IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 11, 2018, accepted November 5, 2018, date of publication November 12, 2018, date of current version January 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2881019

Enhanced Association With Supervoxels in Multiple Hypothesis Tracking

HAO SHENG[1,2,3], (Member, IEEE), XINYU ZHANG[1], YANG ZHANG[1], YUBIN WU[1], JIAHUI CHEN[1], AND ZHANG XIONG[1,3]

[1] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[2] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China
[3] Shenzhen Key Laboratory of Data Vitalization, Research Institute in Shenzhen, Beihang University, Beijing 100191, China

Corresponding author: Yang Zhang (yang.zhang@buaa.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFC0803700, in part by the National Natural Science Foundation of China under Grant 61872025, in part by the Macao Science and Technology Development Fund under Grant 138/2016/A3, in part by the Open Fund of the State Key Laboratory of Software Development Environment under Grant SKLSDE-2017ZX-09, in part by the Project of Experimental Verification of the Basic Commonness and Key Technical Standards of the Industrial Internet network architecture, and in part by the Technology Innovation Fund of China Electronic Technology Group Corporation.

ABSTRACT Remarkable progress has been made in the field of multi-object tracking. Although tracking-by-detection has recently became one of the most popular frameworks, it still has one main drawback: this approach relies heavily on the quality of detection. Thus, the missing detections caused by partial occlusion usually lead to fragment problem. To address this problem, this paper introduces supervoxels to represent objects with partial occlusion, even for missing detections. We first extract superpixels of the foreground, and then our proposed supervoxel consists of spatial-temporal sequences of superpixels. The supervoxels represent tracklets at the image level, so it is robust for initial detection. Then, we incorporate supervoxels into multiple hypotheses tracking by considering the enhanced association with supervoxels (EAS). Moreover, we propose a detection refinement method based on EAS. As our approach allows us to handle partial occlusion problems, we achieve remarkable results in crowded scenes. Finally, our experiments on both MOT15 and MOT16 benchmarks show that our EAS is competitive with the state-of-the-art trackers.

INDEX TERMS Enhanced association, multiple object tracking, partial occlusion, supervoxel.

I. INTRODUCTION

Multiple object tracking is a vital study in computer vision, which focuses on recovering spatiotemporal trajectories of objects from videos. Tracking-by-detection is the most popular framework [1] in multiple object tracking. It independently detects objects from each frame using an offline trained detector [2], [3]. It reduces the search space relative to the very large solution space of global searching. In addition, it converts object tracking into a data association problem, *i.e.*, assigning detections to the appropriate objects and associating those detections with a consistent trajectory. Despite the remarkable progress in this field in recent years, multiple object tracking remains a major challenge in partial occlusion scenes.

Partial occlusion occurs frequently while targets are moving, which leads to tracking errors, such as fragments and false joints. This is mainly because the algorithm is not robust enough to utilize partial information or to associate inaccurate detections of partially occluded objects. In human vision, objects are associated, even partially visible objects, according to the motion and appearance information because partial information estimates the complete object position. Hence, we propose a more robust approach to exploit partial information and maintain a stable association with objects in the partial occlusion scenes.

To sufficiently represent the partial information, a superpixel [4] is proposed to segment the image into pixel regions (*i.e.*, superpixel). The superpixel accurately represents an object's partially visible region with similar features. It can be accurately correlated, based on the stability of the object moving between neighboring frames. Therefore, it also accurately associates the corresponding objects. In this paper, a supervoxel is defined as a spatial-temporal continuous sequence of superpixels, which indicates the tracklet of a partial region of an object, as shown in Figure 1. Then, we propose a novel enhanced association with supervoxels (EAS)

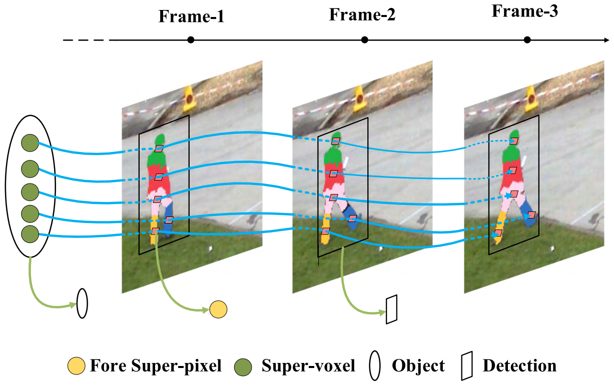


FIGURE 1. Continuous superpixels are colored as the foreground in the adjacent 3 frames. An object(pedestrian) contains 5 superpixels in each frame. Superpixels in the same color compose 5 supervoxels in the above image.

approach for multiple object tracking to solve the partial occlusion problem. First, based on the association of superpixels, a superpixel association tree is built to obtain the candidate supervoxels. The tree structure allows EAS to search for more candidate superpixel sequences and has great local continuity. The supervoxel has a remarkable effect in representing partially occluded objects and associating objects among frames. Then, the objects' positions are accurately estimated, and the pairwise edges association between detections is enhanced with the supervoxel relationship on the MHT framework. Overall, partially occluded objects are joined into trajectories effectively, which reduces false negatives in tracking.

We evaluate our approach on a set of standard public datasets. The experimental results show that EAS is better than other approaches in partial occlusion scenes and it achieves competitive results compared with state-of-the-art approaches.

II. RELATED WORK

Multi-object tracking has been studied for many years and considerable research work has been performed in this field. A segmentation-based tracking algorithm was proposed in recently. Segmentation technology can obtain more accurate foreground information for objects and provide more partial information for multiple object tracking. This section briefly reviews the most important milestones in this field, as well as the research conducted on the segmentation algorithm in the related domain.

A. MULTIPLE OBJECT TRACKING

Most tracking approaches fall into two categories: online [5], [6] or offline [7]–[9]. The online approach uses information from past frames to estimate the current state recursively which is commonly applied in time-critical scenes. The offline approach uses global information from all the frames of a video sequence in order to achieve higher accuracy.

Online algorithms do not use global information to associate trajectories, which eventually leads to an accumulation of errors. Breitenstein *et al.* [10] used a particle filters approach and interpolated missing or inaccurate detections caused by objects with nonlinear motion. Breitenstein *et al.* [11] later adopted particle filtering to approximate more complex multi-modal posteriors. Yamaguchi *et al.* [12] proposed the agent-based behavior model. Wu *et al.* [13] compared online tracking approaches.

Offline algorithms, on the other hand, use information from all frames for tracking. Network flow based algorithms [14] define each detection as a node in the network and solve the tracking problem by calculating the min-cost network flow. Mclaughlin *et al.* [7] linked distant trackers based on motion information to better address long-term occlusions and missing detections. Kim *et al.* [8] introduced dummy nodes to overcome missed detections. Conditional random field (CRF) algorithms [9], [15] generalize the global CRF costs to assign label to detections. However, the failure of trajectory association, which is caused by object occlusion, results in trajectory fragmentation and ID-switch in the above approaches.

