Robust Online Multiobject Tracking With Data Association and Track Management

Seung-Hwan Bae, *Student Member, IEEE,* and Kuk-Jin Yoon, *Member, IEEE*

***Abstract*— In this paper, we consider a multiobject tracking problem in complex scenes. Unlike batch tracking systems using detections of the entire sequence, we propose a novel online multiobject tracking system in order to build tracks sequentially using online provided detections. To track objects robustly even under frequent occlusions, the proposed system consists of three main parts: 1) visual tracking with a novel data association with a track existence probability by associating online detections with the corresponding tracks under partial occlusions; 2) track management to associate terminated tracks for linking tracks fragmented by long-term occlusions; and 3) online model learning to generate discriminative appearance models for successful associations in other two parts. Experimental results using challenging public data sets show the obvious performance improvement of the proposed system, compared with other state-of-the-art tracking systems. Furthermore, extensive performance analysis of the three main parts demonstrates effects and usefulness of the each component for multiobject tracking.**

***Index Terms*— Online multi-object tracking, tracking-by-detection, data association, track management, online learning, track existence probability, particle filtering, affinity model, surveillance system.**

I. Introduction

MULTI-OBJECT tracking is to find the locations and sizes of multiple objects and conserve their IDs in image sequences. It is important for many applications such as a surveillance system, human machine interfaces, motion capture, and a medical system. In complex scenes, the multi-object tracking problem is still challenging due to frequent occlusions and complex interactions between objects. Recently, development of detectors [10], [31] allows us to obtain reliable detections, and this leads to the prosperity of the tracking-by-detection approach [7], [17], [26], [40].

The tracking-by-detection approach can be categorized into two classes: batch tracking and online tracking systems. Once a set of detections is collected by temporal sliding window search for all frames, the detections are gradually connected based on data association in the batch tracking systems (*or* detection-association-based systems) [2], [28], [40]. A hierarchical association framework [17] is developed to produce longer tracklets (*i.e.* trajectory fragments) at each level gradually. [23] explicitly models object interaction such as mutual occlusion and spatial layout consistency, and simultaneously optimizes trajectories using dynamic programming. [40] and [28] solve a global data association problem using a min-cost flow algorithm in a network flow. In [8], short tracklets are merged into longer ones by finding maximum weighted independent sets in a graph of detection pairs. [37] develops an online-learned CRF model and links tracklets by minimizing an energy function. The batch tracking systems can handle detection errors and tracking failure caused by occlusions, and show high accuracy and robustness even in complex scenes.

However, the batch tracking systems are unsuitable for real-time tracking applications. They require detection responses of future frames beforehand and accompany enormous computation to generate optimized trajectories — in order to construct longer trajectories, an iterative linking process is performed until maximizing the predefined association cost. It implies that identities of the tracklets can be changed by linking results at each iteration.

On the other hand, online tracking systems [7], [26], [33] can be applied for time critical applications since they sequentially build trajectories based on frame-by-frame association without the iterative associations. However, the online systems are likely to produce fragmented trajectories under occlusions when detections of occluded objects are not available or inaccurate. Moreover, they suffer from template drift when motions and appearances rapidly change. As a result, the performance of the online systems is significantly degraded in complex scenes where objects are frequently occluded, and their appearances are quite similar.

In this paper, we propose an online tracking system, which can robustly track multiple objects even in complex scenes but also be suitable for online tracking applications. Basically, we develop our system based on the Bayesian approach as done in previous online tracking systems [7], [18], [26], [33] to sequentially estimate the states of objects (*i.e.* position, size, and ID) with online provided detections at each frame. The proposed system, however, integrates three main parts to tackle the problems of previous online tracking systems as shown in Fig. 1.

In order to correctly assign detections with tracks under partial occlusions, the visual tracking part associates online detections with existing tracks by evaluating track existence probabilities as well as the likelihoods of them, and updates

Manuscript received August 13, 2013; revised January 29, 2014; accepted April 11, 2014. Date of publication April 29, 2014; date of current version May 20, 2014. This work was supported in part by the Basic Science Research under Grant 2012R1A1A1010871 and in part by the Global Frontier Research and Development under Grant 2013M3A6A3075453 Programs through the Ministry of Science, ICT and Future Planning, National Research Foundation of Korea. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jesus Malo.

The authors are with the School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju 500-712, Korea (e-mail: bshwan@gist.ac.kr; kjyoon@gist.ac.kr).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. It consists of a video that shows tracking results using the proposed systems for several sequences. The total size of the videos is 45 MB. Contact kjyoon@gist.ac.kr for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2320821

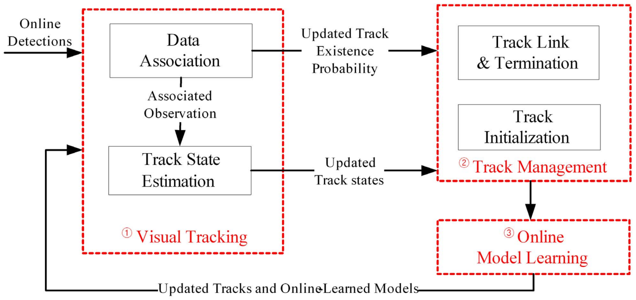


Fig. 1. The proposed framework for online multi-object tracking.

states of tracks with associated detections. However, it is still difficult to track objects when no detection is available for a long time. In this case, the track management part terminates tracks with low existence probabilities and associates the terminated tracks with other tracks or detections belong to the same objects so as to link them. For successful association in the other two parts, the online model learning part incrementally learns discriminative appearance models with updated tracking results. It allows us to distinguish between tracked objects and background, but also between interacting objects (*i.e.* closely spaced objects).

The proposed system produces long trajectories without future frames and any iterative optimization in complex scenes. We extensively evaluate the performance of our system and the key parts of the system using challenging tracking datasets. The main contributions of this work include: (1) a fully automated online tracking system to track objects robustly through severe occlusions; (2) a Bayesian tracking method based on a novel data association to update states of tracks with online detections; (3) a track management method to associate fragmented tracks; (4) an online learning method to learn discriminative appearance models of tracked objects.

The rest of the paper is organized as follows. We first discuss related works in Section II. Then, we formulate an online multi-object tracking problem in Section III. We propose an online multi-object tracking system in Section IV and provide some experimental results in Section V. We briefly discuss the effectiveness of main parts of the system in Section VI and finally conclude the paper in Section VII.

II. Related Work

A tracking-by-detection approach finds states and IDs of multiple objects with detections of pre-trained detectors [10], [31]. The tracking-by-detection approach often fails in complex scenes when the detection responses are unreliable (e.g., false positive and missing detections, inaccurate detections of object locations and sizes). To track multiple objects in online under difficult situations, various online tracking systems [7], [26], [33], [34] are developed.

Based on the Bayesian framework [18], [26], the online tracking systems perform two recursive procedures to estimate states of multiple objects at each frame. In prediction estimated states up to previous frames are propagated using the object dynamic model. The predicted states are updated by evaluating the likelihood between the predicted states and associated observations using the observation model.

[26] designs a particle filtering-based framework and carefully guides multiple trackers with detections provided by a boosted object detector. [9] extends the boosted particle filter [26] using an independent particle set for each target to improve robustness under occlusions. For single and multi-target tracking, [35] extends the PDA and JPDA algorithms [4] with the concept of target existence. To enhance detection and tracking performance, [31] exploits an edgelet-based part model for describing appearances of objects. For tracking in multi-dimensional state space, [27] develops the scatter search particle filter by embedding scatter search metaheuristic [15] into the particle filtering [18]. However, all of these approaches suffer from template drift when motions and appearances dramatically change since they only rely on outputs of offline-trained detectors.

In order to solve the drift problems under occlusions, [7] uses the continuous confidence map by combining a pre-trained detector and an online-trained classifier outputs. To handle partial occlusions in the detection and tracking stages, [34] employs deformable part models for describing appearances of objects. Although they show improved performance in many scenarios, both approaches are prone to produce fragmented trajectories under long-term occlusions since any tracking linking process is not adopted.

To resolve both partial and long-term occlusions, the proposed online tracking system takes advantages of batch and online tracking approaches. As similar to other online tracking approaches, our visual tracking part is designed based on the Bayesian approach for online tracking, but a novel data association with a track existence probability is incorporated to assign detections into tracks more correctly under partial occlusions. Subsequently, our track management part performs track-to-track association to link fragmented tracks under long-term occlusions as similar to tracklet association in batch tracking systems.

Lately, in an attempt to combine the both approaches, [38] has presented a tracking system with discriminative part-based models. However, it is significantly different from ours since their system is designed based on the batch tracking framework. Therefore, tracks (*or* trackelts) are generated by globally associating detections of the entire sequences. It indicates that their system is not suitable for online tracking applications. The two-stage tracking system proposed by [32] is also similar to ours. They produce locally optimized tracks by associating observations with tracks and globally optimized tracks by associating fragmented tracks. They use the greedy method for local association, whereas we employ a novel data association. In addition, they use the predefined appearance model, but our online learning part updates discriminative appearance models with online tracking results. As a result, our system is able to distinguish between different objects well, even though the appearances of the objects frequently change.

III. Online Multi-Object Tracking Formulation

In image sequences, we denote the state of the i-th object at frame t as $\mathbf{x}_t^i = \left(p_t^i, s_t^i\right)$, where p_t^i and s_t^i are the position and size of the object. A set of states of all objects existing at frame t is denoted as $\mathbf{X}_t$. Then, the state of the i-th object and the states of all objects up to frame t can be represented as $\mathbf{x}^i_{t_1^i:t_2^i}$ and $\mathbb{X}_{1:t} = \left\{\mathbf{x}^i_{t_1^i:t_2^i} | 1 \le t_1^i \le t_2^i \le t, i = 1, \cdots, M_t\right\}$, where M_t is the total number of objects up to time t. t_1^i and t_2^i are the time stamps of start- and end- frame of the i-th object.

