J Supercomput
https://doi.org/10.1007/s11227-017-2202-8

Online multi-person tracking assist by high-performance detection

Weixin Hua[1,2] · **Dejun Mu**[1] · **Zhigao Zheng**[3] · **Dawei Guo**[1]

Abstract Detection plays an important role in improving the performance of multi-object tracking (MOT), but most recently MOT works mainly focus on association algorithm and usually ignore the detections. To assist in associating object detections and to overcome detection failures, in this paper, we explore the low-rank-based foreground detection method to refine the detections and show it can significantly lead a better tracking result in online multi-object tracking. Firstly, the low-level pixel information from low-rank foreground segmentation and high-level detection responses from object detector are combined to form an overcomplete detections set, which serves as input for the tracking-by-detection-based multi-object tracking. Then, the predicted object location in online tracking as a prior to feedback for the foreground segmentation in sparse approximation for future frames can improve the foreground detection performance. Finally, to effectively solve the data association problem in online MOT, two-step data association relies on tracklet confidence is used to associate the detections and generate long trajectories since the existing trajectories provide a reliable history to support their presence in current frame. The experimental results in public pedestrian tracking datasets show that our detection optimization strategy can help to improve the tracking performance compared with several state-of-the-art multi-object trackers, with improved recall, precision, FP, FN and MOTA, MOTP results.

✉ Zhigao Zheng
zhengzhigao@hust.edu.cn

1 Shenzhen Research Institute, Northwestern Polytechnical University, ShenZhen 518000, China

2 School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

3 School of Computing Science and Technology, Huazhong University of Science and Technology, Wuhan, China

Published online: 30 November 2017

Keywords Multi-object tracking · Foreground segmentation · Two-step data association

1 Introduction

Although numerous multiple objects tracking algorithms have been proposed in recent years, the MOT tracking stills a challenging problem in computer vision field. The performance for the existing MOT algorithms shows some limitation in handling long-term occlusion, abrupt motion change of objects and un-reliable detections in complex scenes [1].

Tracking-by-detection (TBD) method has proven to be the state-of-the-art MOT algorithms. Most existing tracking-by-detection-based MOT works focus on data association strategy and usually ignore the detections where the set of detections are extracted by an object detector independently in each frame. Then, the detections serve as input to a tracker in tracking-by-detection-based MOT framework. An obvious shortcoming of TBD-based MOT algorithm is that most of the information in image sequence is simply ignored by threshold weak detection response. The trackers mainly concern how correctly associate these detections to build object trajectories over time.

TBD-based multi-object tracking techniques use the results of an object detector to guide the tracking process. The process of TBD-based multi-object tracking can be divided into two parts: the first one is detecting moving objects in each frame; the second one is associating the detections to build trajectories for objects over time. However, existing object detectors usually suffer from detection errors, which may mislead the trackers. Thus, improving the performance of the existing detections will consequently enhance the performance of detection-based tracking. Therefore, the detections play an important role for improving the performance of multi-object tracking. Since the sparse approximation and low-rank decomposition methods are the state-of-the-art foreground detection and background modeling methods, in this paper, we explore the low-rank based foreground detection method to refine the detections and associate every pixel to a special target or classifier it as background. By this strategy, we can use the low-level pixel information to deal with occlusion or absence of detections since the low-rank-based foreground detection and background modeling method give unique IDs for every pixel in image sequence. We show that our detection optimization strategies lead a better tracking result and significantly decrease the number of false positive (FP) and false negative (FN), improving recall and precision results. The main contributions of our work are:

1. To handle missing detections, a lot of image information, including high-level detector responses and low-level pixel information with low-rank-based foreground detection, are used to construct an overcomplete set of detections to guide data association during tracking.
2. To handle false detections caused by low-rank-based foreground detection, we implement a feedback loop for moving object detection, which passes the predicted object locations as a prior to guide foreground detection in sparse approximation for future frames. Meanwhile, the prior knowledge that the tracked objects are

pedestrians used to filter the detections by pre-trained pedestrian mask, which is helpful for segmenting each target and eliminating false detections.
3. The trajectories for existing objects provide a reliable history to support their presence in current frame; therefore, we solve the data association problem by implementing two-step data association based on tracklets confidence.

The rest of paper is organized as follows. The related work is described in Sect. 2. The proposed algorithm is presented in Sect. 3. Section 4 is the experimental results, and conclusion is presented in Sect. 5.

2 Related work

Numbers of multi-object tracking approaches have been presented in recent years. In this section, we mainly review closely related work on tracking-by-detection and segmentation-based method.

Multi-object tracking methods based on tracking-by-detection (TBD) framework are the state-of-the-art methods due to significantly development of the object detectors [2,3]. The main task for this kind of tracking can be split into two parts. The first one is obtaining a set of independent target detections. The second one is associating those target detections, achieving unique identities for each target and connecting them into consistent trajectories by performing data association. Due to the huge number of possible assignments between detections, the space of possible trajectories grows exponentially with the number of targets and increases frames. As a result, several association optimization methods have been proposed to improve association problem in MOT [4–9], such as, linear programming [6], dynamic programming [7], energy minimization [8,9], the maximum-weight independent set [4] and so on. The most existing association optimization methods are devised either for batch tracking or online tracking. The batch-based tracking builds object trajectories by globally associating past and future detections in the entire video [7,9–11]. While the online-based tracking builds multiple trajectories with frame-by-frame association up-to-now detections [12–18].

As is known, most existing TBD methods rely on a standard detector to obtain a set of detections as the input for MOT algorithm. The input detections are used for guiding the tracking process. However, even the state-of-the-art object detectors only work reliably up to a certain degree of occlusion, its suffers miss-detection or false detection in complex scenes with partial occlusion. In [19], Yu et al. prove that the detection plays a central role in data association-based MOT and high-performance detection can significantly improve the performance of multi-object tracking in both online and offline tracking. In [20], Esther et al. point out that the detections provided by object detector may themselves be noisy, containing false positives and misaligned detection bounding boxes; then, they address this issue by integrating level-set segmentation with top-down shape information in MOT. In [21], Leibe et al. consider object detection and space time trajectory estimation as a coupled optimization problem, the trajectory hypotheses are fed back to guide object detection. With the combined model selection framework, their multi-object tracking method can revise its decision and recover from mistake associations, which can cope with large-scale background

changes. Tang et al. [22] argue that the detectors used for tracking are typically trained independently from the tracker; then, they have proposed a joint multi-person detector method for tracking, which significantly improves both detection accuracy as well as tracker performance. In [23], Milan et al. introduce superpixel segmentation into tracking by detection; in their work, they have presented a joint tracking and segmentation framework for multiple objects tracking, which show significantly improved results of recover fragment trajectories in crowded sequences.