To solve the association problem of detections among frames, it is necessary to measure the similarity among detections. Some typical similarity measurement approaches take the feature of whole objects as the basis of the association. Spatio-temporal constraints [15], [16] are commonly used in current tracking algorithms. To avoid making mistakes, detections are only associated in closer frames and regions. Appearance is an important feature in measuring similarities. For instance, Mclaughlin *et al.* [7] used color histogram information to calculate similarity; Kim *et al.* [8] and Sadeghian *et al.* [17] introduced convolutional neural network features. Li *et al.* [18] reviewed popular visual tracking methods based on deep learning features. Milan *et al.* [19] and Mclaughlin *et al.* [7] utilized motion information. However, when the tracking objects are partially occluded, these approaches are most likely ineffective because the partial information is directly used to compare the whole object, which is obviously sub-optimal.

Moreover, some tracking approaches [9], [20] use partial information of partially occluded objects to solve the occlusion problem. Some use partial region information obtained through segmentation technology [21], optical flow point matching [22] and other methods to describe the partial information.

B. SEGMENTATION APPROACH IN TRACKING

Segmentation technology assigns labels to pixels according to predefined features and then divides pixels into homogeneous regions. In this case, segmentation provides a basis for the extraction and association of partial information in multi-object tracking.

In tracking, some approaches use partial information of partially occluded objects to solve the occlusion problem, with image segmentation and the background

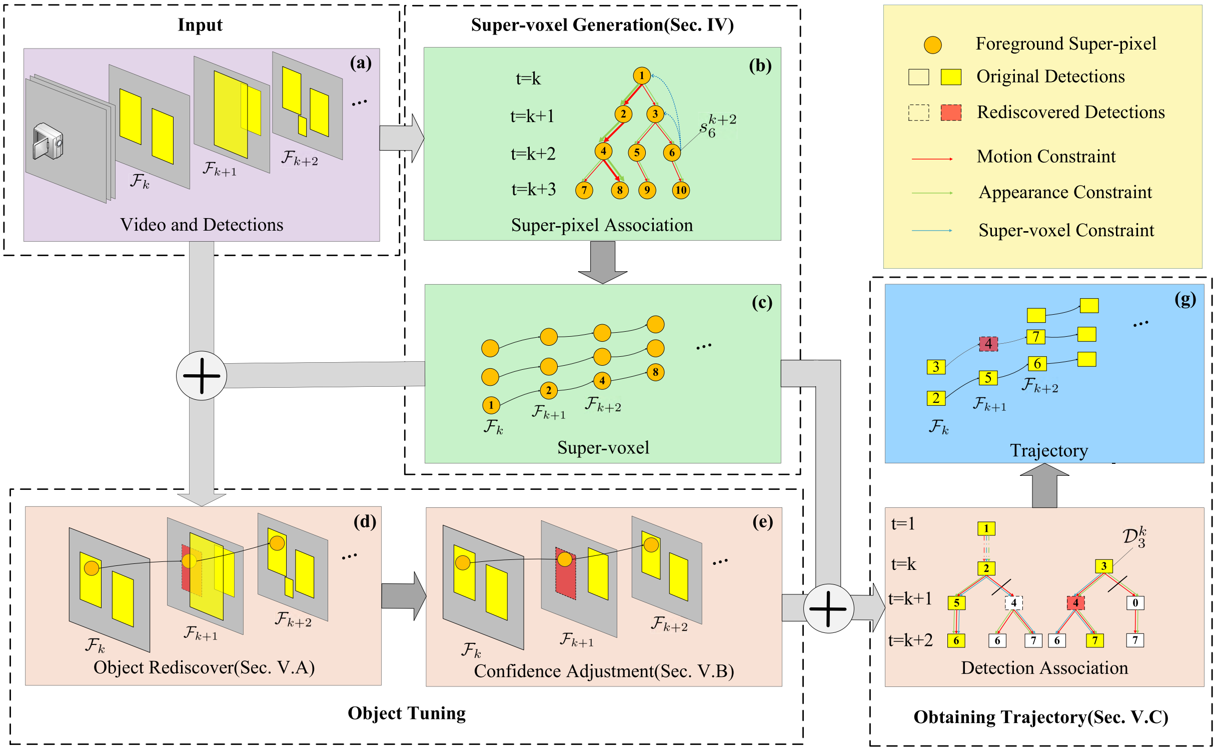


FIGURE 2. Enhanced association with supervoxels(EAS) approach.

model approach. Fragkiadaki and Shi [21] computed a figure-ground segmentation of the video and then assigned repulsive forces between foreground trajectories that belong to various interconnected components in tracking. Wen *et al.* [20] integrated the multi-part tracking and segmentation model [23] into a unified energy optimization framework to handle the object tracking and video segmentation task. In the multi-cut framework, Keuper *et al.* [22] obtained outstanding tracking results based on optical flow point tracking. CRF based on superpixels [9] uses low-level information and associates superpixels with specific objects, giving each superpixel the same label as objects or classifying it as part of the background. However, the above approach is not stable or robust enough in partial occlusion scenes for multiple object tracking, which makes it difficult to obtain accurate tracking results.

In partial object connections between adjacent frames, superpixels have excellent properties. They represent a partial region with similar features of the object, which overcomes the lack of semantics of methods that are based on the point trajectory [21], [22]. Superpixels were first proposed by Ren and Malik [4] and gradually applied to computer vision, pattern recognition and other related fields [9], [24]. Achanta *et al.* [23] proposed a simple linear iterative clustering(SLIC) superpixel generation algorithm based on K-means clustering. When segmentation is applied to the video, it needs to consider the temporal factor. Chang *et al.* [25] presented temporal superpixels (TSP), which use optical flow information to evolve and accurately maintain the labels between adjacent frames. However, the current segmentation algorithms cannot maintain the continuity and stability of the superpixel in long-term video sequences.

Taking the above-mentioned challenges into account, the enhanced association with supervoxels(EAS) approach is proposed with the strong spatial-temporal continuity of supervoxels in tracking. It is robust for solving the problems of partial occlusion and achieves a remarkable result because of the association enhancement of the detections.

III. OVERVIEW

To sufficiently represent partial information, we utilize spatial-temporal continuous sequences of superpixels as a supervoxel. A novel enhanced association with supervoxel (EAS) approach for multiple object tracking is presented to solve the partial occlusion problem. Much more partially occluded objects are effectively joined into trajectories that reduce false negatives in tracking. Multiple hypothesis tracking(MHT) [26] is a classic tracking framework that has been further studied in recent years [8], [27]–[29]. We conduct our approach in multiple hypothesis tracking (MHT) framework, which is a popular tracking-by-detection approach.