Given detection responses, we denote an observation of the i-th object at frame t as $\mathbf{z}^i_t$ and denote an observation set of the i-th object collected up to frame t as $\mathbf{z}^i_{t_1^i:t_2^i}$, and all observations collected up to frame t are denoted as $\mathbb{Z}_{1:t}$. Given the set $\mathbb{Z}_{1:t}$, our goal is then to find the optimal states of all objects $\mathbb{X}_{1:t}$. This problem can be formulated to find $\mathbb{X}_{1:t}$ by maximizing $p(\mathbb{X}_{1:t}|\mathbb{Z}_{1:t})$ as the maximum a posterior (MAP) formulation:

$$\hat{\mathbb{X}}^*_{1:t} \triangleq \underset{\mathbb{X}_{1:t}}{\operatorname{argmax}}\ p\left(\mathbb{X}_{1:t}|\mathbb{Z}_{1:t}\right) \qquad (1)$$

Here, it is impossible to globally optimize Eq. (1) using brute force search. Thus, we tackle the problem by recursively updating $p(\mathbf{X}_t|\mathbb{Z}_{1:t})$ based on the sequential Bayesian approach [18], [26]. Given detections $\mathbf{z}^i_{t_1^i:t}$ of the i-th object, $\mathbf{x}^i_t$ is estimated by two recursive procedures as

$$\text{Predict: } p(\mathbf{x}^i_t|\mathbf{z}^i_{t_1^i:t-1}) = \int p\left(\mathbf{x}^i_t|\mathbf{x}^i_{t-1}\right) p(\mathbf{x}^i_{t-1}|\mathbf{z}^i_{t_1^i:t-1}) d\mathbf{x}^i_{t-1},$$
$$\text{Update: } p(\mathbf{x}^i_t|\mathbf{z}_{t_1^i:t}) \propto p\left(\mathbf{z}^i_t|\mathbf{x}^i_t\right) p(\mathbf{x}^i_t|\mathbf{z}^i_{t_1^i:t-1}). \qquad (2)$$

When the observation set $\mathbf{z}^i_{t_1^i:t}$ corresponding to the i-th object is provided, the state of the object $\mathbf{x}^i_t$ can be updated well by Eq. (2). In most multi-object tracking scenarios, however, it is not easy to reveal origins of observations since the detections are often unreliable (*e.g.* false positive and missing detections, and inaccurate detections for object locations and sizes) and detections of other objects exist. Therefore, a data association method is usually required to correctly match observations with corresponding tracks.

Suppose that we have a set of tracks $\mathbf{X}_t = \left\{\mathbf{x}^i_t\right\}_{i=1}^{M_t}$ and a set of observations $\mathbf{Z}_t = \left\{\mathbf{z}^l_t\right\}_{l=1}^{L_t}$ at frame t, where M_t and L_t are the number of tracks and observations at frame t, respectively. Let denote an event that the l-th observation is associated with the i-th track as $\Theta^i_{t,l}$. Then, a pairwise association problem is formulated as

$$\hat{\Theta}^i_{t,l} \triangleq \underset{\Theta^i_{t,l}}{\operatorname{argmax}}\ P\left(\Theta^i_{t,l}|\mathbf{Z}_t\right). \qquad (3)$$

To solve Eq. (3), greedy and Hungarian methods [1] are mostly used. Here, it is assumed that observations are conditionally independent given the object state $\mathbf{x}^i_t$. Then, the likelihood $p\left(\mathbf{Z}_t|\mathbf{X}_t\right)$ can be expressed as

$$p\left(\mathbf{Z}_t|\mathbf{X}_t\right) = \prod_{i=1}^{M_t}\prod_{l=1}^{L_t} p\left(\mathbf{z}^l_t|\mathbf{x}^i_t\right), \qquad (4)$$

and an association score matrix S can be defined as

$$S = \left\{s_{i,l}\right\}_{M_t \times L_t}, \quad s_{i,l} = -\log\left(p(\mathbf{z}^l_t|\mathbf{x}^i_t)\right). \qquad (5)$$

In the greedy method [1], the track-observation pairs having the minimum score in the association matrix S are selected in ascending order until no further valid pair is available. On the other hand, the association pairs minimizing a total cost of the matrix S are determined in the Hungarian method [1]. In both methods, unreliable association pairs with low matching scores are usually removed by thresholding.

IV. Proposed Multi-Object Tracking System

In this section, we discuss the proposed multi-object tracking system consisting of three main parts; visual tracking, track management, and online model learning.

A. Overall Structure

We discuss the overall framework of the proposed system illustrated in Fig. 1 as follows:

Visual tracking: At each frame, object hypotheses are detected using a pre-trained detector and used as an input of our system. Based on the proposed data association, the provided detections are associated with existing tracks and existence probabilities of tracks are updated. Then, track states are estimated with the associated detections using particle filtering.

Track management: Existing tracks with the low existence probabilities are terminated. Terminated tracks are associated with other tracks or detections to link them. A new track is initialized using observations which are not associated with any tracks.

Online model learning: Discriminative appearance, shape and motion models of describing tracked objects are learned by updated tracking results.

In the next sections, we present the details of each part.

B. Visual Tracking Based on a Novel Data Association With Track Existence Probability

In complex scenes, many objects are often close to each other and/or detections are inaccurate, and the likelihood $p\left(\mathbf{z}^l_t|\mathbf{x}^i_t\right)$ between a track and an observation is unreliable. In this case, an incorrect association pair might have higher association scores than the correct pair, and it gives rise to ID switch problems. For accurate association even in complex scenes, we propose a novel association method. Our method is motivated from [24], and we extend their works to be successfully applied for visual multi-object tracking. A main difference is that the association score Eq. (5) is computed with a track existence probability[1] as well as the likelihood, compared with the greedy [7], [26] and the Hungarian [17], [32] association methods using only the likelihood. Moreover, it allows us to consider a possibility that an associated observation could be originated from other objects and scene clutters when calculating the association score. Once the association pair is determined for each track, we update states of tracks using particle filtering.

[1] A probability of that a track exists at frame t.

1) Data Association With Track Existence Probability: Let us denote a track existence probability as $p(\chi_t^i)$, where χ_t^i is an event that the i-th track exists at frame t. Given the observation set $\mathbb{Z}_{1:t}$, we define a posterior association problem by considering the existence probability, and the problem of Eq. (3) can be reformulated as follows:

$$\hat{\Theta}_{t,l}^i \triangleq \underset{\Theta_{t,l}^i}{\operatorname{argmax}}\ P\left(\Theta_{t,l}^i, \chi_t^i | \mathbb{Z}_{1:t}\right). \tag{6}$$

Based on the Bayesian rule, the posterior data association probability (Posterior DA) is represented as follows:

$$\underbrace{P\left(\Theta_{t,l}^i, \chi_t^i | \mathbb{Z}_{1:t}\right)}_{Posterior DA} = \underbrace{p\left(\mathbf{Z}_t | \Theta_{t,l}^i, \chi_t^i, \mathbb{Z}_{1:t-1}\right)}_{Observation\ density} \cdot$$
$$\underbrace{P\left(\Theta_{t,l}^i, \chi_t^i | \mathbb{Z}_{1:t-1}\right)}_{Prior DA} / p(\mathbf{Z}_t | \mathbb{Z}_{1:t-1}), \tag{7}$$

where the first and second terms are an observation density function and a prior data association probability. Now, we derive the prior data association and the observation density function to achieve the posterior association probability.

Prior data association. In a prior association step, we approximately compute the association probability. Then, the probability $P_{t,l}^i = P(\Theta_{t,l}^i, \chi_t^i | \mathbb{Z}_{1:t-1})$ is expressed as

$$P_{t,l}^i = P\left(\Theta_{t,l}^i | \chi_t^i, \mathbb{Z}_{1:t-1}\right) \cdot P\left(\chi_t^i | \mathbb{Z}_{1:t-1}\right), \tag{8}$$

where $P(\chi_t^i | \mathbb{Z}_{1:t-1})$ means the propagation of the existence probability $P\left(\chi_{t-1}^i | \mathbb{Z}_{1:t-1}\right)$ updated up to frame $t-1$, and is computed using the first-order Markov chain model:

$$P\left(\chi_t^i | \mathbb{Z}_{1:t-1}\right) = \triangle_{11} \cdot P\left(\chi_{t-1}^i | \mathbb{Z}_{1:t-1}\right) + \triangle_{21} \cdot \left(1 - P\left(\chi_{t-1}^i | \mathbb{Z}_{1:t-1}\right)\right), \tag{9}$$

with transition probability $\triangle_{11} \equiv P\left(\chi_t^i | \chi_{t-1}^i\right)$ and $\triangle_{21} \equiv P\left(\chi_t^i | \tilde{\chi}_{t-1}^i\right)$, where $\tilde{\chi}_t^i$ represents the non-existence of the track i at frame t ($\triangle_{11} = 0.9$ and $\triangle_{21} = 0.1$ are set in our experiments).

As it requires knowledge of true states of other objects to compute the first term of Eq. (8), we approximate it as $\Omega_{t,l}^i = P(\Theta_{t,l}^i | \chi_t^i, \mathbb{Z}_{1:t-1}) \approx P(\Theta_{t,l}^i | \chi_t^i, \mathbb{Z}_{1:t},$ single track) by assuming that there exists one track only. Then, this term can be represented with the likelihood $p(\mathbf{z}_t^l | \mathbf{x}_t^i)$ between the track i and the observation l [2].

$$\Omega_{t,l}^i \approx \frac{p(\mathbf{z}_t^l | \mathbf{x}_t^i)}{\rho_l^i} / \sum_{j=1}^{L_t} \frac{p(\mathbf{z}_t^j | \mathbf{x}_t^i)}{\rho_j^i},$$
$$\text{where } \rho_l^i = \sum_{\sigma=1, \sigma \neq i}^{M_t} P\left(\chi_t^\sigma | \mathbb{Z}_{1:t-1}\right) p(\mathbf{z}_t^l | \mathbf{x}_t^\sigma). \tag{10}$$

By substituting Eq. (9) and Eq. (10) into Eq. (8), we calculate the prior data association probability:

$$P_{t,l}^i = \frac{p(\mathbf{z}_t^l | \mathbf{x}_t^i)}{\rho_l^i} / \sum_{j=1}^{L_t} \frac{p(\mathbf{z}_t^j | \mathbf{x}_t^i)}{\rho_j^i} \cdot P\left(\chi_t^i | \mathbb{Z}_{1:t-1}\right). \tag{11}$$

Observation density function. To derive the observation density function, we consider the following two events:

- Event 1: Observation l is originated from other objects.
- Event 2: Observation l is originated from a clutter.

