IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 20, 2019, accepted November 8, 2019, date of publication November 13, 2019,
date of current version November 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2953276

Online Multi-Object Tracking With GMPHD Filter and Occlusion Group Management

YOUNG-MIN SONG[1], (Student Member, IEEE), KWANGJIN YOON[1], YOUNG-CHUL YOON[2], KIN CHOONG YOW[3], (Senior Member, IEEE), AND MOONGU JEON[1], (Senior Member, IEEE)

[1]School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea
[2]LG Electronics, Seoul 07796, South Korea
[3]Faculty of Engineering and Applied Science, University of Regina, Regina, SK S4S 0A2, Canada

Corresponding author: Moongu Jeon (mgjeon@gist.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) through the Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis under Grant 2014-0-00077, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) through the Real world object recognition based on deep learning and domain adaptation under Grant 2019R1A2C2087489.

ABSTRACT In this paper, we propose an efficient online multi-object tracking method based on the Gaussian mixture probability hypothesis density (GMPHD) filter and occlusion group management scheme where a hierarchical data association is utilized for the GMPHD filter to reduce the false negatives caused by missed detection. The hierarchical data association consisting of two modules, detection-to-track and track-to-track associations, can recover the lost tracks and their switched IDs. In addition, the proposed grouping management scheme handles occlusion problems with two main parts. The first part, ''track merging'' can merge the false positive tracks caused by false positive detections from occlusions. The occlusion of the false positive tracks is usually measured with some metric. In this research, we define the occlusion measure between visual objects, as sum-of-intersection-over-each-area (SIOA) instead of the commonly used intersection-over-union (IOU). The second part, ''occlusion group energy minimization (OGEM)'' prevents the occluded true positive tracks from false ''track merging''. Each group of the occluded objects is expressed with an energy function and an optimal hypothesis will be obtained by minimizing the energy. We evaluate the proposed tracker in benchmarks such as MOT15 and MOT17 which are public datasets for multi-person tracking. An ablation study in training dataset reveals not only that ''track merging'' and ''OGEM'' complement each other, but also that the proposed tracking method shows more robust performance and less sensitiveness than baseline methods. Also, the tracking performance with SIOA is better than that with IOU for various sizes of false positives. Experimental results show that the proposed tracker efficiently handles occlusion situations and achieves competitive performance compared to the state-of-the-art methods. In fact, our method shows the best multi-object tracking accuracy among the online and real-time executable methods.

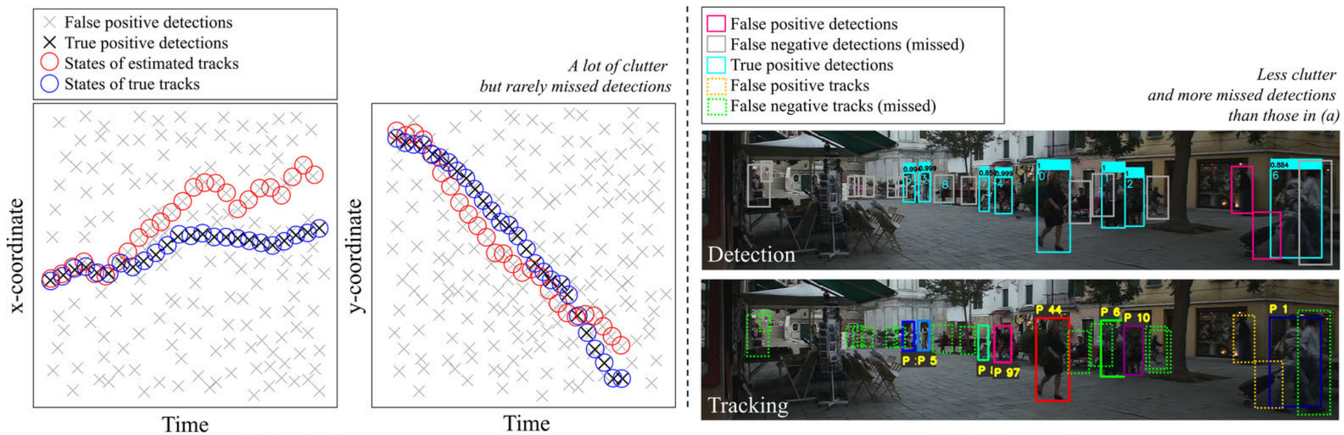
INDEX TERMS Multiple object tracking, GMPHD filter, hierarchical data association, occlusion handling, energy minimization.

I. INTRODUCTION

Multi-object Tracking (MOT) has become one of key techniques for intelligent video surveillance [5], [6] and autonomous vehicle systems [7] in the last decade.

In the view of the processing pipeline, many state-of-the-art MOT methods [13]–[52] have exploited the tracking-by-detection paradigm because of its representative advantages (utility and modularity) that developers only need to implement a tracking module and make it receive detection modules' results as inputs. However, object detectors are still imperfect in the presence of various illumination conditions (too bright or dark), occlusions between objects, and high similarity between foregrounds (objects) and backgrounds as described in Figure 1(b), although deep neural networks (DNN) based detectors such as FRCNN [10], SDP [11],

The associate editor coordinating the review of this manuscript and approving it for publication was John See.