In our work, we introduce the low-rank-based foreground segmentation method into multi-object tracking, which can refine the detections provided by object detector. Then, we use high-level detection responses and low-level pixel information from low-rank foreground detection to formulate an overcomplete set as detection responses. The overcomplete detection responses act as the input for TBD to guide data association during multi-object tracking. Meanwhile, the predicted object location in online tracking as a prior to feedback for foreground detection in sparse approximation for future frames. To effectively associate the detections to form long trajectories, we use the tracklets confidence to guide the two-step data association during tracking since the existing trajectories provide a reliable history to support their presence in current frame. As the proposed method relies on an overcomplete detection set as initial state space for tracking, which used the low-rank foreground detection to guide multi-object tracking. However, the Principal Component Pursuit (PCP) based algorithms are considered to be the state-of-the-art methods for video background modeling and foreground detection [24,25], we exploit the incremental Principal Component Pursuit [26] online segment foreground objects from the background and associate every pixel to 'foreground' as a special target or classifier it as 'background'.

3 The proposed method

3.1 Problem formulation

For a video sequence, the goal of our online MOT is to form an overcomplete detection responses by combining high-level object detector and low-level segmentation. Then progressively associate the detections frame-by-frame to form long trajectories within the sequence. Here, the DPM detector [2] detects objects at each frame; thus, the high-level object detections at each frame are represented as Z^D. Then, we use the low-level detection results Z^s from the low-rank-based foreground segmentation to refine Z^D. By this way, a lot of missing detections can be recovered. Meanwhile, with the prior knowledge that the tracked objects are pedestrian, we use the pre-trained pedestrian mask to filter the detections. In this way, the false detections such as trees and obstacles in background can be effectively removed. Next, the tracking process implemented by two-step data association relies on tracklet confidence. The first stage of data association progressively built long trajectories by associating reliable tracklets with detections. While the second stage of data association is reassignment problem for un-reliable tracklets. By associating the un-reliable tracklets with reliable tracklets or detections, it can recover fragmented tracklets to link with others, which is beneficial

for building optimal tracklets. After above process, the object states and trajectories are updated with the associated results.

3.2 Low-level foreground segmentation

Since our algorithm relies on an overcomplete detection set as initial state space for tracking, the first step for the proposed method is to segment the moving objects from the background. This can be regarded as a video background modeling problem. As the Principal Component Pursuit (PCP) based algorithms are considered to be the state-of-the-art methods for video background modeling [24,25], in our work, we follow the work in [26] online segment foreground objects from the background and associate every pixel to 'foreground' as a special target or classifier it as 'background'.

The most existing PCP methods such as [27–30] are batch methods with huge computation cost, which are unsuitable for online MOT. Therefore, we proposed a new PCP method that can process one frame at a time and introduce it in online MOT. In our work, we take the predicted object locations in online MOT as a prior to implement a feedback loop for moving object detection, which is beneficial for improving foreground detection performance in sparse approximation for future frames.

Given a video sequence D, we decompose the input image D into a low-rank matrix B and a binary sparse matrix E, where the sparse outlier $E = S + G$, consisted by moving objects S and noise G. Then, the PCP optimization problem can be formulated as follows:

$$\min_{L,E} \frac{1}{2} \|D - B - E\|_F^2 + \lambda_1 \|B\|_* + \lambda_2 \|E\|_1 \tag{1}$$

where $D = [d_1, d_2, \ldots, d_n] \in \mathbb{R}^{m \times n}$ is a matrix of n images and each of video frame denoted as d_t, $t \in \{1, 2, \ldots, n\}$, is a vectorized image. $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_*$ is the nuclear norm, and $\|\cdot\|_1$ is ℓ_1 norm indicates the number of nonzero entries of a matrix. λ_1 and λ_2 are two regularization parameters.

For the input observed image D, $D = B + E$, B and E represent background region and foreground region, respectively. In the background region, $E_{ij} = 0$; then, the $D_{ij} = B_{ij}$. Otherwise, in the foreground regions, $E_{ij} = 1$; then, we can represent the input image as $D_{ij} = E_{ij}$ due to the background is occluded by foreground, where i and j are the pixel index for a given image. Instead of solving (1) directly, we introduce prior motion information of the foreground objects to improve the foreground detection since the low-rank approximation is specially applied in online MOT. In online MOT, the dynamic status of the objects can be predicted in advance by motion model such as Kalman filter or particle filter. In our work, the predicted object location is fed to guide the foreground detection in sparse approximation for future frames. Therefore, the optimization problem in Eq. (1) can be rewritten as follows:

$$\min_{L,S} \frac{1}{2} \|D_k - B_k - E_k\|_F^2 + \lambda_1 \|B_k\|_* + \lambda_2 \|W \circ E_k\|_1 \quad \text{s.t. } \operatorname{rank}(B_k) \leq t \tag{2}$$

where W is a mask matrix valued in [0, 1], indicating the locations of the foreground objects. This mask is predicted via motion model used in online MOT. In our work,

we simply use the Kalman filter to predict object state. $\circ$ represents the Hadamard product of the matrix.

Then, an incremental approach is used to solve Eq. (2) by implementing the two-step alternating minimization as follows:

$$B_t^{k+1} = \arg\min_{L_t} \left(\left\| D_t - B_t - E_t^k \right\|_F^2 \right), \quad \text{s.t. } \operatorname{rank}(B_t) \leq t \tag{3}$$

$$E_t^{k+1} = \arg\min_{E_t} \left\| D_t - B_t^{k+1} - E_t \right\|_F^2 + \lambda_2 \left\| W_k \circ E_t \right\|_1 \tag{4}$$

where $W_t = \{w_{ij}\}_{i,j=1}^n$ is the pixel value matrix of the predicted position by Kalman during online tracking, valued in [0, 1]. $D_t = [D_{t-1}, d_t]$, $B_t = [B_{t-1}, b_t]$ and $E_t = [E_{t-1}, e_t]$. D_t, B_t and E_t represent the input image set, background set and the set of sparse outlier up to tth frame, respectively. d_t, b_t and e_t are the input image, background and sparse outlier in tth frame, respectively.

For current frame, the input observed image is d_t, then the formed image set is denoted as D_t and the corresponding background be B_t. We need to compute the foreground mask S_t from the sparse outlier $E_t = D_t - B_t$. With the predicted object locations in online MOT as a prior to implement MRF algorithm to obtain the foreground mask [31], the energy function of foreground mask S_t is defined as follows:

$$\lambda_2 \sum_{t \in \nu} S_t(ij) + \gamma \sum_{(ij,kl) \in \varepsilon} \left| w_{ij} s_i - w_{kl} s_j \right| \tag{5}$$

where $g = (\nu, \varepsilon)$ is the graph in MRF, ν is the set of vertices pixel for the observed image, ε is the set of edges that connect neighboring pixels.