To model the first event, we define the probability that the observation l comes from the σ-th object among all other objects excluding the i-th object using the prior association probability Eq. (11) as:

$$\begin{aligned} Q_l^{i,\sigma} &\triangleq P_{t,l}^\sigma \prod_{w=1, w \neq i, w \neq \sigma}^{M_t} \left(1 - P_{t,l}^w\right) \\ &= \frac{P_{t,l}^\sigma}{1 - P_{t,l}^\sigma} \prod_{w=1, w \neq i}^{M_t} \left(1 - P_{t,l}^w\right). \end{aligned} \tag{12}$$

To model the second event, we introduce a clutter density $\varrho_{t,l}$ for the event that the observation comes from a clutter (*i.e.* $\mathbf{z}_t^l$ becomes a false positive) [3]. In addition, we calculate a probability that the observation l does not originate from one of $M_t - 1$ potential objects excluding the i-th object

$$Q_l^{i,0} \triangleq \prod_{\sigma=1, \sigma \neq i}^{M_t} \left(1 - P_{t,l}^\sigma\right). \tag{13}$$

By considering the both events, we can approximately evaluate a density function for an observation that does not generate from the i-th object:

$$\begin{aligned} &p\left(\mathbf{z}_t^l | \tilde{\Theta}_{t,l}^i, \mathbb{Z}_{1:t-1}\right) \\ &\approx \varrho_{t,l} Q_l^{i,0} + \sum_{\sigma=1, \sigma \neq i}^{M_t} p(\mathbf{z}_t^l | \mathbf{x}_t^\sigma) Q_l^{i,\sigma} \\ &= Q_l^{i,0} \left(\varrho_{t,l} + \sum_{\sigma=1, \sigma \neq i}^{M_t} \frac{P_{t,l}^\sigma}{1 - P_{t,l}^\sigma} \cdot p(\mathbf{z}_t^l | \mathbf{x}_t^\sigma) \right). \end{aligned} \tag{14}$$

Now, let us define a scatterer density meaning the event that the l-th observation originates from a clutter or other objects using the second term in Eq. (14) and denote it as

$$\Phi_{t,l}^i \triangleq \underbrace{\varrho_{t,l}}_{clutter} + \underbrace{\sum_{\sigma=1, \sigma \neq i}^{M_t} p(\mathbf{z}_t^l | \mathbf{x}_t^\sigma) \cdot \frac{P_{t,l}^\sigma}{1 - P_{t,l}^\sigma}}_{other\ target}. \tag{15}$$

Given that l-th observation comes from the i-th object, the observation density function becomes

$$p\left(\mathbf{z}_t^l | \Theta_{t,l}^i, \chi_t^i, \mathbb{Z}_{1:t-1}\right) = p(\mathbf{z}_t^l | \mathbf{x}_t^i)$$

and the observation density function

$$p\left(\mathbf{Z}_t | \Theta_{t,l}^i, \chi_t^i, \mathbb{Z}_{1:t-1}\right)$$

[2] The likelihood model will be discussed in the section IV-D1 and the likelihood score is evaluated by the observation model Eq. (28).

[3] It relies on the false positive rate of the used detector. When the false positive rate decreases, the density also decreases.

is represented as

$$\begin{aligned} p\left(\mathbf{Z}_t|\Theta^i_{t,l}, \chi^i_t, \mathbb{Z}_{1:t-1}\right) &= p\left(\mathbf{z}^l_t|\Theta^i_{t,l}, \chi^i_t, \mathbb{Z}_{1:t-1}\right) \\ &\quad \times p\left(\mathbf{Z}_t \backslash z^l_t|\Theta^i_{t,l}, \chi^i_t, \mathbb{Z}_{1:t-1}\right) \\ &= p(\mathbf{z}^l_t|\mathbf{x}^i_t) \prod_{j=1, j\neq l}^{L_t} Q^{i,0}_j \Phi^i_{t,j} \\ &= \frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\Phi^i_{t,j} Q^{i,0}_l} \rho^{i,0}_t, \end{aligned}$$

$$\text{where } \rho^{i,0}_t \triangleq p\left(\mathbf{Z}_t|\Theta^i_{t,0}, \mathbb{Z}_{1:t-1}\right) \approx \prod_{j=1}^{L_t} Q^{i,0}_j \Phi^i_{t,j}, \tag{16}$$

where $\Theta^i_{t,0}$ represents an event that no observation originates from the i-th object.

Posterior data association. To derive the posterior association probability, we substitute the prior probability Eq. (11) and the observation density Eq. (16) into Eq. (7) and denote the normalization term with $c_t = p(\mathbf{Z}_t|\mathbb{Z}_{1:t-1})$.

$$\begin{aligned} &P\left(\Theta^i_{t,l}, \chi^i_t|\mathbb{Z}_{1:t}\right) \\ &= c_t^{-1} \frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\Phi^i_{t,j} Q^{i,0}_j} \rho^{i,0}_t \cdot \frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\rho^i_l} \Big/ \sum_{j=1}^{L_t} \frac{p(\mathbf{z}^j_t|\mathbf{x}^i_t)}{\rho^i_j} \cdot P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right) \\ &\approx c_t^{-1} \frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\Phi^i_{t,j}} \rho^{i,0}_t \cdot P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right). \end{aligned} \tag{17}$$

In a similar manner, the posterior probability for the non-association event of the i-th object $P\left(\Theta^i_{t,0}|\mathbb{Z}_{1:t}\right)$ can be expressed using the Bayesian rule

$$\begin{aligned} &P\left(\Theta^i_{t,0}|\mathbb{Z}_{1:t}\right) \\ &= p\left(\mathbf{Z}_t|\Theta^i_{t,0}, \mathbb{Z}_{1:t-1}\right) \cdot P\left(\Theta^i_{t,0}|\mathbb{Z}_{1:t-1}\right) / p(\mathbf{Z}_t|\mathbb{Z}_{1:t-1}), \\ &= c_t^{-1} \rho^{i,0}_t \cdot \left\{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)\right\} \end{aligned} \tag{18}$$

Since data association events are mutually exclusive,

$$P\left(\Theta^i_{t,0}|\mathbb{Z}_{1:t}\right) + \sum_{l=1}^{L_t} P\left(\Theta^i_{t,l}, \chi^i_t|\mathbb{Z}_{1:t}\right) = 1$$

and the $c_t = p(\mathbf{Z}_t|\mathbb{Z}_{1:t-1})$ then can be derived as

$$\begin{aligned} c_t &= \rho^{i,0}_t \cdot \left\{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)\right\} \\ &\quad + \sum_{l=1}^{L_t} c_t^{-1} \frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\Phi^i_{t,l}} \rho^{i,0}_t P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right), \\ &= \rho^{i,0}_t \left\{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right) \cdot \Psi^i_t\right\}, \\ &\text{where, } \Psi^i_t = \left(1 - \sum_{l=1}^{L_t} \frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\Phi^i_{t,l}}\right). \end{aligned} \tag{19}$$

Algorithm 1 The algorithm for learning a discriminative appearance model

```
**Input** : Training samples $B^+$ and $B^-$ from Eq. (30).
**Output**: Updated strong classifier $H_i(f^i) = \sum_{k=1}^{K} \alpha_k h_k(f^i)$.
// Initialize a strong classifier
Initialize weights $\{w_l\}_{l=1}^{N}$
for $k = 1$ to $K$ do
    Make $\{w_l\}_{l=1}^{N}$ a distribution ;
    Train a weak classifier $h_k$ ;
    Set an error $r = \sum_{l=1}^{N} w_l \left|h_k(f^i_l) - y_l\right|$;
    Set a weak classifier weight $\alpha_k = 0.5\log\left(\frac{1-r}{r}\right)$
    Update weak example weights:
    $w_l = w_l \exp\left(\alpha_k \left|h_k(f^i_l) - y_l\right|\right)$;
end
// Update the strong classifier
for $k = 1$ to $T$ do
    //Select $T$ best weak classifiers and update their weights;
    Make $\{w_l\}_{l=1}^{N}$ a distribution ;
    Select $h_k(f^i)$ with the minimum error $r$ from
    $\left\{h_1(f^i), ..., h_K(f^i)\right\}$;
    Update $r$ and $\alpha_k$;
    Remove $h_k(f^i)$ from $\left\{h_1(f^i), ..., h_K(f^i)\right\}$;
end
for $k = T + 1$ to $K$ do
    Make $\{w_l\}_{l=1}^{N}$ a distribution ;
    Train a weak classifier $h_k$ ;
    Compute $r$ and $\alpha_k$;
    Update example weights $\{w_l\}_{l=1}^{N}$;
end
```

Finally, we express the posterior data association probabilities $P\left(\Theta^i_{t,l}, \chi^i_t|\mathbb{Z}_{1:t}\right)$ Eq. (17) and $P\left(\Theta^i_{t,0}|\mathbb{Z}_{1:t}\right)$ Eq. (18) by combining Eq. (19) as follows

$$\begin{aligned} P\left(\Theta^i_{t,0}|\mathbb{Z}_{1:t}\right) &= c_t^{-1} \rho^{i,0}_t \cdot \left\{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)\right\} \\ &= \frac{\rho^{i,0}_t \left\{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)\right\}}{\rho^{i,0}_t \left\{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right) \cdot \Psi^i_t\right\}} \\ &= \frac{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)}{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right) \cdot \Psi^i_t} \end{aligned} \tag{20}$$

$$\begin{aligned} P(\Theta^i_{t,l}, \chi^i_t|\mathbb{Z}_{1:t}) &= c_t^{-1} \frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\Phi^i_{t,j}} \rho^{i,0}_t \cdot P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right) \\ &= \left(\frac{p(\mathbf{z}^l_t|\mathbf{x}^i_t)}{\Phi^i_{t,l}}\right) \cdot \left(\frac{P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)}{1 - P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right) \cdot \Psi^i_t}\right), \end{aligned} \tag{21}$$

We present an updated track existence probability with the posterior association probabilities Eq. (20) and Eq. (21)

$$\begin{aligned} P\left(\chi^i_t|\mathbb{Z}_{1:t}\right) &= P\left(\Theta^i_{t,0}, \chi^i_t|\mathbb{Z}_{1:t}\right) + \sum_{l=1}^{L_t} P\left(\Theta^i_{t,l}, \chi^i_t|\mathbb{Z}_{1:t}\right) \\ &\approx \frac{\left(1 - \Psi^i_t\right) \cdot P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)}{1 - \Psi^i_t \cdot P\left(\chi^i_t|\mathbb{Z}_{1:t-1}\right)}. \end{aligned} \tag{22}$$

Using Eq. (21) we compute the score matrix S, Eq. (5), and solve the association problem using the Hungarian method [1]. When evaluating the posterior association and track existence probabilities, all parameters are automatically calculated in the procedure except for the observation model $p(\mathbf{z}^l_t|\mathbf{x}^i_t)$ and the clutter density $\varrho_{t,l}$. We design the observation model using Eq. (28) with several cues including appearance, shape, and

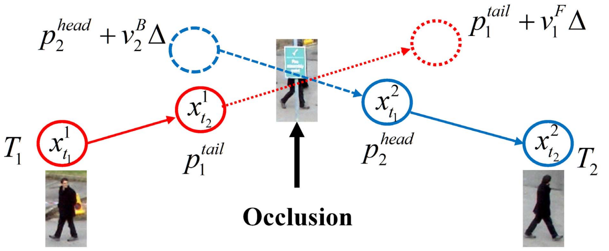


Fig. 2. Evaluation of the motion affinity between T_1 and T_2 under occlusion.

motion. The detail description of the clutter density is given in Section V-A.2.

