Online Multi-Object Tracking via Structural Constraint Event Aggregation

Ju Hong Yoon
KETI
jhyoon@keti.re.kr

Chang-Ryeol Lee
CV Lab., GIST
crlee@gist.ac.kr

Ming-Hsuan Yang
UC Merced
mhyang@ucmerced.edu

Kuk-Jin Yoon
CV Lab., GIST
kjyoon@gist.ac.kr

Abstract

Multi-object tracking (MOT) becomes more challenging when objects of interest have similar appearances. In that case, the motion cues are particularly useful for discriminating multiple objects. However, for online 2D MOT in scenes acquired from moving cameras, observable motion cues are complicated by global camera movements and thus not always smooth or predictable. To deal with such unexpected camera motion for online 2D MOT, a structural motion constraint between objects has been utilized thanks to its robustness to camera motion. In this paper, we propose a new data association method that effectively exploits structural motion constraints in the presence of large camera motion. In addition, to further improve the robustness of data association against mis-detections and false positives, a novel event aggregation approach is developed to integrate structural constraints in assignment costs for online MOT. Experimental results on a large number of datasets demonstrate the effectiveness of the proposed algorithm for online 2D MOT.

1. Introduction

Multi-object tracking (MOT) aims to estimate object trajectories according to the identities in image sequences. Recently, thanks to the advances of object detectors [6, 24], numerous tracking-by-detection approaches have been developed for MOT. In this type of approaches, target objects are detected first and tracking algorithms estimate their trajectories using 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., [22, 27]). On the other hand, online MOT algorithms estimate object trajectories using only detections from the current as well as past frames (e.g. [4]), and online MOT algorithms 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 similar objects (e.g., pedestrians, faces, and vehicles). 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 2D MOT in scenes acquired from moving cameras, observable motion cues are complicated by global camera movements and not always smooth or predictable. In other words, 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 updated to take camera motion into account. While significant advances on batch (or semi-online) trackers have been made (e.g., [5, 14, 20, 28]), online MOT using motion constraints from detection results has not yet been much explored.

In this paper, we propose a new data association method for effectively exploiting the structural motion constraints between objects for *online 2D* MOT, which considers camera motion as well as ambiguities caused by the frequent mis-detections. The structural constraints are represented by the location and velocity differences between objects. Using these 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 mis-detections as shown in Figure 1, we propose the event aggregation approach which considers the structural constraints and assignment events.

We incorporate the proposed data association and the structural constraints into a two-step online 2D MOT framework, which consists of two data association steps. In the first step, by using the proposed structural constraint event

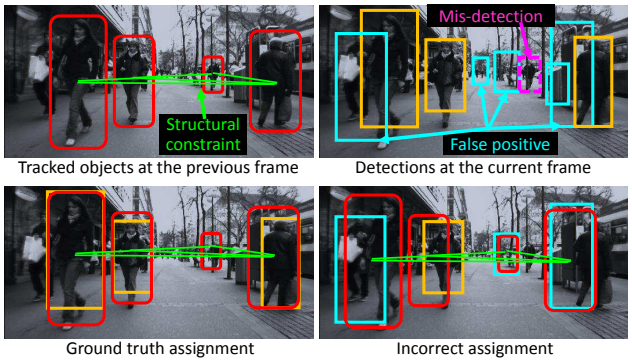

Figure 1. An example of structural constraint ambiguity: The tracked objects and their correct detections are represented by the red box and the yellow box, respectively. The overlap ratio costs of the ground truth assignment and the incorrect assignment are similar due to mis-detections and multiple false positive detections.

aggregation, even under large camera motion or fluctuations, we can robustly estimate continuously tracked objects where structural constraints are sufficiently reliable due to consecutive updates at each frame. 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 from the first step. We demonstrate the merits of the proposed algorithm for online MOT using a large number of challenging datasets.

2. Related Work

We review related MOT methods that utilize the structural motion constraints. Numerous MOT methods directly utilize the first or the second order motion models to locate objects [1, 4, 15]. However, those 2D independent motion models do not work properly under unpredictable camera motion, especially when tracking methods do not exploit the visual information from future frames.

Pellegrini *et al.* [21] and Leal-Taixé *et al.* [18] use social force models which 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.* [13] propose to exploit the relative distance between feature points for single object tracking and reduce tracking drifts caused by drastic appearance changes. In [7], a mutual relation model is proposed to reduce tracking errors when target objects undergo appearance changes. To reduce ambiguities caused by similar appearances in MOT, motion constraints between objects are used along with object appearance models using the structured support vector machines [30]. Unlike the aforementioned methods [7, 13, 30], our method exploits structural constraints to solve the online 2D MOT problem with the frame-by-frame data association that assigns objects to correct detections.

Yang and Nevatia [28] use conditional random field 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.* [29] exploit structural spatial information in terms of relative motion to handle camera motion. This method basically assumes that the camera motion is small and smooth to guarantee that at least a few objects are well predicted and tracked by linear motion models. Different from the aforementioned methods, the proposed method aggregates structural constraints along with assignment events taking abrupt camera motion and ambiguities caused by mis-detections into account for online MOT.

