Multimed Tools Appl (2018) 77:23167–23191
https://doi.org/10.1007/s11042-017-5530-z

Online multiple objects tracking with detection reliability prior constraint

Honghong Yang[2,3] · **Li He**[1]

Received: 23 May 2017 / Revised: 23 October 2017 / Accepted: 11 December 2017 /
Published online: 22 January 2018

Abstract Multi-object tracking (MOT) is one popular topic in computer vision. It remains a challenging problem in complex scenes, especially of objects with similar appearance. In this case, many existing data association strategies, which link detections among consecutive frames according appearance and motion cues, may fail to track due to unreliable detections or confused appearance and motion. To solve this problem, this paper proposed a novel online multi-object tracking method with detection reliability prior constraint. Our method integrates the trajectory estimation and detection-prediction association into a unified framework. The detection reliability prior constraint is built with the Hankel matrix from object motion model. When we build the Hankel matrix, we adaptively select a set of previous frames to predict object states and calculate the associated weights between detections and candidate objects. Data association in MOT then is estimated by maximum a posteriori (MAP) in a Bayesian framework, accompanied with both previous trajectory and the current detection reliability. Experimental results using synthetic dataset and four public challenging datasets demonstrate that, the proposed method has a good tracking performance compared with the state-of-the-art multi-object trackers.

Keywords Online multi-object tracking · Data association · Trajectory estimation; detection reliability · Local associated motion constraint

✉ Li He
heli9903@mail.nwpu.edu.cn

[1] School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China

[2] Department of Computing Science, University of Alberta, T6G2E8, Edmonton, Alberta, Canada

[3] Department of Automation, Northwestern Polytechnical University, Xi'an 710072, China

1 Introduction

Multi-object tracking (MOT) is a challenging task in computer vision. The aims of MOT are to simultaneously identify multi-objects and estimate their trajectories from clutter scenes. It has a wide application scope, ranging from surveillance, traffic safety to automotive driver assistance systems and robotics. Several challenges, such as occlusion, mis-detection, false detection, camera motion in complex scenes or similar appearance with occlusion, make MOT still a tough problem [29].

Due to the development of object detectors [13, 38], tracking-by-detection (TBD) methods always show state-of-the-art performance in recent years [2, 3, 7, 20, 31, 34, 39, 41, 44]. The TBD methods can be roughly categorized into a couple classes, batch tracking and online tracking.

The batch tracking methods solve the data association problem in MOT using the forward-backward information from entire sequence [2, 7, 12, 31, 34]. They first build short tracklets by linking detections frame by frame. Then the short tracklets are globally associated to form the long trajectories. Many global association methods have been proposed in recent years. However, the performance of batch tracking methods has some limitations. One is that the batch tracking requires the detections of future frames for the entire sequence. Using detections from the whole video will need enormous computation because it has to iteratively link short tracklets to construct the optimized trajectories. The iterative optimized linking process in batch tracking implies that tracklets may change their links at each iteration. Link change may result in the ambiguity of close targets identification, especially of objects with similar appearance. A global optimal matching, in this case, always relies on pair-wise point matching in consecutive frames. Pairwise matching, however, sometimes may fail to find the correct matching due to the ambiguities among competing candidates. Hence, the global optimal matching, achieved in the iterative optimized linking process, may change their linking results at each iteration with non-unique or incorrect pairwise matching in consecutive frames [18, 37]. The other problem in batch tracking is that it is not suitable for time-critical applications due to huge computational burden in global optimization. Comparing to the batch tracking methods, the online tracking methods are more suitable for real time applications since they only use the detections from recent frames to build trajectories [3, 20, 39, 41, 43, 44]. There is no iterative optimization process and the tracking results are outputted on the fly based on the up-to-present detections. However, the online methods are not robust under occlusion, in which case the online tracking may produce short fragment trajectories. This is because the data association in online MOT without iterative associations, when the detections are partly reliable with possible false positive and missing detections, the association result is inaccurate. Hence, in terms of tracking accuracy, the batch tracking methods are more accurate than the online tracking methods due to available future information and iterative associations can be used to tackle detection errors and tracking failures. In this paper, we pay attention to online MOT tracking and aim to improve the performance of online MOT.

In detection-based MOT, data association plays an important role for robust tracking. Both the appearance model and motion model are typically used to solve data association in online MOT [3, 20, 26, 39, 41–44]. The detection-based MOT achieve a good performance in many situations such as pedestrian tracking or vehicle tracking, where the objects follow a simple motion model and their appearance can help to distinguish objects from each other. However, most of the existing online MOT methods adopt the first or second order motion model to describe object dynamic states in the current frame [44], which indicates that the current object

states heavily depend on previous one or two frames only. This kind of state estimation performs well when one object is detected in continuous frames. In addition, data association becomes more and more complicated when multiple targets with similar-looking appearance, as shown in Fig. 1. It is difficult to distinguish objects based on color and shapes (Fig.1a, b). Moreover, when detections are unavailable for several frames due to occlusion or mis-detection, the object states predicted by motion model may be unreliable.

To resolve the above mentioned problems, in this paper, a novel online MOT focusing on multi-objects with similar appearances is proposed. The framework of the proposed method is shown in Fig. 2. With the detections up to the present frame, the data association in online MOT is solved by maximum a posteriori (MAP) with trajectory estimation and detection reliability. The trajectory estimation, on one hand, is solved based on Bayesian framework with the number of involved frames being selected adaptively. On the other hand, detection reliability, computed by tracklet dynamic estimation and detection-prediction association in continuous sequences, is sequentially introduced into trajectory estimation stage. The detection-prediction association in consecutive frames is used to filter out unreliable association among the trajectories and the detections according to local associated motion constraint. The local associated motion constraint in this paper is built by the predicted object states and the detections with the Hankel matrix. The Hankel matrix based object states predication is beneficial to recover short fragment tracklets in online MOT because it takes long history object states into consideration. In addition, the MAP framework allows the trajectory estimation and detection reliability interact with each other in a sequential manner, which facilitates online multi-object tracking. Experimental results on synthetic dataset and four public-available challenging datasets confirm the superiority of the proposed method.

The main contributions of this paper are:(1) Data association problem in online MOT is solved by MAP in a Bayesian framework with previous trajectory and the current detection reliability. (2) The detection reliable prior is computed by tracklet dynamic estimation and detection-prediction association in continuous sequences. By MAP the trajectory estimation and detection reliable prior interact with each other in a sequential manner. (3) The Hankle

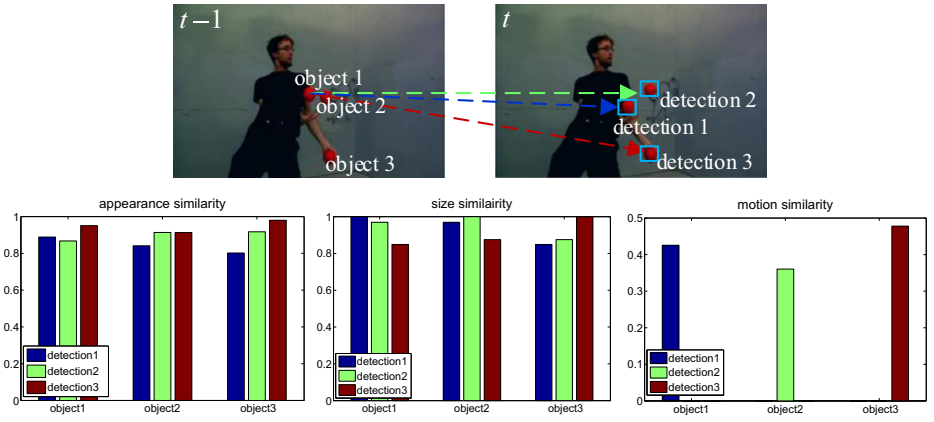


Fig. 1 A difficult scene with similar-looking appearance of multi-objects in two consecutive frames. Their color and size provide quite a few cues for matching, but motion can help to distinguish the objects

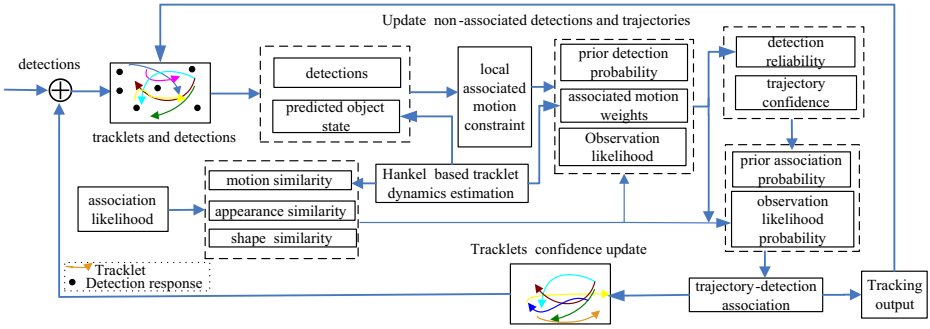


Fig. 2 The tracking framework of the proposed method

matrix based object dynamic motion estimation is used to measure the association weights between detections and the predict object states. Such estimation is beneficial to recover short fragment tracklets and improve the correctness of the association among the trajectories and the detections.

2 Related works

Numerous multi-object tracking approaches have been proposed in recent years [2, 3, 7, 12, 20, 26, 31, 34, 39, 41–44]. In this section, we mainly introduce related MOT methods.