Motivation and benefits. The proposed association method is motivated by the idea that a track with higher existence probability is firstly considered to be associated with detections when several tracks have similar likelihoods with detections. In other words, a more reliable detection is likely to originate from a track with high existence probability rather than a track with low existence probability. From extensive evaluation and comparison with different association methods as shown in Table III, we found this strategy allows us to determine the track-to-observation (or detection) pair more accurately, and improves tracking precision and reduces the number of ID switch. In addition, as discussed in [24], it reduces the computation complexity. Unlike JPDA [4], [35] and MHT [30] with exponential computation complexity, the computation complexity of the proposed method linearly increases with the number of tracks and the number of detections. Moreover, the track existence probability can be utilized for the basis of track initialization, termination, and linking as discussed in [21].

2) Track State Estimation: Once an associated observation is determined for each track using the proposed association method, the states of the track is estimated by Eq. (2). In our case, we exploit particle filtering [18] although several versions of Bayesian filtering methods exist: with weighted samples, the prediction, update, and resampling steps are recursively performed. The dynamic model $p\left(\mathbf{x}_t^i|\mathbf{x}_{t-1}^{i,n}\right)$ represents the temporal correlation of track states between consecutive frames, where n is a sample index. Four parameters consisting of 2D positions and sizes (width and height) to handle translation and scale change is transformed according to the Gaussian distribution $p\left(\mathbf{x}_t^i|\mathbf{x}_{t-1}^{i,n}\right) \sim \mathcal{N}(\mathbf{x}_t^i; \mathbf{x}_{t-1}^{i,n}, Q_t^i)$, where Q_t^i is a diagonal covariance matrix whose elements are the variances of the parameters, and more discussed in Section V-A.2. Sample weights $w_t^{i,n} \propto w_{t-1}^{i,n} \cdot p\left(\mathbf{z}_t^l|\mathbf{x}_t^{i,n}\right)$ are evaluated using the likelihood Eq. (28). Given N samples, the state of each object is determined by averaging the states of samples with their weights as $\hat{\mathbf{x}}_t^i = \sum_{n=1}^N w_t^{i,n}\mathbf{x}_t^{i,n} / \sum_{n=1}^N w_t^{i,n}$.

C. Track Management

In this section, we describe an automatic track management method for track initialization, link and termination.

1) Track Link and Termination: Once long-term occlusion occurs, it is extremely difficult in tracking an occluded object because the observation of the track could not be available and the appearance of the occluded object is significantly different from the updated reference model of the object. To deal with the long-term occlusion, we associate the fragmented tracks to link them. We consider a terminated track as a fragmented track or a complete track. To determine whether the track is terminated or not, we employ the track existence probability Eq. (22). When the probability is below to a termination threshold ($\theta = 0.65$ in our experiment), we consider it as the terminated track.

Now, we propose a track-to-track association method to link fragmented tracks, also determine whether the terminated tracks are complete trajectories or not at once. Let denote sets of terminated tracks, existing tracks and detections as $\{T_i^{(-)}\}_{i=1}^{n_a}$, $\{T_j^{(+)}\}_{j=1}^{n_b}$ and $\{y_j\}_{j=1}^{n_l} \subseteq \mathbf{Z}_t$, where n_a, n_b and n_l are the number of the elements of each set, respectively. Since data association events are mutually exclusive, we only consider the detection y_j which is not associated with any existing tracks.

Following association events are considered:

- $T_i^{(-)}$ associates with $T_j^{(+)}$;
- $T_i^{(-)}$ associates with y_j;
- $T_i^{(-)}$ is a complete (ended) trajectory.

Let us define a cost matrix for all events as follows:

$$S_{(n_a)\times(n_b+n_l+n_a)} = \left[A_{n_a\times n_b} \quad B_{n_a\times n_l} \quad C_{n_a\times n_a}\right], \qquad (23)$$

where $A = [a_{ij}]_{n_a\times n_b}$ and $B = [b_{ij}]_{n_a\times n_l}$ represent the first and second events, $a_{ij} = -\log\left(\Lambda(T_i^{(-)}, T_j^{(+)})\right)$ and $b_{ij} = -\log\left(\Lambda(T_i^{(-)}, y_j)\right)$ are the association scores computed by the track affinity and the observation models, Eq. (24) and Eq. (28). $C = diag\left\{c_1, ..., c_{n_a}\right\}$ represents the third event, and $c_i = -\log\left(1 - P_E(T_i^{(-)})\right)$ is a probability that the track is to be complete or not, where the existence probability of the terminated track $P_E(T_i^{(-)})$ is computed by Eq. (22).

Once the cost matrix is constructed, we can determine the optimal association pairs which are subject to minimizing the total cost of the matrix S using the Hungarian algorithm [1]. We then obtain an optimal assignment matrix as $O^* = \left[o_{ij}\right]_{n_a\times(n_b+n_l+n_a)}$. For each pair $o_{i,j} = 1$, we execute following operations:

- If $j \leq n_b$, link the tail of $T_i^{(-)}$ to the head of $T_j^{(+)}$ and allocate the label of $T_i^{(-)}$ into the linked track;
- If $n_b < j \leq n_b + n_l$, the states of $T_i^{(-)}$ is updated with y_j and $P_E(T_i^{(-)})$ is updated by Eq. (22);
- If $j > n_b + n_l$, $P_E(T_i^{(-)})$ is propagated by Eq. (9).

In the second and third operations, the existence probability of the terminated track is adaptively updated. If the existence probability of the terminated track exceeds to the termination threshold, the terminated track is recovered to the existing track. If not, we consider the track as the complete trajectory.

2) New Track Initialization: By applying the offline-trained detector at each frame, we can obtain object hypotheses with the detection responses. In order to find new object hypotheses, we search continuous and consistent detection responses having both overlapped areas and similar sizes within temporal sliding windows that are not already associated with any existing tracks. In our implementation, we link detections when the ratio of an overlapped area over an union area of detections is more than 0.5. If more than two detections are overlapped in neighboring frames, we associate them with the maximum ratio based on the greed algorithm [1]. When the object hypotheses are associated in T_{init} subsequent frames, we generate a new track with associated hypotheses (in our experiments, $T_{init} = 5$).

D. Online Model Learning

In this section, we present track affinity and observation models used in visual tracking and track management. In particular, we discuss an online learning method for learning a discriminative appearance model.

1) Track Affinity and Observation Models: In this paper, a track

$$T_i = \left\{ \mathbf{x}^i_{t^i_1:t^i_2} \right\}$$

is represented using several cues $T_i = \{A_i, S_i, M_i\}$, where A_i, S_i and M_i are appearance, shape and motion models, respectively (for clarity, we omit the time index t). To compute the affinity between two tracks T_i and T_σ we propose a track affinity model as follows:

$$\Lambda\left(T_i, T_\sigma\right) = \Lambda^A\left(T_i, T_\sigma\right) \cdot \Lambda^S\left(T_i, T_\sigma\right) \cdot \Lambda^M\left(T_i, T_\sigma\right), \quad (24)$$

The affinity score is computed based on similarities of appearance, shape and motion models. The appearance affinity is computed as follows:

$$\Lambda^A\left(T_i, T_\sigma\right) = \frac{1}{1+\exp(-H(X))}, \quad H(X) = H_i(f^\sigma) + H_\sigma(f^i), \quad (25)$$

where f^i and f^σ are appearance descriptors extracted from the tail (*i.e.* the last refined position) and the head (*i.e.* the first refined position) of the track i and the track σ as shown in Fig. 2. An online learning method to train discriminative appearance models H_i and H_σ and explanation of the appearance descriptor are given in the next section IV-D2.

The shape affinity between tracks is evaluated with their heights and widths. Here, the ratio between w_i $(or\ w_\sigma)$ and h_i $(or\ h_\sigma)$ is not fixed and can be different for each object and for each frame. The shape affinity is defined as

$$\Lambda^S\left(T_i, T_\sigma\right) = \exp\left(-\frac{1}{2}\left\{\frac{|h_i - h_\sigma|}{h_i + h_\sigma} + \frac{|w_i - w_\sigma|}{w_i + w_\sigma}\right\}\right). \quad (26)$$

As depicted in Fig. 2, the motion affinity of T_i and T_σ with frame gap Δ is evaluated using a tail of T_i and a head of T_σ based on a linear motion assumption as

$$\begin{aligned}\Lambda^M\left(T_i, T_\sigma\right) = \ & \mathcal{N}\left(p_i^{tail} + v_i^F \Delta;\, p_\sigma^{head}, \Sigma\right) \cdot \\ & \mathcal{N}\left(p_\sigma^{head} + v_\sigma^B \Delta;\, p_i^{tail}, \Sigma\right), \end{aligned} \quad (27)$$

The difference of the predicted position with the velocity and the refined position is assumed to follow the Gaussian distribution. Using the Kalman filtering [4] the forward velocity v_i^F and refined positions are estimated from the head to the tail of T_i, while the backward velocity v_σ^B and refined positions are estimated from the tail to the head of T_σ. Thus, the four dimensional state vector consisting of positions and velocities along with x and y coordinates are predicted and updated in the Kalman filtering process. We use the constant velocity model [4], [24] as a motion dynamic model and predict last estimated states using the model. Then, we update the predicted states with the associated detection determined by the association method described in the section IV-B.1. Thanks to the filtering, we evaluate the motion affinity with the estimated forward and backward motions of tracks, but also make trajectories more smooth with the refined positions.

