



Structural Constraint Data Association for Online Multi-object Tracking

Ju Hong Yoon¹ · Chang-Ryeol Lee² · Ming-Hsuan Yang³ · Kuk-Jin Yoon⁴

Received: 20 October 2016 / Accepted: 4 April 2018 / Published online: 27 April 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Online two-dimensional (2D) multi-object tracking (MOT) is a challenging task when the objects of interest have similar appearances. In that case, the motion of objects is another helpful cue for tracking and discriminating multiple objects. However, when using a single moving camera for online 2D MOT, observable motion cues are contaminated by global camera movements and, thus, are not always predictable. To deal with unexpected camera motion, we propose a new data association method that effectively exploits structural constraints in the presence of large camera motion. In addition, to reduce incorrect associations with mis-detections and false positives, we develop a novel event aggregation method to integrate assignment costs computed by structural constraints. We also utilize structural constraints to track missing objects when they are re-detected again. By doing this, identities of the missing objects can be retained continuously. Experimental results validated the effectiveness of the proposed data association algorithm under unexpected camera motions. In addition, tracking results on a large number of benchmark datasets demonstrated that the proposed MOT algorithm performs robustly and favorably against various online methods in terms of several quantitative metrics, and that its performance is comparable to offline methods.

Keywords Multi-object tracking · Data association · Structural constraint

Communicated by Robert T. Collins.

✉ Kuk-Jin Yoon
kjyoon@kaist.ac.kr

Ju Hong Yoon
jhyoon@keti.re.kr

Chang-Ryeol Lee
crlee@gist.ac.kr

Ming-Hsuan Yang
mhyang@ucmerced.edu

¹ Korea Electronics Technology Institute, Seongnam, Gyeonggi-do, Republic of Korea

² Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

³ Electrical Engineering and Computer Science, University of California at Merced, Merced, CA 95344, USA

⁴ Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

1 Introduction

Multi-object tracking (MOT) aims to estimate object trajectories according to the identities in image sequences. Recently, thanks to the advances in object detectors (Wang et al. 2015; Dollar et al. 2014), numerous tracking-by-detection approaches have been developed for MOT. In this type of approach, target objects are detected first and tracking algorithms estimate their trajectories using the detection results. Tracking-by-detection methods can be broadly categorized into online and offline (batch or semi-batch) tracking methods. Offline MOT methods generally utilize detection results from past and future frames. Tracklets are first generated by linking individual detections in a number of frames and then iteratively associated to construct long trajectories of objects in the entire sequence or in a time-sliding window with a temporal delay (e.g., Xing et al. 2009; Pirsiavash et al. 2011). On the other hand, online MOT algorithms estimate object trajectories using only the detections from the current as well as past frames (e.g., Breitenstein et al. 2011), and they are more applicable to real-time applications such as advanced driving assistant systems and robot navigation.

In MOT, object appearances are used as important cues for data association, which solves the assignment problems of detections to detections, detections to tracklets, and tracklets to tracklets. However, appearance cues alone are not sufficient to discriminate multiple objects, especially for tracking objects with similar appearances (e.g., pedestrians, faces, and vehicles). For that reason, tracking-by-detection methods typically exploit motion as well as appearance cues, and use certain (e.g., linear or turn) models to describe the object movements. However, for online two-dimensional (2D) MOT in scenes acquired from moving cameras, observable motion cues are complicated by global camera movements and are not always smooth or predictable. Even when the individual object motion model is updated with consecutive detections, it is not reliable enough to predict the next location of an object when the camera moves severely. The situation becomes worse when objects are not correctly detected since, without correct detections, object motion models cannot be properly updated to take camera motion into account. Furthermore, self-motion information alone is not discriminative enough to disambiguate between objects and uncertain detections. To handle the aforementioned problems, we propose to utilize the structural constraint information between objects, which is represented by the relative positions and velocity differences between objects. This constraint information is robust under the unexpected global camera motion, and provides more discriminative motion cues to reduce mis-matches. While significant advances in batch (or semi-online) trackers have been made (e.g., Yang and Nevatia 2014; Milan et al. 2014; Kim et al. 2015; Choi 2015), online MOT using structural constraints from detection results has not yet been explored much.

In this paper, we propose a new data association method for effectively exploiting the structural constraints between objects for *online 2D* MOT, which considers unexpected global camera motions as well as ambiguities caused by the uncertain detections (e.g., false positives and negatives). In this work, we consider the unexpected global camera motions caused by translational motion, pitch motion, or yaw motion of a camera, which are very common when we record video sequences by using a camera equipped on a moving platform. Using the structural constraints, we introduce a new cost function, which takes global camera motion into account to associate multiple objects. In addition, to reduce the assignment ambiguities caused by false negatives and positives, as shown in Fig. 1, we propose an event aggregation method, which fuses data association costs along the assignment event.

The proposed MOT framework consists of two data association steps. In the first step, by using the proposed structural constraint event aggregation method, we robustly estimate continuously tracked objects where structural constraints are sufficiently reliable because of the consecutive updates at

each frame even under large global camera motions or fluctuations. In the second step, we infer and recover the missing objects between frames to alleviate the problems of mis-detection from detectors. Using the structural constraints of objects between frames, we can re-track the missing ones from the tracked objects in the first step.

Some preliminary results of this work are presented in Yoon et al. (2016). In this paper, we describe and analyze the proposed structural constraint event aggregation algorithm in depth. We reorganize the main body of the paper with intensive modifications to describe each of MOT modules in detail. In addition, we propose the data association solution, which estimates the assignments between objects and detections in a more exhaustive manner. To validate the effectiveness of the proposed data association algorithm, we present additional quantitative and qualitative evaluations.

2 Related Work

We introduce representative MOT algorithms that focus on motion models, which can be categorized based on the types of used motion models: independent motion or structural motion as considered in this work. In addition, we also review the closely related data association algorithms that are used for the MOT problems.

Numerous MOT methods directly utilize the first- or the second-order independent motion models to locate objects (Kim et al. 2012; Bae and Yoon 2014). Takala et al. (2007) propose to measure the directional smoothness and speed of each object based on the current location and the past trajectory to track multiple objects. In Yang and Nevatia (2012), the nonlinear motion patterns of each object and the entry/exit maps are trained by exploiting past and future object trajectories. In Breitenstein et al. (2011), object velocities are used to construct confidence maps of future trajectories for tracking. However, those 2D independent motion models do not work properly under unpredictable camera motions, especially when the tracking methods do not exploit the visual information from future frames.

We review related MOT methods that utilize the structural motion constraints. Pellegrini et al. (2009) and Leal-Taixé et al. (2011) use social force models that consider pairwise motion (such as attraction and repulsion) and visual odometry to obtain 3D motion information for tracking multiple objects. Different from the proposed online 2D MOT algorithm, this method requires 3D information to project objects and detections on the top-view plane for association. In addition, this method does not consider scenes with large camera motion. Grabner et al. (2010) exploit the relative distance between feature points for single object tracking and reduce tracking drifts caused by drastic appearance changes. In Duan et al. (2012), a mutual relation model is proposed to reduce

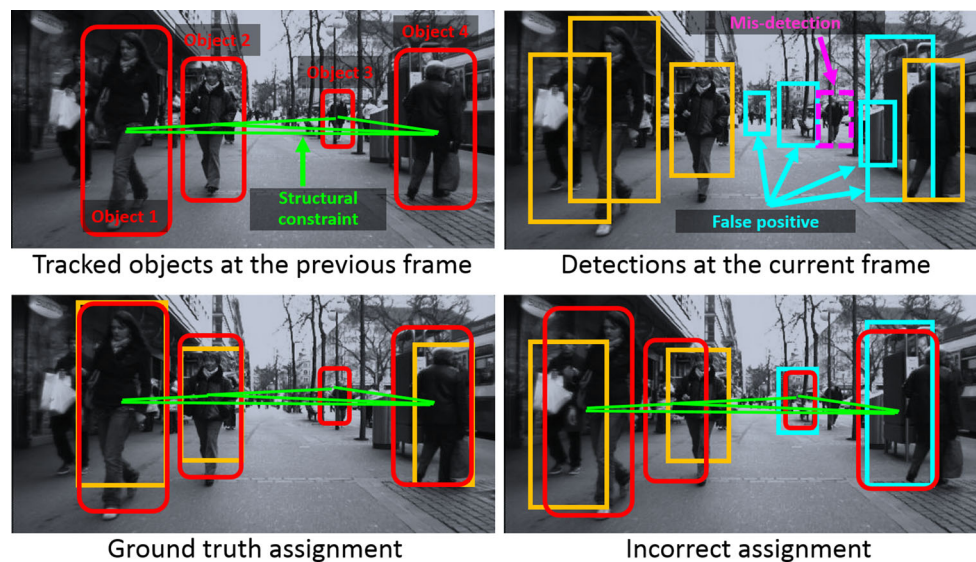


Fig. 1 An example of structural constraint ambiguity: The tracked objects and their correct detections are represented by the red and the yellow boxes, respectively. The overlap ratio costs of the ground truth

assignment (bottom left) and the incorrect assignment (bottom right) based on the structural constraint are similar because of mis-detections and multiple false-positive detections (Color figure online)

tracking errors when the target objects undergo appearance changes. To reduce ambiguities caused by similar appearances in MOT, Zhang and van der Maaten (2013) utilize motion constraints between objects along with object appearance models with structured support vector machines. Unlike the aforementioned methods in Grabner et al. (2010), Duan et al. (2012), Zhang and van der Maaten (2013), our method exploits structural constraints to solve the online 2D MOT problem with a frame-by-frame data association that assigns objects to correct detections. Yang and Nevatia (2014) use a conditional random field model for MOT in which the unary and binary terms are based on linear and smooth motion to associate past and future tracklets in sliding windows. Recently, Yoon et al. (2015) develop a method based on structural spatial information of relative moving objects to handle large camera motion. This method basically assumes that the camera motion is small and smooth such that at least a few objects are well predicted and tracked by linear motion models. Therefore, when the object motion prediction fails because of large camera motions, the method is not able to track objects with the structural information. Different from the aforementioned methods, the proposed method effectively deals with large and abrupt camera motions by utilizing anchor assignments, and alleviates ambiguities caused by mis-detections and false positives by applying the event aggregation algorithm.

Recently, data association methods commonly used in radar and sonar target tracking have been applied to vision-based multi-object tracking. Among them, multiple hypothesis tracking (MHT) and probabilistic data association (JPDA) methods perform robustly and accurately in tracking multi-

ple targets using radar or sonar sensors (Bar-Shalom and Li 1995; Blackman and Popoli 1999). Kim et al. (2015) modify the MHT method by effectively utilizing the visual appearance information and achieved the state-of-the-art results. The MHT method constructs object trajectories by considering all possible data associations throughout the given frames. Therefore, it can effectively reduce the incorrect matching between objects and detections. However, it is difficult to apply the MHT method to online tracking problems because it requires the future frame information as mentioned above. Rezatofghi et al. (2015, 2016) propose an algorithm to significantly reduce the computational complexity of JPDA to make it 10 times faster while producing comparable performance. Other data association approaches for the vision-based MOT can be found in Betke and Wu (2016) including the network flow based data association (Wu and Betke 2016).

Similar to the JPDA and MHT methods, our algorithm considers all possible assignments between objects and detections at each frame (although the MHT method prunes and merges possible associations). However, unlike the data association methods proposed in Kim et al. (2015), Rezatofghi et al. (2015), we incorporate the structural constraints into the data association framework to deal with large camera motions and uncertain detections. To effectively utilize constraint information, we design new data association cost functions and introduce the event aggregation method, which can resolve ambiguities caused by uncertain detections (e.g., false positives and false negatives). Since the proposed data association method is different from the JPDA method, we cannot utilize the fast solution from

Rezatofighi et al. (2015). Hence, we propose two fast solutions to solve the proposed data association with structural constraint costs.

3 Online Multi-object Tracking with Structural Constraints

3.1 Problem Formulation

The trajectory of an object is represented by a sequence of states denoting the position, velocity, and size of an object in the image plane with time. We denote a state vector of the target i at frame t as $\mathbf{s}_t^i = [x_t^i, y_t^i, \dot{x}_t^i, \dot{y}_t^i, w_t^i, h_t^i]^\top$, where (x_t^i, y_t^i) represents the position, $(\dot{x}_t^i, \dot{y}_t^i)$ represents the 2D velocity, and (w_t^i, h_t^i) represents the size, respectively, and the set of the target states at frame t as \mathcal{S}_t ($\mathbf{s}_t^i \in \mathcal{S}_t$) with its index set as $i \in \mathcal{I}_t \triangleq \{1, \dots, N\}$. To deal with large camera motion, we utilize a structural constraint information, which is described by the location and velocity difference between two objects as

$$\begin{aligned} \mathbf{e}_t^{i,j} &= [\chi_t^{i,j}, \xi_t^{i,j}, \dot{\chi}_t^{i,j}, \dot{\xi}_t^{i,j}]^\top \\ &= [x_t^i - x_t^j, y_t^i - y_t^j, \dot{x}_t^i - \dot{x}_t^j, \dot{y}_t^i - \dot{y}_t^j]^\top. \end{aligned} \quad (1)$$

Here, $(\dot{\chi}_t^{i,j}, \dot{\xi}_t^{i,j})$ denotes the velocity difference to consider objects moving with different tendencies. The set of structural constraints for the object i is represented by $\mathcal{E}_t^i = \{\mathbf{e}_t^{i,j} | \forall j \in \mathcal{I}_t\}$, and the set of all structural constraints at frame t is denoted by $\mathcal{E}_t = \{\mathcal{E}_t^i | \forall i \in \mathcal{I}_t\}$. We denote a detection k at frame t as $\mathbf{d}_t^k = [x_{d,t}^k, y_{d,t}^k, w_{d,t}^k, h_{d,t}^k]^\top$ and the set of detections at frame t used for MOT as \mathcal{D}_t ($\mathbf{d}_t^k \in \mathcal{D}_t$) with its index set as $k \in \mathcal{K}_t \triangleq \{0, 1, \dots, K\}$, where 0 is included to stand for mis-detected objects. Without loss of generality, we remove the time index t for simplicity in the following sections.