We represent our method in Figure 2. As shown in Figure 2(a), our input data are video sequences and the detections are obtained through a public detector. Each detection has a confidence to describe the probability that it belongs to pedestrians. We use the detection provided by the

MOT15 [30] and MOT16 [31] datasets which came from different detectors. This verifies the robustness of our algorithm.

Spatial-temporal continuous sequences of superpixels constitute supervoxel as shown in Figure 1. Each frame is segmented into several superpixels with asegmentation algorithm. Each superpixel is given a score by a fore-background model where we use the SVM (Support Vector Machine) method based on the superpixel's color [9]. Foreground superpixels are used to associate partial regions between adjacent frames and have strong stability. In addition, foreground superpixels that are not located in the detections form the missing superpixel set. The superpixels in the missing set may be potential objects and are used to estimate position based on the relationship with context detection.

We produce a supervoxel(Figure 2(b) and 2(c)) by building a superpixel association tree(Sec. IV-B), with the foreground superpixels as the root node. We utilize the supervoxel association information between the missing foreground superpixel and the original detections(obtained by the detector) to rediscover objects(Figure 4(d)) as described in Sec. V-A. We find that there is a problem with the confidence of the detection set, so we use the relationship among detections to determine height perceptive and perform confidence refinement(Figure 4(e)) as described in Sec. V-B. The association between detections is a very important step. The features usually include appearance and motion information. We introduce the supervoxel association cost(Figure 2(f)) to measure the similarity between the detections.

Finally, as shown in Figure 2(f), the association of detections maintains multiple track trees, and delay data association decisions by keeping multiple hypotheses active until the data association ambiguities are resolved, where every detection is a vertex and associations between detections form edges. At each frame, the track trees are updated from observations by adding to the existing tree and creating a new tree for each observation, where each branch in the tree is scored with three types of linking constraints(including appearance, motion and supervoxel).

Following the formulation [8], we use the log likelihood ratio (LLR) between the target hypothesis and the score can be computed recursively [32]. Supervoxel cost is introduced to the score calculation by pairwise edge association (Sec. V-C). The final trajectories are found by solving a maximum weighted independent set problem(Figure 2(f)) to obtain the more complete tracking result(Figure 2(g)). A large number of occlusion objects are added to the trajectory.

IV. SPATIAL-TEMPORAL SUPERVOXEL

When objects are partially occluded, low-level superpixel evidence [9] sufficiently represents partial information and is relatively stable in spatial and temporal associations. We build a superpixel association tree to obtain candidate supervoxels by associating superpixels with motion and appearance(color) information. The supervoxel has a prominent effect in representing partially occluded objects and associating objects among frames. The generation of the supervoxel is described in this section.

A. DEFINITION OF SUPERVOXEL

In the tracking-by-detection framework, the effective detection set $\mathcal{D} = \{\mathcal{D}_i^t\}$ is the basis of tracking. Each detection $\mathcal{D}_i^t$ has position information, index i, the frame t that is located and a confidence C_i^t to describe the probability that it belongs to pedestrians. Traditional detectors generally use fully visible objects as training data, which results in partially occluded objects being given very low confidence or even being lost. Therefore we propose a supervoxel that has strong spatial-temporal continuity in tracking. It is robust for discovering and associating the partially occluded object.

In this paper, we define the spatial-temporal continuous sequence of superpixels [4] as a supervoxel. The foreground supervoxel (*i.e.*, included in the pedestrian) indicates the tracklet of a partial region of an object. To obtain the supervoxel, we first use the TSP algorithm([25]) to divide the image into superpixels by frame $\mathcal{F}_i = \{s_i^t\}$. A robust fore-background model is designed to find foreground superpixels based on SVM classification of superpixel color features. It is insensitive to camera movement or pedestrians standing still. The fore-background model gives a score $\mathcal{J}_i^t$ for each superpixel and the superpixels with the higher scores tend to be the foreground. Missing foreground superpixel set F_M forms the starting nodes for all the supervoxels.

A superpixel association tree is built and used to obtain candidate supervoxels. For a certain foreground superpixel s_i^t, a tree T is constructed with the associations of the superpixel in different frames. The tree is extended with s_i^t as the root node. As shown in Figure 2(b), when the superpixel s_3^{k+1} is looking for association nodes s_6^{k+2} in the next frame, the information of the ancestor node s_1^k and the parent s_3^{k+1} are considered. It evaluates the quality of the nodes s_6^{k+2} in a comprehensive manner. A candidate supervoxel is a branch $\mathcal{V}_i = \{s_{i_0}^t, s_{i_1}^{t+1}, \ldots, s_{i_m}^{t+m}\}$ in T. The optimal supervoxel is obtained by constructing proper functions to evaluate these candidate supervoxels. The details are discussed in the next section. As shown in Figure 2.(b), sequence $\{s_1^k, s_1^{k+1}, s_4^{k+2}, s_8^{k+3}\}$ is the best supervoxel selected by the evaluation function.

As shown in Figure 1, after the segmentation algorithm [25], the image is segmented into multiple superpixels, where the colored regions indicate the foreground superpixels. Based on the association and evaluation methods mentioned above, five supervoxels are finally formed. The supervoxels effectively connect the partial regions of the pedestrians.

B. SUPER-VOXEL EVOLUTION

As discussed in related work, the superpixels associated with the original label(obtained by the segmentation algorithm) cannot remain stable and accurate in long-interval frames, so we build a superpixel association tree to obtain supervoxels (Figure 2(b) and 2(c)). The superpixel association process can be either forward or backward. The following shows the approach of forward association, but we use all directions in the experiment.

The color and position information are selected to represent the superpixel and supervoxel. These features are used to evaluate the similarity of superpixels when associating adjacent frames. These features are synthetically considered to form an evaluation function $\mathcal{Q}$ for measuring the candidate supervoxels. The evaluation score of the supervoxel reflects the similarity of two objects for the supervoxel's starting and ending nodes located in the two objects. We use ϕ to represent the features between adjacent frames and use ψ to represent the features of the supervoxel. For a superpixel association tree T, a foreground superpixel s_i^t is used as a root node. When the tree is extended in the temporal dimension, it needs to select the appropriate superpixels as the association node in the next frame. For a pending association node s_i^t, multiple similar nodes s_j^{t+1} are added to the tree.

We measure color differences in the Lab color space because of its perceptual accuracy [33]. We define the mean of pixel color in the superpixel s_i^t as superpixel color $Lab(s_i^t)$. In adjacent frames, the cost of color ϕ_{Lab} is defined as:

$$\phi_{Lab}(s_i^t, s_j^{t+1}) = \|Lab(s_i^t) - Lab(s_j^{t+1})\|_2. \tag{1}$$

It describes the color difference between s_i^t and s_j^{t+1}.