The observation model $p\left(\mathbf{z}_t^l | \mathbf{x}_t^i\right)$ evaluates the likelihood between an associated observations -> $\mathbf{z}_t^l$ and a track

$$T_i = \left\{ \mathbf{x}^i_{t^i_1:t^i_2} | 1 \leq t^i_1 \leq t^i_2 \leq t \right\}.$$

The likelihood is evaluated using same cues as the affinity models but modified as follows

$$\begin{aligned}\Lambda^A\left(T_i, \mathbf{z}_t^l\right) &= \frac{1}{1+\exp(-H(X))}, \quad H(X) = H_i(f^z), \\ \Lambda^S\left(T_i, \mathbf{z}_t^l\right) &= \exp\left(-\frac{1}{2}\left\{\frac{|h_i - h_z|}{h_i + h_z} + \frac{|w_i - w_z|}{w_i + w_z}\right\}\right), \\ \Lambda^M\left(T_i, \mathbf{z}_t^l\right) &= \mathcal{N}\left(p_i^{tail} + v_i^F \Delta;\, p_z, \Sigma\right), \end{aligned} \quad (28)$$

Note that we evaluate the likelihood using all estimated states T_i up to frame t rather than instant states $\mathbf{x}_t^i$ at frame t. f^z is the appearance descriptor extracted from the location of $\mathbf{z}_t^l$. The forward motion model is only exploited to evaluate the motion affinity.

2) Online-Learned Discriminative Appearance Model: To learn discriminative appearance models of tracks, we collect training samples from tracking results at each frame and use them for online appearance learning. In single object tracking the discriminative appearance models [3], [16] are trained to grow discrimination power between the tracked object and scene background around the object. On the other hand, in multi-object tracking, the appearance models should distinguish well not only between the objects and the background but also between different objects.

To learn the discriminative appearance models effectively for multi-object tracking, we consider two situations for each object: a non-interacting object and an interacting object. In this paper, we define some objects to be interacting when they are closely located and/or (partially or fully) occluded each other. To determine whether the object i is interacting or not, we determine a set of objects D_i^σ interacting with the object i using the mahalanobis distance as follows:

$$D_i^\sigma = \left\{ d^\sigma_{i,t} | \left(p_t^\sigma - p_t^i\right)^T \left(\mathbf{S}_t^i\right)^{-1} \left(p_t^\sigma - p_t^i\right) \leq \gamma \right\}_{\sigma=1,\sigma\neq i}^{M_t}, \quad (29)$$

where p_t^i and p_t^σ are the positions along with x and y coordinates of the tracked objects. $S_t^i = \text{diag}[(w_t^i)^2\ (h_t^i)^2]$ is determined by the width w_t^i and the height h_t^i of the object i at frame t.

When the object is not interacting with any object $D_i^\sigma = \emptyset$, we collect N^+ positive samples from image patches extracted

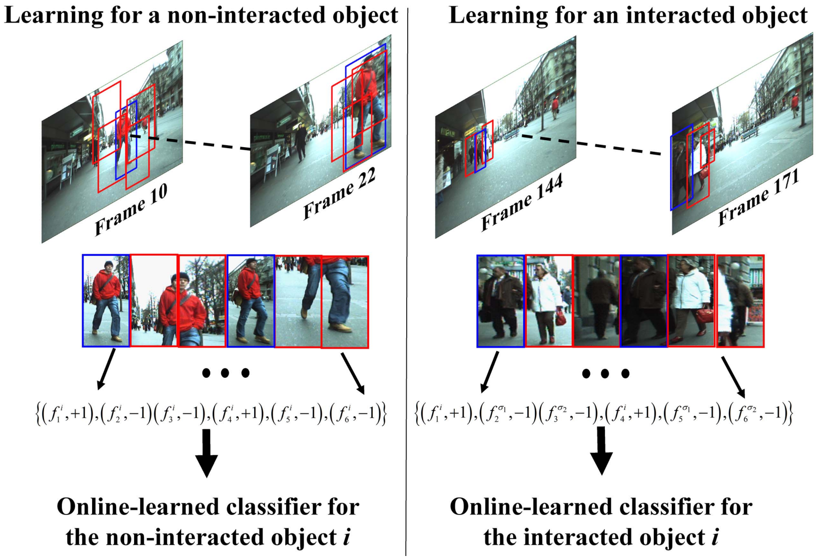

Learning for a non-interacted object
Frame 10
Frame 22
$\{(f_1^i,+1),(f_2^i,-1)(f_3^i,-1),(f_4^i,+1),(f_5^i,-1),(f_6^i,-1)\}$
Online-learned classifier for the non-interacted object *i*
Learning for an interacted object
Frame 144
Frame 171
$\{(f_1^i,+1),(f_2^{\sigma_1},-1)(f_3^{\sigma_2},-1),(f_4^i,+1),(f_5^{\sigma_1},-1),(f_6^{\sigma_2},-1)\}$
Online-learned classifier for the interacted object *i*

Fig. 3. Our sample collection strategy for learning discriminative appearance models for non-interacting and interacting objects. Blue and red rectangles mean positive and negative samples, respectively.

within positions and sizes of the object up to current frames. On the other hand, N^- negative samples are collected from the image patches at different locations around the object for discrimination between the object and background. In our case, we extract 10 negative sample patches at random locations to be overlapped with the positive sample patch less than 50%. However, we use negative sample patches from the positive samples of other interacting objects when the object is interacting with other objects. Fig. 3 shows our sample collection strategy for learning appearance models of the non-interacting and interacting objects.

At the image patch, an appearance descriptor by concatenating several feature histograms consisting of the HSV color histogram, the histogram of oriented gradient (HOG) [10], the local binary pattern (LBP) [25] is generated for capturing color, shape, and texture properties;

$$f_l^i = \left[f_{hsv_l}^i, f_{lbp_l}^i, f_{hog_l}^i \right] \in \Re^d.$$

Then, positive and negative samples of the track i at frame t are constructed by

$$\begin{array}{ll} \text{Positive samples:} & \\ & B^+ = \left\{ f_{l^+}^i, +1 \right\}, \ l^+ = 1, \ldots, N^+, \\ \text{Negative samples:} & \\ \text{(Non-interacting)} & B^- = \left\{ f_{l^-}^i, -1 \right\}, \ l^- = 1, \ldots, N^-, \\ \text{(Interacting)} & B^- = \left\{ f_{l^+}^\sigma, -1 \right\}, \ \sigma = 1, \ldots, M_t, \ \sigma \neq i. \end{array} \tag{30}$$

where l^+ and l^- are indexes of positive and negative samples, respectively. Once training samples are collected, we learn a discriminative appearance model of each track at frame t using ensemble learning [3]. The overall learning algorithm is described in Algorithm 1.

V. Experimental Results

As described in Algorithm 2, the proposed system has been implemented in MATLAB. More detailed explanation to implement our system is given in Sec. V-A. The datasets and evaluation metrics for performance evaluation are explained in Sec V-B and Sec. V-C. We first evaluate performance of our system by comparing with other state-of-the-art systems in Sec. V-D and speed of our system in Sec. V-E. In addition, we show how the main parts (*i.e.* visual tracking, track management and online model learning) affect the overall performance of our system in Sec. V-F.

Algorithm 2 The overall algorithm for the proposed multi-object tracking system

```
Input  : A detection set $\mathbf{Z}_t$, a terminated track set
         $\{T_i^{(-)}\}_{i=1}^{n_a}$ and an existing track set $\{T_j^{(+)}\}_{j=1}^{n_b}$ at
         frame $t-1$, where each track with 3-models
         $\{A_i, S_i, M_i\}$
Output: Updated terminated tracks $\{T_i^{(-)}\}_{i=1}^{n_a}$, existing
        tracks $\{T_j^{(+)}\}_{j=1}^{n_b}$ and 3-models $\{A_i, S_i, M_i\}$
Track-to-observation association: Associate all $T_j^{(+)}$ with
  $\mathbf{Z}_t$ and update $P_E(T_j^{(+)})$ as shown in Sec. IV-B1 ;
Tracking: For all $T_j^{(+)}$ with associated observations, do
  update states using particle filtering as shown in Sec. IV-B2 ;
Link and termination: Associate $T_i^{(-)}$ with $T_j^{(+)}$ and $Y_j$ as
  shown in Sec. IV-C1 ;
Initialization: Generate new tracks as shown in Sec. IV-C2 ;
Model learning: Learn $\{A_i, S_i, M_i\}$ with tracking results of
  $T_i^{(-)}$ and $T_j^{(+)}$ as shown in Sec. IV-D ;
```

A. Implementation

1) Detection: For VS-PETS 2009 and ETHMS datasets, we have used the public available detections provided by [2] and [20], [37], respectively. Thus, we can compare tracking performance of our system and [2] with same detections as in Table. II. Also, the comparison results between our system and [20], [37] with identical detections are reported in Table. II. Since public detections for CAVIAR and Hockey datasets are not provided, we have used the multiscale pedestrian detector provided by the [13]. A main reason is that the detector can be operated in almost real time (∼ 5fps on 640×480 images) by avoiding constructing finely sampled image pyramid: They exploit the gradient histogram extracted at a single scale to approximate feature responses at nearby scales. We have not manipulated setting parameters in the detector code [13] except for the image upscaling levels. For CAVIAR dataset, we have tuned the upscaling level to 2, but we have increased the level to 8 in order to detect small hockey players for the Hockey dataset.

2) System Parameters: All parameters have been found experimentally, and most remained identical parameters for all datasets. From the extensive experiments, we observe that the most parameters do not much affect the overall performance of our system. However, the clutter density $\varrho_{t,l}$ is a crucial parameter since the combination of the likelihood term $p(\mathbf{z}_t^l|\mathbf{x}_t^i)$ and clutter density $\varrho_{t,l}$ Eq. (15) determines the posterior data association probability Eq. (21) and posterior track existence probability Eq. (22). The data association probability and track existence probability are decreasing when likelihood score is fixed, but clutter density is increasing. When the clutter density is extremely high, a track could be quickly terminated and not associated with any observation. In our experiment, we have set the clutter density to 0.1 for all experiment although the clutter density can be accurately modeled using scene structure information (*e.g* entrances, exits and occluders).

In order to calculate data association and track existence probabilities, the likelihood model $p(\mathbf{z}_t^l|\mathbf{x}_t^i)$ Eq. (28) is also essential. However, the most parameters (*i.e.* the positions, sizes and velocities) of appearance, motion and shape models are automatically determined by tracking results. The covariance for forward and backward motions have been set to $\Sigma = \text{diag}[25^2\ 75^2]$ and fixed for all experiments.

The initial object size has been determined by averaging associated detections for 5 frames. The initial particle positions are drawn from a normal distribution with a standard deviation $q = 2 \cdot scale_h$ pixels, centered at the position of the last associated detection. The standard deviations for the position and size noises have been set to $q = 1.5 \cdot scale_h$ and $q = 2 \cdot scale_h$ pixels, receptively. The $scale_h = 2.5$ has been determined for a height of 180 pixels and automatically tuned by the estimated height of the tracked object. For all dataset, 150 samples are used in particle filtering process. To deal with abrupt motion change, we increased the standard deviation of the position to $q = 3 \cdot scale_h$ pixels for ETHMS datasets.