Data association plays an important role for robust tracking in detection-based MOT. Both the appearance and motion models are typically used to solve data association in MOT. Several features, such as color histogram, HOG, haar-like feature, sparse feature and deep feature, are designed to describe appearance changes. In [46], a multitask shared sparse regression framework is proposed to represent the input image at different levels. In [35], a CNN based feature representation method with an adaptive hedge method is proposed for constructing robust appearance model of the object. In [19], Zhang et al. proposed a new object detection framework by using high-level feature representation and extended their work in [45] by high-level convolutional feature and visually similar neighbors. In [9], Chen et al. proposed a robust object tracking method based on subspace learning-based appearance model with sparse feature representation. In [21], Hong et al. employed the Integrated Correlation Filter (ICF) to improve the single object tracking performance. In terms of motion cue of the object, there are many MOT methods directly exploit Kalman or particle filter to locate objects. These methods typically use the first or second order motion models to predict object states, which indicate that the current object states heavily rely on previous one or two frames. The simple motion model performs well in short durations but show limitations in sequences with long-term occlusion, complex motion or cluttered scenes. Data association methods like JPDAF [17] and MHT [10, 36] have been proposed to link the short tracklets and generate long trajectories. Since the search space grows exponentially with the number of frames, both JPDAF and MHT are less effective for long–time association. To overcome this limitation, a variety of data association approaches have been developed to consider pairwise association of detections in consecutive frames as an optimization task based on Hungarian algorithm [25], K-shortest paths (KSP) [5], the Linear Programming [23], the Quadratic Boolean Programming [27], the Markov Chain Monte Carlo (MCMC) [32] and the maximum weight-independent set [8].

However, the pairwise based data association methods only consider pairs of detections and set a pairwise interframe edge costs. These algorithms do not have a good performance when appearance constraint of the closely moved multi-objects is weak. The motion model constraint like Kalman or particle filter, which heavily rely on the former frame motion information to predict the current object, also does not provide useful information to distinguish similar-looking appearance objects. In addition, merely depend on the previous one or two frames kinestate to predict the current motion state is insufficient. In [9], Chen et al. point out that the particle filter is an approximate nonlinear Bayesian filter, which is used to get suboptimal solution of posterior probability for object state with observation. In [11], Collins shows that higher-order motion constraint has a major effect on improve the quality of data association in MOT, especially when multi-objects with similar appearance. In [41], nonlinear motion patterns and robust appearance model are learned for each of object to better explain direction changes and construct more robust motion affinities between tracklets. In [14], both individual and mutual relation models are introduced in MOT to build graph model, but the mutual relation model only works when the objects move in the same direction. In [42], the pairwise relative motion model is introduced as an additional term to construct CRF energy function. Most recently, the notion of relative motion network proposed in [44], which is designed to improve the data association performance by utilizing the relative spatial constraint between objects. In [20], a structural motion constraint among objects has been utilized to assist data association against unreliable detections in online MOT. Bae and Yoon [3] exploited trajectory confidence constraint and incremental linear discriminate appearance to assist their two step data association. Then they extended their work in [4] by introducing a track existence probability into data association. However, these methods exploit prior information into two separate stages, either in the detection or association stage. In addition, the pairwise motion constraint in those methods are building based on the position information no more than three consecutive frames. However, the occlusion or mis-detection always appears in multiple frames, often more than three consecutive frames in practical scene. In this paper, we propose a novel association based multi-objects tracking method to combine trajectory estimation and detection prior together for better enhancement the tracking performance. However, both our work and [43] are online multi-object tracking methods based on maximizing a posteriori estimation with sequential prior knowledge. The major differences of them are: (1) the way to compute the detection prior; (2) how to combine the detection prior into MAP estimation during online multi-object tracking; In our work, the detection reliability is sequentially introduced into trajectory estimation stages by using Bayesian framework [22, 33], which is different from prior in [43]. Our work takes the local associated motion constraint and associated weight to refine the detections. The associated weight is calculated by Hankel matrix based dynamic motion model with the number of involved frames' instead of manually fixing the order of motion model. In addition, the MAP framework allows the trajectory estimation and detection reliable prior interact with each other in a sequential manner, which facilitates online multi-object tracking. While in [43], the multi-object tracking is solved by two MAP estimation problems: object detection and trajectory-detection association. In their detection refinement with MAP estimation stage, the posterior detection probability computed by combining the observation likelihood function and the prior detection probability. The prior detection probability in their work is computed based on the spatio-temporal consistency assumption with the Kalman filter to predict the object states. Based on the object states, they build density map with the position constraint of the object.

3 Online tracking with detection reliability under local associated motion constraint

3.1 Problem formulation

The essential problem for MOT is data association, which implements the task of matching detections in one frame to a set of previous trajectories with corresponding detections. Let $\mathbb{X}_t = \{\mathrm{x}_t^1, \cdots, \mathrm{x}_t^N\}$ and $\mathbb{Z}_t = \{z_t^1, \cdots, z_t^M\}$ be the set of object detections and predicted object states at frame t. Denote the set of detections, trajectories and predicted object states up to frame t as $\mathbb{X}_{1:t}$, $\mathbb{T}_{1:t}$ and $\mathbb{Z}_{1:t}$, respectively. For online MOT, the trajectories $\mathbb{T}_{1:t}^j$ of object j up to frame t can be represented as $\mathbb{T}_{1:t}^j = \{\mathrm{x}_k^j | 1 \le t_s^j \le k \le t_e^j \le t\}$, where t_s^j and t_e^j are the start and end frame of a tracklet. Then, the online MOT problem can be solved within the Bayesian framework by maximizing the joint posterior probability over $\mathbb{X}_{1:t}$ and $\mathbb{T}_{1:t-1}$ given the predicted object states $\mathbb{Z}_{1:t}$ as follows:

$$\begin{aligned}\langle \mathbb{T}_{1:t}, \mathbb{X}_{1:t} \rangle &= \underset{\mathbb{T}_{1:t-1}, \mathbb{X}_{1:t}}{\operatorname{argmax}} p(\mathbb{T}_{1:t-1}, \mathbb{X}_{1:t} | \mathbb{Z}_{1:t}) \\ &= \underset{\mathbb{T}_{1:t-1}, \mathbb{X}_{1:t}}{\operatorname{argmax}} \underbrace{p(\mathbb{T}_{1:t-1} | \mathbb{X}_{1:t}, \mathbb{Z}_{1:t})}_{\textit{ttrajectory estimation}} \underbrace{p(\mathbb{X}_{1:t} | \mathbb{Z}_{1:t}, \mathfrak{R}_t)}_{\textit{detection reliablity estimation}}\end{aligned} \tag{1}$$

The first term is the trajectory estimation, which is used to generate current trajectories $\mathbb{T}_t$ conditioned on $\mathbb{Z}_t$, for pairwise associations between $\mathbb{T}_{t-1}$ and $\mathbb{X}_t$. The second term is the posterior probability for detection reliability estimation between $\mathbb{X}_t$and $\mathbb{Z}_t$. Due to the huge number of possible combination of $\mathbb{T}_{1:t-1}$ and $\mathbb{X}_{1:t}$, the space of possible trajectories grows exponentially over time. As a result, it is often difficult to optimize Eq. (1) exhaustively. Therefore, we decompose Eq. (1) into two estimation stages with local associated motion constraint $\mathfrak{R}_t$, the local associated motion constraint will be detailed described in section 3.2.

3.2 Local associated motion constraint

By the fact that the object states in two consecutive frames should not change drastically, the detections in frame t are more likely to appear around the predicted location according to the existing trajectories using tracklet dynamic estimation model, as will be shown later. Then, a local associated motion constraint (LAMC, denoted as $\mathfrak{R}_t$) is built, to represent the affinity between detections and predicted object states, based on two constraints as follows:

$$\begin{aligned}&\left\| y_{z_t^j} - y_{\mathrm{x}_t^i} \right\| < 0.5 \sqrt{\left(w_{z_t^j}\right)^2 + \left(h_{z_t^j}\right)^2} \\ &\exp\left(-\left(\frac{h_{\mathrm{x}_t^i} - h_{z_t^j}}{h_{\mathrm{x}_t^i} + z_t^j} + \frac{w_{\mathrm{x}_t^i} - w_{z_t^j}}{w_{\mathrm{x}_t^i} + w_{z_t^j}}\right)\right) > \tau_s\end{aligned} \tag{2}$$

where $y_{\mathrm{x}^i}$and y_{z^j} are the positions of detection i and object j, respectively, (w, h) are the weight and height of one object.

The first constraint in Eq. (2) is location constraint, means that one detection is considered for tracking only if it is located closed to the predicted object location. The second constraint in Eq. (2) is size constraint, reflects the fact that both the detected object and the predicted object have similar size. We empirically set $\tau_s = 0.7$, if the size change and location change of the

predicted object state and the detection are satisfied the size constraint and location constraint in Eq. (2) the association assignment $d^{i,j}=1$, which indicates that the *i_th*detection is associated with the *j_th* object. Otherwise, $d^{i,j}=0$, which means there is no association between $x^i$and z^j. We use the association constraint in Eq. (2) to filter out the unreliable association between detections and predictions. Consequently, the total number of possible assignments between $\mathbb{Z}_t$ and $\mathbb{X}_t$ is thus reduced. Figure 3 is an example to illustrate the local associated motion constraint.

When there exist M trajectories in frame $(t-1)$ and N detections in frame t, the LAMC $\mathfrak{R}_t$is defined as

$$\begin{aligned}\mathfrak{R}_t &= \cup_{j=1}^{M}\mathfrak{R}_t^j \\ \mathfrak{R}_t^j &= \left\{(i,j)|d_t^{i,j}=1, 1\le i\le N\right\}\end{aligned} \tag{3}$$

with the fact that the linked edges represent the affinity between object states and detections in frame t. The LAMC build between z^j and x^i in $\mathfrak{R}_t^j$ only if $d^{i,j}=1$at frame t. Since the affinities between objects and detections are different, the associated motion weight $\theta_t^{(i,j)}$ is represented as follows:

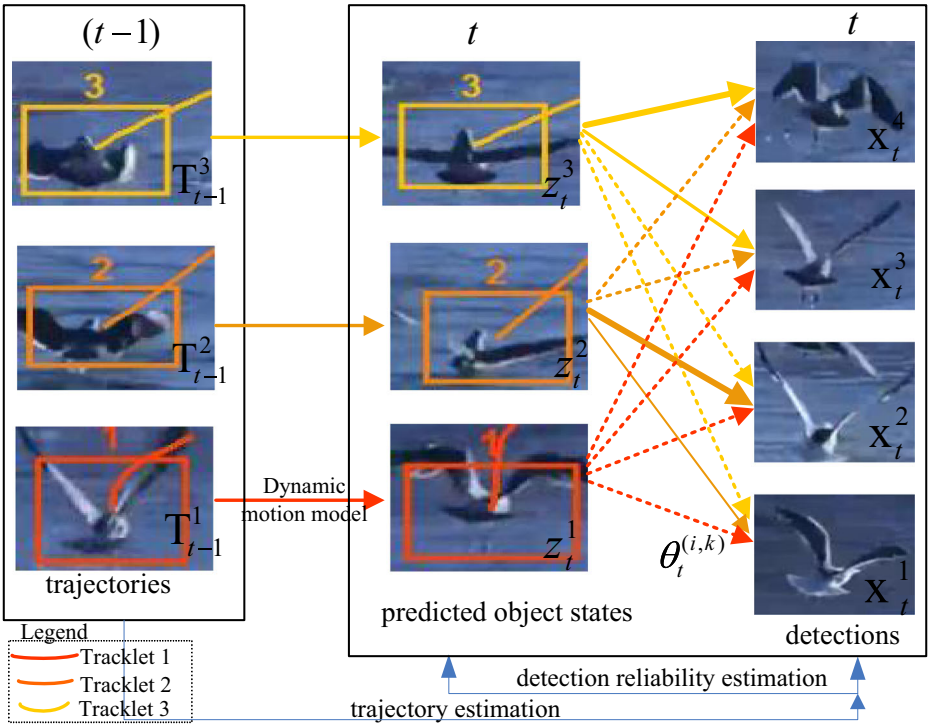


Fig. 3 An example to illustrate the associated motion constraint among three objects in frame $t-1$ and t. Each box color denotes a unique target ID. The association for each object state and detection is determined by Eq. (2). The solid arrow represents the object is associated with the detection, $d^{i,j}=1$. While the dotted arrow denotes there is no association between the detection and the object state, $d^{i,j}=0$. The thickness of solid arrow represents the association strength between the detection and the object state, which is determined by associated motion weight in Eq. (11)

$$\theta_t^j = \left\{ \theta_t^{(i,j)} | (i,j) \in \Re_t^j, 1 \le i \le N \right\}$$
$$\sum_{(i,j)\in\Re_t^j} \theta_t^{(i,j)} = 1 \tag{4}$$

where the initial associated motion weights $\theta_t^{(i,j)} = \frac{1}{|\Re_t^j|}$, and $|\Re_t^j|$ is the cardinality of an association set.

3.3 Detection reliability under local associated motion constraint

With the assumption that each object state is independent, the posterior detection probability for detection x_t^i and the predicted object states set $\mathbb{Z}_{1:t}$ in Eq. (1) under local associated motion constraint are defined as follows:

$$p\left(\mathrm{x}_t^i | \mathbb{Z}_{1:t}, \Re_t^j\right) = \sum_{(i,j)\in\Re_t^i} \theta_t^{(i,j)} p\left(\mathbb{Z}_t | \mathrm{x}_t^i, \Re_t^j\right) p\left(\mathrm{x}_t^i | \mathbb{Z}_{1:t-1}, \Re_t^j\right) \tag{5}$$

In Eq. (5), the posterior detection probability takes the associated motion weights $\theta_t^{(i,j)}$ into consideration. The prior detection probability $p\left(\mathrm{x}_t^i | \mathbb{Z}_{1:t-1}, \Re_t^j\right)$ is approximated by recursively procedure from the sequential Bayesian approach [33] under LAMC.

The observation likelihood $p\left(\mathbb{Z}_t | \mathrm{x}_t^i, \Re_t^j\right)$ in Eq. (5) is the association probability between z_t^j and x_t^i under LAMC, which is defined as follows:

$$p\left(\mathbb{Z}_t | \mathrm{x}_t^i, \Re_t^j\right) = p_0\left(E_t^{(i,j)}\right) + \sum_j p_j\left(E_t^{(i,j)}\right) p\left(z_t^j | \mathrm{x}_t^i, \Re_t^j\right) \tag{6}$$

where $p_j\left(E_t^{(i,j)}\right)$ represents the association probability between the *j_th* object state and the *i_th*detection and $p_0\left(E_t^{(i,j)}\right)$ denotes the not-associated probability. $p\left(z_t^j | \mathrm{x}_t^i, \Re_t^j\right)$ is the likelihood between z_t^j and x_t^i.

Similarly to [44], the likelihood function is computed by using appearance, shape and motion cues as follows:

$$p\left(z_t^j | \mathrm{x}_t^i, \Re_t^i\right) = p_a\left(z_t^j | \mathrm{x}_t^i\right) p_s\left(z_t^j | \mathrm{x}_t^i\right) p_m\left(z_t^j | \mathrm{x}_t^i, \Re_t^j\right) \tag{7}$$

where p_a, p_s and p_m are appearance, size and motion similarity, respectively, which are defined as

$$\begin{aligned} p_a &= \exp\left(-\sum_{b=1}^{B} \sqrt{\mathrm{H}^b\left(z_t^j\right)\mathrm{H}^b\left(\mathrm{x}_t^i\right)}\right) && \text{(a)} \\ p_s &= \exp\left(-\left(\frac{h_{\mathrm{x}_t^i} - h_{z_t^j}}{h_{\mathrm{x}_t^i} + h_{z_t^j}} + \frac{w_{\mathrm{x}_t^i} - w_{z_t^j}}{w_{\mathrm{x}_t^i} + w_{z_t^j}}\right)\right) && \text{(b)} \\ p_m\left(z_t^j | \mathrm{x}_t^i, \Re_t^j\right) &= \frac{S\left(z_t^j\right) \cap S\left(\mathrm{x}_t^i\right)}{S\left(z_t^j\right) \cup S\left(\mathrm{x}_t^i\right)}_{(i,j)\in\Re_t^j} && \text{(c)} \end{aligned} \tag{8}$$

where $\mathrm{H}^b\left(\mathrm{x}_t^i\right)$,$\mathrm{H}^b\left(z_t^j\right)$are the color histogram of the *i_th*detection and the *j_th*predicted object state, respectively. *b* denotes the *b_th* bin and *B* is the number of bins. Here, we use $B = 64$ bins for each HSV color space. In terms of shape similarity in Eq. (8)

(b), (h_x, h_z), $(w_x, w_z)$are the height and width for detection xand object z. The motion similarity in Eq. (8) (c) is computed by PASCAL score [16], where $S(\bullet)$ is the area of the z_t^j and x_t^i.

The associated weight $\theta_t^{(i,j)}$ is calculated by tracklet dynamics estimation proposed in [12]. Since target motion can be formed as a sequence of piecewise linear regression and the order of regression can be estimated from the positions of the object in previous frames. Therefore, the trajectory for an object can be represented by an ordered sequence of dynamic measurements as follows:

$$y_t = \sum_{i=1}^{m} a_i y_{t-i}, m \le l, t \ge s+m \tag{9}$$

where y is the set of positions of a trajectory, a_i is the regression coefficient, l is the length of the trajectory, m is the order of regression model and s represents the start frame of a trajectory. According to [39], the order m of the regression model equals to the rank of corresponding Hankel matrix, $m = rank(H_{\mathbb{T}_i})$, where $H_{\mathbb{T}_i}$is the Hankel matrix with $n \ge m$columns.

$$H_{\mathbb{T}_i} = \begin{bmatrix} y_s & y_{s+1} & \cdots & y_{s+n-1} \\ y_{s+1} & y_{s+2} & \cdots & y_{s+n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{t-n+1} & y_{t-n} & \cdots & y_t \end{bmatrix} \tag{10}$$

where $n = l_i - \lceil l_i/3 \rceil + 1$, $l_i = t - s + 1$.l_i is the length of tracklet $\mathbb{T}_i$, $\mathbb{T}_i$ is a tracklet from frame s.

Then the associated motion weight $\theta_t^{(i,j)}$ is described as follows:

$$\theta_t^{(i,j)} = \frac{rank\left(\mathrm{H}(\mathbb{T}_i) + \mathrm{rank}\left(\mathrm{H}\left(\mathbb{T}_j\right)\right)\right)}{rank\left(H\left(\mathbb{T}_{\mathrm{ij}}\right)\right)} - 1 \tag{11}$$

where$\mathbb{T}_{ij} = \left[\mathbb{T}_i, \alpha_i^j, \mathbb{T}_j\right]$ is the joint tracklet with gap α_i^j between $\mathbb{T}_i$ and $\mathbb{T}_j$. If x_t^i and z_t^j are belong to the same trajectory, $\mathbb{T}_{ij}$ can be approximated by one relatively low order regression. Otherwise, $\mathbb{T}_{ij}$ is approximated by a higher order regression than the regression of each single tracklet.

The above dynamic motion model uses an *m_th* order sequence to predict object states in current frame. By using the Hankel matrix to estimate the order of the motion model, instead of manually fixing the order of motion model in many existing works, our strategy is beneficial to recover short fragment tracklet and significantly reduce errors in online MOT. This is because the *m_th*order dynamic motion model takes long trajectory motion cue into consideration rather than heavily rely on one or two frames in previous works. Simultaneously, in online MOT, the order of a trajectory is estimated several times with new data. Therefore, the dynamic motion model used in this paper can reduce the ambiguity when a target is undetected in one or more successive frames or two detections are erroneously linked.

After achieving the posterior detection probability, the association detection-prediction pairs are determined by

$$C_t^{i,j} = -\ln\left\{p\left(\mathbb{Z}_t | x_t^i, \mathfrak{R}_t^j\right)\right\} \tag{12}$$

If $C_t^{i,j} < \tau$, then x_t^i is associated with the z_t^j. The corresponding assignment index thus is $\gamma_t^{i,j} = 1$. The association probability is defined as $p_j\left(E_t^{<i,j>}\right) = \frac{\gamma_t^{i,j}}{|\mathfrak{R}_t^j|}$ and the not-associated probability as $p_0\left(E_t^{<i,j>}\right) = 1 - \sum_{j=1}^{|\mathfrak{R}_t^j|} p_j\left(E_t^{<i,j>}\right)$, where $|\mathfrak{R}_t^j|$ denotes the number of detection-prediction pairs.