We define the coordinate mean of pixels in the superpixel as the position $(x_i^t, y_i^t)^T$ of s_i^t. In adjacent frames, the cost of distance ϕ_{Dis} is defined as:

$$\phi_{Dis}(s_i^t, s_j^{t+1}) = \|(x_i^t, y_i^t)^T - (x_j^{t+1}, y_j^{t+1})^T\|_2. \tag{2}$$

This formula describes the distance between s_i^t and s_j^{t+1} in space.

There are some necessary constraints when nodes are associated with the current superpixel node. The superpixel node should be in a neighboring location with similar color. We formulate an extended function $\mathcal{E}(s_i^t, s_j^{t+1})$ to evaluate the similarity of superpixels in the adjacent frame:

$$\mathcal{E}(s_i^t, s_j^{t+1}) = \sum_{k=-2}^{0} \phi_{Lab}(s_i^{t+k}, s_j^{t+1}) + \phi_{Dis}(s_i^{t+k}, s_j^{t+1}), \tag{3}$$

where k indicates that we consider the parent node and the ancestor node in the association. A detailed description of each factor is introduced as above. Considering that the position and appearance of the pedestrian does not suddenly change between adjacent frames, we select two superpixel as the association nodes in the adjacent frame.

For a candidate supervoxel, it is a continuous series of superpixels and needs integrity features to describe it. Acceleration information is a good choice for describing the stability. The cost of color is defined as:

$$\psi_{Lab\_A}(\mathcal{V}_i) = \sum_{k=1}^{\tau-1} \|L\_A_{i_k}^{t+k}\|_2, \tag{4}$$

which indicates the sum of the color acceleration. The acceleration is represented by the second order color difference of the superpixel:

$$L\_A_{i_k}^{t+k} = Lab(s_{i_{k-1}}^{t+k-1}) - 2Lab(s_{i_k}^{t+k}) + Lab(s_{i_{k+1}}^{t+k+1}) \tag{5}$$

Temporally, distance information is expressed as acceleration of the candidate supervoxel. The pixel coordinates of pedestrians in the video are altered as they move. The velocity of the pedestrians remains stable, which means that the acceleration of the superpixel that belongs to an object is fairly small. Therefore, we define the acceleration cost for candidate supervoxels. The cost of acceleration for the supervoxel is defined as:

$$\psi_{Dis\_A}(\mathcal{V}_i) = \sum_{k=1}^{\tau-1} \|D\_A_{i_k}^{t+k}\|_2 \tag{6}$$

The acceleration is represented by the second order difference of the superpixel position:

$$\begin{aligned} D\_A_{i_k}^{t+k} &= (x_{k-1}^{t+k-1}, y_{k-1}^{t+k-1})^T \\ &\quad - 2(x_k^{t+k}, y_k^{t+k})^T + (x_{k+1}^{t+k+1}, y_{k+1}^{t+k+1})^T. \end{aligned} \tag{7}$$

In association tree T, each branch is a candidate supervoxel, and the branches form a set $\mathcal{V} = \{\mathcal{V}_i\}$. The reliable supervoxel usually has a very small rate of change in color and speed, *i.e.*, a low acceleration. We formulate an evaluation function $\mathcal{Q}(\mathcal{V}_i)$ to obtain the supervoxel score of $\mathcal{V}_i$:

$$\mathcal{Q}(\mathcal{V}_i) = \psi_{Lab\_A}(\mathcal{V}_i) + \psi_{Dis\_A}(\mathcal{V}_i). \tag{8}$$

where the detailed description of each factor is introduced as above. Through the evaluation function, we extract the all candidate supervoxel score from the superpixel tree.

V. ENHANCED ASSOCIATION WITH SUPERVOXELS FOR MHT

This section introduces the supervoxel that handles the partial occlusion to solve the tracking problem based on the MHT framework. We use the supervoxel to represent the partially occluded object(Sec V-A) and enhance the pairwise edge associations in tracking(Sec V-C), which results in a remarkably effective approach in the partial occlusion scenes.

A. OBJECT REDISCOVERY WITH SUPERVOXEL

Tracking-by-detection is a popular framework and we use the framework in this paper. Thus, the effective detection set has a significant impact on complete object tracking. As explained in Sec. II, the traditional detector is not very good at detecting partially occluded objects. We find the partially occluded objects based on the fore-background model and estimate the object's position through the association between the supervoxel and the original detections. We describe the detailed process of rediscovering objects as follows.

Each superpixel s_i^t in the missing superpixel set F_M may be a potential object. We use this superpixel as the source node for forward and backward association. Then, the superpixel is

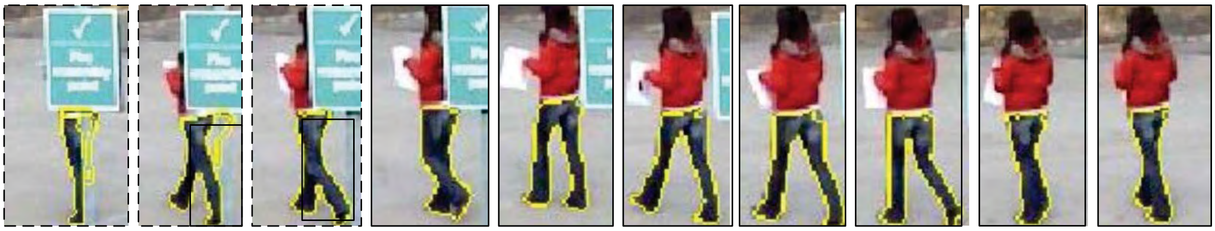
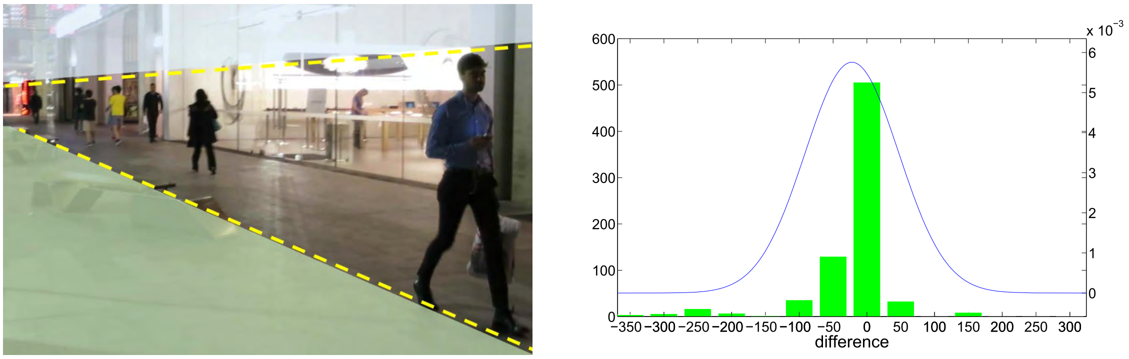


FIGURE 3. These figures are 10 detections in continuous frames, where solid rectangular boxes are the detections produced by the detector and dashed rectangular boxes are the rediscovered detections produced by our approach. The regions surrounded by yellow curves form a supervoxel.