B. Dataset

We evaluate the performance of the proposed multi-object tracking system with four challenging datasets: the CAVIAR [11], the VS-PETS 2009 benchmark [12], the Hockey [26], and the ETH mobile scene (ETHMS) [14] datasets. The datasets are separated into two types: static scenes (CAVIAR and VS-PETS 2009) and moving scenes (ETHMS and Hockey).

The CAVIAR dataset includes 26 video sequences of a corridor in a shopping center taken by a single camera with frame size of 384×288. For fair comparison with other systems, we select 20 video sequences as reported in [20] and use the ground truth provided in [11].

TABLE I

PERFORMANCE COMPARISON BETWEEN OUR SYSTEM AND OTHER STATE-OF-ART TRACKING SYSTEMS FOR THE HOCKEY DATASET

Sequence	System	MOTP	MOTA	FN	FP	IDS
Hockey	Our system	59.3 %	80.3 %	18.8 %	0.2 %	0
	Conf. Map [7]	57.0 %	76.5 %	22.3 %	1.2 %	0
	Boost PF [26]	57.0 %	76.5 %	22.3 %	1.2 %	0
	Max. Ind. [8]	60.0 %	79.7 %	19.5 %	1.1 %	0

The VS-PETS 2009 dataset is captured from multiple static cameras, and we only used tracking sequences captured from view-1 with frame size of 768×576. We exploit PETS S2.L1 and PETS S2.L2 sequences in the dataset. PETS S2.L2 is relatively more difficult than PETS S2.L1 since the crowded density of PETS S2.L2 is much higher and frequent occlusions are occurred by object interactions.

The hockey dataset includes additional difficulties to track objects. Most of all, the scene is taken from a moving camera, while CAVIAR and PETS dataset are captured from a static camera. Thus, there exists abrupt motion change of players caused by the object and camera motions. Moreover, the appearances and sizes between players are significantly similar. Therefore, in associating detections with the tracks, the motion cue is more distinctive than appearance and shape cues.

The ETHMS dataset is taken by a pair of cameras on a moving stroller. We use Sunny day and Bahnhof sequences captured in busy streets scenes for comparison with [20], [29], [37]. The dataset is much more difficult to track objects than the Hockey dataset since the dataset is captured from front view cameras on the ground plane, while the Hockey dataset is captured from the top view camera. Thus, the motions and sizes of pedestrians change more abruptly, and severe occlusions occur. These challenges make data association more difficult. Though in the dataset left and right views are provided, we only employ the left view images in tracking objects without using depth and ground plane information.

C. Evaluation Metric

For quantitative evaluation for tracking performance, we utilize not only the CLEAR MOT metrics [6] but also metrics used in [19] and [20]:

- **MOTP** (↑): Multi-object tracking precision, intersection areas between bounding boxes of tracking results and ground truth over union areas of the bounding boxes.
- **MOTA** (↑) $= 1 - \frac{\sum_t m_t + fp_t + mme_t}{\sum_t g_t}$: Multi-object tracking accuracy, where m_t, fp_t, mme_t, and g_t are the number of misses, of false positives, of mismatches and total objects, respectively, at frame t.
- **FN** (↓) $= \frac{\sum_t m_t}{\sum_t g_t}$: The ratio misses in the sequences over the total number of objects present in all frames.
- **FP** (↓) $= \frac{\sum_t fp_t}{\sum_t g_t}$: The ratio false positives over the total number of objects present in all frames.
- **IDS** (↓): Identity switches, the number of times that an output track changes its matched GT identity.

TABLE II

PERFORMANCE COMPARISON BETWEEN OUR SYSTEM AND OTHER STATE-OF-ART SYSTEMS FOR THE CAVIAR, PETS AND ETHMS DATASETS

Dataset (Sequence)	System	MOTP	MOTA	IDS	GT	MT	PT	ML	FG	REC	PRE	FAF
CAVIAR (20 Seq.)	Our system	89.15%	87.78%	5	143	127/89.01%	16/10.99%	0/0.00 %	12	90.05%	96.05%	0.097
	PIRMPT [20]	—	—	4	143	86.00%	13.30%	0.70 %	4	88.10%	96.60%	0.082
	OLDAM [19]	—	—	11	143	84.60%	14.70%	0.70 %	11	89.40%	96.90%	0.085
	Hybrid [22]	—	—	11	143	84.60%	14.00%	1.40 %	17	89.00%	—	0.157
	Part Trk. [31]	—	—	17	140	75.70%	17.90%	6.40 %	35	75.20%	—	0.281
	Net. Flow[40]	—	—	15	140	85.70%	10.70%	3.60 %	20	76.40%	—	0.105
	Hier. DA [17]	—	—	12	143	78.30%	14.70%	7.00 %	54	86.30%	—	0.186
PETS (S2.L1)	Our system	69.72%	80.34%	3	23	23/100%	0/0.00%	0/0.00%	2	99.04%	90.24%	0.180
	OLMOAP [36]	—	—	0	19	89.50%	10.50%	0.00 %	9	91.80%	99.00%	0.05
	⋆ Energy Min. [2]	80.20%	91.60%	11	23	91.30%	4.35%	4.35 %	6	92.40%	98.40%	0.07
	PIRMPT [20]	—	—	1	19	78.90%	21.10%	0.00 %	23	89.50%	99.60%	0.02
	Conf. Map [7]	56.03%	79.70%	—	—	—	—	—	—	—	—	—
	Prob. Tracking [39]	53.80%	76.00%	—	—	—	—	—	—	—	—	—
	Global. Opt [5]	60.08%	66.00%	—	—	—	—	—	—	—	—	—
PETS (S2.L2)	Our system	63.43%	63.89%	28	74	45/60.81%	28/37.84%	1/1.35 %	51	68.29 %	88.83%	1.68
	⋆ Energy Min. [2]	59.40%	56.90%	99	74	37.84%	45.95%	16.22 %	73	65.50%	89.80%	1.43
	Conf. Map [7]	51.30%	50.00%	—	—	—	—	—	—	—	—	—
ETHMS (Bahnhof & Sunny Day)	Our system	62.43%	64.96%	13	125	88/70.40%	34/27.20%	3/2.40 %	45	80.82%	90.20%	0.642
	⋆ PIRMPT [20]	—	—	11	125	51/53.58%	37/38.95%	7/7.37%	23	76.80%	86.60%	0.891
	⋆ Online CRF [37]	—	—	11	125	68.00%	24.80%	7.20 %	19	79.00%	90.40%	0.637
	MT-TBD [29]	—	—	45	125	62.40%	29.60%	8.00%	69	78.70%	85.50%	—

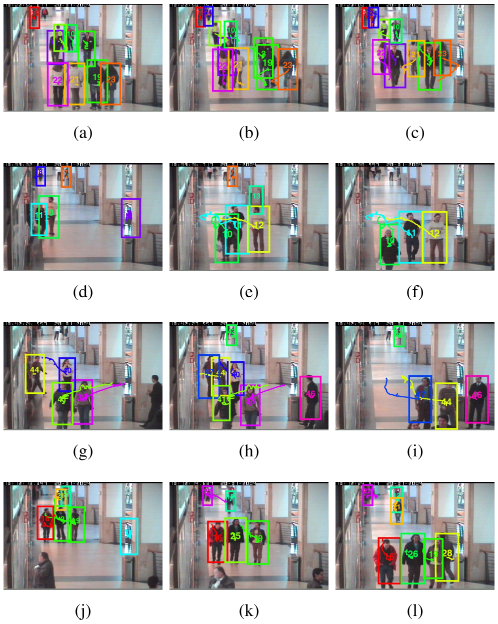

(a) (b) (c)
(d) (e) (f)
(g) (h) (i)
(j) (k) (l)

Fig. 4. From top to bottom, tracking results for CAVIAR - OneStopMoveEnter1, ShopAssistant1, ShopAssistant2 and ThreePastShop1 sequences are shown. (a) Frame #774. (b) Frame #866. (c) Frame #949. (d) Frame #1369. (e) Frame #1465. (f) Frame #1521. (g) Frame #3358. (h) Frame #3383. (i) Frame #3568. (j) Frame #936. (k) Frame #1092. (l) Frame #1214.

- **GT** : The number of objects in the ground truth.
- **MT** (↑): The number of mostly tracked objects, which the trajectories are tracked for more than 80%.
- **ML** (↓): The number of mostly lost objects, which the trajectories are tracked for less than 20%.
- **PT** $= 1 - MT - ML$: The number of partially tracked objects.
- **FG** (↓): Fragments, the number of times that a ground truth trajectory is interrupted.
- **REC** (↑): Recall, the number of correctly matched detections divided by the total number of detections in ground truth.
- **PRE** (↑): Precision, the number of correctly matched detections divided by the total number of output detections.
- **FAF** (↓): False alarm per frame, the number of false alarms per frame.