3.4 Data association with detection reliability constraint

In online MOT, suppose we have found Mtrajectories $\mathbb{T}_{t-1} = \left\{\mathrm{T}_{t-1}^j\right\}_{j=1}^M$in frame $(t-1)$ and N detections $\mathbb{X}_t = \left\{\mathrm{x}_t^i\right\}_{i=1}^N$ in frame t. The pairwise trajectory-detection association between $\mathbb{T}_{t-1}$ and $\mathbb{X}_t$ to generate current trajectories $\mathbb{T}_t$ is formulated by Bayesian rule as follows:

$$p(< \mathbb{T}_t, \mathbb{X}_t > | \mathbb{X}_t, \mathbb{T}_{1:t-1}) = \frac{p(\mathbb{X}_t | < \mathbb{T}_t, \mathbb{X}_t >, \mathbb{T}_{t-1}) p(< \mathbb{T}_t, \mathbb{X}_t > | \mathbb{T}_{t-1})}{p(\mathbb{X}_t | \mathbb{T}_{t-1})} \tag{13}$$

where $p(< \mathbb{T}_t, \mathbb{X}_t > | \mathbb{X}_t, \mathbb{T}_{1:t-1})$ is the posterior association probability representing the detections assigned to the exist trajectories. $p(\mathbb{X}_t | < \mathbb{T}_t, \mathbb{X}_t >, \mathbb{T}_{t-1})$is the observation likelihood between the detections$\mathbb{X}_t$and the trajectories $\mathbb{T}_{t-1}$. $p(< \mathbb{T}_t, \mathbb{X}_t > | \mathbb{T}_{t-1})$ is the prior association probability. $p(\mathbb{X}_t | \mathbb{T}_{t-1})$ is the transition density, which is estimated by dynamic motion model of the object.

The prior association probability $p(< \mathbb{T}_t, \mathbb{X}_t > | \mathbb{T}_{t-1})$ is computed by two cues. The first one is the detection reliability, as described in Section 3.3. The second one is the trajectory confidence, which is used to measure the reliability of an existing trajectory T^j as follows:

$$conf\left(\mathrm{T}^j\right) = \left(\frac{1}{l_j} \sum_{t \in \left[t_s^j, t_e^j\right], d^{j,k}=1} \Omega_{t-1}^j\right) \times \exp\left(-\beta \cdot \frac{W}{l_j}\right) \tag{14}$$

where l_j is the length of tracklet T^j. Ω_{t-1}^j is the posterior association probability for trajectory T^j at frame $(t-1)$ computed by Eq. (13). $W = t - t_s^j - l_j$ is the number of frames the object j is missing. β is a control parameter.

Combining the detection reliability and trajectory confidence, the prior association probability is approximated as follows:

$$\begin{aligned} p(< \mathbb{T}_t, \mathbb{X}_t > | \mathbb{T}_{t-1}) &= \prod_{j=1}^{M} p\left(< \mathrm{tT}^j, \mathrm{x}^i > | \mathbb{T}_{t-1}\right) \\ p\left(< \mathrm{tT}^j, \mathrm{x}^i > | \mathbb{T}_{t-1}\right) &= \frac{\delta(\mathrm{x}^i)}{\sum_{i=1}^{M} \delta(\mathrm{x}^i)} \times conf\left(\mathrm{T}^j\right) \end{aligned} \tag{15}$$

where $\delta(\mathrm{x}^i) = p\left(\mathrm{x}_t^i | \mathbb{Z}_{1:t}, \mathfrak{R}_t^j\right)$ is the posterior detection probability computed in Eq. (5).

With the fact that the observation likelihood probability $p(\mathbb{X}_t | < \mathbb{T}_t, \mathbb{X}_t >, \mathbb{T}_{t-1})$ in Eq. (13) is the probability of detections association with the trajectories, then we have:

$$p(\mathbb{X}_t| < \mathbb{T}_t, \mathbb{X}_t >, \mathbb{T}_{t-1}) = \prod_{i=1}^{N} p\left(\mathrm{x}^i| < \mathrm{tT}^j, \mathrm{x}^i >, \mathbb{T}_{t-1}\right) \tag{16}$$

By considering the probability that a detection x^i is originated from an existing trajectory T^j or x^i is a false positive detection, the likelihood $p(\mathrm{x}^i| < \mathrm{tT}^j, \mathrm{x}^i >, \mathbb{T}_{t-1})$ can be computed as follows:

$$\begin{aligned} p\left(\mathrm{x}^i| < \mathrm{tT}^j, \mathrm{x}^i >, \mathbb{T}_{t-1}\right) &= p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{i0}| < \mathrm{tT}^j, \mathrm{x}^i >, \mathbb{T}_{t-1}\right) \\ &+ \sum_i p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{ij}| < \mathrm{tT}^j, \mathrm{x}^i >, \mathbb{T}_{t-1}\right) \\ &= p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{i0}|\mathbb{T}_{t-1}\right) + \sum_i p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{ij}|\mathbb{T}_{t-1}\right) \end{aligned} \tag{17}$$

where $p(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{i0}|\mathbb{T}_{t-1})$ denotes the probability that none of the existing trajectories is associated with the *i_th* detection and $p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{ij}|\mathbb{T}_{t-1}\right)$ represents the probability that the *i_th* detection associated with the *j_th* trajectory. With the fact that the predicted object state for a trajectory in frame t is estimated by the existing tracklet, the pairwise detection-prediction association probability $p_j\left(E_t^{<i,j>}\right)$ and $p_0\left(E_t^{<i,j>}\right)$ computed in section 3.3 are used to compute the association probability $p(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{i0}|\mathbb{T}_{t-1})$ and $p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{ij}|\mathbb{T}_{t-1}\right)$.

$$\begin{aligned} p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{ij}|\mathbb{T}_{t-1}\right) &= p\left(\mathrm{x}^i| < \mathrm{tT}^j, \mathrm{x}^i >_{ij}, \mathbb{T}_{t-1}\right) p\left(< \mathrm{tT}^j, \mathrm{x}^i >_{ij}|\mathbb{T}_{t-1}\right) = p\left(\mathrm{x}^i|\mathrm{T}^j\right) p_j(E_t) p_{\Phi_{ij}} \\ p\left(\mathrm{x}^i, < \mathrm{tT}^j, \mathrm{x}^i >_{i0}|\mathbb{T}_{t-1}\right) &= p\left(\mathrm{x}^i| < \mathrm{tT}^j, \mathrm{x}^i >_{i0}, \mathbb{T}_{t-1}\right) p\left(< \mathrm{tT}^j, \mathrm{x}^i >_{i0}|\mathbb{T}_{t-1}\right) = p_0(E_t) \prod_{j=1}^{M} \left(1 - p_{\Phi_{ij}}\right) \end{aligned} \tag{18}$$

where $p(\mathrm{x}^i|\,\mathrm{T}^j)$ is the association likelihood between x^i and T^j, which is computed by Eq. (7), $p(\mathrm{x}^i|\,\mathrm{T}^j) = p_m(\mathrm{x}^i|\,\mathrm{T}^j) p_s(\mathrm{x}^i|\,\mathrm{T}^j) p_a(\mathrm{x}^i|\,\mathrm{T}^j)$. $p_{\Phi_{ij}}$ is the prior association probability computed in Eq. (15). Then the observation likelihood probability $p(\mathbb{X}_t| < \mathbb{T}_t, \mathbb{X}_t >, \mathbb{T}_{t-1})$ in Eq. (16) can be rewritten as follows:

$$p(\mathbb{X}_t| < \mathbb{T}_t, \mathbb{X}_t >, \mathbb{T}_{t-1}) = \prod_{i=1}^{N} \left\{ p_0(E_t) \prod_{j=1}^{M} \left(1 - p_{\Phi_{ij}}\right) + p\left(\mathrm{x}^i|\mathrm{T}^j\right) p_j(E_t) p_{\Phi_{ij}} \right\} \tag{19}$$

Finally, the data association problem in online MOT is solved by the Hungarian algorithm, in which the association matrix $S_{M \times N} = -\ln\{p(< \mathbb{T}_t, \mathbb{X}_t > |\mathbb{X}_t, \mathbb{T}_{t-1})\}$. The association matrix indicates the cost that the detection x^i is associated with the trajectory T^j. The optimal trajectory-detection pairs are determined by minimizing the total cost in $S_{M \times N}$. According to data association, the final results are achieved by solving the maximizing problem in Eq. (1). Then, the object states and confidence of existing trajectories are updated. The non-associated detections are remained to initialize new potential trajectories, and a new trajectory is found when it grows over five consecutive frames. The non-associated trajectories are terminated if they unassociated in five consecutive frames. The main steps of the proposed online multi-object tracking method are summarized in Algorithm 1.

Algorithm1. The overall algorithm for the proposed online multi-object tracking

Input: A detection set $\mathbb{X}_t$, exist trajectory set $\mathbb{T}_{t-1}$ at frame $(t-1)$ and predicted object state set $\mathbb{Z}_t$

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

Step1: detection reliability estimation based on detection-prediction pair

1. construct local associated motion constraint(LAMC)
2. tracklet dynamic estimation model measure the associated weight $\theta_t^{(i,j)}$ between detection-prediction pair
3. construct posterior detection probability for pairwise detection-prediction

$$p(\mathrm{x}_t^i \mid \mathbb{Z}_{1:t}, \Re_t^j) = \sum_{(i,k)\in\Re_t^j} \theta_t^{(i,k)} p(\mathbb{Z}_t \mid \mathrm{x}_t^i, \Re_t^j) p(\mathrm{x}_t^i \mid \mathbb{Z}_{1:t-1}, \Re_t^j)$$

Step2: The pairwise trajectory-detection association with detection reliability constraint

1. trajectory estimation formulated based on Bayesian framework

$$p(<\mathbb{T}_t, \mathbb{X}_t> \mid \mathbb{X}_t, \mathbb{T}_{1:t-1}) = \frac{p(\mathbb{X}_t \mid <\mathbb{T}_t, \mathbb{X}_t>, \mathbb{T}_{t-1}) p(<\mathbb{T}_t, \mathbb{X}_t> \mid \mathbb{T}_{t-1})}{p(\mathbb{X}_t \mid \mathbb{T}_{t-1})}$$

2. compute the trajectory confidence for each of trajectory as Eq. (14)
3. compute the prior association probability $p(<\mathbb{T}_t, \mathbb{X}_t> \mid \mathbb{T}_{t-1})$ based on detection reliability and trajectory confidence as Eq. (15)
4. The observation likelihood probability $p(\mathbb{X}_t \mid <\mathbb{T}_t, \mathbb{X}_t>, \mathbb{T}_{t-1})$ for pairs trajectory-detection association is computed conditioned on prior detection-prediction pair as Eq. (19)

Step3: construct association matrix $S_{M\times N}$ for trajectory-detection pair. Data association is solved by Hungarian algorithm

Step4: Update trajectory confidence and predicted object state. The non-associated detections are remained to initialize new potential trajectories and non-associated trajectories are remained as potential terminated trajectories.