3. Structural Constraint Event Aggregation

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 the state of an object i at frame t as $\mathbf{s}_t^i = [x_t^i, y_t^i, \dot{x}_t^i, \dot{y}_t^i, w_t^i, h_t^i]^\top$ and the set of the states at frame t as $\mathcal{S}_t$ ($\mathbf{s}_t^i \in \mathcal{S}_t$) with its index set $i \in N_t$. Each structural motion constraint is described by the location and velocity difference between two objects as

$$\begin{aligned}\mathbf{e}_t^{i,j} &= [\chi_t^{i,j}, \upsilon_t^{i,j}, \dot{\chi}_t^{i,j}, \dot{\upsilon}_t^{i,j}]^\top \\ &= [x_t^i - x_t^j, y_t^i - y_t^j, \dot{x}_t^i - \dot{x}_t^j, \dot{y}_t^i - \dot{y}_t^j]^\top.\end{aligned} \tag{1}$$

Here, $(\dot{\chi}_t^{i,j}, \dot{\upsilon}_t^{i,j})$ denotes the velocity difference to consider objects moving with different tendency. The set of structural constraints for the object i is represented by $\mathcal{E}_t^i = \{\mathbf{e}_t^{i,j} | \forall j \in N_t\}$, and the set of all structural constraints at frame t is denoted by $\mathcal{E}_t = \{\mathcal{E}_t^i | \forall i \in N_t\}$.

3.1. Structural constraint cost function

The MOT task can be considered as a data association problem, which finds the correct assignment event between objects and detections. In this paper, the assignment event $a^{i,k} \in \mathcal{A}$ describes the state of assignments between objects and detections. If the detection k is assigned to the object i, the assignment is denoted by $\{a^{i,k} = 1\}$. Otherwise, it is denoted by $\{a^{i,k} = 0\}$. For data association, the dissimilarity cost between objects and detections is computed based on the cost function. The best assignment event is then estimated by minimizing total assignment costs. In this section, we introduce a new cost function that considers the structural motion constraints between objects.

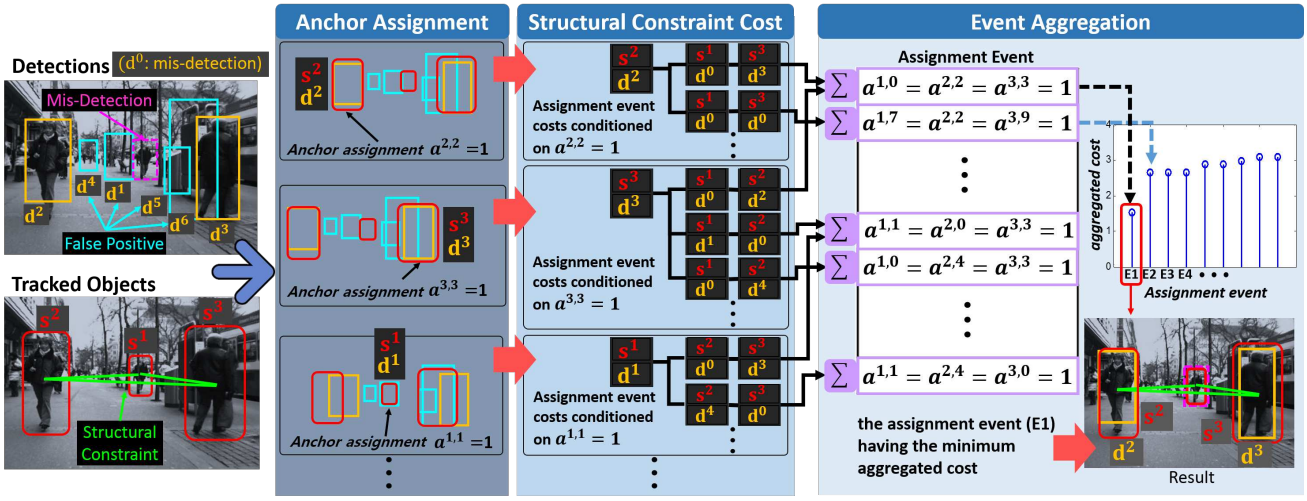

Figure 2. Structural constraint event aggregation (Algorithm 1). The tracked objects and their detections are represented by red boxes and yellow boxes, respectively, and the green lines connecting objects denote structural constraints. Black boxes represent assignments. $\mathbf{d}^0$ stands for the case of mis-detections. As shown in this figure, in the anchor assignment $a^{2,2}$ of the object 2 and the detection 2, we move the object 2 to align the center location of the object 2 with that of the detection 2. Then, in the structural constraint cost, we compute the assignment 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 the different anchor assignment. Σ represents the summation of the structural constraint costs.

We denote a detection k resulting from detectors 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 the detections at frame t used for MOT as $\mathcal{D}_t$ ($\mathbf{d}_t^k \in \mathcal{D}_t$) with its index set as $k \in M_t$. Without loss of generality, we remove the time index t for simplicity in the following sections. Since each object is assigned with at most one detection, the structural constraint cost function with the assignment event $\mathcal{A}$ is described by

$$\hat{\mathcal{A}} = \arg\min_{\mathcal{A}} C(\mathcal{S}, \mathcal{E}, \mathcal{D}),$$
$$\text{s.t.} \sum_{\substack{i\in N \\ k\neq 0}} a^{i,k} \leq 1 \;\wedge \sum_{k\in M\cup\{0\}} a^{i,k} = 1 \;\wedge \sum_{i\in N} a^{i,0} \leq |N|, \tag{2}$$

where each assignment is a binary value $a^{i,k} = \{0,1\}$, $a^{i,k} \in \mathcal{A}$, $k \in M \cup \{0\}$, and $a^{i,0}$ stands for the case of mis-detected objects. Hence, the sum of $a^{i,0}$ along i is equal to the number of objects $|N|$ when all objects are mis-detected.