As the sparse outlier E_t is combination of foreground objects S and noise G. In most cases, the noise comes from dynamic background such as illumination changes and shadow moving. The noise almost has no influence on background estimation because they are also belonged to the moving parts. However, they can cause false alarm during foreground detection. Therefore, we separate S_t from the noise with the help of the foreground mask W_t predicted by motion model in online tracking. Then, the objective function to minimize the energy over S_t via sparse outliers E_t can be rewritten as follows:

$$\begin{aligned} &\min_S \frac{1}{2} \|E\|_F^2 + \lambda_2 \|S_t\|_1 + W_t \|\varphi(S_t)\|_1 \\ &= \min_{t:S} \sum_{ij, s_{ij} \in \{0,1\}} (D_{ij} - B_{ij})^2 + \lambda_2 \sum_{ij} w_{ij} s_{ij} + \gamma \sum_{(ij,kl) \in \varepsilon} \left| w_{ij} s_i - w_{kl} s_j \right| \end{aligned} \tag{6}$$

By using graph-cut [32] to solve the optimization problem in Eq. (6), the foreground pixels of the input image are achieved, the results of Eq. (6) is the binary foreground mask S_t.

As described above, we make a conclusion in algorithm 1 to show how to use the low-rank-based method achieves the low-level foreground segmentation.

Algorithm1. Low-rank based foreground segmentation

Input: observed image D, regularization parameter λ_1, λ_2, number of outerLoops, eigenvalue tolerance τ, W = predicted objects location,

Output: background B_t, sparse outliers E_t and binary foreground mask S_t

Initialization: D = input video, W, initial rank $t = 1$

for $t = 1$ to n

Initial solution ($k = 0$)

$$L_t^0 = \arg\min_{L}(\|(D - B)\|_F^2), \text{s.t.rank}(B) \le t$$

$$S_t^0 = \arg\min_{S} \|(D - B - S)\|_F^2 + \lambda_2 \|W_t \circ S\|_1$$

for $k = 1, 2, \cdots, outerLoops$

if $\frac{v_{rank}}{\sum_{k=1}^{rank} v_k} > \tau$ then $++t$ (v:singular values from $k-1$)

solve $B_t^{k+1} = \arg\min_{L_t}(\|D_t - B_t - E_t^k\|_F^2), \text{s.t.rank}(B_t) \le t$

$$E_t^{k+1} = \arg\min_{E_t} \|D_t - B_t^{k+1} - E_t\|_F^2 + \lambda_2 \|W_k \circ E_t\|_1$$

end

solve Eq. (6)

end

For the initialization process where $t = 1$, the input image $D_1 = d_1$. In the first frame of the video, we can consider the detections Z_1^D provided by object detector are the truly objects in scenes, and simply assumes $W_1 = ones(d_1)$, $L_1 = D_1 - Z_1^D$.

3.3 Object shape for removing false detections

With the low-rank-based foreground segmentation in Sect. 3.2, we achieve the foreground motion objects $S_t = \{s_t^1, \ldots, s_t^{N_s}\}$ for any frame, which can be regarded as low-level detections $Z_t^s = S_t$. The DPM detector provides high-level object detections at tth frame, which can be represented as $Z_t'^D = \{z_t'^1, \ldots, z_t'^{N_D}\}$. Then, the overcomplete detection responses constructed by foreground segmentation results and high-level object detections can be represented as follows:

$$Z_t = \{z_t^1, \ldots, z_t^N\}$$

$$z_t = \begin{cases} z_t'^D & \text{if } \theta \ge 0.6 \\ s_t^s & \text{else} \end{cases}, \quad \theta(s_t^s, z_t'^D) = \frac{\mathrm{O}(s_t^s) \cap \mathrm{O}(z_t'^D)}{\mathrm{O}(s_t^s) \cup \mathrm{O}(z_t'^D)} \tag{7}$$

where $\theta(\cdot)$ is the position-size affinity between the low-level detections and the high-level object detections. $\mathrm{O}(s_t^s) \cap \mathrm{O}(z_t'^D)$ and $\mathrm{O}(s_t^s) \cup \mathrm{O}(z_t'^D)$ are the intersection and union area of s_t^s and $z_t'^D$, respectively.

For each frame, the detections Z_t are constructed by foreground segmentation results and high-level object detections. However, the detections contain some false detection, such as shadow moving, wave trees and illumination changes. Then, we use the prior knowledge that the tracked objects are pedestrians to refine the detections Z_t. The pre-trained pedestrian mask M is achieved by averaging multiple annotated silhouettes as [23], which contained 2111 annotation pedestrians from the UrbanStreet dataset [33]. With the pedestrian mask M, the final detections Z_t for every frame are achieved by comparing the shape difference between the detections and the pedestrian

mask. If the difference is small than a pre-defined threshold (set it 0.6 in our work), it can be regarded as a real detection for pedestrian. After this step, a lot of false detection can be effectively removed.

3.4 Data association based on the Tracklet confidence

For current frame t, the final detection responses Z_t are achieved after implementing the processes of Sects. 3.2 and 3.3. $X_t = \{x_t^1, \ldots, x_t^M\}$ and $Z_t = \{z_t^1, \ldots, z_t^N\}$ represent the object states and detections in current frame t, respectively. The state of x_t^i and detection of z_t^j are represented as $x_t^i = \{p_t^i, s_t^i, v_t^i\}$ and $z_t^j = \{p_t^j, s_t^j, v_t^j\}$, respectively, where p_t, s_t and v_t^i are the position, size and velocity of an object. For online MOT, the trajectories $T_{1:t}^i$ of object i up to frame t can be represented as $T_{1:t}^i = \{x_k^i | 1 \leq t_s^i \leq k \leq t_e^i \leq t\}$, where t_s^i and t_e^i are the start and end frame of a tracklet. The set of object trajectories in current frame t is represented as $\mathbf{T}_t$.

Tracklet confidence As defined in [15], a tracklet confidence reflects how well a tracklet matches the real trajectory of the object. With considering the length of a tracklet, the missing detections and appearance affinity between a tracklet and associated detections, the tracklets confidence is defined as follows:

$$\text{conf}(T^i) = \left(\frac{1}{L_i} \sum_{t \in [t_s^i, t_e^i]} p(T^i, z^j) \right) \times \exp\left(-\beta \cdot \frac{M}{L_i} \right) \tag{8}$$

where L_i is the length of tracklet T^i, $p(T^i, z^j)$ is the appearance affinity. $M = t - t_s^i - L$ is the number of missed detections for object i. β is a control parameter. Since the tracklets confidence with value in 0–1, a trajectory with $\text{conf}(T) > 0.5$, it is regarded as reliable tracklets; otherwise, it is regarded as an un-reliable tracklets, which is likely to drift or fragment.