4 Experiments

In this section, to demonstrate the effectiveness of the proposed online MOT method, thirteen state-of-the-art multi-object tracking algorithms are used to compare with the proposed MOT. Five of them are online algorithms (RMOT [44], SCEA [20], CMOT [3], TSML [39], MDP [40]) and eight are batch methods (SMOT [12], CEM [31], KSP [5], DCT [2], DTLE [30], GOGA [34], JPDA_100 [18], MHT_DAM [24]). For fair comparisons, we have used the reported results in their paper or achieve the results by using the source codes provided by the authors with default parameters on the four public datasets. In addition, we performed our proposed method on synthetic dataset to demonstrate the robustness of the proposed method against noise and missing detections.

4.1 Implementation details

The proposed online MOT algorithm performed on MATLAB 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 $\tau_s = 0.7$in Eq. (2), $\tau = 2$ in Eq. (12), $\beta = 2$ in Eq. (14).

4.2 Synthetic data results

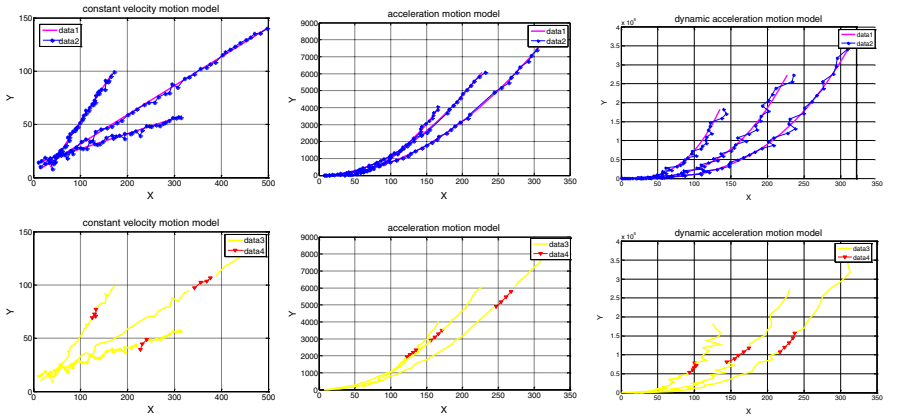
In our first test, we estimate the robustness of our method against noise and missing observations. We construct synthetic data from three models: a) constant velocity motion model, b) acceleration motion model and c) dynamic acceleration motion model, as shown in Eq. (20).

$$[x, y] = a_3 t^3 + a_2 t^2 + a_1 t^1 + a_0 \tag{20}$$

where $a_0$to $a_3$are all random coefficients. In Eq. (20), we set $a_3 = a_2 = 0$to form a linear motion model, use$a_3 = 0$to represents a constant acceleration motion, and set a_0 to a_3 with non-zero random values for a dynamic acceleration motion model.

We then add Gaussian noises on observations with zero mean and 0.3 variance. In our test, we first generate a sequence with 45 frames and manually eliminate frame 31 to 35, to verify the robustness of the proposed methods with missing observations.

The synthetic data association results of our method are shown in Fig. 4. In Fig. 4, data1 (magenta line) denote the ground truth trajectories, data2 (blue star) denote the noised observations, data3 (yellow solid line) denote the association results by the proposed method and data4 (red downward triangle) denote the missing observations estimated by tracklet dynamic estimation model. We can see from Fig. 4a~c that, the proposed data association method effectively overcomes noise interference and accurately associates observations. In addition, the tracklet dynamic estimation model can effectively estimate missing data, a property improving the fragment tracklet association.



(a) constant velocity motion model (b) acceleration motion model (c) dynamic acceleration motion model

Fig. 4 The association results for synthetic dataset

4.3 Real-world datasets and evaluation metrics

For the performance evaluation, four public available challenging datasets, SMOT (Similar Multi-Object Tracking) dataset [12], PETS2009 dataset [15], TUD dataset [1] and MOT 2015 dataset, are used. The SMOT dataset is a very challenging dataset and specially designed for multi-object tracking with similar-looking appearance. It is including various multiple targets, the number of objects from 3~80, the length of sequences from 130~1285, both including non-rigid and rigid objects. We adopt five sequences from SMOT, the Salmon, Juggling, Acrobats, Seagulls and Crowd sequences, for evaluation. In PETS2009 dataset, the widely used S2.L1 and S2.L2 sequences are included for evaluation, in which the sequences show outdoor surveillance scenes with many pedestrians. The numbers of the tracked objects are 19 and 43, and the length of S2.L1 and S2.L2 sequences are from 436 to 795. We adopt the Campus, Crossing and Stadtmitte sequences from TUD dataset for evaluation. The challenges for TUD dataset is severely occlusions with low viewpoint, the number of tracked objects in Campus, Crossing and Stadtmitte are from 7~12, the lengths are range from 71 to 201. The MOT 2015 dataset is a latest MOT dataset, it contains 11 changeling sequences with occlusion, clutter background, scale and shape changes, moving camera and stationary camera, the length of sequences in this dataset from 187 ~ 1194, the number of tracked object from 12 ~157. For fair comparisons, we use the public available detections provided by detector. For SMOT dataset, we use the public available detections provided by [12]. The public detections for PETS 2009 and TUD dataset are provided by [31]. The detections for MOT 2015 datasets is provided by https://motchallenge.net/data/2D_MOT_2015/.

We use the common CLEAR performance metrics, including MOTP, MOTA, FP, FN, IDS, [6] for quantitative evaluation. The multiple object tracking precision (MOTP↑) evaluates the average overlap rate between true and estimated bounding boxes. The multiple objects tracking accuracy (MOTA↑) indicates the accuracy composed of false positives (FP↓), false negatives (FN↓) and identities switches (IDS↓). In addition, some metrics defined in [28], the percentage of mostly tracked (MT↑), mostly lost (ML↓) or partially tracked (PT) trajectories, the number of ground truth trajectory fragments by tracking result (FM↓), the percentage of correctly matched objects with ground truth objects (Recall), as well as the percentage of correctly matched targets with detect results (Precision), are used to evaluate the performance of MOT. (↑ denotes the higher score is the better results, and ↓means that lower is better).

4.4 Results and discussion

Tables 1, 2, 3 and 4 and Figs. 5, 6, 7, 8 and 9 show the quantitative results of the proposed method and the state-of-the-art tracking methods on SMOT, PETS2009, TUD and MOT 2015 datasets, respectively. For all metrics, the best scores are shown in red and the batch multi-object tracking methods are marked with star. Some sample results from SMOT, PETS2009, TUD and MOT 2015 datasets are shown in Figs. 5, 6, 7 and 8. Fig. 9 shows plots for Recall, Precision, MT, PT, MOTA and MOTP scores for all videos of SMOT, PETS2009, TUD and MOT 2015 datasets.

Results for SMOT dataset Slamon sequence has three skiers racing down a slalom with complex zig-zag motion. They frequently move closely with each other and one of the skier escapes out of the field for a long time. Slamon sequence is also accompanied with camera motion and frequently zooming. Due to the tracklet dynamic estimation model used in this

Table 1 Performance comparison between state-of-the-art methods and ours on SMOT dataset

DATASET	METHOD	RECALL (%)↑	PRECISION (%)↑	MT (%)↑	PT (%)	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA (%)↑	MOTP (%)↑
SMOT	RMOT	99.9	69.5	100	0	0	784	2	112	1	49.90	91.90
-Salmon	SCEA	92.3	92.3	100	0	0	138	138	0	28	84.60	99.10
600frames	CMOT	90.5	100	100	0	0	0	9	1	1	99.40	94.60
	SMOT*	32.4	98.6	0	100	0	8	1210	23	23	30.70	100
	Ours	98.2	98.8	100	0	0	0	0	2	0	95.6	98.70
SMOT	RMOT	56.7	75.4	0	100	0	72	169	41	33	27.70	92.90
-Juggling	SCEA	76.9	97.7	33.3	66.7	0	7	90	30	11	67.40	87.80
130frames	SMOT*	87.2	98.8	66.7	33.3	0	4	50	7	5	84.40	100
	Ours	83.6	98.2	66.7	33.3	0	6	64	6	6	80.50	93.20
SMOT	RMOT	39.9	76.4	0	80	20	96	467	18	11	25.20	77.00
-Acrobats	SCEA	91.6	94.8	80	20	0	39	65	2	12	86.40	98.50
156frames	CMOT	91.7	96.5	80	20	0	26	57	7	5	88.40	89.80
	SMOT*	83.8	99.2	80	20	0	5	126	4	2	82.60	100
	Ours	96.7	98.2	100	0	0	0	26	2	2	96.40	97.90
SMOT	RMOT	96.3	39.8	95.8	4.2	0	6708	171	188	9	53.70	88.70
-Seagulls	SCEA	93.5	93.6	95.8	4.2	0	294	297	1	48	87.10	99.80
581frames	CMOT	93.1	96.9	100	0	0	143	86	6	6	94.90	96.90
	SMOT*	90.6	99.5	87.5	12.5	0	23	433	17	5	89.70	100
	Ours	85.3	99.8	91.4	0	8.4	6	678	1	1	85.10	99.70
SMOT	RMOT	99.8	72	97.5	0	2.5	8245	32	116	6	60.40	98.40
-Crowd	SCEA	96.2	96.4	95	2.5	2.5	772	800	0	78	92.60	99.80
1285frames	CMOT	93.7	97.3	97.5	0	2.5	573	274	11	6	96.00	99
	SMOT*	82.0	99.7	75	11.25	13.75	55	3817	47	12	81.50	100
	Ours	98.7	99.8	96.25	0	3.75	33	284	1	0	98.50	99.70

↑ denotes the higher the better; ↓means the lower the better

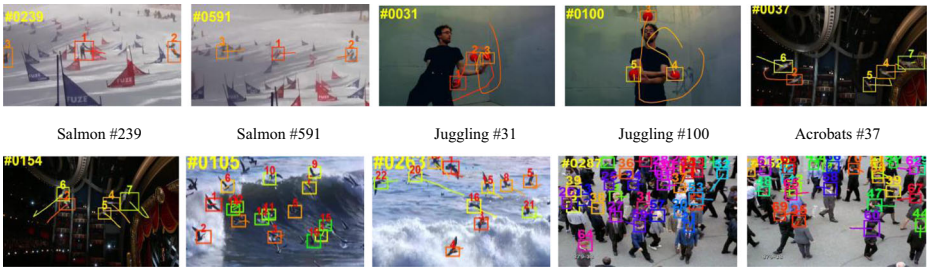


Fig. 5 Sample tracking results of the proposed method on SMOT dataset sequences. At each frame, objects are attached with bounding boxes and ID labels with different colors

paper can effectively estimate missing data, the proposed method achieves the best performance except MOTA metric compared to the competing trackers.