To deal with large camera motion, we first set an anchor assignment by associating the object i and the detection k as shown in Figure 2. Anchor assignment $a^{i,k}$ makes the center location of the object i coincide with that of the detection k. Based on the anchor assignment and the structural constraint $\mathcal{E}^i$, we conduct all possible assignment events between the remaining objects and detections. By doing this, the structural constraint cost evades the error caused by the global camera motion. Based on this concept, the proposed structural constraint cost function is formulated by

$$C(\mathcal{S}, \mathcal{E}, \mathcal{D}) = \sum_{i\in N}\sum_{k\in M}\Big(a^{i,k}\,\Omega_{i,k} + \sum_{\substack{j\in N \\ j\neq i}}\sum_{\substack{q\in M\cup\{0\} \\ q\neq k}} a^{j,q}\,\Theta_{i,k}^{j,q}\Big), \tag{3}$$

where the subscripts i,k denote the index for costs computed based on the anchor assignment $a^{i,k}=1$, and the cost of the anchor assignment is represented by

$$\Omega_{i,k} = F_s(\mathbf{s}^i, \mathbf{d}^k) + F_a(\mathbf{s}^i, \mathbf{d}^k). \tag{4}$$

Here, we compute the size and appearance costs as

$$F_s(\mathbf{s}, \mathbf{d}) = -\ln\left(1 - \frac{|h-h_d|}{2(h+h_d)} - \frac{|w-w_d|}{2(w+w_d)}\right),$$
$$F_a(\mathbf{s}, \mathbf{d}) = -\ln\sum_{b=1}^{B}\sqrt{p^b(\mathbf{s})p^b(\mathbf{d})}, \tag{5}$$

where (w,h) and (w_d,h_d) denote width and height of an object and a detection, respectively. In addition, $p^n(\mathbf{s})$ and $p^n(\mathbf{d})$ denote the histogram of an object and a detection, respectively. b is the bin index and B is the number of bins. From the anchor position, we calculate the cost of the structural constraint which is described by

$$\Theta_{i,k}^{j,q} = \begin{cases} F_s(\mathbf{s}^j, \mathbf{d}^q) + F_a(\mathbf{s}^j, \mathbf{d}^q) \\ \quad + F_c(\mathbf{s}^j, \mathbf{e}^{j,i}, \mathbf{d}^k, \mathbf{d}^q) & \text{if } q \neq 0 \\ \tau & \text{if } q = 0 \end{cases}, \tag{6}$$

where we empirically set the cost τ to some non-negative value (e.g., 4 in this work) for the case of mis-detected objects, $\mathbf{d}^0$. The constraint cost is formulated by

$$F_c(\mathbf{s}^j, \mathbf{e}^{j,i}, \mathbf{d}^k, \mathbf{d}^q) = -\ln\left(\frac{\text{area}\left(\mathbf{B}(\mathbf{s}^{j,k}) \cap \mathbf{B}(\mathbf{d}^q)\right)}{\text{area}\left(\mathbf{B}(\mathbf{s}^{j,k}) \cup \mathbf{B}(\mathbf{d}^q)\right)}\right), \quad (7)$$
$$\mathbf{s}^{j,k} = [x_d^k, y_d^k, 0, 0]^\top + [\chi^{j,i}, \upsilon^{j,i}, w^j, h^j]^\top.$$

Here, we determine the position of the object j by the position of the detection k and the structural constraint $\mathbf{e}^{j,i}$. The constraint cost is measured by using the overlap ratio [9] of the object bounding box and the detection bounding box to compute a normalized cost since it automatically compensates bias errors caused by the size of objects.

3.2. Event aggregation

Based on the different anchor assignments, we obtain different costs due 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 with the different anchor assignments. Compared to conventional one-to-one matching process for the data association as shown in Figure 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 Figure 2. This aggregation process is described by

$$C(\mathcal{A}) = \sum_{\substack{i \in N, k \in M \\ a^{i,k}=1}} \left(a^{i,k}\Omega_{i,k} + \sum_{\substack{j \in N \\ j \neq i}} \sum_{\substack{q \in M \cup \{0\} \\ q \neq k}} a^{j,q}\Theta_{i,k}^{j,q} \right), \quad (8)$$

where $\mathcal{A} \subset \mathcal{A}_{\text{all}}$ and $\mathcal{A}_{\text{all}}$ denotes all possible assignment events. Finally, we select the best assignment event having the minimum aggregated cost as

$$\hat{\mathcal{A}} = \arg\min_{\mathcal{A}} \left(\frac{C(\mathcal{A})}{\Delta}\right), \Delta = \sum_{i \in N, k \in M} a^{i,k}, \quad (9)$$

where Δ denotes the normalization term that is equal to the number of the anchor assignments from the same assignment event $\mathcal{A}$.

3.3. Assignment event initialization and reduction

Since considering all of assignment events is not computationally efficient, we propose a simple but effective reduction approach. First, we adopt the simple gating technique [2] before conducting the structural constraint event aggregation. This approach is widely used in the MOT literature. We roughly remove the negligible assignments based on two conditions as

$$\left(\|\mathbf{p}^i - \mathbf{p}_d^k\| < \sqrt{(w^i)^2 + (h^i)^2}\right) \wedge \left(\exp\left(-F_s(\mathbf{s}^i, \mathbf{d}^k)\right) > \tau_s\right), \quad (10)$$

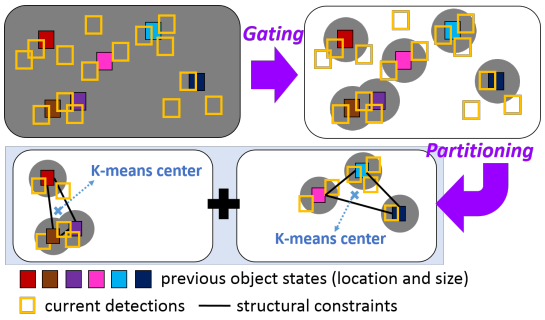

Figure 3. Assignment event reduction concept: The gating and the partitioning reduces the number of assignment events. Gray circles represents the assignment region reduced by the gating. The objects are grouped based on the K-means center.