Stepwise construct trajectories with reliable tracklets This stage implements data association between reliable tracklets and detections to build long trajectories. Assume there are h reliable tracklets $T_{t-1}^{(hi)}$ defined in Eq. (8) in $t-1$ frame and m detections in frame t, then, the set of reliable tracklets can be represented as $\mathbf{T}_{t-1}^H$, and the set of detections in frame t is Z_t. Consequently, the pairs of association in frame t are represented as $\{(T_{t-1}^{(hi)}, z_t^j)\}$, where $T_{t-1}^{(hi)} \in \mathbf{T}_{t-1}^H$ and $z_t^j \in Z_t$. The affinity matrix $C_{h \times n}$ can be defined as follows:

$$\begin{aligned} C &= [c_{ij}]_{h \times n}, c_{ij} = -\log\left(p(T_{t-1}^{(hi)}, z_t^j) \right) \\ p(T_{t-1}^{(hi)}, z_t^j) &= p_a(T_{t-1}^{(hi)}, z_t^j) p_s(T_{t-1}^{(hi)}, z_t^j) p_m(T_{t-1}^{(hi)}, z_t^j) \end{aligned} \tag{9}$$

where $p(T_{t-1}^{(hi)}, z_t^j)$ is the affinity score. $p_a(\cdot)$, $p_s(\cdot)$ and $p_m(\cdot)$ are the appearance, position-size and motion affinity of the input tracklet–detection pair, respectively. The input pairs are determined by Hungarian algorithm [34] to achieve the minimum cost of $C = [c_{ij}]_{h \times n}$. If the association cost for input pair is smaller than a pre-defined threshold $-\log(\eta)$, z_t^j is matched with the $T_{t-1}^{(hi)}$. Then, the object states such as position, size and tracklet confidence values are updated with the matched detections.

Consequently, the set of unmatched reliable tracklets and detections are represented as $T_t^{H_U}$ and Z_t^U, respectively.

Reassignment association for un-reliable tracklets This stage solves the association problem for un-reliable tracklets. As is known, the un-reliable tracklets with low-confidence are trend to be fragmented, they may be associated with unmatched reliable tracklets or detections which are not associated in stepwise construct trajectories stage, or may be terminated. Assume there are q un-reliable tracklets $T_{t-1}^{(li)}$ defined in Eq. (8) in $t-1$ frame and q' unmatched detections and h' unmatched reliable tracklets in frame t, then the set of un-reliable tracklets, unmatched detections and unmatched reliable tracklets can be represented as $\mathbf{T}_{t-1}^L$,$\mathbf{T}_t^{H_U}$ and Z_t^U. Consequently, for current frame t, the input association between un-reliable tracklets and unmatched detections is represented as $\{(T_{t-1}^{(li)}, z_t^j)\}$, where $T_{t-1}^{(li)} \in \mathbf{T}_{t-1}^L$ and $z_t^j \in Z_t^U$. The affinity matrix $A_{q\times q'}$ between un-reliable tracklets and unmatched detections is equal to Eq. (9). The fragmented tracklets association between un-reliable tracklets and unmatched reliable tracklets to generate longer trajectories can be represented as $\{(T_{t-1}^{(li)}, T_t^{(hj)})\}$, where $\mathbf{T}_{t-1}^{(li)} \in T_{t-1}^L$ and $T_{t-1}^{(hi)} \in \mathbf{T}_t^{H_U}$, the affinity matrix for this case can be defined as follows:

$$\begin{aligned} V &= [v_{ij}]_{q\times h'},\ v_{ij} = -\log(p(T_{t-1}^{(li)}, T_t^{(hj)})) \\ p(T_{t-1}^{(li)}, T_t^{(hj)}) &= p_a\left((T_{t-1}^{(li)}, T_t^{(hj)})\right) p_s\left((T_{t-1}^{(li)}, T_t^{(hj)})\right) \end{aligned} \tag{10}$$

In Eq. (10), we only consider the appearance and position-size affinity for the input association pairs. This is because the motion affinity based on simplified motion model for fragmented tracklets is not reliable. The fragmented tracklets often caused by occlusion or abrupt motion. If the occlusions last long-term period or the tracked object changes its motion frequently, the confidence value of tracklet decreases, then the reliable tracklets are converted to un-reliable tracklet. Consequently, the motion dynamics for the object is changed drastically, it cannot be estimated by simplified motion model like Kalman or Markov.

The probability for un-reliable tracklets to be terminated can be defined as follows:

$$U = \mathrm{diag}[u_1, \ldots, u_q], \quad u_i = -\log(1 - \mathrm{conf}(T^{i(lo)})) \tag{11}$$

Then for current frame t, the affinity matrix in reassignment association stage can be defined as follows:

$$R_{(q+q')\times(h+q)} = \begin{bmatrix} A_{q\times q'} & -\log(\eta)_{q'\times h'} \\ U_{q\times q} & V_{q\times h'} \end{bmatrix} \tag{12}$$

where η is a threshold equal in stepwise association stage. Equation (12) is solved by Hungarian algorithm. After the association, the object states and the value of tracklets confidence are updated with their associated results.

Association affinity model The association affinity model is used to measure the affinity between the input association pairs. For any frame t, the set of object states can be represented as $X_t = \{x_t^1, \ldots, x_t^M\}$, where $x_t^i = \{p_t^i, s_t^i, v_t^i\}$, p_t, s_t and v_t are the position, size and velocity of an object. The appearance, position-size and motion

affinities used in our work are defined as follows:

$$
\begin{aligned}
p_a(X,Y) &= \exp\left(-\sum_{b=1}^{B}\sqrt{\mathrm{H}^b(X)\mathrm{H}^b(Y)}\right)\\
p_s(X,Y) &= \exp\left(-\left(\frac{h_X-h_Y}{h_X+h_Y}+\frac{w_X-w_Y}{w_X+w_Y}\right)\right)\\
p_m(X,Y) &= \begin{cases}\mathbb{N}\left(p_X^{\text{tail}}+v_X^{\text{tail}}\Delta t-p_Y^{\text{head}},\Sigma_p\right) & \text{if } Y \text{ represents detection}\\ \mathbb{N}\left(p_X^{\text{tail}}+v_X^{\text{tail}}\Delta t-p_Y^{\text{head}},\Sigma_p\right)\\ \mathbb{N}\left(p_Y^{\text{head}}-v_Y^{\text{head}}\Delta t-p_X^{\text{tail}},\Sigma_p\right) & \text{else}\end{cases}
\end{aligned}
\tag{13}
$$

where X and Y can be tracklet or detection. For appearance affinity $p_a(X|Y)$, we adopt color histogram with $\mathbf{B} = 64$ for each HSV color space. The shape affinity $p_s(X|Y)$ is measured by height (h_X, h_Y) and width (w_X, w_Y) of the objects. In terms of the motion affinity $p_m(X|Y)$ between the tail of tracklet X (the last refined position) and Y head (the first refined position of tracklet or detection) with frame gap Δt is assumed to satisfy Gaussian model $\mathbb{N}$, p and v are the positions and velocities for the head part or tail part of T.