Juggling sequence is a very challenging scene with juggler performing three ball alternating tricks by adding artistic motions. The combined motion of balls, juggler and camera with similar appearance of balls makes it even hard for a human to keep track on the balls. The SMOT method, achieves the best result on this sequence due to its global iterative optimization and dynamic motion model. The proposed method shows a second best results on this challenging sequence.

The main difficult in Acrobats sequence is the acrobats dressed same and lineup in air with occlusions. The proposed method achieves the best performance in all terms except MOTA. The data association method used in this paper can effectively overcome weak appearance cue and rely on the dynamic motion model to accurately associate observations. SMOT tracker has the highest MOTA = 100%, which is 2.1% higher than the proposed method.

Seagulls sequence shows an extremely difficult scene where a flock of seagulls take off at sea with similar appearance, spatial close-moving, frequent occlusion and clutter background. CMOT tracker gives the best MOTA in this challenging sequence, which is 9.8% higher than the propose method. SMOT tracker achieves the best MOTP with 100%, 0.3% higher than our method.

The Crowd sequence is an over-crowded surveillance scene with frequent occlusions and close-moving pedestrians. The data association strategy and the tracklet dynamic estimation model used in this paper can help to improve the fragment tracklet association. Hence, our method achieves the highest MOTA and precious, with low MT, FP, IDS and FM. The qualitative tracking results are shown in Fig. 5.

Results for PETS2009 dataset PETS09-S2.L1 and PETS09-S2.L2 are the most widely used multi-pedestrian tracking sequences. The datasets provided multi-view from different camera angles. We only use the first view of each sequence in our experimental. The S2.L1 sequence is a medium crowed scene with pedestrian frequently changing their motion

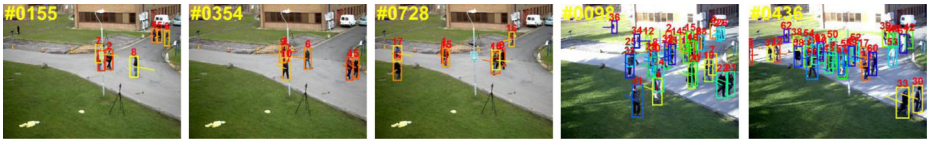
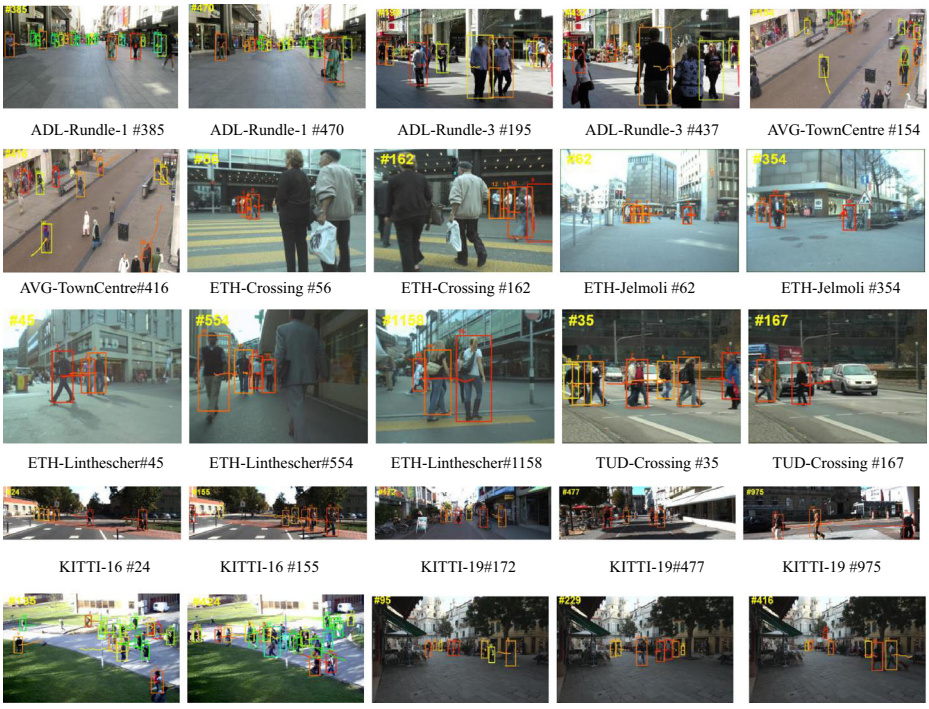


Fig. 6 Sample tracking results of the proposed method on PETS2009 dataset sequences. At each frame, objects are attached with bounding boxes and ID labels with different colors

TUD-campus #36 TUD-crossing #44 TUD-crossing #177 TUD-Stadtmitte #65 TUD-Stadtmitte #171

Fig. 7 Sample tracking results of the proposed method on TUD dataset sequences. At each frame, objects are attached with bounding boxes and ID labels with different colors

directions. Many state-of-the-art methods, both batch tracking and online tracking methods, are included for fair comparison. Table 2 shows that the batch based tracking, such as CEM and DCT methods, achieve the best MOTA in this sequence. The proposed method achieves fairly good results in terms of six metrics: MOTA MOTP, recall, precious, few ID switches and few fragments. PETS-S2.L2 sequence is a highly crowd scene with frequently pedestrian occlusion and illumination changes. Table 2 shows that batch based tracking, GOGA and DCT methods, give the best precision and MOTP, respectively. Our method achieves high recall, MOTP and high ration of MT with relatively low ID switches. Comparing with the online tracking methods, the batch based methods give more satisfied results. This is because the available information of following-up frames and, consequently, a globally iterative optimal association can be used in batch based trackers to effectively tackle detection errors and tracking failures. However, as one of the online tracker which focuses on objects only up to



ADL-Rundle-1 #385 ADL-Rundle-1 #470 ADL-Rundle-3 #195 ADL-Rundle-3 #437 AVG-TownCentre #154

AVG-TownCentre#416 ETH-Crossing #56 ETH-Crossing #162 ETH-Jelmoli #62 ETH-Jelmoli #354

ETH-Linthescher#45 ETH-Linthescher#554 ETH-Linthescher#1158 TUD-Crossing #35 TUD-Crossing #167

KITTI-16 #24 KITTI-16 #155 KITTI-19#172 KITTI-19#477 KITTI-19 #975

PETS09-S2L2 #135 PETS09-S2L2 #424 Venice-1 #95 Venice-1 #229 Venice-1 #416

Fig. 8 Sample tracking results of the proposed method on 11 testing video sequences of the MOT 2015 dataset. At each frame, objects are attached with bounding boxes and ID labels with different colors

Table 2 Performance comparison between state-of-the-art methods and ours on PETS09 dataset

DATASET	METHOD	RECALL (%)↑	PRECISION (%)↑	MT (%)↑	PT (%)	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA (%)↑	MOTP (%)↑
PETS2009_S2L1	RMOT	91.0	63.0	89.5	10.5	0	2485	417	41	78	36.7	65.3
795 frames	SCEA	59.1	59.2	10.5	84.25	5.25	1890	1903	3	175	18.4	90.4
	CMOT	87.7	74	73.7	16.5	0	1435	572	37	108	56	65.4
	TSML	96	97.7	94.7	5.3	0	–	–	18	21	93.4	86.4
	DCT*	–	–	100	0	0	–	–	10	8	95.9	78.7
	CEM*	92.4	98.4	91.3	4.4	4.3	59	302	11	6	90.6	80.2
	KSP*	83.8	96.3	73.9	17.4	8.7	126	641	13	22	80.3	72
	Ours	92.6	92.7	89.5	10.5	0	341	343	0	52	85.3	91.2
PETS2009_S2L2 436 frames	RMOT	52.7	48	6.5	83.7	9.8	4326	563	570	548	47	65.4
	SCEA	46.5	72	0	83.7	16.3	1859	5504	760	718	21.1	59.9
	CMOT	52.8	69.2	6.9	79.1	14.0	2422	4860	230	396	27	66.5
	SMOT*	–	–	0	76.2	23.8	–	–	251	514	34.4	70
	DCT*	–	–	–	–	–	–	–	–	–	89.3	56.4
	CEM*	65.5	89.8	11.9	73.8	14.3	622	2881	150	165	44.9	70.2
	KSP*	26.8	92.1	9.5	36.4	54.1	193	6117	22	38	24.2	60.9
	DTLE*	65.1	92.4	25.6	72.1	2.3	549	3592	167	–	58.1	59.8
	GOGA*	49	95.4	9.5	67.5	23.0	199	4257	137	–	45	64.1
	JPDA100*	69.8	87.2	25.6	72.1	2.3	1051	3108	143	–	58.2	59.5
	Ours	75.2	76	41.8	53.6	4.6	2443	2555	28	382	51.2	71.2

current frame, the proposed method achieves comparable results with the batch based tracking methods. Some sampled visual results of S2.L1 and S2.L2 of the proposed method are shown in Fig. 6.