where $\mathbf{p}^i$ and $\mathbf{p}_d^k$ represent the position of the object i and the detection k, respectively, and (w^i, h^i) denotes the size of the 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, we propose a partitioning approach that splits the structural constraints to handle a large number of objects and detections as shown in Figure 3. The assignments of objects and detections in different paritions are set to $a^{i,k} = 0$. For the partition p, we generate all possible assignment events $\mathcal{A}^p \subset \mathcal{A}_{\text{all}}^p$ based on the condition in (2). The structural constraint event aggregation is carried out for each partition. The final assignment event is then obtained by merging the assignment event results from each partition. 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 splits the partition possibly to have the same number of objects. Here, we use the center location as a partitioning condition. As shown in Figure 3, P K-means centers are obtained, and the objects located close to each K-means center are then gathered in the same partition. Another reduction approach [23] can be alternatively modified and applied to our structural constraint event aggregation. The main steps of the proposed structural constraint event aggregation (SCEA) method is summarized in Algorithm 1.

4. Two-Step Online MOT via SCEA

We adopt a two-step approach for effectively exploiting the structural constraints between objects for online 2D MOT. Since the structural constraints of objects tracked in the previous frame have been also updated with their corresponding detections, their constraints are more robust than mis-detected objects. This allows us to more robustly and accurately assign detections to tracked objects. The overall process of the proposed online MOT via SCEA is described in Algorithm 2.

Data: objects $\mathcal{S}$, detections $\mathcal{D}$, structural constraints $\mathcal{E}$
Result: assignment event $\mathcal{A}$
begin
Step 1: Initializing possible assignment events (Section 3.3)
· Remove negligible assignments by using the gating ((10)).
· Divide objects, structural constrains, and detections into the subset $\mathcal{S}^p \subset \mathcal{S}, \mathcal{E}^p \subset \mathcal{E}, \mathcal{D}^p \subset \mathcal{D}$ by the partitioning (Fig. 3).
· Generate all possible assignment events of each partition $\mathcal{A}^p_{\text{all}}$ from the $\mathcal{S}^p$ and $\mathcal{D}^p$ based on the condition in (2).
Step 2: Aggregating assignment event costs ((8) and (9))
$\mathcal{A} = \phi$;
for $p = 1 : P$ **do**
$$C(\mathcal{A}^p) = \sum_{\substack{i \in N^p, k \in M^p \\ a^{i,k}=1}} \left(a^{i,k}\Omega_{i,k} + \sum_{\substack{j \in N^p \\ j \neq i}} \sum_{\substack{q \in M^p \cup \{0\} \\ q \neq k}} a^{j,q}\Theta^{j,q}_{i,k} \right),$$
$$\hat{\mathcal{A}}^p = \arg\min_{\mathcal{A}^p} \left(C(\mathcal{A}^p)/\Delta \right), \mathcal{A}^p \subset \mathcal{A}^p_{\text{all}};$$
$\mathcal{A} := \mathcal{A} \cup \hat{\mathcal{A}}^p$;
end
end

Algorithm 1: Structural Constraint Event Aggregation.

We denote the set of tracked objects in the previous frame by $\mathcal{S}^w$, and their structural information is represented by $\mathcal{E}^w$. Using $\mathcal{S}^w$, $\mathcal{E}^w$, and the current detections $\mathcal{D}$, we conduct the first data association via the SCEA introduced in Section 3. Then, we obtain the new assignment event $\hat{\mathcal{A}}^w$ from which we store the position of associated detections for the object i as $\mathbf{s}^i_1 = [x^k_d, y^k_d]^\top$, $\mathbf{s}^i_1 \in \mathcal{S}^w_1$ if $a^{i,k} = 1$, and the set of associated object index is represented by $i \in N^w$. In the second step, similar to [13, 29], we recover missing objects, which are not associated with any detections in the previous frame but re-detected in the current frame. The recovery process is conducted by using the tracked objects in the first step and their structural constraint information as described in Figure 4. The mis-detected objects are denoted by $\mathcal{S}^m$, and the structural constraints between mis-detected objects and tracked objects are represented by $\mathcal{E}^m$. Using $\mathcal{S}^m$, $\mathcal{E}^m$, and $\mathcal{S}^w_1$, we recover the re-detected objects as

$$\begin{aligned} \hat{\mathcal{A}}^m &= \arg\min_{\mathcal{A}} C(\mathcal{S}^m, \mathcal{E}^m, \mathcal{S}^w_1, \tilde{\mathcal{D}}), \\ &\text{s.t.} \sum_{i \in N^m} a^{i,q} = 1 \ \wedge \ \sum_{q \in \tilde{M}} a^{i,q} = 1, \end{aligned} \tag{11}$$

where N^m denotes the set of the mis-detected object index, and $\tilde{M}$ represents the index set of the detections $\tilde{D}$. Here, detections $\tilde{D}$ contains the not-assigned detections in the first step and dummy detections $\mathbf{d}^0$ for the case of missing objects. The structural constraint cost function for missing objects is defined as

$$C(\mathcal{S}^m, \mathcal{E}^m, \mathcal{S}^w_1, \tilde{\mathcal{D}}) = \sum_{i \in N^m} \sum_{q \in \tilde{M}} a^{i,q} \, \Phi^{i,q}$$
$$\Phi^{i,q} = \begin{cases} F_s(\mathbf{s}^i, \mathbf{d}^q) + F_a(\mathbf{s}^i, \mathbf{d}^q) \\ \quad + F_r(\mathbf{s}^i, \mathcal{E}^m, \mathcal{S}^w_1, \mathbf{d}^q) & \text{if } q \neq 0 \\ \tau & \text{if } q = 0 \end{cases}, \tag{12}$$