As described above, after finishing the association, the non-associated detections are regarded as candidate tracks and when the candidate track is associated in five consecutive frames, it is converted into a new track. Meanwhile, the non-associated trajectories are terminated if they are unassociated in five consecutive frames. The main steps of the proposed multi-object tracking method are summarized in Algorithm 2.

Algorithm2. The overall algorithm for the proposed online multi-object tracking

Input: Object state set X_t and the set of detection responses Z_t in frame t

Output: updated trajectory set $\mathbf{T}_t$

Step1: Low-level foreground segmentation

1. background modeling and sparse outlier estimation based on PCP method;
2. foreground mask extraction with the prior information predicted by Kalman filter;

Step2: construct overcomplete detection set Z_t with the results of low-rank based foreground segmentation and high-level object detection responses;

Step3: remove false detections with object shape mask;

Step4: data association optimization based on the tracklets confidence

1. calculate the tracklet confidence value according to Eq. (8), bipartition tracklets into reliable tracklets and un-reliable tracklets;
2. stepwise construct trajectories with reliable tracklets $\mathbf{T}_{t-1}^{H}$ and detections Z_t;
3. reassignment association the un-reliable tracklets $\mathbf{T}_{t-1}^{L}$ with un-matched reliable tracklets $\mathbf{T}_t^{H_U}$ or un-associated detections Z_t^U;

Step5: construct association matrix $C_{l\times n}$ and $R_{(q+q')\times(l+q)}$ as shown in Eq. (9) and (10), respectively. Implement the Hungarian algorithm to achieve the optimal input association pairs;

Step6: update the object states and the tracklets confidence values with the associated results. The non-associated detections regard as candidate tracks and the un-associated trajectories are remained as potential terminated trajectories.

4 Experiments

4.1 Dataset and detections

The performance of the proposed multi-object tracking method is evaluated on two typical public pedestrian datasets: PETS 2009 and TUD dataset. In terms of the performance evaluation on PETS 2009 dataset, three sequences S1.L1, S2L1 and S3MF sequences from S1, S2 and S3 dataset are used, those pedestrian sequences with different crowd density contain numerous occlusions. For the TUD dataset, the crossing and campus sequences are used for performance evaluation. The TUD dataset is a challenging sequence due to the videos captured on streets with a very low camera angle in close range, which easily cause frequently full occlusion among pedestrians. In these sequences, there exist various challenges such as occlusion, clutters background, scale and pose changes. In terms of the high-level object detections provided by object detector for those sequences, we use the DPM detector [2].

4.2 Parameters setting, evaluation metrics and baseline methods

The proposed online MOT algorithm is performed on MATLAB 2014b with an Intel Core i7 8GHz PC. The average run time is about 15 fps without any code optimization and parallel programming. In the experimental, we empirically set $\lambda_1 = \lambda_2 = 0.01$ in Eq. (1), $\gamma = 0.002$ in Eq. (5),$\beta = 2$ in Eq. (8), $\eta = 0.4$ for data association.

For quantitative evaluation, the common CLEAR performance metrics [35], MOTP ↑, MOTA ↑, FP ↓, FN ↓ and IDS ↓, as well as GT, MT ↑, ML ↓, PT, Recall ↑ and Precision ↑ defined in [36] are used. Here, the arrow ↑ represents the higher scores are the better results, while ↓ means that lower scores are the better tracking results.

The state-of-the-art MOT methods [37] such as RMOT [16], SCEA [17], CMOT [15], and TSML [18] are used to compare the proposed methods. All of those trackers are online tracking methods. For fair comparisons, we use the publicly available source codes or the results reported in their published papers. For fair comparisons, we use the public available detections.

4.3 Quantitative analyses

First, the illustrative foreground segmentation results for improving the detection responses in PETS 09 dataset are shown in Figs. 1, 2 and 3. As shown in Figs. 1, 2 and 3, the detections provided by DPM detector are shown in red bounding boxes, which include missing detections and false detections. The missing detections and false detections are indicated with blue arrows and dark arrows, respectively, as shown in Figs. 1a, c, 2a, c and 3a, c. While the exemplar overcomplete detection sets, which are constructed by the foreground motion objects segmentation and DPM detector after performing the process of using the object shape prior to remove the false detections, are shown in Figs. 1b, d, 2b, d and 3b, d, the overcomplete detection sets are indicated with yellow bounding boxes, as shown in Figs. 1b, d, 2b, d and 3b, d. The proposed method with low-rank sparse approximation can eliminate the missing detections in

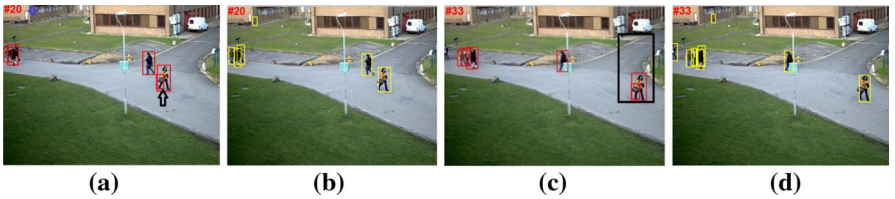

Fig. 1 The detection results in PETS'09 S1L1: **a** and **c** are the detection results provided by DPM detector in frame 20 and 33, respectively, including miss-detection and false detection as shown in blue and dark arrows, respectively. **b** and **d** are the overcomplete detection responses set constructed by our method (color figure online)

Fig. 2 The detection results in PETS'09 S2L1: **a** and **c** are the detection results provided by DPM detector in frame 118 and 476, respectively, including miss-detection and false detection as shows in blue and dark arrows, respectively. **b** and **d** are the overcomplete detection responses set constructed by our method (color figure online)

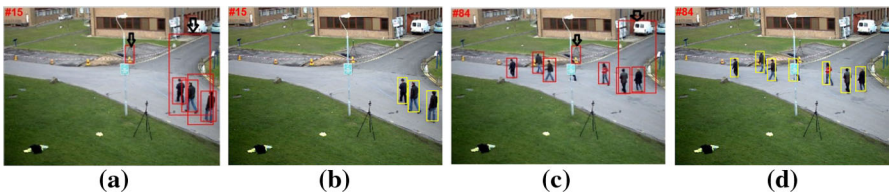

Fig. 3 The detection results in PETS'09 S3MF: **a** and **c** are the detection results provided by DPM detector in frame 15 and 84, respectively, including false detections as shown in dark arrow. **b** and **d** are the overcomplete detection responses set constructed by our method (color figure online)

DPM detector, and the pre-trained pedestrian mask in Sect. 3.3 can help to remove the false detections. By the proposed method, we can build a high-performance detection response set, which is beneficial for improving tracking performance.