The MOT task can be considered as a data association problem, which finds the correct assignments between objects and detections. In this paper, we define the assignment event as

$$\mathcal{A} = \{a^{i,k} | i \in \mathcal{I}, k \in \mathcal{K}\}. \quad (2)$$

Here, $a^{i,k} = \{0, 1\}$; when the detection k is assigned to the object i , the assignment is denoted by $\{a^{i,k} = 1\}$. Otherwise, it is denoted by $\{a^{i,k} = 0\}$. The assignment event satisfies the following two conditions that (1) each detection is assigned to at most one object and (2) each object is assigned at most one detection.

In the data association, dissimilarity costs between objects and detections are computed, and then the data association cost is obtained by summing the dissimilarity costs based on an assignment event. The best assignment event is estimated

by selecting one of candidate assignment events that has the minimum data association cost. In this process, to achieve the robust data association under large camera motion, we incorporate the structural constraints \mathcal{E} into the data association cost function as follows.

$$\begin{aligned} \hat{\mathcal{A}} &= \arg \min_{\mathcal{A}} C(\mathcal{A}, \mathcal{S}, \mathcal{E}, \mathcal{D}), \\ \text{s.t. } \sum_{\substack{i \in \mathcal{I} \\ k \neq 0}} a^{i,k} &\leq 1 \wedge \sum_{k \in \mathcal{K}} a^{i,k} = 1 \wedge \sum_{i \in \mathcal{I}} a^{i,0} \leq |\mathcal{I}|, \end{aligned} \quad (3)$$

where $a^{i,0}$ stands for the case of mis-detected objects. Hence, the sum of $a^{i,0}$ along i is equal to the number of objects, $|\mathcal{I}|$, when all objects are mis-detected. We solve the data association in (3) via the structural constraint event aggregation in Sect. 4 and the structural constraint object recovery in Sect. 5.

3.2 Proposed MOT Framework

Similar to other online MOT methods, the main part of the proposed MOT framework is also the data association module that finds the best assignments between objects and detections in each frame. In this work, we adopt a two-step data association method to utilize structural constraints more efficiently and effectively. The proposed data association consists of two parts, i.e., the structural constraint event aggregation (SCEA) and structural constraint object recovery (SCOR).

When new detections come, the SCEA estimates the assignments between well-tracked objects and detections as shown in the second block of Fig. 2. Here, the well-tracked objects are the objects that are detected and tracked at the previous frame. We denote the set of well-tracked objects, \mathcal{S}_w , by

$$\mathbf{s}^t \in \mathcal{S}_w, \quad i \in \mathcal{I}_w, \quad (4)$$

where \mathcal{I}_w denotes the set of well-tracked object indices. The structural constraints for the well-tracked object i are denoted by

$$\mathcal{E}_w = \{\mathcal{E}_w^i | \forall i \in \mathcal{I}_w\}, \quad \text{where } \mathcal{E}_w^i = \{\mathbf{e}^{i,j} | \forall j \in \mathcal{I}_w\}. \quad (5)$$

After we find the best matches between well-tracked objects and detections via the SCEA method, the SCOR module finds the assignments between missing objects and unassigned detections using the updated positions of the well-tracked objects and structural constraints between the well-tracked objects and the missing objects as show in the third block of Fig. 2. We denote the set of missing objects, \mathcal{S}_m , by

$$\mathbf{s}^t \in \mathcal{S}_m, \quad i \in \mathcal{I}_m, \quad (6)$$

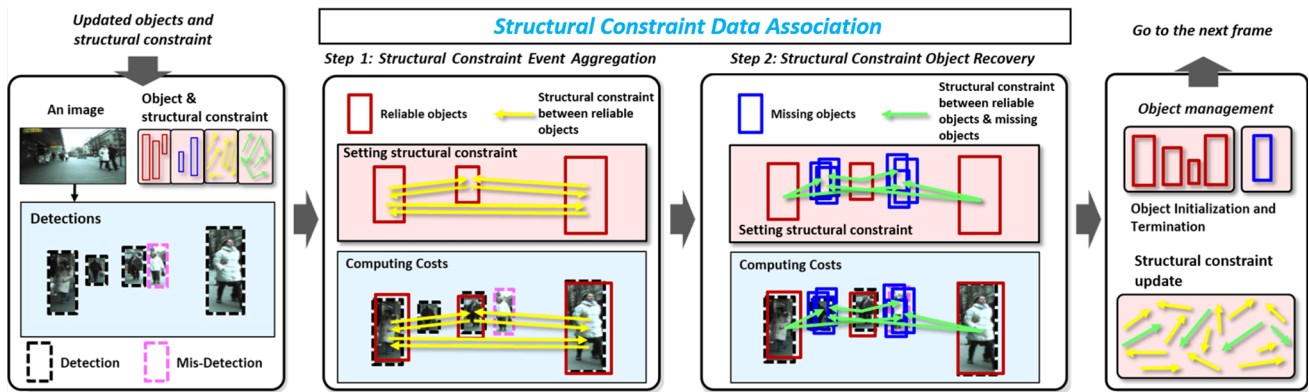


Fig. 2 Proposed online MOT framework with two-step frame-by-frame data association (Algorithm 1). Each step utilizes a different type of objects and structural constraints

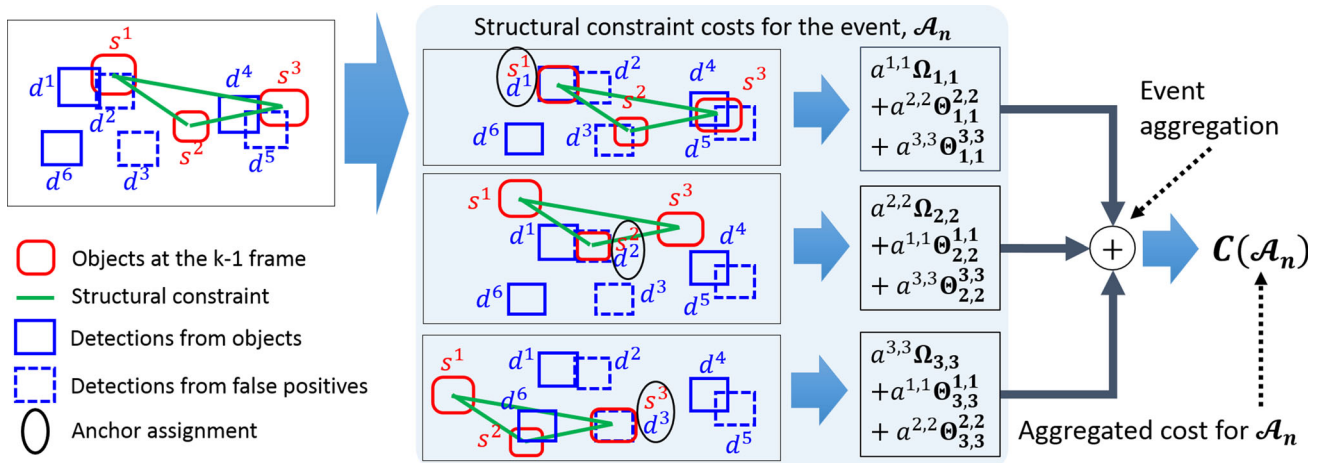


Fig. 3 An example of a structural constraint cost based on the assignment event \mathcal{A}_n in which $a^{1,1} = a^{2,2} = a^{3,3} = 1$ and the other assignments are set to 0. In the first row, object 1 is assigned to detection 1, and the position of object 1 is moved to the position of detection 1. Afterward, the positions of the other objects are determined by the

structural constraint. Then, we compute the cost of \mathcal{A}_n based on the anchor assignment $a^{1,1} = 1$. By changing this anchor assignment, we obtain three different costs, although they have the same assignment event

where \mathcal{I}_m denotes the set of missing object indices. The structural constraints between the missing object i and other well-tracked objects are denoted by

$$\mathcal{E}_m = \{\mathcal{E}_m^i | \forall i \in \mathcal{I}_m\}, \text{ where } \mathcal{E}_m^i = \{e^{i,j} | \forall j \in \mathcal{I}_w\}. \quad (7)$$

After we obtain the assignment events via the data association, we update the states of objects and their structural constraints with assigned detections. In the object termination, the objects that are not assigned with any detection for a certain frame are classified and removed, and their corresponding structural constraints are also removed. Not-assigned detections are used to initialize new objects in the object initialization. After initialization, their structural constraints are also initialized with the well-tracked objects. The proposed MOT conducts the aforementioned procedures at each frame. detail from Sects. 4–6.

4 Structural Constraint Event Aggregation (SCEA)

In this section, we introduce the SCEA, which exploits the structural constraint information based on assignment events. One simple example of the SCEA method based on the n -th assignment event \mathcal{A}_n is illustrated in Fig. 3. The overall procedure of the SCEA method considering all possible assignment events is also demonstrated in Fig. 4.

4.1 Structural Constraint Cost

To deal with unexpected global motions caused by translational motion, pitch motion, or yaw motion of a camera, we first select one assignment, $a^{i,k} = 1$ from \mathcal{A} as an anchor assignment, and then we make the center position of the corresponding object i coincide with that of detection k . On the

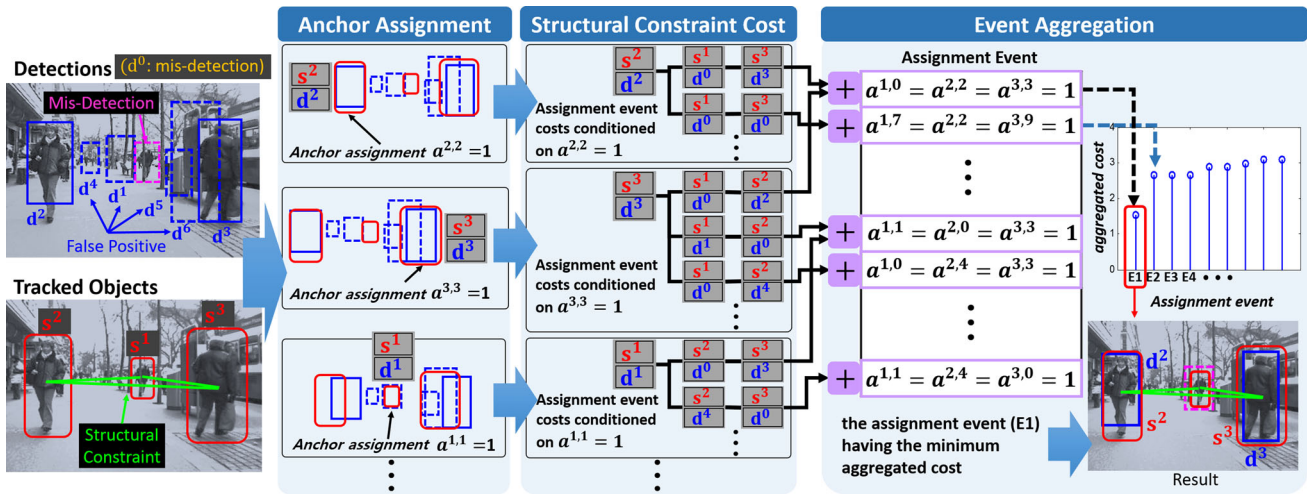


Fig. 4 The overall framework of the structural constraint event aggregation (Algorithms 2 and 3). The tracked objects and their detections are represented by red boxes and blue boxes, respectively. The blue boxes with the dash line are false positive detections. The green lines connecting the objects denote the structural constraints. Black boxes represent assignments. \mathbf{d}^0 stands for the case of mis-detections. As shown in this figure, in the anchor assignment $a^{2,2}$ of object 2 and detection 2, we move object 2 to align its center location with that of detection 2. Then, in the structural constraint cost computation, we compute the assign-

ment costs of other objects and detections based on their structural constraints. From the different anchor assignments, the structural constraint costs for the same assignment event are computed. For instance, the costs of the assignment event ($a^{1,0} = a^{2,2} = a^{3,3} = 1$) are obtained from the anchor assignments $a^{2,2} = 1$ and $a^{3,3} = 1$, respectively. The event aggregation fuses these structural constraint costs having the same assignment event but with different anchor assignments. + represents the summation of the structural constraint costs (Color figure online)

Algorithm 1: Online MOT via Structural Constraint Data Association.