Data: tracked objects $\mathcal{S}^w$, structural constraints of tracked objects $\mathcal{E}^w$, mis-detected objects $\mathcal{S}^m$, structural constraints between tracked objects and mis-detected objects $\mathcal{E}^m$, detections $\mathcal{D}$
Result: Trajectories of the targets
for *video frame* f **do**
Step 1: Data association via SCEA
· $\hat{\mathcal{A}}^w = \text{SCEA}(\mathcal{S}^w, \mathcal{E}^w, \mathcal{D})$; (Section 3 and Algorithm 1)
· $\mathcal{S}^w_1 = \{\mathbf{s}^i_1 = [x^k_d, y^k_d]^\top | a^{i,k} = 1, \forall i \in N^w, \forall k \in M\}$;
Step 2: Recovery of mis-detected objects
· $\hat{\mathcal{A}}^m = \text{Recovery}(\mathcal{S}^m, \mathcal{E}^m, \mathcal{S}^w_1, \tilde{\mathcal{D}})$; ((11) and (12))
· $\hat{\mathcal{A}} = \hat{\mathcal{A}}^w \cup \hat{\mathcal{A}}^m$;
Step 3: Update
· Current tracking result:
$\mathcal{S}^w = \{\mathbf{s}^i := KF(\mathbf{s}^i, \mathbf{d}^k) | a^{i,k} = 1, \forall i \in N^w \cup N^m, \forall k \in M\}$ with Kalman filter $KF()$.
· Object management (Section 4)
· Structural constraint update (Section 4)
end

Algorithm 2: Two-Step Online MOT via SCEA

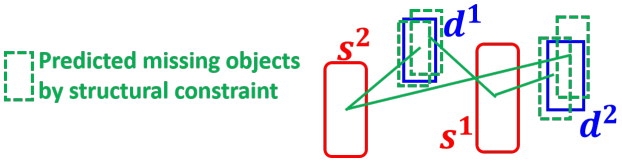

Figure 4. Recovery of missing objects. From the tracked objects ($\mathbf{s}^1$ and $\mathbf{s}^2$) and the structural constraints (the green lines), we recover missing objects when they are re-detected (detection $\mathbf{d}^1$ and $\mathbf{d}^2$). By doing this, we can continuously keep the identity of the missing objects under camera motion and occlusions.

where $\tau = 4$ in this work. We recover the missing object i from the set of tracked objects using their structural constraint. The constraint cost is therefore formulated as

$$\begin{aligned} F_r(\mathbf{s}^i, \mathcal{E}^m, \mathcal{S}^w_1, \mathbf{d}^q) &= -\ln\left(\frac{\text{area}(\mathbf{B}(\mathbf{s}^{i,\gamma}) \cap \mathbf{B}(\mathbf{d}^q))}{\text{area}(\mathbf{B}(\mathbf{s}^{i,\gamma}) \cup \mathbf{B}(\mathbf{d}^q))} \right), \\ \mathbf{s}^{i,\gamma} &= [(\mathbf{s}^\gamma_1)^\top, 0, 0]^\top + [\chi^{i,\gamma}, \upsilon^{i,\gamma}, w^i, h^i]^\top \\ \gamma &= \arg\max_{j \in N^w} \frac{1}{\|[\dot{\chi}^{i,j}, \dot{\upsilon}^{i,j}]\|}. \end{aligned} \tag{13}$$

Here, N^w denotes the index of tracked objects at the first step, and the reliability of 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 into account the motion coherence between objects, $\|[\dot{\chi}^{i,j}, \dot{\upsilon}^{i,j}]\|$. To solve (11), we reformulate (11) in a matrix form as

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

where the matrices are obtained by $\Phi^{det} = [\Phi^{i,q}], \forall i \in N^m, \forall q \in \tilde{M}$ and $\Phi^0 = \text{diag}[\Phi^{i,0}], \forall i \in N^m$. The off-diagonal etries of Φ^0 are set to ∞. We then apply the Hungarian algorithm [16] to get the assignment event having the minimum cost.

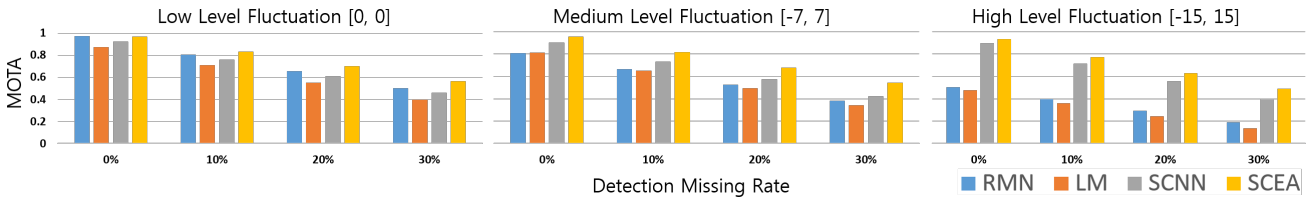

Figure 5. Data association performance according to different levels of camera motion fluctuation and detection missing rates. MOTA $= 1 - \frac{\text{false negative+false positive+mis-match}}{\text{ground truth}}$. The numbers ([0, 0],[-7, 7],[-15, 15]) represent the range of the different levels of camera motion fluctuation noise in terms of pixel. The missing rate of the detections is set to 0%, 10%, 20%, and 30%. The proposed SCEA shows the best overall performance. We analyze the performance of each method in detail in Section. 5.1.