To better show the effectiveness of the segmentation algorithm with low-rank sparse approximation and prove how well it improves the tracking performance, the background model and extracted foreground objects based on the low-rank sparse approximation, as well as tracking results rely on the overcomplete detections set construct in our work on PETS 2009 dataset are shown in Figs. 4, 5 and 6. The quantitative results of the proposed method and the competing state-of-the-art tracking methods are shown in Table 1 where the best results are shown in bold.

From Figs. 4, 5, 6 and Table 1, we can see the proposed method provides a good performance in PETS 2009 dataset. Compared with the competing multi-object tracking methods like RMOT, SCEA and CMOT trackers, the proposed method has improved recall and precision scores, the number of FN and FP is also reduced. In S1L1 sequence,

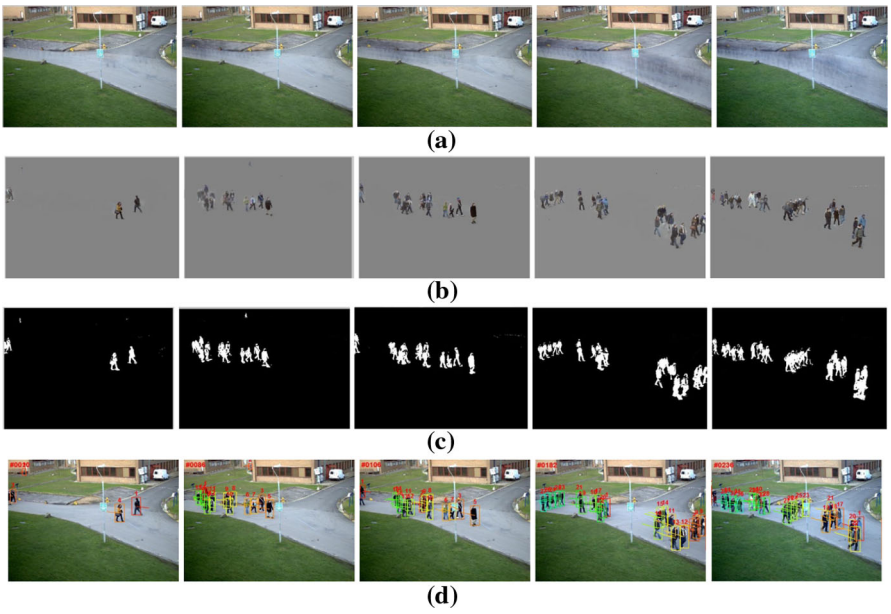

Fig. 4 The tracking results of the proposed method in PETS'09 S1L1: **a** background modeling results, **b** sparse outliers, **c** extracted foreground objects, **d** tracking results of the proposed method

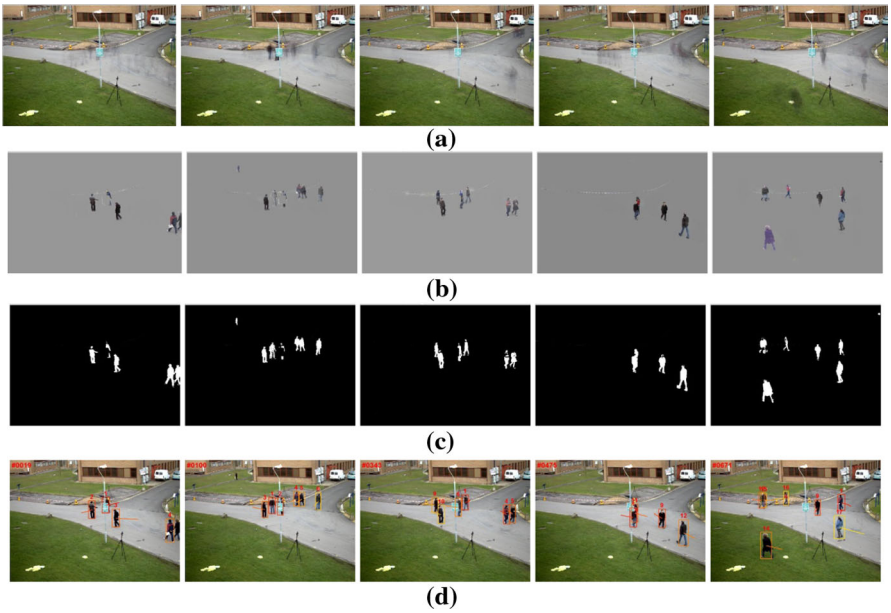

Fig. 5 The tracking results of the proposed method in PETS'09 S2L1: **a** background modeling results, **b** sparse outliers, **c** extracted foreground objects, **d** tracking results of the proposed method

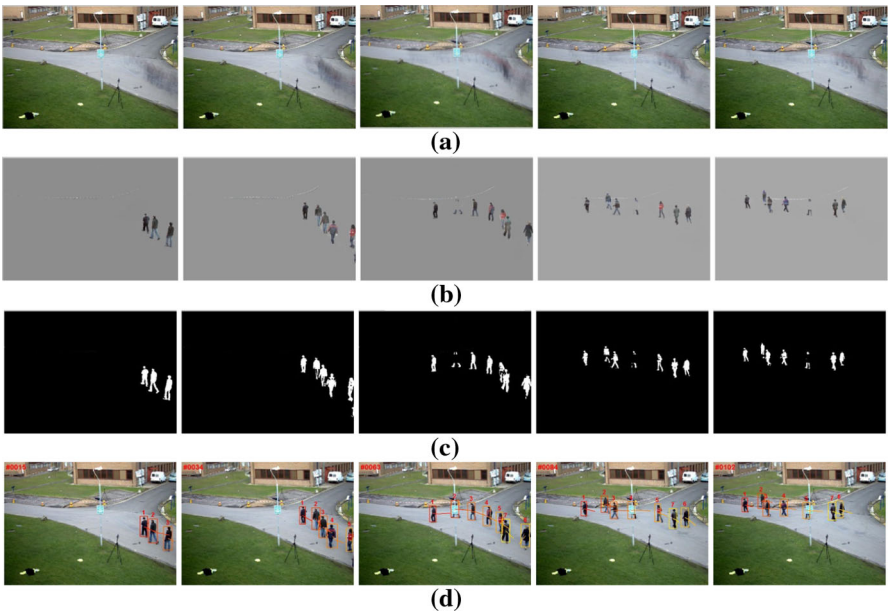

Fig. 6 The tracking results of the proposed method in PETS'09 S3MF: **a** background modeling results, **b** sparse outliers, **c** extracted foreground objects, **d** tracking results of the proposed method

the recall and precision scores of the proposed method reach 89.5 and 99.1%, respectively, which are almost 34 and 22.4% higher than the competing online trackers like RMOT, SCEA and CMOT. The FN and FP reduce to 41 and 521, respectively, which are also far smaller than the competing trackers. The number of FN and FP for the competing trackers is all large than 500 and 1700, respectively. With the improved detection responses, the MOTA and MOTP of the proposed method are also improved, with 88 and 96.6%, respectively. It is higher than the competing online trackers at least 54 and 30%, respectively, which highlight the proposed method that improved the tracking performance in some extent. In S2L1 sequence, the tracking performance of the proposed method is also better than the competing methods, with 92.6% recall and 92.7% precision. The number of FN and FP is 341 and 343, respectively. The MOTA and MOTP reach 85.3 and 99.3%, respectively. In S3MF sequence, recall and precision of the proposed method are 90.5 and 99.3%, respectively. The MOTA and MOTP are 93.4 and 98.9%, respectively. The high scores of the recall, precision, MOTA and MOTP and small number of FN and FP for the proposed method prove the effectiveness of the proposed high-performance detections, which can help to improve the tracking performance.