Data: tracked objects S_w , structural constraints of tracked objects \mathcal{E}_w , mis-detected objects S_m , structural constraints between tracked objects and mis-detected objects \mathcal{E}_m , detections \mathcal{D}

Result: Trajectories of the targets

```

1 for video frame  $f$  do
2   Step 1: Structural constraint event aggregation
3   ·  $\mathcal{A}_w = \text{SCEA}(S_w, \mathcal{E}_w, \mathcal{D})$ ; (Section 4: Algorithm 2 or Algorithm 3)
4   ·  $S_w := \{s_i = \mathbf{d}^k | a^{i,k} = 1, \forall i \in \mathcal{I}_w, \forall k \in \mathcal{K}\}$ ;
5   Step 2: Structural constraint object recovery
6   ·  $\mathcal{A}_m = \text{SCOR}(S_m, \mathcal{E}_m, S_w, \tilde{\mathcal{D}})$ ; (Section 5)
7   ·  $\mathcal{A} = \mathcal{A}_w \cup \mathcal{A}_m$ ;
8   Step 3: Update
9   · Update  $S_w$  and  $S_m$  via object management (Section 6.1)
10  · Update  $\mathcal{E}_w$  and  $\mathcal{E}_m$  Structural constraint update (Section 6.2)
11  · Current tracking result:
12   $S_w = \{s^i | a^{i,k} = 1, \forall i \in \mathcal{I}_w \cup \mathcal{I}_m, \forall k \in \mathcal{K}\}$ .

```

basis of the anchor assignment and the structural constraint, we determine the positions of other objects, as illustrated in Fig. 3. By doing this, we prevent the structural constraint cost from incurring the large prediction error caused by the global camera motion, and all motion costs in the data association are computed based on the structural constraints ignoring objects' own motion information. On the basis of this concept, we formulate the proposed structural constraint cost function that consists of an anchor cost and linked costs. Note

that in the SCEA, we only consider well-tracked objects and their notation are described in (4) and (5). With those notations, the structural constraint cost function is formulated as

$$C(\mathcal{A}, S_w, \mathcal{E}_w, \mathcal{D})|_{a^{i,k}} = a^{i,k} \Omega_{i,k} + \sum_{\substack{j \in \mathcal{I}_w \\ j \neq i}} \sum_{\substack{q \in \mathcal{K} \\ q \neq k}} a^{j,q} \Theta_{i,k}^{j,q}, \quad (8)$$

where the subscripts i and k denote the indices for costs computed based on the anchor assignment $a^{i,k}$, and the anchor cost is computed by

$$\Omega_{i,k} = F_s(s^i, \mathbf{d}^k) + F_a(s^i, \mathbf{d}^k). \quad (9)$$

Here, we do not consider the motion cost but utilize the size and appearance costs as

$$F_s(s^i, \mathbf{d}^k) = -\ln \left(1 - \frac{|h^i - h_d^k|}{2(h^i + h_d^k)} - \frac{|w^i - w_d^k|}{2(w^i + w_d^k)} \right),$$

$$F_a(s^i, \mathbf{d}^k) = -\ln \sum_{b=1}^B \sqrt{p^b(s^i) p^b(\mathbf{d}^k)}, \quad (10)$$

where (w^i, h^i) and (w_d^k, h_d^k) denote width and height of object i and detection k , respectively. In addition, $p^n(s^i)$ and $p^n(\mathbf{d}^k)$ denote their histogram information, respectively. b is the bin index and B is the number of bins. On the basis of

the anchor position, we calculate the linked cost based on the structural constraints, formulated by

$$\Theta_{i,k}^{j,q} = \begin{cases} F_s(\mathbf{s}^j, \mathbf{d}^q) + F_a(\mathbf{s}^j, \mathbf{d}^q) + F_c(\mathbf{s}^j, \mathbf{e}^{j,i}, \mathbf{d}^k, \mathbf{d}^q) & \text{if } q \neq 0 \\ \tau & \text{if } q = 0 \end{cases}, \quad (11)$$

$$F_c(\mathbf{s}^j, \mathbf{e}^{j,i}, \mathbf{d}^k, \mathbf{d}^q) = -\ln \left(\frac{\text{area}(\mathbf{B}(\mathbf{s}^{j,k}) \cap \mathbf{B}(\mathbf{d}^q))}{\text{area}(\mathbf{B}(\mathbf{s}^{j,k}) \cup \mathbf{B}(\mathbf{d}^q))} \right), \quad (12)$$

$$\mathbf{s}^{j,k} = [x_d^k, y_d^k, 0, 0]^\top + [\chi^{j,i}, \upsilon^{j,i}, w^j, h^j]^\top.$$

In (11), we empirically set the cost τ to a non-negative value (e.g., 4 in this work) for the case of mis-detected objects, \mathbf{d}^0 . In (12), we determine the position of object j by the anchor position (i.e., the position of detection k) and the structural constraint $\mathbf{e}^{j,i}$. The cost function $F_c(\cdot)$ is computed by using the overlap ratio (Everingham et al. 2010) of the the detection bounding box and ground truth. The reason for using different metrics is that the overlap ratio compensates bias errors caused by object sizes and is less sensitive to the distance error.

4.2 Event Aggregation

On the basis of the different anchor assignments, we obtain different costs owing to the different sizes of detections and detection noises even if the assignment event \mathcal{A} is the same. Hence, we aggregate all the costs that have the same assignment event but different anchor assignments. Compared to conventional one-to-one matching process for the data association, as shown in Fig. 1, this process significantly reduces ambiguity caused by false positives near objects, mis-detections, and constraint errors since we can measure the cost of each assignment event several times according to the number of corresponding anchor assignments, as described in Fig. 4. With (8), the event aggregation process is formulated by

$$C(\mathcal{A}, \mathcal{S}_w, \mathcal{E}_w, \mathcal{D}) = \frac{1}{\Delta} \sum_{i \in \mathcal{I}} \sum_{\substack{k \in \mathcal{K} \\ a^{i,k}=1}} C(\mathcal{A}, \mathcal{S}_w, \mathcal{E}_w, \mathcal{D})|_{a^{i,k}}, \Delta \\ = \sum_{i \in \mathcal{I}, k \in \mathcal{K}} a^{i,k}, \quad (13)$$

where Δ denotes the normalization term that is equal to the number of anchor assignments selected from the assignment event \mathcal{A} . Finally, we select the best assignment event \mathcal{A}_w having the minimum aggregated cost as

$$\mathcal{A}_w = \arg \min_{\mathcal{A}} C(\mathcal{A}, \mathcal{S}_w, \mathcal{E}_w, \mathcal{D}), \quad \mathcal{A} \subset \mathcal{A}_{\text{all}}. \quad (14)$$

Here, \mathcal{A}_{all} denotes all possible assignment events.

With \mathcal{A}_w , we preliminary update the well-tracked objects with an assigned detection by $\hat{\mathbf{s}}^i = \mathbf{d}^k$ if $a^{i,k} = 1$. These

updated object states are used to find the missing objects with the structural constraints via the SCOR in Sect. 5.

4.3 Solution for the SCEA

The computational cost for the JPDA method grows significantly as the number of objects and detections increase because the number of possible assignment events also increase substantially. To address this issue, the fast JPDA method have been recently proposed (Rezatofighi et al. 2015). However, it does not incorporate the structural constraints and the event aggregation into data association. In this work, we design two effective methods, i.e., partitioning-based and exhaustive combinatorial enumeration methods, to compute the best assignment events between objects and detections.

4.3.1 Solution Based on Simple Partitioning

The partitioning-based solver is simple and intuitive, and it consists of two steps, i.e., gating and partitioning as shown in Fig. 5. First, we adopt the simple gating technique (Bar-Shalom and Li 1995) before conducting the structural constraint event aggregation. This method is widely used in the MOT literature. Note that since we consider large motion changes of objects in this work, we set the gate size large enough. We roughly remove the negligible assignments based on two conditions as

$$\left(\|\mathbf{p}^i - \mathbf{p}_d^k\| < \sqrt{(w^i)^2 + (h^i)^2} \right) \wedge \\ \left(\exp \left(-F_s(\mathbf{s}^i, \mathbf{d}^k) \right) > \tau_s \right), \quad (15)$$

where \mathbf{p}^i and \mathbf{p}_d^k represent the position of object i and detection k , respectively, and (w^i, h^i) denotes the size of object i . We empirically set $\tau_s = 0.7$. If the above conditions are satisfied, $a^{i,k} = 1$. Otherwise, the assignment is set to $a^{i,k} = 0$, and this assignment is not considered for tracking at the current frame. Second, the partitioning method constructs subgroups to handle a large number of objects and detections, as shown in Fig. 5. The assignments of objects and detections in different partitions are set to $a^{i,k} = 0$. For partition p , we generate all possible assignment events $\mathcal{A}^p \subset \mathcal{A}_{\text{all}}^p$ based on the condition in (3). The SCEA method is carried out for each partition. The final assignment events are then obtained by merging the best assignment events from all partitions. In this work, we empirically set the maximum number of objects in each partition to 5. The number of partitions is determined by $P = \lceil \text{the number of objects}/5 \rceil$, and we then make the partitions possibly to have the same number of objects. Here, we use the center locations from K-means clustering for a partitioning condition. As shown in Fig. 5, P centers are obtained from K-means clustering, and the objects located

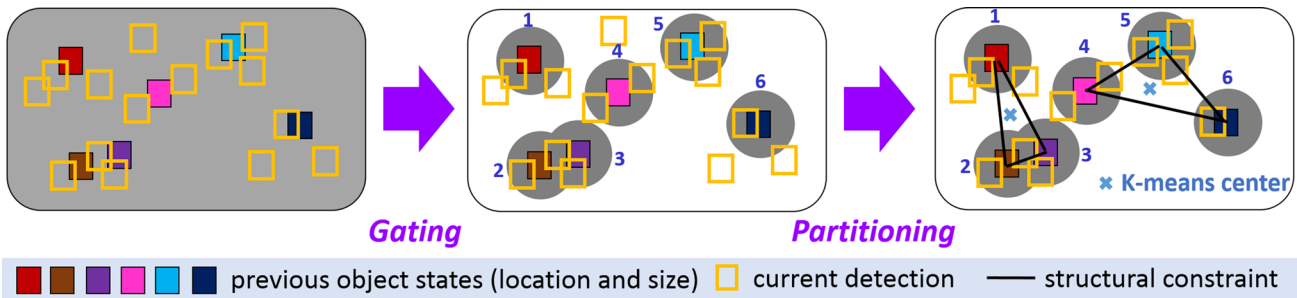


Fig. 5 Gating and partitioning methods for assignment event reduction: The gating and the partitioning reduce the number of assignment events. Gray circles represent the assignment regions reduced by the gating. The objects are grouped based on the K-means clustering

The numbers represent group indices. When objects in different groups have the same detection as the second group and the third group, the detection can be duplicately allocated in the partitioning. Therefore, true associations can be located together

Algorithm 2: SCEA based on partitioning.

Data: well-tracked objects S_w , structural constraints \mathcal{E}_w , detections \mathcal{D}

Result: assignment event \mathcal{A}_w

```

1 begin
2   Step 1: Partitioning (Section 4.3.1)
3   · Removing negligible assignments by using the gating
4     technique ((15)).
5   · Dividing objects, structural constraints, and detections into
6     subgroups  $S_w^p \subset S_w$ ,  $\mathcal{D}^p \subset \mathcal{D}$  by partitioning (Fig. 5).
7   · Generating all possible assignment events  $\mathcal{A}_{all}^p$  of each
8     partition from  $S_w^p$  and  $\mathcal{D}^p$  based on the index sets  $(\mathcal{I}_w^p, \mathcal{K}^p)$ 
9     and the condition in (3).
10  Step 2: Aggregating assignment event costs ((14))
11   $\mathcal{A}_w = \phi$ ;
12  for  $p = 1 : P$  do
13     $C(\mathcal{A}^p, S_w^p, \mathcal{E}_w, \mathcal{D}^p) =$ 
14       $\frac{1}{\Delta} \sum_{\substack{i \in \mathcal{I}_w^p, k \in \mathcal{K}^p \\ a^{i,k}=1}} (a^{i,k} \Omega_{i,k} + \sum_{\substack{j \in \mathcal{I}_w^p \\ j \neq i}} \sum_{\substack{q \in \mathcal{K}^p \\ q \neq k}} a^{j,q} \Theta_{i,k}^{j,q}),$ 
15     $\Delta = \sum_{a^{i,k} \in \mathcal{A}^p} a^{i,k},$ 
16     $\hat{\mathcal{A}}^p = \arg \min_{\mathcal{A}^p} (C(\mathcal{A}^p, S_w^p, \mathcal{E}_w, \mathcal{D}^p)), \mathcal{A}^p \subset \mathcal{A}_{all}^p;$ 
17     $\mathcal{A}_w := \mathcal{A}_w \cup \hat{\mathcal{A}}^p;$ 
18  end
19 end
20 return  $\mathcal{A}_w$ 

```

close to each K-means center are then clustered in the same partition. Note that when objects in different groups have the same detection after gating as the second group and the third group in Fig. 5, the detection can be duplicately allocated in the partitioning. Therefore, true associations are located together. The main steps of the proposed partitioning method are summarized in Algorithm 2.

4.3.2 Solution Based on Exhaustive Combinatorial Enumeration

Different from the partitioning method described above, we can consider all possible combinations of the objects. The

subgroup generated by the K-means clustering-based partitioning can be considered as one specific case among all possible combinations. When we make each subgroup consisting of c objects, the number of all possible combinations G is obtained by the binomial coefficient as

$$G = \binom{|\mathcal{I}|}{c} = \frac{|\mathcal{I}|!}{c!(|\mathcal{I}| - c)!}, \tag{16}$$

where the cardinality of the set of object indices, $|\mathcal{I}|$, denotes the number of tracked objects. The subgroups generated by the exhaustive combinatorial enumeration are represented by $\mathcal{S}^g, g = 1, \dots, G$. We apply the SCEA method to each subgroup and all the results, i.e., assignment event matrix \mathcal{A}^g , are merged together by

$$\mathcal{A} = \frac{1}{N_{max}} \sum_{p=1}^G \mathcal{A}^g. \tag{17}$$

Here, N_{max} normalizes the assignment event matrix \mathcal{A} . N_{max} is the number that represents how many times each object belongs to subgroups.

$$N_{max} = \frac{|\mathcal{I}|!}{c!(|\mathcal{I}| - c)!} \times \frac{c}{|\mathcal{I}|}, \tag{18}$$

where c is the number of objects in each subgroup as described above. As the value of c increases, this method estimates more optimal assignment event. However, the computational complexity also increases significantly. We evaluate the performance according to the different values of c in Sect. 7.