From $\hat{\mathcal{A}}^w$ and $\hat{\mathcal{A}}^m$, we update the final tracking result as $\mathcal{S}^w = \{\mathbf{s}^i = KF(\mathbf{s}^i, \mathbf{d}^k) | a^{i,k} = 1, \forall i \in N^w \cup N^m, \forall k \in M\}$ with the Kalman filter $KF(\cdot)$ [25] for smoothing, and the index set is represented by N^w. After the update, other not-assigned objects are collected again in the set $\mathcal{S}^m$, and their index set is denoted by N^m.

Structural constraint update: After tracking, we update the structural constraints between objects with their corresponding detections based on the same approach proposed in [29], using $\mathbf{z}_t^{i,j} = [x_{d,t}^i, y_{d,t}^i]^\top - [x_{d,t}^j, y_{d,t}^j]^\top$ as an observation, where $[x_{d,t}^i, y_{d,t}^i]^\top$ represents the location of a detection assigned to the object i. We assume that the structural constraint change follows piece-wise linear motion model. With the observation $\mathbf{z}_t^{i,j}$, we indirectly update the structural constraint variations by using the standard Kalman filter [25]. The structural constrains of missing objects are simply based on the linear motion model.

Object management: For any MOT method, an object initialization and termination steps are typically required to manage targets according to their statuses. In this work, objects are initialized in a way similar to [4]. Here, we use the distance and the appearance between two detections as an initialization cue. If the distances between a detection in the current frame and unassociated detections in the past a few (e.g., 4) frames are smaller than a certain threshold, we then initialize this detection as a new object. The structural constraint between the new object and all other objects are then generated by (1), where their initial variation is set to $\dot{\chi}_t^{i,j} = \dot{v}_t^{i,j} = 0$. On the other hand, we simply delete or terminate objects if they are not associated with any detections for two frames.

5. Experiments

In this section, we present the experimental evaluation of the proposed online MOT algorithm and comparison against the state-of-the-art methods especially for the scenes acquired from moving cameras. For reproducibility, we will open source codes of the structural constraint cost aggregation at `cvl.gist.ac.kr/project/scea.html`.

5.1. Performance validation

To show the effectiveness of each component of the proposed method, we utilize the synthetic datasets which are generated based on the ground truth of the ETH sequences (Bahnhof, Sunnyday, and Jelmoli sequences) [8]. We apply the different levels of motion fluctuation noises and detection missing rate as shown in Figure 5. The low level fluctuation represents the original camera motion in the ETH sequences where the camera moves smoothly. The medium level fluctuation and the high level fluctuation represent fluctuation noises synthetically generated by the uniform distribution within $[-7, 7]$ and $[-15, 15]$ pixels, respectively. In addition, for all scenarios, we include at most 10 false detections per each frame. To measure the accuracy of the data association, the number of true positives, false positives, false negative, and mis-matches are counted per each frame.

Data association evaluation: The performance of four different data association approaches is shown in Figure 5. The relative motion network (RMN) approach [29] performs well under the low level fluctuation as this assumes accurate linear prediction of the well-tracked objects under smooth camera motion. The linear motion (LM) method is a baseline method where the data association is carried out without the structural constraints or event aggregation. It is similar to the joint probabilistic data association (JPDA) in that both methods consider the assignment events. A fast and efficient version of JPDA has been recently proposed and applied to the vision-based MOT in [23]. As the fluctuation increases, the performance of the LM method is degraded due to large camera motion where the linear motion model dose not work well. The structural constraint nearest neighbor (SCNN) is a data association method with the structural constraint cost function but without event aggregation. Due to the structural constraint cost function, the SCNN can deal with the large camera motion. However, since the structural constraint costs are obtained by the local nearest neighbors, the performance of the SCNN shows limited performance caused by the ambiguities as discussed

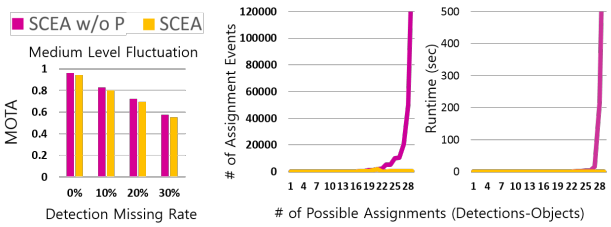

Figure 6. Comparison of the SCEA (with the partitioning) and the SCEA without the partitioning (the SCEA-w/o-P).

in Section 1 and shown in Figure 1. Figure 5 demonstrates that the SCEA performs better than other approaches since it robustly deals with large fluctuations based on the structural constraint cost function, and it can efficiently reduce ambiguities by aggregating costs of the same events computed based on the different anchor assignment.

Efficiency of the event reduction: In the experiments, the same event reduction techniques described in Section 3.3 are applied to the LM, SCNN, and SCEA methods for computational efficiency. Here, the gating technique has long been applied to MOT, and without this, the data association is computationally intractable when considering all possible assignment events as pointed out in [2]. For that reason, we only evaluate the efficiency of the partitioning technique using the SCEA method with and without partitioning (SCEA-w/o-P). Even with the gating technique, the SCEA-w/o-P method becomes computationally intractable when more than a certain number of objects or detections are given as shown in Figure 6. For this reason, the Sunnyday sequence, the Jelmoli sequence, and the roughly half of the Bahnhof sequence (i.e., frame #0-#450) are used for the evaluation. Figure 6 shows that the SCEA method is more applicable to online MOT thanks to the low computational complexity with similar performance to the SCEA-w/o-P.