The illustrative tracking results of the proposed method on TUD dataset are shown in Fig. 7, the quantitative results of the proposed method and the competing state-of-the-art tracking methods on TUD dataset are shown in Table 2, where the best results are indicated in bold.

From Fig. 7 and Table 2, we can see the proposed method can robustly track the objects even with frequent occlusion. Since our work follows the online tracking-by-

Table 1 Performance comparison between the proposed method and other state-of-the-art methods in PETS'09 dataset

Dataset	Method	Recall (%) ↑	Precision (%) ↑	GT	MT (%)↑	PT (%)↑	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA (%)↑	MOTP (%)↑
PETS'09 S1L1	RMOT	55.4	70.8	44	27.3	44.7	28	881	1714	66	110	30.8	65.2
	SCEA	47	76.7	44	20.5	43.2	36.3	548	2040	86	112	30.5	66
	CMOT	53.4	74.2	44	27.3	43.2	29.5	713	1792	43	134	33.7	64.8
	Ours	**89.5**	**99.1**	44	**72.7**	27.3	**0**	**41**	**521**	**35**	**20**	**88**	**96.6**
PETS'09 S2L11	RMOT	56.7	75.4	19	89.5	10.5	0	2485	417	41	78	36.7	65.3
	SCEA	59.1	59.2	19	10.5	84.25	5.25	1890	1903	3	175	18.4	90.4
	CMOT	87.7	74	19	73.7	26.3	0	1435	572	4	**4**	69.59	**83.04**
	TSML	–	–	19	94.7	5.3	0	–	–	18	21	**86.4**	93.4
	MHMHT	–	–	19	**94.7**	5.3	0	–	**–**	–	–	65.81	92.59
	Ours	**92.6**	**92.7**	19	89.5	10.5	0	**341**	**343**	0	52	85.3	**99.2**
PETS'09 S3MF	RMOT	73.4	87.2	7	28.6	71.4	0	165	165	0	24	46.8	65.8
	SCEA	78.7	81.5	7	57.1	42.9	0	111	132	8	13	59.5	69.3
	CMOT	77.1	78	7	57.1	42.9	0	135	142	2	5	55	68.7
	Ours	**96.5**	**96.9**	7	**100**	0	0	**19**	**22**	**0**	**4**	**93.4**	**98.9**

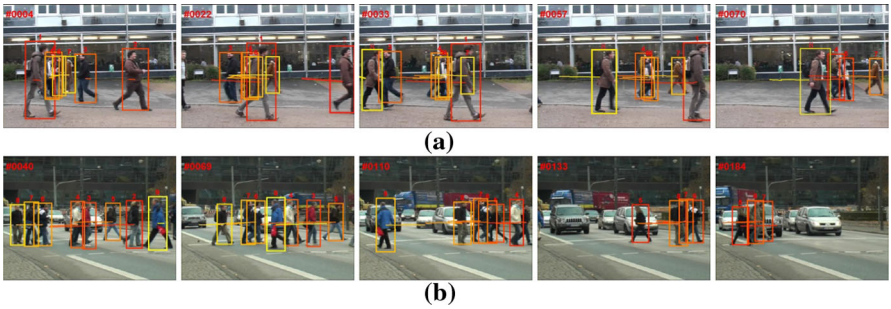

Fig. 7 Tracking results of the proposed method on TUD dataset. At each frame, the tracked objects are indicated with color bounding boxes, the recent trajectories and IDs of the tracked object are indicated with color lines and red numbers. **a** TUD-campus, **b** TUD-crossing (color figure online)

detection framework, the performance of detection has a critical influence on tracking performance. To overcome the limitation of detections provided by detector with number of missing detections and false detections, in our work, we have proposed a low-rank-based sparse approximation method to effectively online build the background model and achieve the foreground segmentation result, which can be beneficial for eliminating the missing detections. Meanwhile, to eliminate the false detections, we also use the pre-trained pedestrian mask. With the prior knowledge that the tracked objects are the special pedestrian, we can use the pre-trained pedestrian mask to eliminate the false detection in some extent. Depending on the overcomplete detections set provided in our work, the proposed method shows better performance compared to other competing methods, the recall and precision of the proposed method in TUD-campus and TUD-crossing sequences are all larger than 98 and 99%, respectively. While the recall and precision for the competing methods are all smaller than the proposed method. The number of the FN and FP for the proposed method is 2, 2 and 13, 9 in those two sequences, respectively. With the good performance detections, the MOTA and MOTP for the proposed method reach 98.5 and 99.2% in TUD-campus sequence and 97.7 and 99.8% in TUD-crossing sequences, respectively.

5 Conclusion

In this paper, an online multi-object tracking method is proposed depend on high-performance detections. To overcome the detection failures and construct a high-performance detection set, we have proposed the low-rank-based foreground detection method to refine the detections, which can help for eliminating the missing detections. Meanwhile, we use a pre-trained pedestrian mask to remove the false detections. With the high-performance detection set, the multi-object tracking implemented by two-step data association relies on tracklet confidence. The first stage of data association progressively built long trajectories by associating reliable tracklets with the detections. While the second stage of data association is the reassignment association problem for un-reliable tracklets. By associating the un-reliable tracklets with reliable tracklets or detections, it can recover fragmented tracklets to be tracked. Representa-

Table 2 Performance comparison between the proposed method and other state-of-the-art methods in TUD dataset

