In this paper, we have proposed a novel framework for multi-target tracking in dual-camera surveillance systems. Information of the heterogeneous cameras is fused into a tracking-by-detection framework to improve the multi-target tracking performance, based on the learned discriminative appearance models in both cameras. To achieve this goal, an efficient online multi-target tracking algorithm is introduced to generate reliable tracklets. A max-entropy target selection strategy is proposed to reduce the ambiguity of multi-target tracking. Experiments in an outdoor scene show the significant improvement produced by our proposed method compared with the state-of-the-art multi-target trackers.

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

References

1. Horaud, R., Knossow, D., Michaelis, M.: Camera cooperation for achieving visual attention. Mach. Vis. Appl. **16**(6), 1–2 (2006)
2. 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)
3. Chen, C.H., Yao, Y., Page, D., Abidi, B., Koschan, A., Abidi, M.: Heterogeneous fusion of omnidirectional and PTZ cameras for multiple object tracking. IEEE Trans. Circuits Sys. Video Technol. (TCSVT) **18**(8), 1052–1063 (2008)
4. Senior, A.W., Hampapur, A., Lu, M.: Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration. In: IEEE Workshops on Application of Computer Vision (WACV), pp. 433–438 (2005)
5. Zhou, X., Collins, R.T., Kanade, T., Metes, P.: A master-slave system to acquire biometric imagery of humans at distance. In: ACM SIGMM International Workshop on Video Surveillance, pp. 113–120 (2003)
6. Alberto, D.B., Dini, F., Grifoni, A., Pernici, F.: Uncalibrated framework for on-line camera cooperation to acquire human head imagery in wide areas. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 252–258 (2008)
7. Bernardin, K., Van, F., Stiefelhagen, R.: Automatic person detection and tracking using fuzzy controlled active cameras. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
8. Cui, Y., Samarasckera, S., Huang, Q., Greiffenhagen, M.: Indoor monitoring via the collaboration between a peripheral sensor and a foveal sensor. In: IEEE Workshop on Visual Surveillance, pp. 2–9 (1998)
9. Yao, Y., Abidi, B., Abidi, M.: Fusion of omnidirectional and PTZ cameras for accurate cooperative tracking. In: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 46–46 (2006)
10. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recogn. Lett. **34**(1), 3–19 (2013)
11. Kuo, C.H., Huang, C., Nevatia, R.: Multi-target tracking by on-line learned discriminative appearance models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 685–692 (2010)
12. Kuo, C.H., Nevatia, R.: How does person identity recognition help multi-person tracking? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), PP. 1217–1224 (2011)
13. Bae, S.H., Yoon, K.J.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1218–1225 (2014)
14. Schindler, K.: Continuous energy minimization for multi-target tracking. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(1), 51–65 (2014)
15. Wu, Z., Radke, J.: Keeping a pan-tilt-zoom camera calibrated. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **35**(8), 1994–2007 (2013)
16. Huang, C., Wu, B., Nevatia, R.: Robust object tracking by hierarchical association of detection responses. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 788–801. Springer, Heidelberg (2008)
17. Munkres, J.: Algorithms for the assignment and transportation problems. J. Soc. Ind. Appl. Math. **5**(1), 32–38 (1957)

18. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Mach. Learn. **37**(3), 297–336 (1999)
19. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **36**(8), 1532–1545 (2014)