For ease of understanding, we illustrate the proposed method with one example. Suppose that we have a set of objects as $\mathcal{I} = \{1, 2, 3, 4\}$ and each subgroup is supposed to be formed with three objects based on the exhaustive combinatorial enumeration. As $c = 3$, the number of all possible subgroups is computed by the binomial coefficient as $\frac{|\mathcal{I}|!}{3!(|\mathcal{I}|-3)!} = \frac{4!}{3!(4-3)!} = 4$ and all possible subgroups are

{1, 2, 3}, {1, 2, 4}, {1, 3, 4}, and {2, 3, 4}. The SCEA method is applied to each of 4 subgroups and four assignment event matrices $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3,$ and \mathcal{A}_4 are obtained. Further assume that there exist six detections and the best assignment event matrix for each subgroup is obtained as

$$\mathcal{A}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \mathcal{A}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathcal{A}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \mathcal{A}_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

The sum of all assignment event matrices is computed as $\mathcal{A} = \sum_{n=1}^4 \mathcal{A}_n$ and we normalize this matrix by N_{max} as

$$\mathcal{A} = \frac{1}{N_{max}} \sum_{n=1}^4 \mathcal{A}_n = \frac{1}{3} \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}.$$

For the case of object 4, we select the detection 6 that receives more votes than others. Note that, when the normalized vote is below the threshold, we do not assign that detection to the object.

Although the partitioning method performs favorably as shown in Fig. 10, this exhaustive combinatorial enumeration method generates more robust results. In addition, since this voting scheme for the SCEA method considers all possible subgroups, it alleviates the problem of being trapped into local optima and helps to find better solutions. The main steps of the proposed method are described in Algorithm 3. Similarly to the partitioning method, it is also only executed when the number of objects is larger than a threshold (it empirically set to 5 in this work). As shown on line 4 of Algorithm 3, we first make all possible subgroups consisting of c objects based on the exhaustive combinatorial enumeration. Then, for each subgroup p , we generate all possible assignment events $\mathcal{A}^g \subset \mathcal{A}_{all}^g$ based on the condition in (3). We obtain the best sub-assignment event $\hat{\mathcal{A}}^g$ by running the SCEA on each subgroup, and we update the vote of assignments based on $\hat{\mathcal{A}}^g$ as on line 13 of Algorithm . After we obtain all the votes, we normalize the voting assignment as on line 15 of Algorithm 3 and then we set the assignment indicator $a^{i,k}$ to 0 if the normalized value is below the threshold η^1 as on line 16 of Algorithm 3.

¹ η is set to 0.5 in our experiments.

Algorithm 3: SCEA based on exhaustive combinatorial enumeration.

Data: well-tracked objects \mathcal{S}_w , structural constraints \mathcal{E}_w , detections \mathcal{D}

Result: assignment event \mathcal{A}_w

```

1 begin
2   Step 1: Subgroup generation (Section 4.3.2)
3   · Removing negligible assignments by using the gating
   technique ((15)).
4   · Generating all possible subgroups consisting of  $c$  objects
   based on the exhaustive combinatorial enumeration.
5   · Subgroups and detections are denoted by
 $\mathcal{S}_w^g \subset \mathcal{S}_w, \mathcal{D}^g \subset \mathcal{D}, g = 1, \dots, G$  where  $G$  is the number of
   subgroups in (16).
6   · Generating all possible assignment events  $\mathcal{A}_{all}^g$  of each
   subgroup from  $\mathcal{S}_w^g$  and  $\mathcal{D}^g$  based on the subgroup index sets
   ( $\mathcal{I}_w^g, \mathcal{K}^g$ ) and the condition in (3).
7   Step 2: SCEA on subgroups and Voting ((14))
8    $\hat{\mathcal{A}} = \{\hat{a}^{i,k} = 0 | i \in \mathcal{I}_w, k \in \mathcal{K}\}$ ; voting assignment
   initialization
9   for  $g = 1 : G$  do
10     $C(\mathcal{A}^g, \mathcal{S}_w^g, \mathcal{E}_w, \mathcal{D}^g) =$ 
         $\frac{1}{\Delta} \sum_{\substack{i \in \mathcal{I}_w^g, k \in \mathcal{K}^g \\ a^{i,k}=1}} (a^{i,k} \Omega_{i,k} + \sum_{\substack{j \in \mathcal{I}_w^g \\ j \neq i}} \sum_{\substack{q \in \mathcal{K}^g \\ q \neq k}} a^{j,q} \Theta_{i,k}^{j,q}),$ 
11     $\Delta = \sum_{a^{i,k} \in \mathcal{A}^g} a^{i,k},$ 
12     $\hat{\mathcal{A}}^g = \arg \min_{\mathcal{A}^g} (C(\mathcal{A}^g, \mathcal{S}_w^g, \mathcal{E}_w, \mathcal{D}^g)), \mathcal{A}^g \subset \mathcal{A}_{all}^g,$ 
13     $\hat{\mathcal{A}} = \{\hat{a}^{i,k} := \hat{a}^{i,k} + 1 | a^{i,k} = 1, a^{i,k} \in \hat{\mathcal{A}}^g\}$ ; Voting
14  end
15   $\mathcal{A} = \{a^{i,k} = \frac{\hat{a}^{i,k}}{N_{max}} | \hat{a}^{i,k} \in \hat{\mathcal{A}}\}$ ;  $N_{max}$  in (18)
16   $\mathcal{A}_w = \{a^{i,k} = 1 | a^{i,k} \geq \eta, a^{i,k} \in \mathcal{A}\} \cup \{a^{i,k} = 0 | a^{i,k} <$ 
         $\eta, a^{i,k} \in \mathcal{A}\}$ ;
17 end
18 return  $\mathcal{A}_w$ 

```

5 Structural Constraint Object Recovery (SCOR)

We adopt two-step data association for effectively exploiting the structural constraints for online 2D MOT. Since the structural constraints of objects tracked in the previous frame have been updated with their corresponding detections, their constraints are more reliable than those of mis-detected objects. This allows us to assign detections to tracked objects more robustly and accurately. In the SCOR, similar to Grabner et al. (2010), Yoon et al. (2015), we recover missing objects, which are not associated with any detections in the previous frame but re-detected in the current frame. To this end, we first find the assignments for the objects tracked at the previous frame, and then we track other missing objects using the updated positions of the tracked objects from the SCEA. The recovery process is conducted by using the updated objects, \mathcal{S}_w , from the SCEA as described in Fig. 6. As described in (6) and (7), the mis-detected objects are denoted by \mathcal{S}_m , and the structural constraints between mis-detected objects and

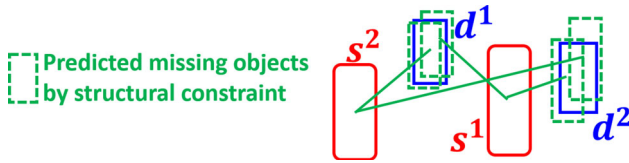


Fig. 6 Concept of structural constraint object recovery: From the tracked objects (s^1 and s^2) and the structural constraints (the green lines), we recover missing objects when they are re-detected (detection d^1 and d^2). By doing this, we can continuously keep the identity of the missing objects under camera motion and occlusions (Color figure online)

tracked objects are represented by \mathcal{E}_m . Using \mathcal{S}_m , \mathcal{S}_w , and \mathcal{E}_m , we recover the re-detected objects as

$$\begin{aligned} \mathcal{A}_m &= \arg \min_{\mathcal{A}} C(\mathcal{S}_m, \mathcal{E}_m, \mathcal{S}_w, \tilde{\mathcal{D}}), \text{ s.t. } \sum_{i \in \mathcal{I}_m} a^{i,q} \\ &= 1 \wedge \sum_{q \in \tilde{\mathcal{K}}} a^{i,q} = 1, \end{aligned} \tag{19}$$

where \mathcal{I}_m denotes the set of mis-detected object indices and $\tilde{\mathcal{K}}$ represents the index set of detections $\tilde{\mathcal{D}}$. Here, a set of detections, $\tilde{\mathcal{D}}$, contains the not-assigned detections in the SCEA and dummy detections \mathbf{d}^0 for the case of mis-detected objects. The structural constraint cost function for missing objects is defined as

$$\begin{aligned} C(\mathcal{S}_m, \mathcal{E}_m, \mathcal{S}_w, \tilde{\mathcal{D}}) &= \sum_{i \in \mathcal{I}_m} \sum_{q \in \tilde{\mathcal{K}}} a^{i,q} \phi^{i,q} \\ \phi^{i,q} &= \begin{cases} F_s(s^i, \mathbf{d}^q) + F_a(s^i, \mathbf{d}^q) + F_r(s^i, \mathcal{E}_m, \mathcal{S}_w, \mathbf{d}^q) & \text{if } q \neq 0 \\ \tau & \text{if } q = 0 \end{cases}, \end{aligned} \tag{20}$$

where the cost τ is a non-negative constant and set to 4 in this work as in (11). We recover the mis-detected object i from the set of tracked objects using their structural constraints. The constraint cost is therefore formulated as

$$\begin{aligned} F_r(s^i, \mathcal{E}_m, \mathcal{S}_w, \mathbf{d}^q) &= -\ln \left(\frac{\text{area}(\mathbf{B}(s^{i,\gamma}) \cap \mathbf{B}(\mathbf{d}^q))}{\text{area}(\mathbf{B}(s^{i,\gamma}) \cup \mathbf{B}(\mathbf{d}^q))} \right), \\ s^{i,\gamma} &= [(s_1^\gamma)^\top, 0, 0]^\top + [\chi^{i,\gamma}, v^{i,\gamma}, w^i, h^i]^\top, \\ \gamma &= \arg \max_{j \in \mathcal{I}_w} \frac{1}{\|\dot{\chi}^{i,j}, \dot{v}^{i,j}\|}, \end{aligned} \tag{21}$$

where \mathcal{I}_w denotes the set of indices of tracked objects from the SCEA. Here, the reliability of the structural constraints between tracked objects and missing objects can be different according to the past motion coherence. To consider this constraint reliability, we select the object moving in the most similar direction and velocity by taking the motion coherence between objects into account, $\|\dot{\chi}^{i,j}, \dot{v}^{i,j}\|$.

To solve (19), we reformulate (19) in a matrix form as

$$\mathbf{C} = \begin{bmatrix} \Phi^{det}_{|N^m| \times |\tilde{\mathcal{M}}|} & \Phi^0_{|N^m| \times |N^m|} \end{bmatrix}, \tag{22}$$

where the matrices are obtained by $\Phi^{det} = [\Phi^{i,q}], \forall i \in \mathcal{I}_m, \forall q \in \tilde{\mathcal{K}}$ and $\Phi^0 = \text{diag}[\Phi^{i,0}], \forall i \in \mathcal{I}_m$. The off-diagonal entries of Φ^0 are set to ∞ . We then apply the Hungarian algorithm (Kuhn 1955) to get the assignment event having the minimum cost.

6 Object and Structural Constraint Management

We obtain the final assignments \mathcal{A} between objects and detections by merging \mathcal{A}_w from the SCEA and \mathcal{A}_m from the SCOR as described on 7 of Algorithm 1. Using estimated assignments, we update both the object states and structural constraints. In the following, we first discuss the update of object states including the object management strategy. We next describe the structural constraint management in Sect. 6.2.

6.1 Object Management

6.1.1 Update of Object States

While the constant velocity model described in Li and Jilkov (2003) is widely adopted in target tracking, it performs well only when a camera is stationary. To account for abrupt and large motion changes, the motion covariance matrix of the Kalman filter (KF) is set to be large in this work. We update the object states by

$$\begin{aligned} \mathcal{S}_w &= \{s^i = KF(s^i, \mathbf{d}^k, \mathbf{F}, \mathbf{Q}, \mathbf{H}, \mathbf{R})\} a^{i,k} \\ &= 1, \forall i \in \mathcal{I}_w \cup \mathcal{I}_m, \forall k \in \mathcal{K}, \end{aligned} \tag{23}$$

where $KF(\cdot)$ represents the Kalman filter with the motion model \mathbf{F} , \mathbf{Q} is the motion covariance matrix, \mathbf{H} describes the observation model, and \mathbf{R} denotes the detection noise covariance matrix. These terms are described by

$$\begin{aligned} \mathbf{F} &= \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0.25\sigma_q^2 & 0.5\sigma_q^2 & 0 & 0 & 0 & 0 \\ 0.25\sigma_q^2 & 0.5\sigma_q^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5\sigma_q^2 & \sigma_q^2 & 0 & 0 \\ 0 & 0 & 0.5\sigma_q^2 & \sigma_q^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_s^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_s^2 \end{bmatrix}, \\ \mathbf{H} &= [\mathbf{I}_2, \mathbf{0}_2, \mathbf{I}_2], \quad \mathbf{R} = \text{diag}(\sigma_x^2, \sigma_y^2, \sigma_w^2, \sigma_h^2), \end{aligned} \tag{24}$$

where \mathbf{F} and \mathbf{Q} are from Li and Jilkov (2003), and \mathbf{I}_n and $\mathbf{0}_n$ denote the n -by- n identity matrix, and the n -by- n zero matrix, respectively. The deviation σ_q describes the possible velocity change over time steps, and σ_s represents the possible size change. The standard deviations of position and size noise of the detection noise are denoted by σ_x , σ_y , σ_w , and σ_h , respectively. In this work, we set those parameters as $\sigma_q = 15$, $\sigma_x = \sigma_y = 3$, and $\sigma_s = \sigma_w = \sigma_h = 15$.