Dataset	Method	Recall (%)↑	Precision (%)↑	GT	MT (%)↑	PT (%)↑	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA (%)↑	MOTP (%)↑
TUD-campus	RMOT	86.9	89.1	7	100	0	0	49	4	15	2	83.2	91.9
	SCEA	89.9	89.9	7	71.4	28.6	0	41	41	**0**	5	79.8	**99.9**
	CMOT	77.3	93.7	7	71.4	28.6	0	21	92	**0**	**0**	72.1	99.1
	Ours	**99.5**	**99.5**	7	**100**	0	0	**2**	**2**	2	**0**	**98.5**	99.2
TUD-crossing	RMOT	95.6	96.5	12	100	0	0	**5**	**1**	2	**0**	**99.2**	97.3
	SCEA	93.6	93.3	12	100	0	0	68	64	**0**	12	86.9	**99.9**
	CMOT	97.3	94.2	12	100	0	0	61	21	2	**0**	91.6	98.6
	Ours	**98.7**	**99.1**	12	100	0	0	9	13	1	**0**	97.7	99.8

tive experimental results on five public pedestrian tracking sequences showed that our detection optimization strategy can help to improve the tracking performance compared with other several state-of-the-art multi-object trackers. However, the run time of our method is about 15 fps, which is still slower than the real-time requirement. How to improve the speed of the proposed method is one of our future works. In the proposed method, we use the high-performance detections to improve the tracking performance. The traditional data association problem in MOT is solved by affinity matrix, which is constructed by appearance, shape and motion models of the tracked objects. The feature used to describe the object has critical role to improve tracking performance; hence, in our future work, we will introduce deep learning into our online MOT. In addition, long-term occlusions pose more challenges for online tracking; then, one of our future researches is occlusion analysis and tries to address the challenges in MOT.

Acknowledgements This work is supported by the Science and Technology Innovation Foundation of Shenzhen (201703063000511), National Natural Science Foundation of China (61672433). The authors would like to thank the valuable comments from the reviewers and editors.

References

1. Luo W, Xing J, Zhang X, Zhao X, Kim TK (2014) Multiple object tracking: a literature review. Eprint Arxiv
2. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32:1627–1645
3. Dollár P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell 36:1532–1545
4. Brendel W, Amer M, Todorovic S (2011) Multiobject tracking as maximum weight independent set. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1273–1280
5. Wen J, Wang J, Zhang Q (2017) Nearly optimal bounds for orthogonal least squares. IEEE Trans Signal Process 65(20):5347–5356
6. Jiang H, Fels S, Little JJ (2007) A linear programming approach for multiple object tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07, pp 1–8
7. Berclaz J, Fleuret F, Turetken E, Fua P (2011) Multiple object tracking using k-shortest paths optimization. IEEE Trans Pattern Anal Mach Intell 33:1806–1819
8. Andriyenko A, Schindler K, Roth S (2012) Discrete-continuous optimization for multi-target tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1926–1933
9. Milan A, Roth S, Schindler K (2014) Continuous energy minimization for multitarget tracking. IEEE Trans Pattern Anal Mach Intell 36:58–72
10. Huang C, Li Y, Nevatia R (2013) Multiple target tracking by learning-based hierarchical association of detection responses. IEEE Trans Pattern Anal Mach Intell 35:898–910
11. Pirsiavash H, Ramanan D, Fowlkes CC (2011) Globally-optimal greedy algorithms for tracking a variable number of objects. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1201–1208
12. Breitenstein MD, Reichlin F, Leibe B, Koller-Meier E, Van Gool L (2011) Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Trans Pattern Anal Mach Intell 33:1820–1833
13. Shu G, Dehghan A, Oreifej O, Hand E, Shah M (2012) Part-based multiple-person tracking with partial occlusion handling. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 1815–1821
14. Kuo C-H, Huang C, Nevatia R (2010) Multi-target tracking by on-line learned discriminative appearance models. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 685–692
15. Bae S-H, Yoon K-J (2014) Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1218–1225

16. Yoon JH, Yang M-H, Lim J, Yoon K-J (2015) Bayesian multi-object tracking using motion context from multiple objects. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), pp 33–40
17. Hong Yoon J, Lee C-R, Yang M-H, Yoon K-J (2016) Online multi-object tracking via structural constraint event aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1392–1400
18. Wang B, Wang G, Chan KL, Wang L (2017) Tracklet association by online target-specific metric learning and coherent dynamics estimation. IEEE Trans Pattern Anal Mach Intell 39:589–602
19. Yu F, Li W, Li Q, Liu Y, Shi X, Yan J (2016) Poi: Multiple object tracking with high performance detection and appearance feature. In: European Conference on Computer Vision, pp 36–42
20. Horbert E, Rematas K, Leibe B (2011) Level-set person segmentation and tracking with multi-region appearance models and top-down shape information. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp 1871–1878
21. Leibe B, Schindler K, Cornelis N, Van Gool L (2008) Coupled object detection and tracking from static cameras and moving vehicles. IEEE Trans Pattern Anal Mach Intell 30:1683–1698
22. Tang S, Andriluka M, Milan A, Schindler K, Roth S, Schiele B (2013) Learning people detectors for tracking in crowded scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1049–1056
23. Milan A, Leal-Taixé L, Schindler K, Reid I (2015) Joint tracking and segmentation of multiple targets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5397–5406
24. Bouwmans T, Zahzah EH (2014) Robust PCA via principal component pursuit: a review for a comparative evaluation in video surveillance. Comput Vis Image Underst 122:22–34
25. Wen J, Zhou Z, Wang J, Tang X, Mo Q (2017) A sharp condition for exact support recovery with orthogonal matching pursuit. IEEE Trans Signal Process 65:1370–1382
26. Rodriguez P, Wohlberg B (2016) Incremental principal component pursuit for video background modeling. J Math Imaging Vis 55:1–18
27. Zhou X, Yang C, Yu W (2013) Moving object detection by detecting contiguous outliers in the low-rank representation. IEEE Trans Pattern Anal Mach Intell 35:597–610
28. Zhou T, Tao D (2011) Godec: Randomized low-rank and sparse matrix decomposition in noisy case. In: International Conference on Machine Learning
29. Javed S, Oh SH, Sobral A, Bouwmans T, Jung SK (2014) OR-PCA with MRF for robust foreground detection in highly dynamic backgrounds. In: Asian Conference on Computer Vision, pp 284–299
30. Wright J, Peng Y, Ma Y et al (2009) Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. In: International Conference on Neural Information Processing Systems. Curran Associates Inc, pp 2080–2088
31. Li SZ (2009) Markov random field modeling in image analysis. Springer, Berlin
32. Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. IEEE Trans Pattern Anal Mach Intell 23:1222–1239
33. Gong H, Sim J, Likhachev M, Shi J (2011) Multi-hypothesis motion planning for visual object tracking. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp 619–626
34. Kuhn HW (2005) The Hungarian method for the assignment problem. Naval Res Logist 52:7–21
35. Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J Image Video Process 2008:246309
36. Li Y, Huang C, Nevatia R (2009) Learning to associate: hybridboosted multi-target tracker for crowded scene. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp 2953–2960
37. Leal-Taixé L, Milan A, Schindler K, Cremers D, Reid I, Roth S (2017) Tracking the trackers: an analysis of the state of the art in multiple object tracking. arXiv preprint arXiv:1704.02781