(a)

(b)

FIGURE 4. (a)This is a perspective in the video scene(MOT16-10); the following shaded area represents the ground plane; the upper shadow area represents the head of the pedestrian and the area above pedestrian. We can see that the pedestrians eventually disappear in the perspective vanishing point. (b)Histogram of distance distribution between detections and fitted plane with top 25% confidence of detection set in the video sequence MOT16-10.

used to form two superpixel correlation trees. We select two supervoxels with the minimum score corresponding to the forward and backward association. The object is discovered again, with the association between $\mathcal{V}_{min}$(a branch with a minimum score by Eq .8) and the contextual original detection set $\{\mathcal{D}_i\}$. As previously mentioned, the following approach is forward association.

Supervoxel $\mathcal{V}_{min}$ includes a sequence of superpixels $\{s_{i_0}^t, s_{i_1}^{t+1}, \cdots, s_{i_\tau}^{t+\tau}\}$, some of which belong to detections. The superpixel that is not located in the detection is a potential object. If $\mathcal{D}_i$ and s_j belong to the same frame, and $Pos(s_j) \in \mathcal{D}_i$(*i.e.*, coordinate mean of the superpixel s_j located in detection $\mathcal{D}_i$), we define $\mathcal{D}_i$ and s_j as matched. Therefore, we obtain a sequence of matching detections $\{\mathcal{D}_{i_{k_1}}^{t+k_1}, \mathcal{D}_{i_{k_2}}^{t+k_2}, \cdots, \mathcal{D}_{i_{k_l}}^{t+k_l}\}_{l \le \tau}$.

The detection sequence $\{\mathcal{D}_{i_{k_1}}^{t+k_1}, \mathcal{D}_{i_{k_2}}^{t+k_2}, \cdots, \mathcal{D}_{i_{k_l}}^{t+k_l}\}$ is discontinuous, so it is used to discover the object in frame $\{t, t+1, \cdots, t+\tau\} \backslash \{t+k_1, t+k_2, \cdots, t+k_l\}$. These missed pedestrian positions are fully filled by polynomial interpolation based on contextual detection. To avoid redundancy, *i.e.*, two observations occupying the same position, we use the NMS (non-maximal suppression) algorithm for rediscovered detection. Figure3 shows that the supervoxel surrounded by the yellow curve connect the partial and complete visually visible pedestrians. The dashed rectangular boxes are the discovered pedestrians produced by our approach.

As shown in Figure 2(f), there are now three kinds of nodes in the tracking tree: the original nodes(obtained by the detector), the rediscovered nodes and the dummy nodes (labeled 0). The dummy nodes are obtained by Kalman filter which represents the completely occluded objects. Both the original and rediscovered nodes are used in tracking.

B. OBJECT CONFIDENCE ADJUSTMENT

By fusing the supervoxel and the original detection, the object rediscovery approach joins the partially occluded objects in the tracking. However, we find that there is a problem with the confidence of the detection set, and some abnormal detections are still given high confidence. We use the relationship among detections to determine height perceptive and perform confidence refinement without using any camera parameters.

According to the perspective principle, the objects in a relatively stable camera have a similar height at the same ground position(Figure 4(a)). In other words, every object has a perspective height in a 2-D image. We use the detections that have the top 25% confidence in the detection set and set them to fit a 3-D plane. Figure 4(b) shows that the difference between the original detection height and the fitted plane falls along a normal distribution with 0 as the approximate mean.

Without loss of generality, we assume that the height of a human obeys normal distribution $\mathcal{N}(h, \sigma^2)$, where $h = 1.7m$ and $\sigma = 0.3$. Hence, for a perspective height $\tilde{h}$, it obeys a normal distribution $\mathcal{N}(\tilde{h}, \tilde{\sigma}^2)$, where $\tilde{\sigma} = \sigma * \tilde{h}/h$. The perspective height is used to refine the confidence of detection. When the height of the detection is abnormal, it is given a low score. The refined function is defined as follows:

$$C_i^{new} = \sqrt{2\pi}\tilde{\sigma} * p(\tilde{h}_i) * C_i, \tag{9}$$

where $p(\tilde{h}_i)$ is the probability density function of normal distribution $\mathcal{N}(\tilde{h}, \tilde{\sigma}^2)$, C_i is the confidence of detection $\mathcal{D}_i$, $\tilde{h}_i$ is the perspective height of detection $\mathcal{D}_i$ and C_i^{new} is the refined confidence.

For a moving camera, the ground in the scene is variable, but the effect is slight for the perspective model. At the same time, to minimize the impact of scene changes, we select 50 frames as a sliding window, *i.e.*, every 50 frames we update a perspective model.

C. PAIRWISE EDGES ASSOCIATION ENHANCEMENT

The original MHT framework uses motion and appearance features to associate detection but the similarity between occluded and completely visible objects is usually low. In Sec. IV-B, evaluation score $\mathcal{Q}(\mathcal{V})$ of supervoxel $\mathcal{V}$ reflects the similarity of two objects and can maintain a strong correlation for the occluded object.

Figure 2(f) shows two detection nodes that satisfy motion constraints(associated by red edges) and supervoxel constraints(associated by blue edges). While the appearance constraints(associated by yellow edges) cannot be satisfied, we still add it to the multiple hypothetical tracking.

For a hypothesis trajectory $\{\mathcal{D}_{i_1}^{t+1}, \mathcal{D}_{i_2}^{t+2}, \cdots, \mathcal{D}_{i_k}^{t+k}\}$, we remove the dummy nodes(obtained by Kalman filter) to obtain the sequence $\{\mathcal{D}_{i_{m_1}}^{t+m_1}, \mathcal{D}_{i_{m_2}}^{t+m_2}, \cdots, \mathcal{D}_{i_{m_l}}^{t+m_l}\}$. The supervoxel cost of the two adjacent detections $(\mathcal{D}_{i_{m_j}}^{t+m_j}, \mathcal{D}_{i_{m_{j+1}}}^{t+m_{j+1}})$ is defined as:

$$SVCost(\mathcal{D}_{i_{m_j}}^{t+m_j}, \mathcal{D}_{i_{m_{j+1}}}^{t+m_{j+1}}) = min\{\mathcal{Q}(\mathcal{V}_i)\} \tag{10}$$

where all $\mathcal{V}_i$ are the supervoxels that go through $(\mathcal{D}_{i_{m_j}}^{t+m_j}, \mathcal{D}_{i_{m_{j+1}}}^{t+m_{j+1}})$, *i.e.*, supervoxel's start and end nodes are located in the two detections.