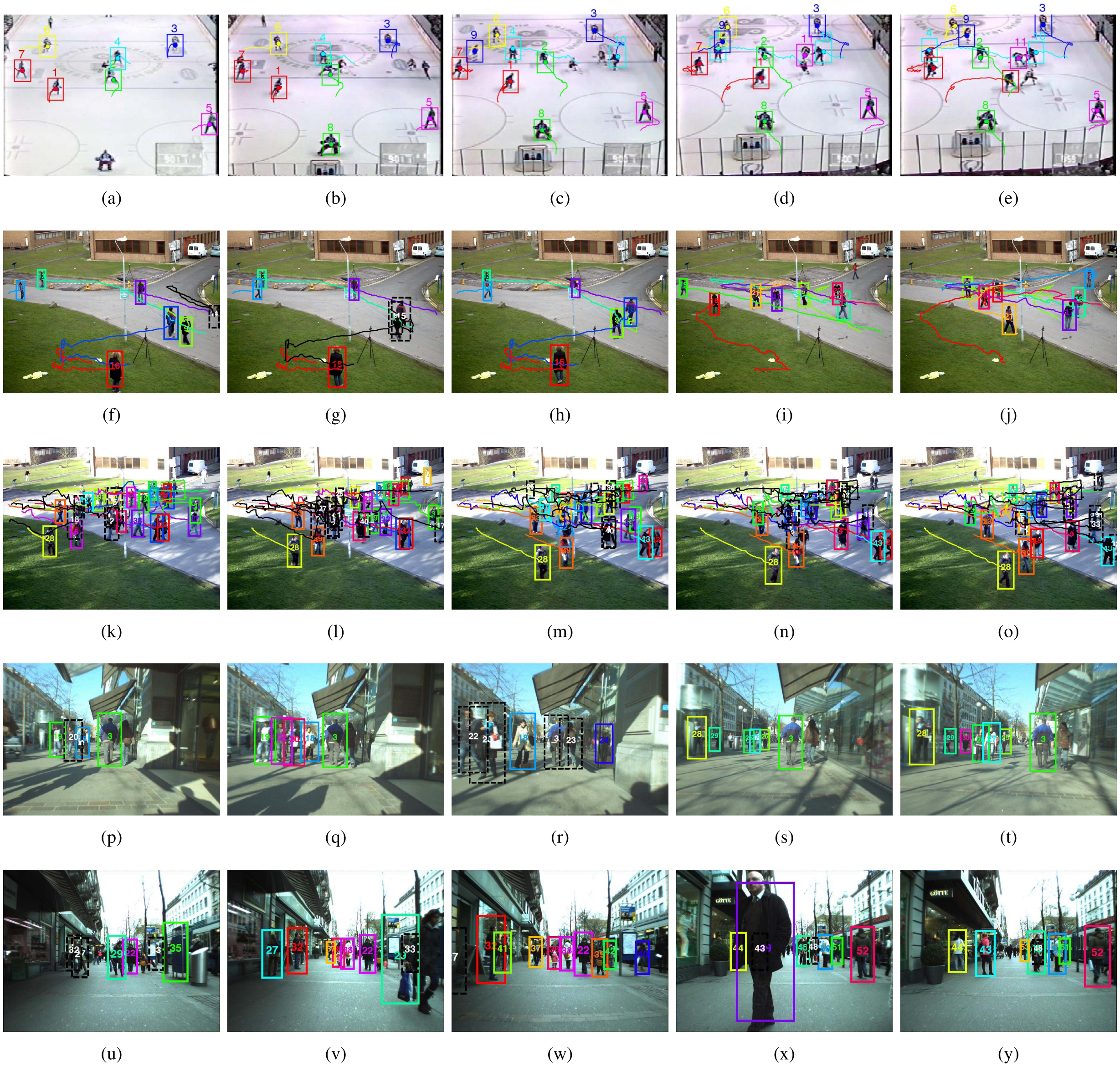
Here, the arrow symbol ↑ represents that higher scores indicate better results, and ↓ means that lower scores indicate better tracking results.

D. Performance Evaluation of the Proposed System

In Table I–II, the performance of our system is highlighted with gray color as in the first rows in each Table. We mark the online tracking systems with blue color, and the best scores for metrics with red color. Also, tracking systems evaluated with the same detections as our system are marked with ⋆.

Results for Hockey dataset. We employ the pedestrian detector provided by [13], and the locations and sizes of objects are manually labeled for all frames because public detections and ground truth are not provided. The comparison results are given in Table I. Thanks to the low FN and FP scores, we achieve the best performance compared to state-of-art tracking systems [7], [8], [26] in terms of MOTA. The performance differences for the MOTP and IDS metrics are negligible. The qualitative tracking results are shown in Fig. 5(a) – 5(e), and our system has successfully tracked motions of players although their motions abruptly change and appearances are similar.

Results for CAVIAR dataset. In Caviar dataset, we obtain detections with the pre-trained pedestrian detector [13] since public detections are not available. The comparison results are shown in Table II. As can be seen, our system achieves significant improvements of the performance; our system achieves the best performance in terms of MT, ML and REC scores. In addition, we reduce IDS by over 50%, compared to [17], [19], [22], [40]. Remarkably, our system achieves the performance improvement without future frames. Some visual tracking results are shown in Fig. 4. Although the pedestrian 1


(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

(p) (q) (r) (s) (t)

(u) (v) (w) (x) (y)

Fig. 5. Tracking results with the proposed system for the Hockey, VS-PETS 2009 and ETHMS datasets. At each frame, states (*i.e.* positions and sizes) and identifies (IDs) of tracked objects are illustrated in color boxes and color numbers. The terminated tracks and their IDs are shown in black boxes and white numbers. Also, for Hockey, PETS-L1 and PETS-L2 datasets (captured in top views), the trajectories are depicted with color lines. (a) Hockey #35. (b) Hockey #52. (c) Hockey #70. (d) Hockey #86. (e) Hockey #100. (f) PETS-L1 #618. (g) PETS-L1 #626. (h) PETS-L1 #630. (i) PETS-L1 #741. (j) PETS-L1 #784. (k) PETS-L2 #67. (l) PETS-L2 #82. (m) PETS-L2 #97. (n) PETS-L2 #101. (o) PETS-L2 #105. (p) ETH-Sunny #224. (q) ETH-Sunny #237. (r) ETH-Sunny #254. (s) ETH-Sunny #298. (t) ETH-Sunny #321. (u) ETH-Bahnhof #471. (v) ETH-Bahnhof #495. (w) ETH-Bahnhof #520. (x) ETH-Bahnhof #612. (y) ETH-Bahnhof #623.

TABLE III

QUANTITATIVE EVALUATION RESULTS OF THREE MAIN PARTS OF OUR SYSTEM. WE TAKE THE AVERAGE OF THE EXPERIMENTAL RESULTS FOR THE FOUR SEQUENCES WITH DIFFERENT EXPERIMENTAL SETTINGS

Dataset (Sequence)	System	MOTP	MOTA	F. Neg.	F. Pos.	IDS	GT	MT	PT	ML	FG
PETS & ETHMS	Our system	64.55%	68.54%	22.96%	8.31%	44	223	156/69.96%	63/28.25%	4/1.79%	98
	Exp. 1-(a)	59.57%	67.67%	20.70%	11.36%	59	223	149/66.82%	69/30.94%	5/2.24%	105
	Exp. 1-(b)	60.53%	67.74%	20.63%	11.33%	62	223	154/69.06%	64/28.70%	4/2.24%	101
	Exp. 2	60.30%	67.73%	20.84%	10.14%	63	223	155/69.51%	64/28.70%	4/1.79%	128
	Exp. 3	58.49%	66.58%	23.26%	9.84%	72	223	132/69.96%	86/28.25%	5/1.79%	102

and 3 are severely occluded by other pedestrians, our system accurately keeps their IDs and estimates locations and sizes as shown in Fig. 4(a) – 4(c).

Results for VS-PETS dataset. The comparison results are shown in Table II. We exploit the same detections and ground truth provided by [2] for both sequences. We observe that the performance of our system outperforms [2] in terms of the most metrics for the PETS-L1 and PETS-L2 sequences, receptively. In particular, the MT and REC scores are greatly improved. However, in return the PRE and FAF scores are deceased and increased, respectively, due to more false positives.

Although different detections and ground truth are used, a large performance gap between our system and other online-tracking systems [7], [39] is shown in terms of the MOTP and MOTA. Compared to batch tracking systems [5], [20], [36], our system shows the improved performance. Remarkably, our system achieves the perfect MT and ML scores for the PETS-L1 sequence. As shown in Fig. 5(f) - 5(o), our system robustly tracks multiple objects and constructs long trajectories although there exist frequent occlusions cased by other object interactions and the scene clutter (streetlight).

Results for ETHMS dataset. The performance of our system is compared in Table II. We exploit same detections and ground truth provided in [20] and [37]. Note that batch tracking systems [20], [37] use detections of the entire sequence at once, but we use detections at each frame only. Nevertheless, we improve the MT scores by more than 16.82% and 2.4% and reduce ML scores by about 4.97% and 4.8%, when comparing [20], [37], separately. Also, our system is far beyond the performance of online tracking system [29] for all metrics. The performance improvement shows the superiority of our online-tracking system.

For Sunny day and Bahnhof sequences, Fig. 5(p) - 5(y) show qualitative evaluation results using our system. In both examples, motion and shape affinity scores are less informative when associating (linking) tracks due to the abrupt change of motions and shapes caused by the camera motion. However, by exploiting our online-learned discriminative appearance models, the proposed system successfully links fragmented tracks. The results also explain the performance improvement in terms of MT, ML and REC scores in Table II. As can be seen, we observe that pedestrian 3 (Fig. 5(p) - 5(t)), 22, and 43 (Fig. 5(u) - 5(y)) are correctly tracked even under long-term occlusions.

E. Speed of the Proposed System

The proposed systems was implemented using the MATLAB on a PC with 3.07 GHz CPU without any parallel programming. The complexity of the implemented system depends on the number of detections and objects. For less crowded scenes such as PETS-L1, ETHMS, and CAVIAR datasets, the run time is 0.2 $\sim$ 0.5 (sec/frame). For very crowded scene such as PETS-L2, the run time is 0.75 $\sim$ 1.0 (sec/frame). The most expensive processing in our system is spent in extracting an appearance descriptor, which are shared by online-learning and tracking. Furthermore, we test the processing time of our system with pre-defined appearance model. By removing the online-learning procedure, the computation cost is reduced by 16%. This results means that computation expense of our tracking system is not significantly increased by our online-learning method.

F. Performance Evaluation of the Main Parts

To verify the effectiveness of the main parts, we compare the performance of our system

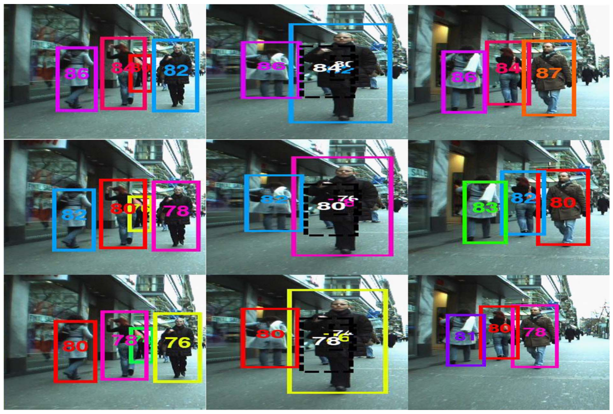


Fig. 6. For ETH-Bahnhof sequence, tracking results with different association methods: From top to bottom, tracking results with the our data association, greedy, and Hungarian methods are depicted.

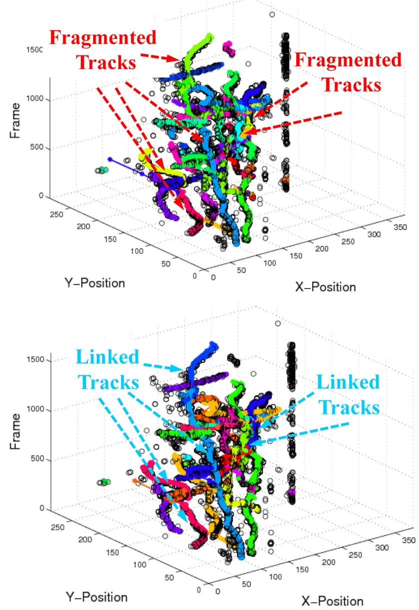


Fig. 7. For the CAVIAR-OneStopMoveEnter1 sequence, detections (black circles) and the estimated trajectories of objects (color lines) up to 1590 frames are depicted. The estimated trajectories without our linking method and with it are shown in the first and second rows, respectively.