We update object appearances incrementally over time. When denoting a target object by a histogram \mathcal{H}_{t-1} , the object appearance is updated by

$$\mathcal{H}_t = (1 - \delta)\mathcal{H}_{t-1} + \delta\hat{\mathcal{H}} \tag{25}$$

where $\hat{\mathcal{H}}$ represents the histogram of the target object in the current frame and δ is a learning rate set to 0.1 in this work. For the mis-detected objects, we do not update their appearances. In addition, we do not update the sizes of missing objects. However, due to the appearance model, we do not keep an object without detections for a long time because it may affect the MOT performance, especially data association.

Figure 7 shows the similarities between the size without update and the ground truth size. The size similarity without update does not change much for several frames. Thus, it does not significantly affect the overall MOT performance if we remove the objects that are not detected for more than a certain number of frames.

6.1.2 Initialization and Termination

After the state update, we terminate the tracks of the objects that are not assigned with any detections for a certain number of frames and initialize new tracks (i.e. objects) from unassociated detections. In this work, objects are initialized in a way similar to the method developed in Breitenstein et al. (2011) using distance and appearance between two detections. If the distances between a detection in the current frame and unassociated detections in the past a few frames are smaller than a certain threshold, we then initialize this detection as a new object. Note that the proposed data association handles initialized objects when the sudden and large camera motion occurs. When the sudden and large camera motion occurs in the initialization process, the conventional initialization method fails to initialize the object for tracking.

6.2 Structural Constraint Management

6.2.1 Update of Structural Constraints

After tracking, we update the structural constraints between objects with their corresponding detections, based on the same method proposed in Yoon et al. (2015), using $\mathbf{z}^{i,j} =$

$[x_d^i, y_d^i]^\top - [x_d^j, y_d^j]^\top$ as an observation where $[x_d^i, y_d^i]^\top$ represents the location of a detection assigned to object i . We assume that the structural constraint change follows the constant-velocity model from Li and Jilkov (2003). Similar to the object state update, we update the structural constraint variations by using the Kalman filter by

$$\begin{aligned} \mathbf{e}^{i,j} &= KF(\mathbf{e}^{i,j}, \mathbf{z}^{i,j}, \mathbf{F}_{sc}, \mathbf{Q}_{sc}, \mathbf{H}_{sc}, \mathbf{R}_{sc}), \text{ if } a^{i,k} \\ &= a^{j,k} = 1. \end{aligned} \tag{26}$$

In (26), the motion model \mathbf{F}_{sc} , motion covariance matrix \mathbf{Q}_{sc} , observation model \mathbf{H}_{sc} , and detection noise covariance matrix \mathbf{R}_{sc} are described by

$$\begin{aligned} \mathbf{F}_{sc} &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Q}_{sc} = \begin{bmatrix} 0.25\sigma_{sc}^2 & 0.5\sigma_{sc}^2 & 0 & 0 \\ 0.25\sigma_{sc}^2 & 0.5\sigma_{sc}^2 & 0 & 0 \\ 0 & 0 & 0.5\sigma_{sc}^2 & \sigma_{sc}^2 \\ 0 & 0 & 0.5\sigma_{sc}^2 & \sigma_{sc}^2 \end{bmatrix}, \\ \mathbf{H}_{sc} &= [\mathbf{I}_2, \mathbf{0}_2], \quad \mathbf{R}_{sc} = \text{diag}(\sigma_x^2, \sigma_y^2). \end{aligned} \tag{27}$$

The motion change standard deviation σ_{sc} for the structural constraints describes possible changes over time steps. Note that different from the object state update, the structural constraints are not significantly affected by the global camera motion. Therefore, we set the small value as $\sigma_{sc} = 1$, which enforces the structural constraints to follow the predefined motion model \mathbf{F}_{sc} . For the observation noise covariance, \mathbf{R}_{sc} , we set the same parameters $\sigma_x = \sigma_y = 3$ used in the object state update. The structural constraints of missing objects are predicted by using the motion model.

$$\mathbf{e}_t^{i,j} = \mathbf{F}_{sc}\mathbf{e}_{t-1}^{i,j}, \text{ if } \neg(a^{i,k} = a^{j,k} = 1). \tag{28}$$

Although the prediction becomes less accurate when it is not updated for a long duration, it performs favorably for several frames as used in the SCOR because it is not significantly affected by the global camera motion as mentioned.

6.2.2 Initialization and Termination

When a new object is initialized, the corresponding structural constraints are generated by (1), where their initial change values are set as $\dot{\chi}_t^{i,j} = \dot{\xi}_t^{i,j} = 0$. On the other hand, we do not initialize the structural constraints between the new objects and mis-detections. Once mis-detected objects are re-tracked via the SCOR method, we initialize their structural constraints. In addition, when the track of an object is terminated, we delete the corresponding structural constraints.

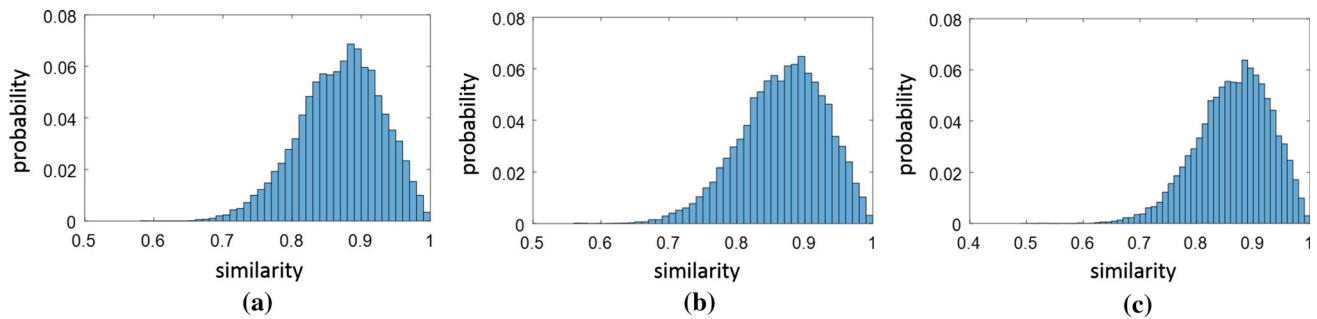


Fig. 7 Similarity distribution without the size update on different frames. The size similarity function in (10) is used without logarithm. **a** After 1 frame. **b** After 3 frames. **c** After 5 frames

7 Experiments

In the experiments, the proposed MOT framework is referred to as the SCEA method for simplicity. We first present the experimental evaluation of the proposed SCEA method and the comparison against the other state-of-the-art online and offline MOT methods. The source codes of the proposed SCEA method will be available at <https://cvl.gist.ac.kr/project/scea.html>.

7.1 Analysis of Data Association Performance

7.1.1 Datasets and Evaluation Metric

To evaluate the proposed data association method, we use the ETH (Bahnhof, Sunnyday, and Jelmoli sequences) (Ess et al. 2008) and KITTI car datasets. The sequences in both datasets are recorded by a camera equipped on a moving platform. We add different levels of motion fluctuation with different detection missing rates for performance evaluation. Figure 8 shows several example frames. The motion fluctuation is synthetically generated by the uniform distribution within $[0, 0]$ (no fluctuation), $[-7, 7]$, and $[-15, 15]$ pixels, respectively. We also set the detection missing rate as 0, 10, 20, and 30%. Therefore, the experimental setup using $[0, 0]$ fluctuation and 0% missing rate is identical to that in the original sequences. In addition, for all scenarios, we include at most 10 false detections for each frame.

The overall performance of a MOT method is affected by many internal tracking modules such as object initialization and termination. To compare the data association algorithms fairly, we use the same input for all data association methods in each frame, and measure the number of true matches and number of false matches between objects and detections in each frame. We measure the data association matching accuracy (ACC) by

$$\text{ACC} = \frac{\text{TM}}{\text{TM} + \text{FM}}, \quad (29)$$

based on the number of true matches (TM) and the number of false matches (FM) in each sequence.

7.1.2 Data Association Methods

We evaluate the performance of data association methods including some variants of the SCEA method. The first method utilizes the constant velocity motion model (i.e., linear motion (LM)) introduced in (24), and its data association is carried out without the structural constraints. We apply the Hungarian method (HM) to solve the data association and name it as the LM-HM method. The second method utilizes the relative motion network (RMN) (Yoon et al. 2015), which describes the relative distance between objects as the structural constraint information. This method assumes the smooth camera motion, and the prediction of well-tracked objects based on their own past motion information is used. Therefore, well-tracked objects have two kinds of predictions from its own motion and other objects with their structural constraints. For the case of mis-detected objects, it only exploits the predictions from well-tracked objects with their structural constraints. This method, referred to as the RMN-HM method, also uses the linear motion model for prediction, and the data association is conducted based on the Hungarian method. We also evaluate the SCEA method based on different solutions, i.e., SCEA with partitioning (the SCEA-P) and the SCEA with the exhaustive combinatorial enumeration (the SCEA-E). In addition, to evaluate the effect of the number of objects in each subgroup, we change the number of object in one subgroup, c , from 2 to 5 (2, 3, 4, 5). Those variants are referred to as the SCEA-E2, SCEA-E3, SCEA-E4, and SCEA-E5 method, respectively.

7.1.3 Comparison of Different Data Association Methods

The performance of different data association methods is presented in Figs. 9 and 10. For the baseline LM-HM, data is associated without using structural constraints. As can be seen, the LM-HM method achieves lower accuracy than other

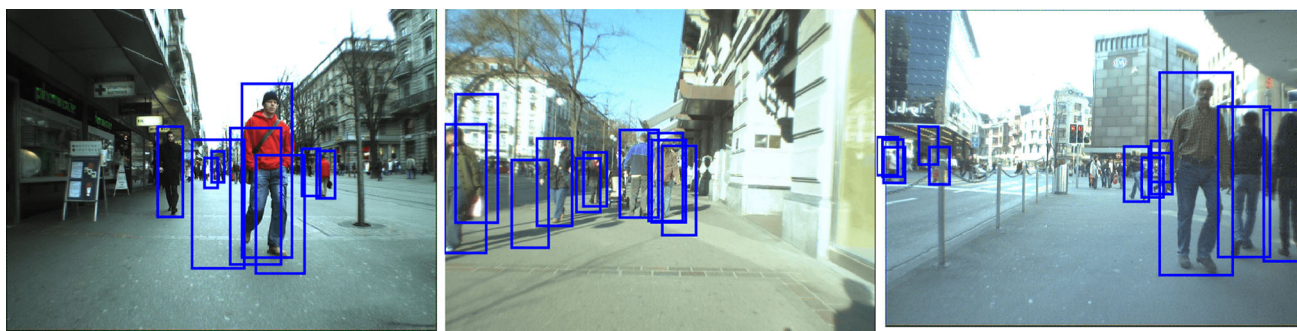


Fig. 8 Examples of the synthetic dataset: the dataset was generated based on the ground truth of the ETH and the KITTI datasets. We applied the different levels of motion fluctuation noises and detection missing rates. In addition, we included at most 10 false positives for each frame

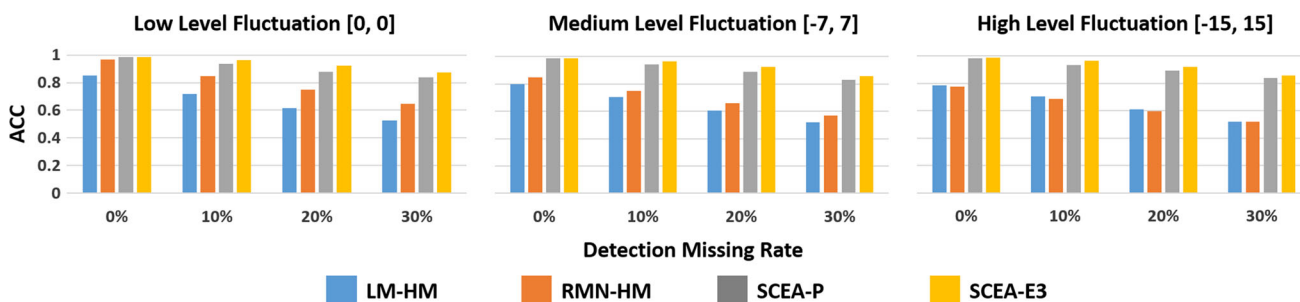


Fig. 9 Matching accuracy (ACC) of the LM-HM, RMN-HM, SCEA-P, and SCEA-E3 under the different levels of camera motion fluctuation and detection missing rates. The numbers $([0, 0], [-7, 7], [-15, 15])$

represent the range of camera motion fluctuation noise in terms of pixels. The missing rate of the detection is set to 0, 10, 20, and 30%. We analyze the performance of each method in detail in Sect. 7.1