Results for TUD dataset TUD-Campus, TUD-Crossing and TUD-Stadtmitte sequences are used to evaluate the performance of pedestrian tracking. The main challenges for these sequences are long occlusions, spatial closely moving targets with low viewpoint. Some sampled tracking results for the proposed method are shown in Fig. 7. The quantitative results for all competing algorithms are shown in Table 3. Comparing with the state-of-the-art methods, the proposed method significantly improves the recall, precision, MOTA and MOTP for TUD-Stadtmitte and TUD-campus sequences. In two out of the total three sequences, our MOTA is higher than the batch based tracking methods like CEM, DCT and KSP.

Results for MOT 2015 dataset The MOT 2015 dataset is a latest MOT dataset. The test set of this dataset contain 11 changeling sequences with occlusion, clutter background, large scale and shape changes, moving camera and stationary camera. Table 4 shows the compare results of our method with the state-of-the-art methods on test sequences in MOT 2015 dataset. Fig. 8 shows some sampled tracking results on the 11 test sequences. As shown in Table 4, our tracker achieve the second best MOTA metric with 32.6% compared with the state-of-the-art methods and achieve the best performance in IDs even though the proposed method works in online mode. The good tracking performance of the proposed method demonstrates the advantages of our method that using Hankel matrix based object states predication and local motion constraint are beneficial to recover short fragment tracklets and reduce the ID switches in online MOT. This is because it takes long history object states to construct dynamic motion model of the objects, which is robust to occlusion and can filter out some unreliable association among the trajectories and the detections with local associated motion constraint.

4.5 Run time performance

In the experiment, given the detection responses, we present the average execution speed of the proposed method and compared trackers by averaging the trackers 5 times for all test sequence in MOT 2015 dataset. The results are show in Table 5 and Fig. 10. The speed of the trackers is measured by frame-per-second (FPS). From the Table 5 and Fig.10, we can see that the proposed method perform well in run time, which has a comparative result with the state-of-the-art methods in MOT. The average speed reaches 15.21 FPS. With further optimization of the code, the speed of the proposed method can improved.

Online multi-object tracking plays an important role in numerous essential applications, such as visual surveillance, traffic safety, autonomous driving and navigation. We test the proposed method on four public available MOT datasets with frequently occlusion, clutter background, large scale and shape changes, non-rigid and rigid objects, moving camera and stationary camera scenes. Although the proposed method shows some good tracking performance in most test sequences, it still needs to improve until it can be used in real applications. The proposed method is online method, the run time is about 15 fps, which is still slower than the real-time requirement. Therefore, how to further optimization the code of our method in order to speed-up the run time is our future work. As known, appearance and motion models are two main cues used in MOT, in our work, we mainly focus on how to use the history object

Table 3 Performance comparison between state-of-the-art methods and ours on TUD dataset

DATASET	METHOD	RECALL (%)↑	PRECISION (%)↑	MT (%)↑	PT (%)	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA (%)↑	MOTP (%)↑
TUD-Stadtmitte	RMOT	57.9	46	20	80	0	784	487	13	27	–	56.8
179frames	SCEA	54.2	63	10	90	0	368	529	6	18	21.9	56.9
	CMOT	57.2	59	10	90	0	459	495	9	33	16.7	56.8
	CEM*	84.7	86.7	70	30	0	92	108	4	3	71.1	65.5
	KSP*	63.1	79.2	10	80	10	117	261	5	15	45.8	56.7
	DCT*	–	–	60	40	0	–	–	4	1	61.8	63.2
	GOGA*	69.1	85.6	40	60	0	134	357	15	–	56.2	61.6
	JPDA_100*	69.8	87.4	40	50	10	116	349	10	–	58.9	59.8
	Ours	88.6	89.5	60	40	0	116	132	2	6	78.5	99.7
TUD-campus	RMOT	99	89.1	100	0	0	49	4	15	2	83.2	91.9
71frames	SCEA	89.9	89.9	71.4	28.6	0	41	41	0	5	79.8	99.9
	CMOT	77.735	93.7	71.4	28.6	0	21	92	0	0	72.1	99.1
	SMOT*	57.3	97.9	42.8	28.6	28.6	5	173	5	2	54.8	99.9
	Ours	99.5	99.5	100	0	0	2	2	2	0	98.5	99.2
TUD-crossing	RMOT	99	99.5	100	0	0	5	1	2	0	99.2	97.3
201frames	SCEA	93.6	93.3	100	0	0	68	64	0	12	86.9	99.9
	CMOT	97.9	94.2	100	0	0	61	21	2	0	91.6	98.6
	SMOT*	74.2	98.8	50	50	0	9	259	16	12	71.7	99.7
	Ours	98.7	99.1	100	0	0	9	13	1	0	97.7	99.8

Table 4 Performance comparison between state-of-the-art methods and ours on MOT 15 dataset

DATASET	METHOD	MT (%)↑	ML (%)↓	FP↓	FN↓	IDS↓	FM↓	MOTA (%)↑	MOTP (%)↑
MOT15	RMOT	5.3	53.3	12,473	36,835	684	1282	18.6	69.6
	SCEA	8.9	47.3	6060	36,912	604	1182	29.1	71.1
	CMOT	3.2	55.8	12,970	38,538	637	1716	15.1	70.5
	SMOT	2.8	54.8	8780	40,310	1148	2132	18.2	71.2
	TSML	14	39.4	7869	31,908	618	959	34.3	71.7
	CEM*	8.5	46.5	14,180	34,591	813	1023	19.3	70.7
	GOGA*	6	40.8	13,171	34,814	4537	3090	14.5	70.8
	JPDA_100*	5	58.1	6373	40,084	365	869	23.8	68.2
	MDP	13	38.4	9717	32,422	680	1500	30.3	71.3
	MHT_DAM	16	43.8	9064	32,060	435	826	32.4	71.8
	Ours	11.5	41.9	8659	31,040	398	1021	32.6	72.3

states to construct the motion model of the object. Then use the dynamic motion model to predict object state and build detection reliability during online tracking. We just use the color histogram to build appearance model of the object, no extra information are used. In recent years, deep convolutional neural networks have shown impressive performance for many tasks. The features from deep convolutional layers are discriminative while preserving spatial and structural information. Hence, one of our further researches is introducing deep learning into our online MOT by learning discriminative appearance model of the object. In addition, occlusion problem and mis-detection are common issues in MOT, in our work, we use the predicted object state to overcome the occlusion and mis-detection. However, the predicted object state heavily relies on dynamic motion model of the object, which is not considering the appearance cue of the object. While in real application, such as urban traffic scenarios, the

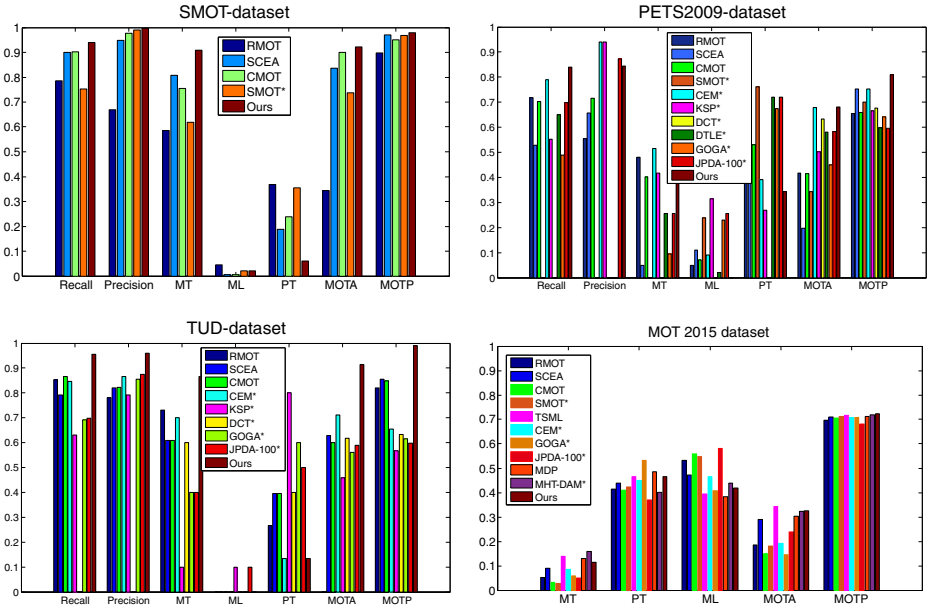


Fig. 9 Evaluation metric results for four public datasets

Table 5 Run time performance (FPS)

Method	RMOT	SCEA	CMOT	SMOT	TSML	CEM	GOGA	JPDA_m	MDP	MHT_DAM	Proposed
MOTA (%)	18.6	26.3	15.1	18.2	34.3	19.3	14.5	23.8	30.3	32.4	32.6
average(FPS)	7.9	6.8	1.7	2.7	6.5	1.1	444.4	32.6	1.1	0.7	15.21

heavy traffic and congestion situation often include serious occlusion. The detection responses provided by detector in this situation always with mis-detection, which will pose more challenges for the proposed method to tackle occlusion and accurately tracking the targets. Therefore, in our further research, we will pay more attention on occlusion analysis in order to better address the challenging caused by occlusions.