5.2. Comparisons with State-of-the-Art Methods

We name the proposed algorithm as SCEA (Online MOT via Structural Constraint Event Aggregation) and evaluate it on a large number of benchmark datasets: 29 sequences from the KITTI dataset [12] and 22 sequences from the MOT Challenge dataset [17]. The datasets contain test sequences from a static camera as well as a dynamic camera. The detections of the KITTI dataset[1] and the MOT Challenge dataset[2] are also provided. Note that, since this work focuses on 2D MOT with a single camera, we do 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.

[1] cvlibs.net/datasets/kitti/eval_tracking.php
[2] motchallenge.net/data/2D_MOT_2015/

Table 1. Comparison to the online trackers on the KITTI dataset.

(a) Car (based on the DPM detections)

	MOTA	MOTP	Rec	Prec	MT	ML	ID	FG	sec(core)	AR
NOMT-HM	**60.2**	78.7	63.8	96.9	**27.0**	30.3	28	**250**	0.09(16)	1.89
RMOT	51.5	75.2	57.2	92.9	15.2	33.5	51	382	**0.01(1)**	3.33
ODAMOT	58.8	75.5	**65.5**	94.6	16.8	**18.9**	403	1298	1(1)	2.78
SCEA	56.3	**78.8**	58.1	**98.9**	20.0	29.3	**17**	468	0.05(1)	2.00

(b) Car (based on the regionlet detections)

	MOTA	MOTP	Rec	Prec	MT	ML	ID	FG	sec(core)	AR
NOMT-HM	74.8	**80.0**	80.6	**96.3**	**38.7**	15.2	109	**371**	0.09(16)	1.78
RMOT	65.3	75.4	80.2	87.7	26.8	**11.4**	215	742	**0.02(1)**	2.56
SCEA	**75.2**	79.4	**81.4**	95.9	**38.7**	12.7	**106**	466	0.06(1)	1.56

(c) Pedestrian (based on the DPM detections)

	MOTA	MOTP	Rec	Prec	MT	ML	ID	FG	sec(core)	AR
NOMT-HM	27.5	68.0	37.1	80.1	**11.3**	51.6	73	743	0.09(16)	2.67
RMOT	**34.5**	68.1	**43.7**	83.2	10.0	**47.4**	81	**692**	**0.01(1)**	1.56
SCEA	33.1	**68.5**	40.1	**85.3**	8.6	**47.4**	**16**	724	0.05(1)	1.67

(d) Pedestrian (based on the regionlet detections)

	MOTA	MOTP	Rec	Prec	MT	ML	ID	FG	sec(core)	AR
NOMT-HM	39.3	71.1	50.4	83.3	**17.2**	42.3	186	870	0.09(16)	2.44
RMOT	43.7	71.0	**53.5**	85.8	16.8	**41.2**	156	760	**0.02(1)**	1.78
SCEA	**43.9**	**71.9**	49.3	**90.7**	14.1	43.3	**56**	**649**	0.06(1)	1.78

We compare the SCEA with the state-of-the-art online MOT methods including MDP [26], TC_ODAL [1], RMOT [29], NOMT-HM [5], and ODAMOT [11]. Here, online methods produce the solution instantly at each frame by a causal approach.

Evaluation metrics: We adopt the widely used evaluation metrics, Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) from [3]. In addition, we also consider the number of mostly tracked (MT), the number of mostly lost (ML), the fragment (FG), the identity switch (ID), the Recall (Rec), and the Precision (Prec) from [19]. The runtime is also considered as a metric in terms of Hz or sec. Motivated by the MOT Challenge evaluation, we also use the average ranking (AR) computed by averaging all metric rankings. Although the AR does not reflect the MOT performance directly, it can be used as a reference to compare overall MOT performance

Benchmark dataset: The KITTI dataset provides two sets of detections, one from the DPM [10] and the other from the regionlet [24]. The regionlet detector generates more accurate detections than the DPM as illustrated on the KITTI website. As shown in Table 1, the AR indicates that the SCEA method performs fairly well compared to other state-of-the-art online trackers. The OMDAMOT method utilizes the additional local detector to deal with missing objects caused by partial occlusions, and the NOMT-HM additionally utilizes the optical flow information to reduce ambiguities caused by similar appearance of objects. Although our method utilizes the information only from detections

Table 2. Comparison to the online trackers on the MOT Challenge dataset (pedestrian sequences). FAF: the average number of false alarms per frame. FP: the number of false positives. FN: the number of false negatives. (The results of the NOMT-HM are from the original paper [5].)

	MOTA	MOTP	FAF	MT	ML	FP	FN	ID	FG	Hz(core)	AR
TC_ODAL	15.1	70.5	2.2	3.2	55.8	12,970	38,538	637	1,716	1.7 (1)	4.30
RMOT	18.6	69.6	2.2	5.3	53.3	12,473	36,835	684	1,282	**7.9 (1)**	3.70
NOMT-HM	26.7	**71.5**	2.0	11.2	47.9	11,162	33,187	637	1,716	11.5 (16)	2.50
MDP	**30.3**	71.3	1.7	**13.0**	**38.4**	9,717	**32,422**	680	1,500	1.1 (8)	2.30
SCEA	29.1	71.1	**1.1**	8.9	47.3	**6,060**	36,912	**604**	**1,182**	6.8 (1)	2.10

Table 3. Comparison to the MDP on the KITTI training dataset.