- Exp. 1-(a): Replace our data association using track existence probability with the greedy association [1].
- Exp. 1-(b): Replace our data association using track existence probability with the Hungarian association [1].
- Exp. 2: Eliminate the linking part in track management.
- Exp. 3: Eliminate the online appearance learning part.

We exploit PETS-L1, PETS-L2, ETH-Bahnhof and ETH-Sunny sequences and provide average results of evaluation metrics for the four sequences in Table III.

1) Visual Tracking: In Exp. 1, we show the effectiveness of the visual tracking part based on our data association with track existence probability by replacing the association with Greedy (Exp. 1-(a)) and Hungarian methods (Exp. 1-(b)). Although the performance difference between the greedy and Hungarian methods is not much, we obtain better results using our data association. In par-

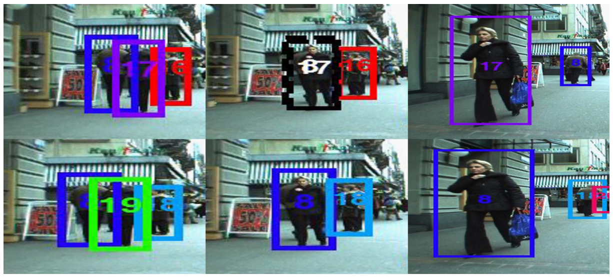


Fig. 8. For the ETH-Bahnhof sequence, tracking results with our online-learning method (top) and without it (bottom).

ticular, our association method allows us to improve the MOTP score and reduce IDS numbers. It implies that the positions and sizes of objects have been more accurately estimated, and IDs of objects are conserved better due to correct association of our method. Fig. 6 shows tracking results using different data association methods. In the first row, two women walking together (ID 84 and 86) are successfully tracked using our data association method even though they are occluded by other pedestrian. On the other hand, their IDs are mismatched using greedy and Hungarian methods as shown in the second and third rows.

2) Track Management: The evaluation results of our system without the track linking method are shown in Exp. 2. As expected, performance of our system is degraded in term of all metrics. Especially, we observe that the number of fragmented tracks has been significantly increased, but also tracking precision has been decreased. Furthermore, estimated trajectories of our system with the linking method and without it are compared in Fig. 7. The experimental results confirm that the fragmented tracks can be accurately linked using our linking method.

3) Online Model Learning: In Exp. 3, we evaluate similarity of appearances using the Bhattacharyya distance of multi-cue histograms in stead of scores of our discriminative classifiers in Eq. (25) and Eq. (28). We observe that the MOTP, IDS and MT scores are severely deteriorated since the discrimination power is reduced. We further compare tracking results in Fig. 8 when using our online-learning method and does not. As shown in the first row, using the online-learning ID 8 of the old man is accurately maintained even under the full occlusion, but its ID is not matched without the learning method in the second row.

VI. Discussion

Each part of our system can be applied for other tracking systems. The visual tracking part based on our data association, which sequentially grows trajectories of objects with online detections, can be used to a tracking algorithm devising data association of online tracking systems [7], [9], [26], [31], [32]. Furthermore, it can be used to link detections for local association in the batch tracking systems [17], [22]. Also, the track management part can be easily incorporated into other tracking systems [7], [9], [33] as a method to link fragmented tracks generated by occlusions. In addition, the online learning part can replace other online learning methods to learn a discriminative appearance model [7], [19], [33].

VII. Conclusion

We have proposed an online multi-object tracking system consisting of three main parts. The visual tracking part based on a novel data association with track existence probability allows us to track objects robustly under partial occlusions. To deal with long-term occlusions, our track management part performs track-to-track association to link fragmented tracks. For successful association in other two parts, the online learning part incrementally updates discriminative appearance models with online tracking results.

Our experimental results using challenging tracking datasets have shown the improved performance of the proposed system, compared to other state-of art tracking systems. We further have demonstrated the effectiveness and usefulness of each part of the system. Indeed, we expect that the main parts can be applied in other tracking systems and used in a wide range of application scenarios. To increase description ability of our system, a part-based appearance model would be beneficial and increase performance of our system. Furthermore, prior knowledge of scene structures allows us to design the clutter density model more accurately, and the well designed model could enhance the performance of data association.

References

[1] R. Ahuja, T. Magnanti, and J. Orlin, *Network Flows*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[2] A. Andriyenko, S. Roth, and K. Schindler, "Continuous energy minimization for multi-target tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 1–15, Jan. 2014.

[3] S. Avidan, "Ensemble tracking," in *Proc. CVPR*, 2005, pp. 494–501.

[4] Y. Bar-Shalom, T. Fortmann, and M. Scheffe, "Joint probabilistic data association for multiple targets in clutter," *Inform. Sci. Syst.*, vol. 24, pp. 843–854, Jan. 1980.

[5] J. Berclaz, F. Fleuret, and P. Fua, "Robust people tracking with global trajectory optimization," in *Proc. CVPR*, 2006, pp. 744–750.

[6] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP J. Image Video Process.*, vol. 2008, no. 1, pp. 1–10, Feb. 2008.

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

[8] W. Brendel, M. Amer, and S. Todorovic, "Multiobject tracking as maximum weight independent set," in *Proc. IEEE CVPR*, Jun. 2011, pp. 1273–1280.

[9] Y. Cai, N. de Freitas, and J. J. Little, "Robust visual tracking for multiple targets," in *Proc. ECCV*, 2006, pp. 107–118.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, Jun. 2005, pp. 886–893.

[11] (2003, Jul. 11). *CAVIAR Dataset* [Online]. Available: http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

[12] (2009). *VS-PETS Dataset* [Online]. Available: http://www.cvg.rdg.ac.uk/PETS2009/a.html

[13] P. Dollár, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. BMVC*, 2010.

[14] A. Ess, B. Leibe, K. Schindler, and L. J. Van Gool, "A mobile vision system for robust multi-person tracking," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.

[15] F. Glover, "A template for scatter search and path relinking," in *Proc. 3rd Eur. Conf. Artificial Evol.*, 1997, pp. 1–51.

[16] H. Grabner and H. Bischof, "On-line boosting and vision," in *Proc. IEEE CVPR*, Jun. 2006, pp. 260–267.

[17] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *Proc. 10th ECCV*, 2008, pp. 788–801.

[18] M. Isard and A. Blake, "Condensation—Conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.

[19] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *Proc. IEEE CVPR*, Jun. 2010, pp. 685–692.

[20] C.-H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?" in *Proc. IEEE CVPR*, Jun. 2011, pp. 1217–1224.

[21] N. Li and X. R. Li, "Target perceivability and its applications," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2588–2604, Nov. 2001.

[22] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: HybridBoosted multi-target tracker for crowded scene," in *Proc. IEEE CVPR*, Jun. 2009, pp. 2953–2960.

[23] H. Jiang, S. Fels, and J. J. Little, "Optimizing multiple object tracking and best view video synthesis," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 997–1012, Oct. 2008.

[24] D. Musicki and B. La Scala, "Multi-target tracking in clutter without measurement assignment," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 44, no. 3, pp. 877–895, Jul. 2008.

[25] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[26] K. Okuma, A. Taleghani, N. de Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *Proc. 8th ECCV*, 2004, pp. 28–39.

[27] J. J. Pantrigo, S. Sánchez, A. S. Montemayor, and A. Duarte, "Multi-dimensional visual tracking using scatter search particle filter," *Pattern Recognit. Lett.*, vol. 29, no. 8, pp. 1160–1174, 2008.

[28] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE CVPR*, Jun. 2011, pp. 1201–1208.

[29] F. Poiesi, R. Mazzon, and A. Cavallaro, "Multi-target tracking on confidence maps: An application to people tracking," *Comput. Vis. Image Understand.*, vol. 117, no. 10, pp. 1257–1272, 2013.

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

[31] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.

[32] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1200–1207.

[33] X. Song, J. Cui, H. Zha, and H. Zhao, "Vision-based multiple interacting targets tracking via on-line supervised learning," in *Proc. IEEE ECCV*, Oct. 2008, pp. 642–655.

[34] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," in *Proc. IEEE CVPR*, Jun. 2012, pp. 1815–1821.

[35] J. Vermaak, S. Maskell, and M. Briers, "Unifying framework for multi-target tracking and existence," in *Proc. IEEE 8th Int. Conf. Inform. Fusion*, Jul. 2005, pp. 250–258.

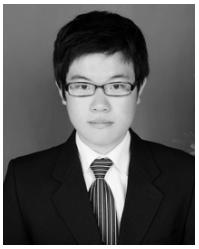
[36] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Proc. IEEE CVPR*, Jun. 2012, pp. 1918–1925.

[37] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proc. IEEE CVPR*, Jun. 2012, pp. 2034–2041.

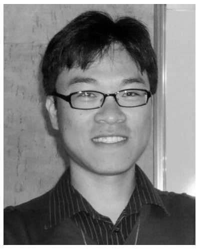
[38] B. Yang and R. Nevatia, "Online learned discriminative part-based appearance models for multi-human tracking," in *Proc. ECCV*, 2012, pp. 484–498.

[39] J. Yang, P. A. Vela, Z. Shi, and J. Teizer, "Probabilistic multiple people tracking through complex situations," in *Proc. CVPRW*, 2009, pp. 79–86.

[40] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.



Seung-Hwan Bae received the B.S. and M.S. degrees in information and communications from Chungbuk National University, Cheongju, Korea, and the Gwangju Institute of Science and Technology, Gwangju, Korea, in 2009 and 2010, respectively, where he is currently pursuing the Ph.D. degree with the Computer Vision Laboratory. His research interests include main research topics in computer vision, such as multiobject tracking, object detection, and medical image analysis.



Kuk-Jin Yoon received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1998, 2000, and 2006, respectively. He was a Post-Doctoral Fellow with the PERCEPTION Team, INRIA, Grenoble, France, from 2006 and 2008, and joined the School of Information and Communications at the Gwangju Institute of Science and Technology, Gwangju, Korea, as an Assistant Professor, in 2008, where he is the Director of the Computer Vision Laboratory. His research interests include main research topics in computer vision, such as stereo, structure-from-motion, multiobject tracking, and SLAM.