5 Conclusion

In this paper, an online detection based multi-object tracking method is proposed. The proposed method splits the data association problem into two related optimized estimation steps, which integrates the trajectory estimation and detection reliability estimation into a unified framework. The trajectory-detection association pairs are achieved by sequentially introducing the previous trajectory and the current detection reliability. To further improve the correctness of association between trajectories and detections, tracklet dynamic estimation model and trajectory confidence are used to recover short fragment tracklet and reduce ambiguity caused by missing detections. In addition, with the local associated motion constraint, the detections are refined and the unreliable associations between tracklets and detections are filtered out. What's more, the MAP framework allows an alternative updating on trajectory estimation and detection reliability estimation in a sequential manner. Compared with the state-of-the-art multi-object tracking algorithms, including both batch tracking and online tracking algorithms, experimental results verify the effectiveness of the proposed

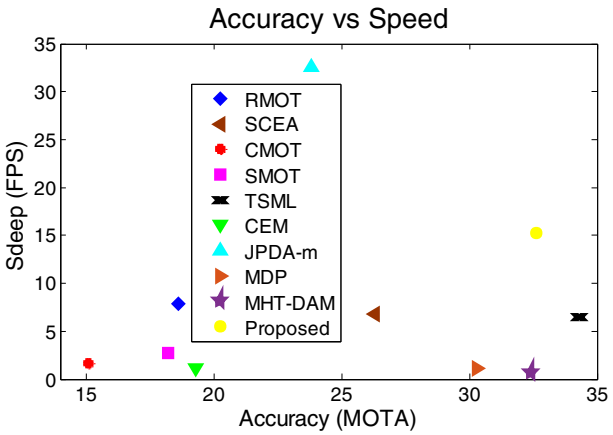


Fig. 10 The performances compare between the proposed method and the stat-of-the-art trackers. Each marker denotes a tracker accuracy and speed measured in FPS, the higher and more right is better

method. However, the proposed method is detection-based tracking method, which heavily relies on object detector, it has limitation under long-term occlusion and the run time of the proposed method is still slower than the real-time requirement. Hence, in our future work, we will focus on speed-up the proposed method and occlusion analysis in order to satisfy real application and better address the challenging caused by occlusions.

Acknowledgment This work was supported in part by National Natural Science Foundation of China (Grant No. 61673125 and 61703115), in part by the Frontier and Key Technology Innovation Special Funds of Guangdong Province (Grant No. 2014B090919002, 2016B090910003 and 2015B010917003) and Program of Foshan Innovation Team of Science and Technology (Grant No. 2015IT100072).

References

1. Andriluka M, Roth S, Schiele B (2008) People-tracking-by-detection and people-detection-by-tracking. In: Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp 1–8
2. Andriyenko A, Schindler K, Roth S (2012) Discrete-continuous optimization for multi-target tracking. In: Computer Vision and Pattern Recognition (CVPR), 2012 I.E. Conference on, IEEE, pp 1926–1933
3. 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
4. Bae S-H, Yoon K-J (2014) Robust online multiobject tracking with data association and track management. IEEE Trans Image Process 23(7):2820–2833
5. Berclaz J, Fleuret F, Turetken E, Fua P (2011) Multiple object tracking using k-shortest paths optimization. IEEE Trans Pattern Anal Mach Intell 33(9):1806–1819
6. Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J Image and Video Processing 2008(1):1–10
7. 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(9):1820–1833
8. Brendel W, Amer M, Todorovic S (2011) Multiobject tracking as maximum weight independent set. In: Computer Vision and Pattern Recognition (CVPR), 2011 I.E. Conference on. IEEE, pp 1273–1280
9. Chen Z, You X, Zhong B, Li J, Tao D (2016) Dynamically modulated mask sparse tracking. IEEE Trans Cybern 99:1–13
10. Chenouard N, Bloch I, Olivo-Marin J-C (2013) Multiple hypothesis tracking for cluttered biological image sequences. IEEE Trans Pattern Anal Mach Intell 35(11):2736–3750
11. Collins RT (2012) Multitarget data association with higher-order motion models. In: Computer Vision and Pattern Recognition (CVPR), 2012 I.E. Conference on. IEEE, pp 1744–1751
12. Dicle C, Camps OI, Sznaier M (2013) The way they move: tracking multiple targets with similar appearance. In: Proceedings of the IEEE international conference on computer vision. pp 2304–2311
13. Dollár P, Appel R, Belongie S, Perona P (2014) Fast feature pyramids for object detection. IEEE Trans Pattern Anal Mach Intell 36(8):1532–1545
14. Duan G, Ai H, Cao S, Lao S (2012) Group tracking: exploring mutual relations for multiple object tracking. Comput Vis ECCV 2012:129–143
15. Ellis A, Shahrokni A, Ferryman JM (2010) Pets2009 and winter-pets 2009 results: a combined evaluation. In: 12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, IEEE, December 7-12, 2009, Snowbird, Utah, USA, pp 1-8
16. Everingham L, Gool C, Williams K et al (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
17. Fortmann T, Bar-Shalom Y, Scheffe M (1980) Joint probabilistic data association for multiple targets in clutter. In: Proc. Conf. on Information Sciences and Systems
18. Hamid Rezatofighi S, Milan A, Zhang Z, Shi Q, Dick A, Reid I (2015) Joint probabilistic data association revisited. In: Proceedings of the IEEE international conference on computer vision, pp 3047–3055
19. Han J, Zhang D, Cheng G, Guo L, Ren J (2015) Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans Geosci Remote Sens 53(6):3325–3337
20. 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

21. Hong Z, Chen Z, Wang C, Mei X, Prokhorov D, Tao D (2015) MUlti-store tracker (MUSTer): a cognitive psychology inspired approach to object tracking. In: Computer Vision and Pattern Recognition (CVPR), June 8-10, 2015, Boston, USA, pp 749–758
22. Isard M, Blake A (1998) Condensation—conditional density propagation for visual tracking. Int J Comput Vis 29(1):5–28
23. Jiang H, Fels S, Little JJ (2007) A linear programming approach for multiple object tracking. In: Computer vision and pattern recognition, 2007. CVPR'07. IEEE Conference on, 2007. IEEE, pp 1–8
24. Kim C, Li F, Ciptadi A, Rehg JM (2015) Multiple hypothesis tracking revisited. In: IEEE international conference on computer vision. pp 4696–4704
25. Kuhn HW (1955) The Hungarian method for the assignment problem. Nav Res Logist Q 2(1–2):83–97
26. Kuo C-H, Nevatia R (2011) How does person identity recognition help multi-person tracking? In: Computer Vision and Pattern Recognition (CVPR), 2011 I.E. Conference on, IEEE, pp 1217–1224
27. Leibe B, Schindler K, Van Gool L (2007) Coupled detection and trajectory estimation for multi-object tracking. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, pp 1–8
28. Li Y, Huang C, Nevatia R (2009) Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: Computer Vision and Pattern Recognition. CVPR 2009. IEEE Conference on, 2009. IEEE, pp 2953-296046. M
29. Luo W, Xing J, Zhang X, Zhao X, Kim T-K (2014) Multiple object tracking: a literature review. arXiv preprint arXiv:14097618
30. Milan A, Schindler K, Roth S (2013) Detection-and trajectory-level exclusion in multiple object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3682–3689
31. Milan A, Roth S, Schindler K (2014) Continuous energy minimization for multitarget tracking. IEEE Trans Pattern Anal Mach Intell 36(1):58–72
32. Oh S, Russell S, Sastry S (2009) Markov chain Monte Carlo data association for multi-target tracking. IEEE Trans Autom Control 54(3):481–497
33. Okuma K, Taleghani A, Nd F, Little JJ, Lowe DG (2004) A boosted particle filter: multitarget detection and tracking. Comput Vis ECCV 2004:28–39
34. Pirsiavash H, Ramanan D, Fowlkes CC (2011) Globally-optimal greedy algorithms for tracking a variable number of objects. In: Computer Vision and Pattern Recognition (CVPR), 2011 I.E. Conference on, IEEE, pp 1201–1208
35. Qi Y, Zhang S, Qin L, Yao H, Huang Q, Lim J, Yang MH (2016) Hedged deep tracking. In: Computer Vision and Pattern Recognition, 2016. CVPR'16. IEEE Conference on, June 27-30, 2016, Las Vegas, USA, pp 4303-4311
36. Reid D (1979) An algorithm for tracking multiple targets. IEEE Trans Autom Control 24(6):843–854
37. Rezatofighi SH, Milani A, Zhang Z, Shi Q, Dick A, Reid I (2016) Joint probabilistic matching using m-best solutions. In: Computer Vision and Pattern Recognition, 2016. CVPR'16. IEEE Conference on, June 27-30, 2016, Las Vegas, USA, pp 136-145
38. Wang X, Yang M, Zhu S (2013) Lin Y Regionlets for generic object detection. In: Proceedings of the IEEE international conference on computer vision, pp 17–24
39. Wang B, Wang G, Chan KL, Wang L (2016) Tracklet association by online target-specific metric learning and coherent dynamics estimation. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(3):589-602
40. Xiang Y, Alahi A, Savarese S (2015) Learning to track: online multi-object tracking by decision making. In: IEEE international conference on computer vision, pp 4705–4713
41. Yang B, Nevatia R (2012) Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In: Computer Vision and Pattern Recognition (CVPR), 2012 I.E. Conference on, IEEE, pp 1918–1925
42. Yang B, Nevatia R (2014) Multi-target tracking by online learning a CRF model of appearance and motion patterns. Int J Comput Vis 107(2):203–217
43. Yang M, Pei M, Shen J, Jia Y (2015) Robust online multi-object tracking by maximum a posteriori estimation with sequential trajectory prior. In: International Conference on neural information processing, Springer, pp 623–633
44. Yoon JH, Yang M-H, Lim J, Yoon K-J (2015) Bayesian multi-object tracking using motion context from multiple objects. In: Applications of Computer Vision (WACV), 2015 I.E. Winter Conference on, 2015. IEEE, pp 33–40
45. Zhang D, Han J, Li C, Wang J, Li X (2016) Detection of co-salient objects by looking deep and wide. Int J Comput Vis 120(2):215–232
46. Zhao S, Yao H, Gao Y, Ji RR, Ding G (2017) Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. IEEE Trans Multimedia 99:1–1

Honghong Yang received her M.Sc. degree in automatic control from Northwestern Polytechnical University, Xi'an, China, in 2014. She is currently working toward the PhD degree in the Department of Automation at Northwestern Polytechnical University and working as a joint PhD student in University of Alberta from September 2016. Her research interests include computer vision, object tracking and detection.

Li He received B.Eng., M.Sc. and Ph.D in Department of Automation at Northwestern Polytechnical University, China, in 2006, 2008 and 2014, respectively. He served as a Postdoctoral Fellow in Department of Computing Science, University of Alberta from 2014 to 2017. He is currently a lecturer in School of Electromechanical Engineering, Guangdong University of Technology. His current research interests include machine learning, image analysis, and computer vision.