(a) Car (except for the training sequences)

	MOTA	MOTP	Rec	Prec	MT	ML	ID	FG
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

(b) Pedestrian (except for the training sequences)

	MOTA	MOTP	Rec	Prec	MT	ML	ID	FG
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

and does not exploit those additional local detector or optical flow information, it shows comparable or better performance compared to the OMDAMOT and the NOMT-HM. The RMOT also uses the structural motion cues between objects to track missing objects robustly. However, the RMOT method does not perform well in the car sequences where large camera panning motion frequently occurs as explained in Section 5.1. Compared to the RMOT, the proposed SCEA algorithm shows much better performance in terms of MOTA, Prec, IDS, and Frag, which indicate the proposed data association method is more accurate than the RMN data association used in the RMOT.

For KITTI pedestrian sequences, the SCEA algorithm achieves better performance in MOTA and in Prec compared to the NOMT-HM, and it also shows better performance in IDS. This is because the optical flow information from pedestrians is less reliable compared to that in the car sequences due to the small size and non-rigid appearance of a pedestrian. In addition, the motion cue (the optical flow) becomes less discriminative when motion of objects is small. In the KITTI dataset, the motion of pedestrians is much smaller than that of cars. Since the SCEA method extracts structural motion information only from detections, its performance is less affected by the object size, appearance, and small motion. As shown in the results on the MOT Challenge dataset (pedestrian sequences, Table 2), the SCEA method performs well compared to other online methods overall. The TC_ODAL utilizes the linear motion model to link the tracklets based on the Hungarian algorithm. For this reason, it shows limited performance under camera motion. The MDP shows better performance in MOTA, MT, ML, and FN 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 other methods. In addition, due to the use of the optical flow for local template tracking, it generates longer trajectories compared to other online methods. However, the SCEA algorithm has advantages over the MDP method in that it does not require any training datasets and it runs 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 pedestrian sequences, we run the MDP with original trained model provided with the original source code by the authors (MDP-MOTC). In addition, we also train the MDP with the KITTI training dataset for car sequences (MDP-KITTI). As shown in Table 3, 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.

6. Conclusion

In online 2D MOT with moving cameras, observable motion cues are complicated by global camera movements and thus not always smooth or predictable. In this paper, we propose a new data association method that effectively exploits structural motion constraints in the presence of large camera motion. In addition, to alleviate data association ambiguities caused by mis-detections and multiple detections, a novel event aggregation approach is developed to integrate structural constraints in assignment event costs for online MOT. Finally, the proposed data association and structural constraints are incorporated into the two-step online 2D MOT algorithm which simultaneously tracks objects and recovers missing objects. Experimental results on a large number of datasets demonstrate the effectiveness of the proposed algorithm for online 2D MOT.

Acknowledgment. This work was supported by the National Research Foundation of Korea(NRF) (No. NRF-2015R1A2A1A01005455) and Institute for Information & communications Technology Promotion(IITP) (No. B0101-16-0552, Development of Predictive Visual Intelligence Technology) grants, and (GK15C0100) and (CISS-2013M3A6A6073718) grants all funded by the Korea government(MSIP). M.-H. Yang is supported in part by the NSF CAREER Grant #1149783, NSF IIS Grant #1152576, and a gift from Adobe.

References

[1] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *CVPR*, 2014. 2, 7

[2] Y. Bar-Shalom and X.-R. Li. *Multitarget-multisensor Tracking: Principles and Techniques*. YBS publishing, Storrs, CT, USA, 1995. 4, 7

[3] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Eurasip Journal on Image and Video Processing*, 2008. 7

[4] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *PAMI*, 2011. 1, 2, 6

[5] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *ICCV*, 2015. 1, 7, 8

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *PAMI*, 2014. 1

[7] G. Duan, H. Ai, S. Cao, and S. Lao. Group tracking: exploring mutual relations for multiple object tracking. In *ECCV*, 2012. 2

[8] A. Ess, B. Leibe, K. Schindler, and L. V. Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008. 6

[9] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010. 4

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 7

[11] A. Gaidon and E. Vig. Online Domain Adaptation for Multi-Object Tracking. In *BMVC*, 2015. 7

[12] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *IJRR*, 2013. 7

[13] H. Grabner, J. Matas, L. J. V. Gool, and P. C. Cattin. Tracking the invisible: Learning where the object might be. In *CVPR*, 2010. 2, 5

[14] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg. Multiple hypothesis tracking revisited. In *ICCV*, 2015. 1

[15] S. Kim, S. Kwak, J. Feyereisl, and B. Han. Online multi-target tracking by large margin structured learning. In *ACCV*, 2012. 2

[16] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955. 5

[17] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. In *arXiv:1504.01942*, 2015. 7

[18] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCVW*, 2011. 2

[19] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In *CVPR*, 2009. 7

[20] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *PAMI*, 2014. 1

[21] S. Pellegrini, A. Ess, K. Schindler, and L. V. Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2

[22] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011. 1

[23] H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid. Joint probabilistic data association revisited. In *ICCV*, 2015. 4, 6

[24] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In *ICCV*, 2013. 1, 7

[25] G. Welch and G. Bishop. An introduction to the kalman filter. Technical report, University of North Carolina at Chapel Hill, 1995. 6

[26] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *ICCV*, 2015. 7

[27] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *CVPR*, 2009. 1

[28] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *CVPR*, 2012. 1, 2

[29] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon. Bayesian multi-object tracking using motion context from multiple objects. In *WACV*, 2015. 2, 5, 6, 7

[30] L. Zhang and L. van der Maaten. Structure preserving object tracking. In *CVPR*, 2013. 2