When we evaluate the quality of multiple hypothetical trajectories, we do not only consider appearance and motion constraints because it is possible to remove appropriate trajectories due to the appearance difference caused by short-term object occlusion or turn-back. We also used the supervoxel cost to evaluate a hypothetical trajectory. For a trajectory that does not satisfy the cost of the appearance, we still retain this hypothetical trajectory if it satisfies the supervoxel cost.

The formula of the supervoxel association cost for a hypothesis of tracking tree is calculated as follows:

$$S_c = \sum_{j=1}^{l-1} SVCost(\mathcal{D}_{i_{m_j}}^{t+m_j}, \mathcal{D}_{i_{m_{j+1}}}^{t+m_{j+1}}). \tag{11}$$

The supervoxel association cost scores reflect the correlation similarity of the hypothetical trajectory.

VI. EXPERIMENTS

We performed several experiments to evaluate our approach. The quantitative results showed the superiority in partial occlusion scenes.

Datasets: We used MOT benchmark in our experiments including MOT15 [30] and MOT16 [31]. The datasets were a combination of multiple sets that included sequences from both the PETS [34] and KITTI [35] datasets. There were 22 challenging video sequences in MOT15 (11 training, 11 test) and 14 sequences in MOT16 (7 training, 7test). The sequences covered several different types of tracking problems, including directions, variable speed, different density of pedestrians and long-term occlusions. The detections are provided by the MOT benchmark.

Parameters: For extending function $\mathcal{E}(s_i^t, s_j^{t+1})$ and the evaluation function $\mathcal{Q}(\mathcal{V}_i)$, each variable is regularized, which solves the problem that the data scale was not uniform. Thus, our approach makes it easier to add more features in the future.

Metrics: We analyzed the integrity of the nodes that constructed the multiple hypothesis tree using the ground truth of tracking as the benchmark. We used comprehensive evaluation metrics recall (Rcll↑), precision (Prcn↑) and F-Score (F-Sc↑, the harmonic mean of precision and recall), where $F\text{-}Sc = 2 * Rcll * Prcn/(Rcll + Prcn)$. We followed the current popular CLEAR MOT [36] metrics to evaluate the tracking performance. The metrics included the multiple object tracking accuracy (MOTA↑) and multiple object tracking precision (MOTP↑). MOTA reflected the tracking accuracy which combined false positives(FP↓), false negatives(FN↓) and identity switches (IDS↓) of the predicted trajectories. MOTP showed the tracking precision, which measured the localization difference between the output trajectories and the ground truth trajectories. The number of mostly tracked objects (MT↑, $>$ 80% overlap) and mostly lost objects (ML↓, $<$ 20% overlap) reflected a temporal coverage of output trajectories. We also used the current popular IDF1 [37](↑), which identified detections over the average number of ground-truth and computed detections. The ↑ indicated the higher the better, while the ↓ indicated the lower the better.

A. EFFECTIVENESS ANALYSIS IN TRACKING

We used our approach on the MOT16 Benchmark and analyzed some intermediate results. Table. 1 shows the detailed quantitative tracking results for each video sequence, in which the **All(ours)** depicted the result of all seven test video sequences.

Compared to the MHT_basic(obtained by running published source code), more accurate detections were added to the trajectory (10988 true positive increase), ensuring that the false positive remained low(only increased by 2701). Rcll increased by 6.0% overall in the seven video sequences and more trajectories were tracked completely. The number of MT rose from 103 to 131 and the ML dropped from 356 to 324. The comprehensive evaluation metric MOTA increased by 4.5%.

TABLE 1. Tracking results of MOT16 benchmark.

Sequence	Rcll	GT	MT	ML	FP	FN	IDs	MOTA	MOTP
MOT16-01	40.1	23	6	10	138	3833	14	37.7	72.7
MOT16-03	59.9	148	39	26	5048	41950	280	54.8	75.9
MOT16-06	54.5	221	46	111	522	5245	50	49.6	74.6
MOT16-07	47.4	54	5	16	569	8580	70	43.5	74.4
MOT16-08	38.2	63	10	21	952	10344	74	32.1	79.7
MOT16-12	47.3	86	15	44	603	4373	17	39.8	78.1
MOT16-14	31.8	164	10	96	537	12606	70	28.5	75.3
All(ours)	52.3	759	131	324	8369	86931	575	47.4	75.9

TABLE 2. Detection results of the MOT15_train and MOT16_train. The bold font indicates better performance on each metric.

Dataset	Approach	Rcll	Prcn	TP	FP	FN	F-Sc
MOT15	[30]	0.52	0.60	22508	14820	20630	0.56
	EAS	0.50	**0.72**	21470	**8513**	21668	**0.59**
MOT16	[31]	**0.26**	0.64	51006	28784	148961	0.37
	EAS	0.25	**0.81**	49611	**11998**	150356	**0.38**

TABLE 3. Association result change in MOT15 and MOT16. 'without' is the original linking result([8]). 'with supervoxel' is the result using our supervoxel association. The count units are in million.

Dataset	without	with supervoxel	promotion
MOT15	2.23	2.43	8.97%
MOT16	5.63	5.93	5.33%

TABLE 4. Tracking results comparison of MOT16 benchmark.

approach	Rcll	MT	ML	FP	FN	MOTA	IDF1
JMC [38]	50.1	118	**301**	6373	90914	46.3	46.3
JPDA_m [39]	28.4	31	512	**3689**	130549	26.2	0.0
NOMT [40]	51.9	**139**	314	9753	87565	46.4	**53.3**
NLLMPa [41]	51.1	129	307	5844	89093	47.6	47.3
GCRA [42]	51.3	98	313	5104	88586	**48.2**	48.6
MHT_DAM [8]	49.6	123	328	6412	91758	45.8	46.1
MHT_basic [8]	46.3	103	356	5668	97919	42.9	\
EAS(ours)	**52.3**	131	324	8369	**86931**	47.4	50.1

1) INTEGRITY ANALYSIS OF TRACKING NODES

We predicted the position of partially occluded pedestrians through a detection rediscovery algorithm(Sec. V-A), which led to an increase in the number of detections. Using perspective theory(Sec. V-B), we obtained a lower confidence score for abnormal detection through our confidence adjustment algorithm, which made the quality evaluation of the detection more accurate.