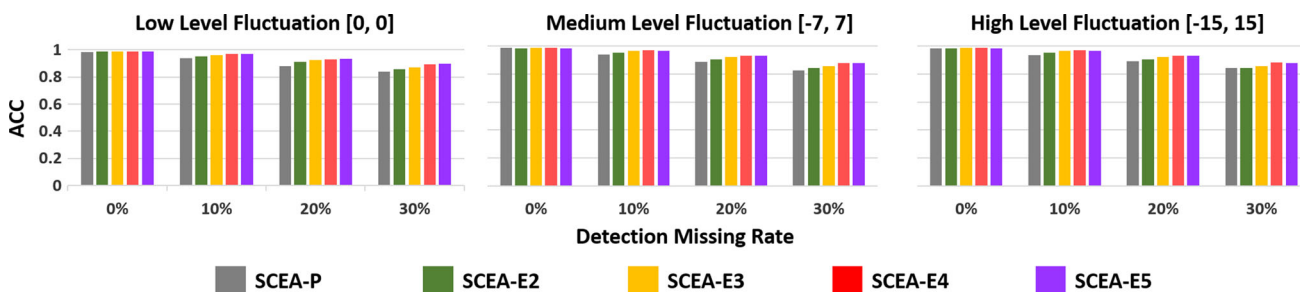


Fig. 10 Matching accuracy (ACC) of the SCEA-P, SCEA-E2, SCEA-E3, SCEA-E4, and SCEA-E5 under the different levels of camera motion fluctuation and detection missing rates. The numbers $([0, 0], [-7, 7], [-15, 15])$ represent the range of camera motion fluctuation

noise in terms of pixels. The missing rate of the detection is set to 0, 10, 20, and 30%. We analyze the performance of each method in detail in Sect. 7.1

methods for the low level fluctuation and 0% missing rate. This can be attributed to that the sequences from the ETH and the KITTI datasets contain significant camera panning motions. For such scenes, the LM-HM method is less accurate as shown in Table 1. In addition, as the missing rate increases, the accuracy of the LM-HM method decreases more than the SCEA method as it is not able to handle ambiguities caused by the uncertain detections. In contrast, the event aggregation method with the structural constraints handles such scenes effectively. Actually, the RMN-HM (or RMOT) method performs well in scenes with low-level fluctuation

in which motion can be predicted with linear models. However, this method does not perform well for scenes with large fluctuations (e.g, fluctuations $[-7, 7]$ and $[-15, 15]$) as the prediction based on structural constraints is not accurate due to self-motion of well-tracked objects.

Figure 9 demonstrates that the SCEA-P and the SCEA-E3 methods perform better than the LM-HM and the RMN-HM methods since the object motion is not used for data association. In addition, these methods efficiently handle ambiguities by aggregating the costs of the same events based on different anchor assignments. As such, these methods

Table 1 Motion accuracy in the data association

Motion type	20	15	10	5
<i>(a) Static or smooth camera motion</i>				
Self motion (Kalman prediction)	0.91	0.88	0.80	0.49
Structural constraint	0.89	0.83	0.74	0.44
<i>(b) Large camera motion</i>				
Motion type	20	15	10	5
Self motion (Kalman prediction)	0.51	0.34	0.17	0.05
Structural constraint	0.87	0.80	0.62	0.26

The numbers (20, 15, 10, and 5) represent the pixel error and the values in the table represent the ratio of the number of pixels whose errors are smaller than the given pixel error. Self motion represents the motion prediction using object's own motion information

perform well even for scenes with high detection missing rate.

Among the SCEA-E3 and SCEA-P methods, the SCEA-E3 method performs more robustly than the SCEA-P method as the fluctuation level and the missing rate increase. The SCEA-P method divides the objects into subgroups based on the spatial information, and the solution of each subgroup may be sub-optimal. In contrast, the SCEA-E3 method exhaustively exploits all possible combinations of objects when each subgroup consists of three objects, and the matching results from all subgroups are summed together. Thus, it can alleviate the effect of the false matching and the local minima problem. However, as the number of objects increases, its computational complexity increases factorially.

We also evaluate the performance of the SCEA method with the different numbers of objects in each subgroup. As shown in Fig. 10, the SCEA method with the different numbers of objects in each subgroup achieve similar performance. In addition, we evaluate the effect of fluctuation noise on these methods. Overall, these methods perform well under different levels of fluctuations. Similar to the results with different missing rates, these methods perform consistently under the different fluctuation levels. However, the performance slightly increases as the number of objects increases. It indicates that the partitioning method generates more locally optimized solutions with the high missing rate.

7.1.4 Speed

All of methods are implemented in MATLAB and the speed is evaluated on single-CPU machine according to the different numbers of objects as shown in Table 2. The LM-HM method performs very fast, and the RMN-HM method also operates in real-time as not all possible associations are considered for data association. The SCEA-P method is more efficient than the SCEA-E method because it uses local spatial partitions. The run time of the SCEA-P method increases in proportion

to the number of objects similar to the LM-HM and RMN-HM methods. Although the run time of the SCEA-E method largely grows as the number of objects increases, the run time can be reduced by using a parallel processor such as GPU or other multi-core processors where each subgroup is solved independently. As shown in Table 3, each subgroup of the SCEA-E3 can be executed in approximately 0.0086 s on average. However, when considering both accuracy and computational complexity, we can say that the SCEA-P method performs more favorably than the SCEA-E method for online MOT as shown in Fig. 10 and Table 2.

7.2 MOT Evaluation Metrics

We adopt the widely used metrics introduced in Bernardin and Stiefelhagen (2008), Li et al. (2009) for performance evaluation. The Multiple Object Tracking Accuracy (MOTA \uparrow) metric shows comprehensive MOT performance by considering false positives, false negatives, and mis-matches over all frames. The Multiple Object Tracking Precision (MOTP \uparrow) metric measures the total error in estimated position for matched pairs over all frames. The mostly tracked (MT \uparrow) targets ratio is the ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span, and the mostly lost (ML \downarrow) targets ratio is the ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span. The fragment (FG \downarrow) metric presents the total number of times that the generated trajectory is fragmented, and the identity switch (ID \downarrow) metric shows the total number of times that the identity of a tracked trajectory switches. The false positive (FP \downarrow) metric presents the number of false positives, and the false negative (FN \downarrow) metric shows the number of missed objects. The Recall (Rec \uparrow) metric measures the rate of correctly tracked objects over the entire sequence based on the ground truth, and the Precision (Prec \uparrow) metric shows the rate of correctly tracked objects over all tracking results of a sequence. The runtime (in second (sec.) \downarrow) or speed (in frames per second (fps) \uparrow) is also considered as a metric.

7.3 Comparisons on Benchmark Datasets

We use three benchmark datasets, ETH,² KITTI,³ and MOT Challenge⁴ datasets, for evaluation and comparison. We note that it is difficult to evaluate MOT systems fairly and thoroughly for the following reasons. First, as most MOT methods consist of several modules in a complicated way, it is difficult to evaluate the performance of each module and the effect of each module on the overall performance

² iris.usc.edu/people/yangbo/downloads.html.

³ cvlibs.net/datasets/kitti/eval_tracking.php.

⁴ motchallenge.net/data/2D_MOT_2015/.

Table 2 Speed (fps) of different data association methods

	# of objects	10	11	12	13	14	15	16	17
Method	LM-HM	474	422	357	261	257	218	138	67
	RMN-HM	209	197	188	136	127	110	98	77
	SCEA-P	17.6	21.1	20.4	16.4	15.1	14.6	11.6	10.0
	SCEA-E2	2.33	2.03	1.75	1.55	1.09	1.05	0.88	0.75
	SCEA-E3	0.83	0.75	0.63	0.35	0.28	0.23	0.22	0.16
	SCEA-E4	0.57	0.36	0.21	0.17	0.10	0.09	0.06	0.04
	SCEA-E5	0.40	0.24	0.14	0.08	0.05	0.03	0.02	0.01

Table 3 Speed (fps) of the SCEA-E3 method

# of objects	10	11	12	13	14	15	16	17
# of combination	120	165	220	286	364	455	560	680
Speed (fps) on each subgroup	100	124	138	100	102	105	123	115
Overall speed	0.83	0.75	0.63	0.35	0.28	0.23	0.22	0.16

directly. Second, MOT methods are evaluated on different metrics and the source codes are often not available. In this section, we focus on analyzing how the proposed data association method facilitates MOT methods in comparison with other alternatives.

7.3.1 ETH Datasets

The SCEA method is evaluated on the ETH dataset (Bahnhof and Sunday) that was recorded by using moving cameras mounted on a mobile platform. The results of the other trackers on both sequences are available in their original papers. As shown in Table 4, our method shows the performance comparable to those of the state-of-the-art methods even though it is an online method based on a simple appearance model. The KalmanSFM and LPSFM methods use social force models, which consider pairwise motion such as attraction and repulsion, and require visual odometry to obtain 3D motion information in the bird's-eye-view maps. The 3D motion information is estimated by the structure-from-motion algorithm. However, the use of odometry information does not alleviate the problem of accumulated motion errors, which adversely affects the effectiveness of the social force models. The MotiCon method shows slightly better performance in terms of Recall, MT, and ML because it utilizes pre-trained motion models in the offline framework. The OnlineCRF method shows good performance in ID and FG because it exploits pairwise motion and the future motion information together in the offline tracking framework. Although our method conducts online 2D MOT without using any pre-trained motion models and 3D motion information, it demonstrates comparable performance against the MotiCon method and shows better Rec, Prec, MT, and ML than the KalmanSFM, LPSFM, OnlineCRF, and CEM methods while keeping lower ID and FG than the KalmanSFM, LPSFM,

MotiCon, and CEM methods. As demonstrated in Sect. 7.1.3, the proposed data association with the event aggregation reduces the mis-matches caused by false positives or other different objects. As a result, it improves Prec. In addition, the SCEA helps to re-track missing objects when they are re-detected via the SCOR. This reduces FG and ID effectively. Overall, the SCEA method performs favorably against other online trackers (e.g., StructMOT, MOT-TBD, and RMOT).

7.3.2 KITTI Datasets

The KITTI dataset contains 29 sequences (Geiger et al. 2013). The datasets contain test sequences from a static camera as well as from a dynamic camera. Note that, since this work focuses on 2D MOT with a single camera, we did not use any other information from stereo images, camera calibration, depth maps, or visual odometry. In addition, we utilize the same detections used for other methods in all experiments for fair comparison. The KITTI datasets provide two sets of detections, one from the DPM (Felzenszwalb et al. 2010) and the other from the Regionlet detector (Wang et al. 2015). The Regionlet detector generates more detections with higher recall than the DPM (Wang et al. 2015), as illustrated on the KITTI website.

We compare the SCEA method with other online MOT methods and offline MOT methods including a semi-online method (e.g., NOMT) in Tables 5 and 6. Here, the online methods produce the solution instantly at each frame by a causal approach. The offline and semi-online MOT methods utilize future frame information in tracking.

Since the offline MOT methods utilize future frame information, they can efficiently resolve the matching ambiguities compared to the online methods including our method. In addition, they can generate longer trajectories by linking tracklets in the data association. Hence, the offline trackers

Table 4 Comparisons to the KalmanSFM and LPSFM methods on the ETH datasets (Bahnhof + Sunnyday)

Setting	Method	Rec ↑	Prec ↑	MT ↑	ML ↓	ID ↓	FG ↓
Offline	*KalmanSFM	72.3	84.1	51.6	5.6	77	206
	*LPSFM	74.1	75.3	55.1	7.9	131	184
	OnlineCRF	79.0	90.4	68.0	7.2	11	19
	MotiCon	83.8	79.7	72.0	4.7	71	85
	CEM	77.3	87.2	66.4	8.2	57	69
Online	StructMOT	78.4	84.1	62.7	7.7	72	5
	MOT-TBD	78.7	85.5	62.4	8.0	69	45
	RMOT	81.5	86.3	67.7	4.8	38	40
	SCEA	82.5	89.6	71.1	5.6	24	32

The results of the other methods are from the original publications, and those of Pellegrini et al. (2009), Leal-Taixé et al. (2011) are from Leal-Taixé et al. (2014). The results obtained by using 3D motion information are marked by “*”. We include the average ranking (AR) metric computed by averaging all metric rankings. Trackers: KalmanSFM (Pellegrini et al. 2009), LPSFM (Leal-Taixé et al. 2011), OnlineCRF (Yang and Nevatia 2014), MotiCon (Leal-Taixé et al. 2014), CEM (Milan et al. 2014), StructMOT (Kim et al. 2012), MOT-TBD (Poiesi et al. 2013), RMOT (Yoon et al. 2015)