In Table. 2, compared to the original detection(obtained by [30] and [31]) with no further processing, our **EAS** approach(Sec. V-A and V-B) obtained better results in the integrity of tracking nodes. While increasing the number of detections, we deleted some detections with low confidence scores in analyzing integrity because our evaluation criteria for detection were more accurate. The detections with low scores tended to be incorrect. We significantly improved the *Prcn* and *F-Sc* by using a detection rediscovery approach while still maintaining *Rcll*. We accurately estimated the position of a large number of partial occlusion detections. We also decreased the *FP* which means that more incorrect detections were deleted using our approach, along with a slight decrease in *TP*.

2) DISTRIBUTION OF ASSOCIATED EDGES

In the association process, we introduced the supervoxel association approach(Sec. V-C). It linked more much partially occluded detection and avoided incorrect elimination of brunch.

As we discussed in Sec. V-C, we used three kinds of edges(motion, appearance, and supervoxel) to associate the object. In tracking, motion(the movement between adjacent frames was limited and had certain regularity) was a necessary constraint, and we associated two objects if appearance or supervoxel was satisfied. In Table. 3 we counted the usage distributions of different edges during constructing the multiple hypothesis tree.

Our supervoxel association approach(Sec. V-C) had better robustness for partially occluded pedestrians. It allowed more partially occluded pedestrians to join the trajectory. As shown in Table. 3, compared to the originally associated edges(linking approach in [8], *i.e.* satisfying space and appearance constraint), our associated edges approach (*i.e.*, considering supervoxel association) improved the effective association of 5% to 8%.

B. COMPARISON WITH STATE-OF-THE-ART

As shown in Table. 4, our approach was compared in the MOT16 Benchmark with JMC [38], JPDA_m [39], NOMT [40], NLLMPa [41], GCRA [42] and MHT_DAM [8]. The last row of Table. 4 presented the tracking results of our **EAS** approach. The tracking result of our approach was significantly higher than the baseline algorithm [8] approach(MHT_basic in Table. 4) on many metrics. Compared to the current popular algorithms, EAS achieved approximate state-of-the-art results. Our approach found and associated more partial occlusion objects. This has greatly ameliorated on Rcll, MT and FN. In terms of various metrics, our algorithms have achieved competitive results.

Figure 5 shows the visual tracking results of the video sequence MOT16-03 by our approach. The pedestrian(with a yellow box) was continuously tracked. However, it failed in the MHT([8]) approach, due to the occlusion and inaccuracy association. In addition, some pedestrians(*e.g.*, with blue boxes in Figure 5) lacked the corresponding detection in the original detection set. It was earlier added to the trajectory using our approach. Many false detections(*e.g.*, with red boxes in Figure 5) did not join trajectories using our approach, through giving appropriate confidence.

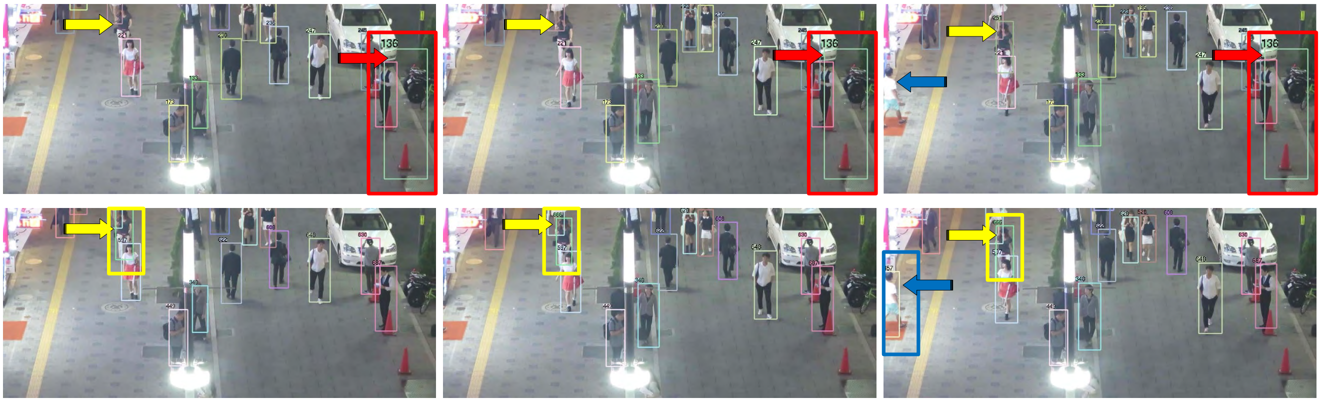


FIGURE 5. The result of tracking on the sequences MOT16-03. The *top* row is the benchmark(MHT) method, and the *bottom* row is the result obtained by our method. The red box indicates the error trajectory under the original confidence description. The yellow box indicates the complete pedestrian trajectory including partial occlusion. The blue box indicates the pedestrian that was earlier added to the trajectory.

TABLE 5. Tracking results comparison to MOT15 benchmark.

approach	Rcll	MT	ML	FP	FN	MOTA	IDF1
ELP [7]	39.2	54	316	7345	37344	25.0	26.2
MHT_DAM [8]	47.8	115	316	9064	32060	32.4	45.3
SegTrack [9]	36.5	42	461	7890	39020	22.5	31.5
JPDA_m [39]	34.8	36	419	6373	40084	23.8	33.8
NOMT [40]	47.0	88	317	7762	32547	33.7	44.6
AMIR15 [43]	39.2	114	**193**	7933	**29397**	**37.6**	46.0
RAR15pub [44]	39.2	93	305	6771	32717	35.1	45.4
AM [45]	43.3	82	313	**5154**	34848	34.3	**48.3**
EAS(ours)	**49.4**	**119**	328	8696	31089	34.6	47.8

To prove effectiveness in different datasets, we also evaluated our approach in the MOT15 Benchmark as shown in Table. 5. Our approach was compared with ELP [7], SegTrack [9], JPDA_m [39], NOMT [40], AMIR15 [43], RAR15pub [44], AM [45] and MHT_DAM [8]. Our method still achieved excellent results in many metrics and obtained approximate state-of-the-art results. The video sequence scene type and detector in MOT15 were different from MOT16. This proved that our **EAS** approach was effective for different detectors.

We consistently outperformed previous algorithms in MT and MOTA, due to the increase in Rcll and the decrease in FN. Our approach identified more image evidence from occluded and neighboring contextual objects, which allowed the tracking model to accurately obtain more trajectories.

VII. CONCLUSION

This paper proposes a novel enhanced association with supervoxels(EAS) method for multiple object tracking in complex scenes where partial occlusion frequently occurs. For object association, a supervoxel with superior ability for resisting occlusion is introduced, which makes full use of contextual information. EAS rediscovers detection for partially occluded objects by utilizing the relationship between existing detections and supervoxels. EAS estimates the perspective height of pedestrians in the video and refines the confidence of detections. The pairwise costs are proposed based on a novel energy function that uses supervoxel association in tracking. We show that our method has the ability to generate longer and more complete trajectories when partial occlusion occurs. Experimental results show the obvious advantages of our approach on tracking partially occluded objects in a set of standard public video sequences.