Table 5 Comparison to the state-of-the-art trackers on the KITTI car tracking

Setting	Method	MOTA	MOTP	Rec	Prec	MT	ML	FP	FN	ID	FG	Sec (core)
<i>(a) Results on the DPM detections</i>												
Offline	TBD	55.07	78.35	56.72	99.30	20.46	32.62	141	15281	31	529	10 (1)
	SSP	57.85	77.64	59.88	98.76	29.38	24.31	266	14222	7	704	0.6 (1)
	CEM	51.94	77.11	55.96	96.09	20.00	31.54	807	15598	125	396	0.09 (1)
	NOMT	66.60	78.17	69.21	98.05	41.08	25.23	492	10981	13	150	0.09 (16)
	LP_S SVM	61.77	76.93	64.78	97.67	35.54	21.69	551	12581	16	422	0.06 (1)
Online	HM	43.85	78.34	45.03	99.67	12.46	39.54	53	19247	12	571	0.01 (1)
	mbodSSP	56.03	77.52	58.52	97.81	23.23	27.23	463	14659	0	699	0.01 (1)
	NOMT-HM	61.17	78.65	64.30	97.83	33.85	28.00	512	12813	28	241	0.09 (8)
	RMOT	52.42	75.18	57.72	93.72	21.69	31.85	1367	14947	50	376	0.01 (1)
	SCEA	57.03	78.84	58.63	99.48	26.92	26.62	109	14653	17	461	0.05 (1)
<i>(b) Results on the Regionlet detections</i>												
Offline	DCO_X	68.11	78.85	78.67	91.99	37.54	14.15	2588	8063	318	959	0.9 (1)
	SSP	72.72	78.55	82.69	92.57	53.85	8.00	2548	6648	185	932	0.6 (1)
	NOMT	78.15	79.46	83.22	96.78	57.23	13.23	1061	6421	31	207	0.09 (16)
	LP_S SVM	77.63	77.80	83.35	96.27	56.31	8.46	1239	6393	62	539	0.01 (1)
Online	mbodSSP	72.69	78.75	80.61	94.10	48.77	8.77	1918	7360	114	858	0.01 (1)
	NOMT-HM	75.20	80.02	80.99	96.45	50.00	13.54	1143	7280	105	351	0.09 (8)
	RMOT	65.83	75.42	80.58	88.09	40.15	9.69	4148	7396	209	727	0.02 (1)
	SCEA	75.58	79.39	81.76	96.00	53.08	11.54	1306	6989	104	448	0.06 (1)

Trackers: TBD (Geiger et al. 2014), SSP (Lenz et al. 2015), CEM (Milan et al. 2014), NOMT/NOMT-HM (Choi 2015), LP_S SVM (Wang and Fowlkes 2015), HM (Geiger 2013), mbodSSP (Lenz et al. 2015), RMOT (Yoon et al. 2015), DCO_X (Milan et al. 2013)

generally result in higher MT and lower ML compared to the online trackers. Since the SCEA is also an online method, it also shows limited performance in generating long object trajectories. It is natural because the SCEA cannot fill out the mis-tracked frames using the future frame information. However, even though the SCEA does not use future frame information in the data association, the SCEA shows competitive accuracy in the data association — the SCEA generates

less false positives (FP) in comparison with offline trackers, and it also improves Prec which represents the ratio of true positives over the sum of false positives and true positives. It is because the aggregation step reduces the matching ambiguities caused by uncertain detections as illustrated by the experiments with the low level fluctuation in Fig. 9.

As shown in Table 5, we compare the SCEA with the HM, RMOT, and mbodSSP on the car tracking sequences. Differ-

Table 6 Comparison to the state-of-the-art trackers on the KITTI pedestrian tracking

Setting	Method	MOTA	MOTP	Rec	Prec	MT	ML	FP	FN	ID	FG	Sec (core)
<i>(a) Results on the DPM detections</i>												
Offline	CEM	27.54	68.48	36.73	80.82	8.93	51.89	2020	14658	96	608	0.09 (1)
	NOMT	36.93	67.75	46.96	82.91	17.87	42.61	2248	12318	34	789	0.09 (16)
	LP_S SVM	33.33	67.38	42.98	82.44	12.37	45.02	2126	13235	72	818	0.06 (1)
Online	NOMT-HM	27.49	67.99	37.30	80.09	15.12	50.52	2153	14559	73	732	0.09 (8)
	RMOT	34.54	68.06	43.94	83.16	14.43	47.42	2065	13008	81	685	0.01 (1)
	SCEA	33.13	68.45	40.32	85.27	9.62	46.74	1616	13849	16	717	0.05 (1)
<i>(b) Results on the Regionlet detections</i>												
Offline	NOMT	46.62	71.45	55.25	87.33	26.12	34.02	1867	10427	63	666	0.09 (16)
	LP_S SVM	43.76	70.48	53.79	84.94	20.62	34.36	2228	10718	73	809	0.01 (1)
Online	NOMT-HM	39.26	71.14	50.38	83.33	21.31	41.92	2355	11523	184	863	0.09 (8)
	RMOT	43.77	71.02	53.64	85.75	19.59	41.24	2075	10790	153	748	0.01 (1)
	SCEA	43.91	71.86	49.52	90.69	16.15	43.30	1183	11746	56	641	0.06 (1)

Trackers: CEM (Milan et al. 2014), NOMT/NOMT-HM (Choi 2015), LP_S SVM (Wang and Fowlkes 2015), RMOT (Yoon et al. 2015)

ent from the SCEA, those methods assume small motion changes of objects, and the car tracking sequences contain frequent unexpected camera motion. When the data association fails due to large motion, the trackers generate more fragments (FG) of object trajectories. Thus, the mbodSSP and the HM show larger FG than the SCEA. Different from them, the RMOT contains a recovery step to re-track missing objects. Hence, it can suppress FG, but due to matching ambiguities caused by false positives, it generates larger ID compared to the SCEA. This trend is also shown in Sect. 7.1.3 with Fig. 9. Due to the inaccurate data association, those trackers show lower Prec in comparison with the SCEA. The SCEA can suppress FG because its data association utilizes the structural constraints in the data association without objects' own motion information, which enables the SCEA to deal with the unexpected camera motions in the data association as validated in Fig. 9 and Table 1. Moreover, the SCEA can further reduce FP because the aggregation method of the data association reduces matching ambiguities caused by uncertain detections. Such characteristic of the SCEA more improves Prec. Qualitative comparisons with the LM-HM and the RMOT are given in Figs. 11, 12, 13, 14. Here, the LM-HM is similar to the HM because both methods utilizes the linear motion model to predict object locations and the Hungarian algorithm for the data association.

Different from other online trackers including the SCEA, the NOMT-HM utilizes optical flow information to extract more explicit object motion information. Thus, it can estimate longer object trajectories and produce greater MT and Rec. Since the NOMT-HM exploits additional information that is not used in other online MOT methods, it is difficult to compare the performance directly. However, we can compare the data association accuracy by comparing the metrics such

as Prec, FP, ID, and FG. These metrics are closely related to the data association accuracy. Due to optical flow information, the NOMT-HM generates less fragments, resulting in small FG. In terms of Prec, FP, and ID, we can find that, even without the optical flow information, the proposed data association shows competitive performance. For the KITTI pedestrian sequences, the SCEA achieves better performance in MOTA and Prec metrics compared to the NOMT-HM. This is because the optical flow information from pedestrians is less reliable compared to that in the car sequences owing to the small size and non-rigid appearance of a pedestrian. In addition, the motion cue (the optical flow) becomes less discriminative when the motion of an object is small. In the KITTI pedestrian tracking dataset, the motion of pedestrians is much smaller than that of cars, and the most of sequences were recorded by a static camera. Since the SCEA method extracts structural motion information only from detections, its performance is less dependent on the object size, appearance, and the magnitude of motion.

The large performance difference of the SCEA according to different detections (DPM and Regionlet) can be explained by the fact that better recall of detections can improve the SCEA considerably. In general, the SCEA is more robust to missing detections than the RMOT, and the performance differences become more obvious when large camera motions occur and better detections are given (as illustrated in Fig. 9, results under 0 or 10% missing rate under the high-level fluctuation). This is because the proposed data association handles large camera motions more effectively, and a high detection rate makes the structural constraints more accurate thanks to the continuous update of the structural constraints. Since the KITTI car dataset contains large camera motion more than the other datasets, the performance gap between

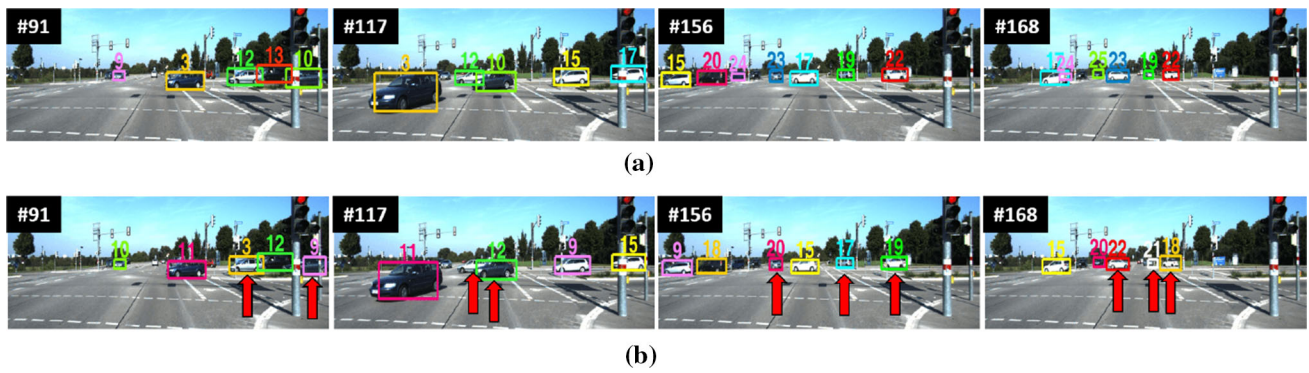


Fig. 11 Qualitative comparisons on the KITTI0011 sequence. **a** SCEA. **b** LM-HM: the baseline method: object 3 at #91 is missed at #117; the label of object 9 at #91 is changed at #117 (9→12); the labels of objects 17, 19 and 20 at #156 are changed at #168 (17→21, 19→18, 20→22)

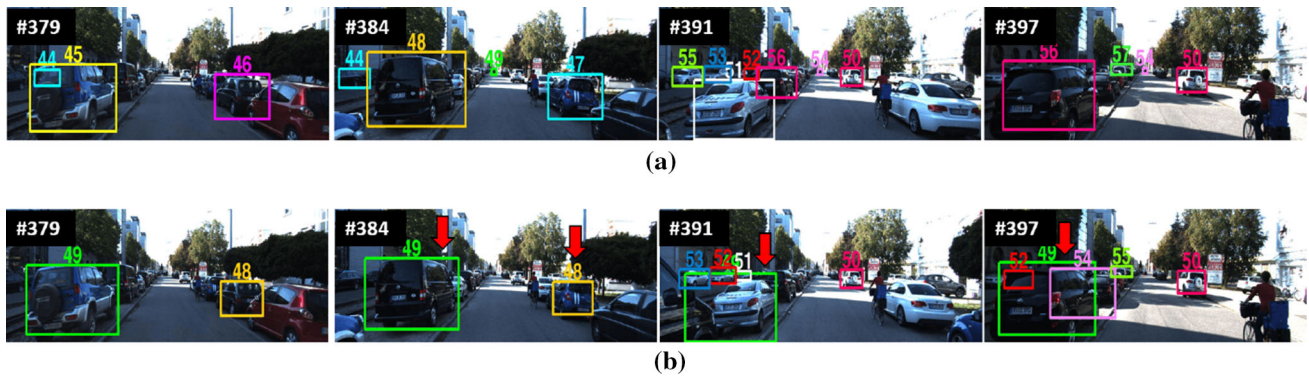


Fig. 12 Qualitative comparisons on the KITTI0014 sequence. **a** SCEA. **b** LM-HM: the label of object 48 at #379 is assigned to a new object at #384; the label of object 49 at #379 is assigned to new objects at #384, #391, and #397



Fig. 13 Qualitative comparisons on the KITTI0011 sequence. **a** SCEA. **b** RMOT (or RMN-HM): the label of object 16 at #78 is changed at #86 (16→10). The labels of object 2 and 10 at #86 are changed at #89 (2→20, 10→19). The label of object 18 at #86 is incorrectly assigned to a new object at #89

the SCEA and the other methods becomes more obvious, especially with the Regionlet detector whose recall performance is much better than that of the DPM. In addition, using the detector with high recall, the performance gap between the SCEA and the offline trackers are alleviated as shown in Table 5.

7.3.3 MOTC Datasets

The MOTC datasets provide more uncertain detections with lower recall compared to the detections of the KITTI dataset. As discussed in Sect. 7.3.2, the offline trackers generally show better performance than the online MOT trackers. However, due to the uncertain detections, some of offline trackers show lower performance than the online trackers.

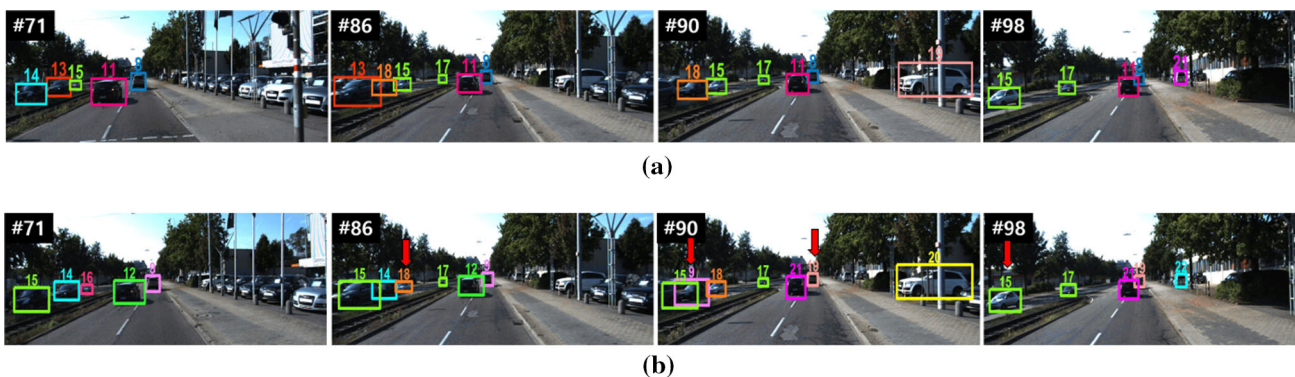


Fig. 14 Qualitative comparisons on the KITTI0014 sequence. **a** SCEA. **b** RMOT (or RMN-HM): the label of object 16 at #71 is changed at #86 (16→18). The label of object 9 at #86 is changed at #90 (9→19). The label of object 9 at #86 is incorrectly assigned to object 15 at #90. The label of object 15 at #90 is incorrectly assigned to object 18 at #98

Table 7 Comparison to the online trackers on the MOT Challenge dataset (pedestrian sequences)

Setting	Method	MOTA	MOTP	FAF	MT	ML	FP	FN	ID	FG	Hz (core)
Offline	CEM	19.3	70.7	2.5	8.5	46.5	14,180	34,591	813	1023	1.1 (1)
	DCO_X	19.6	71.4	1.8	5.1	54.9	10,652	38,232	521	819	0.3 (1)
	SegTrack	22.5	71.7	1.4	5.8	63.9	7890	39,020	697	737	0.2 (1)
	MotiCon	23.1	70.9	1.8	4.7	52.0	10,404	35,844	1018	1061	1.4 (1)
	JPDA_m	23.8	68.2	1.1	5.0	58.1	6373	40,084	365	869	32.6 (1)
	ELP	25.0	71.2	1.3	7.5	43.8	7345	37,344	1396	1804	5.7 (1)
	LP_S SVM	25.2	71.7	1.4	5.8	53.0	8369	36,932	646	849	41.3 (1)
	SiameseCNN	29.0	71.2	0.9	8.5	48.4	5160	37,798	639	1316	52.8 (1)
	MHT_DAM	32.4	71.8	1.6	16.0	43.8	9064	32,060	435	826	0.7 (1)
	NOMT	33.7	71.9	1.3	12.2	44.0	7762	32,547	442	823	11.5 (16)
	Online	TC_ODAL	15.1	70.5	2.2	3.2	55.8	12,970	38,538	637	1716
RMOT		18.6	69.6	2.2	5.3	53.3	12,473	36,835	684	1282	7.9 (1)
NOMT-HM		26.7	71.5	2.0	11.2	47.9	11,162	33,187	637	1716	11.5 (16)
MDP		30.3	71.3	1.7	13.0	38.4	9717	32,422	680	1500	1.1 (8)
SCEA		29.1	71.1	1.0	8.9	47.3	6060	36,912	604	1182	6.8 (1)

We include the average number of false alarms per frame (FAF) as a metric. [The results of the NOMT-HM method are from the original paper (Choi 2015)]. Trackers: CEM(Milan et al. 2014), DCO_X(Milan et al. 2013), SegTrack(Milan et al. 2015), MotiCon(Leal-Taixé et al. 2014), (JPDA_m Rezatofighi et al. 2015), ELP(McLaughlin et al. 2015), LP_S SVM(Wang and Fowlkes 2015), SiameseCNN(Leal-Taixe et al. 2016), MHT_DAM(Kim et al. 2015), NOMT/NOMT-HM(Choi 2015), TC_ODAL(Bae and Yoon 2014), RMOT(Yoon et al. 2015), MDP(Xiang et al. 2015)

Table 8 Comparison to the MDP method on the KITTI training dataset

	MOTA ↑	MOTP ↑	Rec ↑	Prec ↑	MT ↑	ML ↓	ID ↓	FG ↓
<i>(a) Car</i>								
MDP-KITTI	55.0	75.1	60.8	92.3	10.7	40.9	19	118
SCEA	58.8	78.6	61.3	96.5	11.6	32.9	6	100
<i>(b) Pedestrian</i>								
MDP-KITTI	23.8	71.2	49.1	66.4	3.5	36.0	8	204
MDP-MOTC	25.1	71.2	47.8	68.6	3.5	34.9	32	209
SCEA	35.4	73.2	51.5	76.3	7.0	32.6	3	154

The MHT_DAM constructs object trajectories by considering all possible data associations throughout the entire sequence. Therefore, it can effectively reduce the incorrect matching between objects and detections and generate long trajectories. The NOMT utilizes the optical flow information in a semi-online manner and exploits the future frame information up to 30 frames. The SiameseCNN uses deep-features to model the appearance of objects. These offline trackers fill out the missing frames in the trajectory using the future frame information, which helps to improve MT and reduce FN. In addition, with some additional cues such as optical flow information and deep-features, the future frame information much reduces the matching ambiguities caused by uncertain detections. As a result, they generate a small number of false positives. Although the SCEA does not use the future frame information, it shows comparable performance in terms of MOTA and FP. It means that the aggregation method of the proposed data association is robust to uncertain detections, and the proposed data association reduces other mis-matches such as ID and FG further compared to the data association of other online MOT methods (Table 7).

The MDP shows better performance in MOTA, MT, ML, and FN metrics compared to the SCEA. This is because the MDP learns the target state (Active, Tracked, Lost, and Inactive) from a training dataset and its ground truth in an online manner. Therefore, it can initialize and terminate the objects more robustly than the other methods. In addition, owing to the use of optical flow information for local template tracking, the MDP generates longer trajectories compared to the other online methods. However, the SCEA has some advantages over the MDP. First, the proposed data association generates lower FP, ID, and FG. It means that the proposed data association is more robust to uncertain detections. Second, the SCEA does not require any training datasets and it runs much faster because it does not conduct template tracking based on dense optical flow. To show the performance dependency on the training dataset, we compare the SCEA with the MDP on the KITTI dataset. For the pedestrian sequences, we run the MDP with the original trained model provided with the original source code by the authors (MDP-MOTC). In addition, we also train the MDP method with the KITTI training dataset for car sequences (MDP-KITTI). As shown in Table 8, the performance of the MDP depends on the training dataset. Note that the performance of the MDP can be improved further if more training datasets are used.

8 Conclusion

In online 2D MOT with moving cameras, observable motion cues are complicated by global camera movements and, thus, are not always smooth or predictable. In this paper, we proposed a new data association method that effectively exploits

structural constraints in the presence of unexpected camera motion and uncertain detections. In addition, to further alleviate data association ambiguities caused by uncertain detections, e.g., mis-detections and multiple false positives, we developed a novel event aggregation method to integrate the structural constraints in assignment event costs. The proposed data association and structural constraints are incorporated into the online 2D MOT framework, which simultaneously tracks objects and recovers missing objects with structural constraints. Experimental results on a large number of datasets demonstrated and validated the effectiveness of the proposed algorithm for online 2D MOT.

Acknowledgements This work was partially supported by ‘The Cross-Ministry Giga KOREA Project’ grant funded by the Korea government (MSIT) (No. GK18P0200, Development of 4D reconstruction and dynamic deformable action model based hyper-realistic service technology and No. GK18P0300, Real-time 4D reconstruction of dynamic objects for ultra-realistic service). The work was also partially supported by IITP grant funded by the Korea government (MSIP) (2014-0-00059) and Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-TC1603-05. M.-H. Yang is supported in part by the the NSF CAREER Grant #1149783.

References

- Bae, S. H., Yoon, K. J. (2014). Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1218–1225).
- Bar-Shalom, Y., & Li, X. R. (1995). *Multitarget-multisensor tracking: Principles and techniques*. Storrs, CT: YBS Publishing.
- Bernardin, K., & Stiefelwagen, R. (2008). Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 1(1–1), 10.
- Betke, M., & Wu, Z. (2016). Data Association for multi-object visual tracking. Synthesis lectures on computer vision. Morgan & Claypool. <https://books.google.co.kr/books?id=tn0cvgAACAAJ>.
- Blackman, S., & Popoli, R. (1999). *Design and analysis of modern tracking systems*. Norwood: Artech House Radar Library, Artech House.
- Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., & Van Gool, L. (2011). Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9), 1820–1833.
- Choi, W. (2015). Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of IEEE international conference on computer vision* (pp. 3029–3037).
- Dollar, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8), 1532–1545.
- Duan, G., Ai, H., Cao, S., & Lao, S. (2012). Group tracking: Exploring mutual relations for multiple object tracking. In *Proceedings of European conference on computer vision* (pp. 129–143).
- Ess, A., Leibe, B., Schindler, K., & Gool, L. V. (2008). A mobile vision system for robust multi-person tracking. In *Proceedings of IEEE conference on computer vision and pattern recognition*.
- Everingham, M., Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2), 303–338.

- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 1627–1645.
- Geiger, A. (2013). Probabilistic models for 3D urban scene understanding from movable platforms. Ph.D. thesis, KIT.
- Geiger, A., Lauer, M., Wojek, C., Stiller, C., & Urtasun, R. (2014). 3D traffic scene understanding from movable platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5), 1012–1025.
- Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11), 1231–1237.
- Grabner, H., Matas, J., Gool, L. J. V., & Cattin, P. C. (2010). Tracking the invisible: Learning where the object might be. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1285–1292).
- Kim, C., Li, F., Ciptadi, A., & Rehg, J. M. (2015). Multiple hypothesis tracking revisited. In *Proceedings of IEEE international conference on computer vision* (pp. 4696–4704).
- Kim, S., Kwak, S., Feyereisl, J., & Han, B. (2012). Online multi-target tracking by large margin structured learning. In *Proceedings of Asian conference on computer vision* (pp. 98–111).
- Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics*, 2(1–2), 83–97.
- Leal-Taixe, L., Canton-Ferrer, C., & Schindler, K. (2016). Learning by tracking: Siamese CNN for robust target association. In *Proceedings of IEEE conference on computer vision and pattern recognition workshops* (pp. 33–40).
- Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., & Savarese, S. (2014). Learning an image-based motion context for multiple people tracking. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 3542–3549).
- Leal-Taixé, L., Pons-Moll, G., & Rosenhahn, B. (2011). Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *Proceedings of IEEE international conference on computer vision workshop*
- Lenz, P., Geiger, A., & Urtasun, R. (2015). Followme: Efficient online min-cost flow tracking with bounded memory and computation. In *Proceedings of IEEE international conference on computer vision* (pp. 4364–4372).
- Li, X. R., & Jilkov, V. P. (2003). Survey of maneuvering target tracking. Part i. Dynamic models. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4), 1333–1364.
- Li, Y., Huang, C., & Nevatia, R. (2009). Learning to associate: Hybrid-boosted multi-target tracker for crowded scene. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 2953–2960).
- McLaughlin, N., del Rincón, J. M., & Miller, P. C. (2015). Enhancing linear programming with motion modeling for multi-target tracking. In *IEEE winter conference on applications of computer vision* (pp. 71–77).
- Milan, A., Leal-Taixé, L., Schindler, K., & Reid, I. (2015). Joint tracking and segmentation of multiple targets. In *Proceedings of IEEE conference on computer vision and pattern recognition*.
- Milan, A., Roth, S., & Schindler, K. (2014). Continuous energy minimization for multitarget tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1), 58–72.
- Milan, A., Schindler, K., & Roth, S. (2013). Detection- and trajectory-level exclusion in multiple object tracking. In *Proceedings of IEEE conference on computer vision and pattern recognition*.
- Pellegrini, S., Ess, A., Schindler, K., & Gool, L. V. (2009). You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of IEEE international conference on computer vision* (pp. 261–268).
- Pirsiavash, H., Ramanan, D., & Fowlkes, C. C. (2011). Globally-optimal greedy algorithms for tracking a variable number of objects. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1201–1208).
- Poiesi, F., Mazzon, R., & Cavallaro, A. (2013). Multi-target tracking on confidence maps: An application to people tracking. In *CVIU* (pp. 257–1272).
- Rezatofighi, H., Milan, A., Zhang, Z., Shi, Q., Dick, A., & Reid, I. (2015). Joint probabilistic data association revisited. In *Proceedings of IEEE international conference on computer vision* (pp. 3047–3055).
- Rezatofighi, S. H., Milan, A., Zhang, Z., Shi, Q., Dick, A., & Reid, I. (2016). Joint probabilistic matching using m-Best solutions. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 136–145).
- Takala, V., Pietikäinen, M. (2007). Multi-object tracking using color, texture and motion. In *Proceedings of IEEE conference on computer vision and pattern recognition*.
- Wang, S., & Fowlkes, C. (2015). Learning optimal parameters for multi-target tracking. In *Proceedings of the British machine vision conference* (pp. 4.1–4.13).
- Wang, X., Yang, M., Zhu, S., & Lin, Y. (2015). Regionlets for generic object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10), 2071–2084.
- Wu, Z., & Betke, M. (2016). Global optimization for coupled detection and data association in multiple object tracking. *Computer Vision and Image Understanding*, 143, 25–37.
- Xiang, Y., Alahi, A., & Savarese, S. (2015). Learning to track: Online multi-object tracking by decision making. In *Proceedings of IEEE international conference on computer vision* (pp. 4705–4713).
- Xing, J., Ai, H., & Lao, S. (2009). Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1200–1207).
- Yang, B., & Nevatia, R. (2012). Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1918–1925).
- Yang, B., & Nevatia, R. (2014). Multi-target tracking by online learning a CRF model of appearance and motion patterns. *International Journal of Computer Vision*, 107(2), 203–217.
- Yoon, J. H., Lee, C. R., Yang, M. H., & Yoon, K. J. (2016). Online multi-object tracking via structural constraint event aggregation. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1392–1400).
- Yoon, J. H., Yang, M. H., Lim, J., & Yoon, K. J. (2015). Bayesian multi-object tracking using motion context from multiple objects. In *Proceedings of IEEE winter conference on applications of computer vision* (pp. 33–40).
- Zhang, L., & van der Maaten, L. (2013). Structure preserving object tracking. In *Proceedings of IEEE conference on computer vision and pattern recognition* (pp. 1838–1845).