ACKNOWLEDGMENT

The authors would like to thank the support from HAWKEYE Group.

REFERENCES

[1] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. D. Reid, and S. Roth. (Apr. 2017). "Tracking the trackers: An analysis of the state of the art in multiple object tracking." [Online]. Available: https://arxiv.org/abs/1704.02781

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[3] R. Girshick, "Fast R-CNN," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[4] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Oct. 2003, pp. 10–17.

[5] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Dec. 2016, pp. 4705–4713.

[6] I. Leang, S. Herbin, B. Girard, and J. Droulez, "On-line fusion of trackers for single-object tracking," *Pattern Recognit.*, vol. 74, pp. 459–473, Feb. 2018.

[7] N. Mclaughlin, J. M. D. Rincon, and P. Miller, "Enhancing linear programming with motion modeling for multi-target tracking," in *Proc. Appl. Comput. Vis.*, 2015, pp. 71–77.

[8] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4696–4704.

[9] A. Milan, L. Lealtaixé, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5397–5406.

[10] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Robust tracking-by-detection using a detector confidence particle filter," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1515–1522.

[11] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1833, Sep. 2011.

[12] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1345–1352.

[13] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2411–2418.

[14] A. A. Butt and R. T. Collins, "Multi-target tracking by lagrangian relaxation to min-cost network flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1846–1853.

[15] A. Milan, K. Schindler, and S. Roth, "Detection-and trajectory-level exclusion in multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3682–3689.

[16] J. Chen, H. Sheng, C. Li, and Z. Xiong, "PSTG-based multi-label optimization for multi-target tracking," *Comput. Vis. Image Understand.*, vol. 144, pp. 217–227, Mar. 2016.

[17] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proc. ICCV*, Oct. 2017, pp. 300–311.

[18] P. Li, D. Wang, L. Wang, and H. Lu, "Deep visual tracking: Review and experimental comparison," *Pattern Recognit.*, vol. 76, pp. 323–338, Apr. 2018.

[19] A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *Proc. Conf. Artif. Intell.*, 2017, pp. 4225–4232.

[20] L. Wen, D. Du, Z. Lei, S. Z. Li, and M.-H. Yang, "JOTS: Joint online tracking and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2226–2234.

[21] K. Fragkiadaki and J. Shi, "Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2073–2080.

[22] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person reidentification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3539–3548.

[23] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[24] A. Vazquezreina, S. Avidan, H. Pfister, and E. Miller, "Multiple hypothesis video segmentation from superpixel flows," in *Computer Vision–ECCV* (Lecture Notes in Computer Science), vol. 6315, pp. 268–281, 2010.

[25] J. Chang, D. Wei, and J. W. Fisher, "A video representation using temporal superpixels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2051–2058.

[26] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. Autom. Control*, vol. AC-24, no. 6, pp. 843–854, Dec. 1979.

[27] H. Wang, J. Sun, S. Lu, and S. Wei, "Factor graph aided multiple hypothesis tracking," *Sci. China Inf. Sci.*, vol. 56, no. 10, pp. 1–6, 2013.

[28] J. Fu, J. Sun, S. Lu, and Y. Zhang, "Multiple hypothesis tracking based on the shiryayev sequential probability ratio test," *Sci. China Inf. Sci.*, vol. 59, no. 12, pp. 122306-1–122306-11, 2016.

[29] S. Coraluppi and C. Carthel, "Multiple-hypothesis tracking for targets producing multiple measurements," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 3, pp. 1485–1498, Jun. 2018.

[30] L. Leal-Taixé, A. Milan, I. D. Reid, S. Roth, and K. Schindler. (Apr. 2015). "Motchallenge 2015: Towards a benchmark for multi-target tracking." [Online]. Available: https://arxiv.org/abs/1504.01942

[31] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler. (May 2016). "MOT16: A benchmark for multi-object tracking." [Online]. Available: https://arxiv.org/abs/1603.00831

[32] S. S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Norwood, MA, USA: Artech House, 1999.

[33] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[34] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in *Proc. 12th IEEE Int. Workshop Perform. Eval. Tracking Surveill. (PETS-Winter)*, Dec. 2009, pp. 1–6.

[35] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[36] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *J. Image Video Process.*, vol. 2008, p. 1, 2008.

[37] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 17–35.

[38] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person tracking by multicut and deep matching," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 100–111.

[39] S. H. Rezatofighi, A. Milan, Z. Zhang, and Q. Shi, "Joint probabilistic data association revisited," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3047–3055.

[40] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3029–3037.

[41] E. Levinkov *et al.*, "Joint graph decomposition & node labeling: Problem, algorithms, applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1904–1912.

[42] C. Ma *et al.*, "Trajectory factory: Tracklet cleaving and re-connection by deep Siamese Bi-GRU for multiple object tracking," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2018, pp. 1–6.

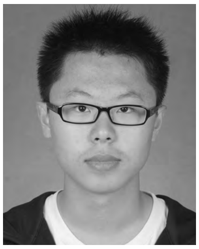
[43] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proc. IEEE Conf. Int. Conf. Comput. Vis.*, Oct. 2017, pp. 300–311.

[44] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 466–475.

[45] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4846–4855.

HAO SHENG received the B.S. and Ph.D. degrees from the School of Computer Science and Engineering, Beihang University, China, in 2003 and 2009, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, pattern recognition, and machine learning.

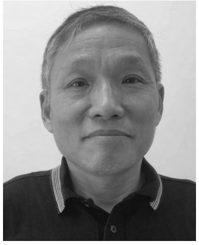
XINYU ZHANG received the B.Sc. and B.S. degrees from the Hebei University of Technology, China, in 2016. He is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Beihang University, China. His research interests include computer vision and multi-object tracking.



YANG ZHANG received the B.S. degree from the School of Advanced Engineering, Beihang University, China, in 2014, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, and he is particularly interested in multiple object tracking.

YUBIN WU received the B.S. degree in computer science from Beihang University, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree. His current research interests include multiple object tracking and 3D reconstruction.

JIAHUI CHEN received the B.S. degree from the School of Advanced Engineering, Beihang University, China, in 2012, where he is currently pursuing the Ph.D. degree. His research interests include computer vision, and he is particularly interested in multiple object tracking.



ZHANG XIONG received the B.S. degree from Harbin Engineering University in 1982 and the M.S. degree from Beihang University, Beijing, China, in 1985. He is currently a Professor with the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, information security, and data vitalization.