(a) Radar/sonar system

(b) Video system

FIGURE 1. **Examples for comparing observations (detection) and states (tracking) between (a) radar/sonar and (b) video system. The radar/sonar sensors receive a lot of clutter (false positive errors) but rarely miss objects (false negative errors), whereas the detector in video data tends to receive a few clutter around the objects and misses more objects than the radar/sonar sensors do. In (b), we used MOT17-01-FRCNN sequences.**

and YOLO [12] have achieved better performance than hand-crafted features based detectors such as ACF [8] and DPM [9] did. Thus, these complex situations generate false positive and missed detections resulting in false positive and missed tracks in 2D video data.

The methods based on tracking-by-detection are categorized into two approaches: offline and online processes. The most different point between two approaches is that whereas the offline process can see the whole time sequences, i.e., whole detection results, at once, the online process can see only the frames from initial time 1 to the current processing time k. In other words, from the system user's perspective, whereas the offline method is suitable for post-processing, the online process is needed for real-time application.

Thus, many offline methods [29], [30], [32], [35], [37], [48] take advantage of the global optimization models. [30], [37], [48] exploit graphical models to solve MOT tasks. Pirsiavash *et al.* [37] designed a min-cost flow network where the nodes and the directed edges indicating observations (detections) and tracklets' hypotheses respectively, form a directed acyclic graph (DAG). The DAG's shortest (min-cost) path can be found with Dijkstra's algorithm. Choi [30] divided the tracking problem into subgraphs and solved each subgraph as conditional random field inferences in parallel. Keuper *et al.* [48] applied vision-based perspective to the proposed graph optimization model. Feature points' trajectories and bounding boxes build low-level and high-level graph models, respectively, and then they find the optimal association results between the two levels' graph models. Rezatofighi *et al.* [32] and Kim *et al.* [29] considered all possible hypotheses for data association. Reference [32] assumed *m*-best solutions and [29] pruned out invalid hypotheses using their own rule, because it involves the exponentially increasing complexity with a tree structure. In addition, Milan *et al.* [35] proposed a sophisticated energy minimization technique considering detection, appearance, dynamic model, mutual exclusion, and target persistence for MOT task in video. Those offline methods have strengths to generate the accurate and refined tracking results but are not suitable for practical real-time applications.

On the other hand, since the online approaches cannot apply the global optimization models, intensive motion analysis and appearance feature learning have been popularly utilized with a hierarchical data association framework and an online Bayesian model [14], [15], [18], [22], [24], [26], [38], [44]. Yoon *et al.* [24] proposed a relative motion analysis model between all objects in a frame, and then improved the work [24] by adding the cost optimization function using context constraints in [22]. Bae and Yoon [26] exploited the incremental linear discriminant analysis (LDA) for appearance learning and presented a tracklet confidence based data association framework. Also, in [14], they improved their previous work [26] by using the DNN based appearance learning instead of the incremental LDA. As DNN has achieved breakthroughs in object classification and detection, some online MOT algorithms have focused on how to adopt the DNN for appearance learning and apply it into their tracking frameworks. Yoon *et al.* [15] exploited the siamese convolutional neural networks (CNN) [53] to train appearance models. They trained the deep appearance networks selectively where only detection responses matched with high confidence between the historical objects are queued in the recent few frames. Then, they combined the trained networks to a simple Bayesian tracking model with a Kalman filter [61]. Chen *et al.* [38] employed a re-identification (Re-ID) model [54] to their tracking framework. They measured the similarity between detection and tracking by calculating the distance between Re-ID feature vectors of the objects. Then, they associated the pairs of detections and tracking objects which madethe sum of the distances minimal. Both approaches [15] and [38] proposed online Bayesian tracking models with conventional DNN models to measure the similarity between the visual objects. These online MOT methods had proposed successful solutions with excellent tracking accuracy but their intensive analysis and learning processes took heavy computing resources and time. Also, even if they just employ conventional DNN models through state-of-the-art GPU processing techniques, the requirement for a lot of computing resources are still inevitable and it makes the trackers difficult to achieve real-time speed. Recently, In a different online MOT research direction, the probability hypothesis density (PHD) filter [2], [3] have been employed as an emerging theory for many online MOT methods [16]–[18], [23], [39], [43]–[45]. That is because Vo *et al.* [2] and Vo and Ma [3] provided not only an online multi-target Bayesian filtering framework but had also approximated the original PHD recursions [1] involving multiple integrals to a closed-form implementation, which alleviated the computational intractability. However, the PHD filter was originally designed for multi-target tracking in radar/sonar systems [4], [64] which have a relatively more false positives and less false negatives detections, while detections in image data (video) systems [8]–[10] have a relatively less false positives and more false negatives detections. Figure 1 describes the different characteristics between the two domains. To handle tracking problems caused by false detections in video data, the online tracking algorithms based on the PHD filter employed appearance learning [18], [44], hierarchical data association [16], [17], [23], and fusing multiple detectors [39], [43], [45]. These PHD filter based online MOT methods are most relevant to our proposal so we will address them in Section II in detail.

These latest MOT research trends motivated our work in terms of three main contributions that we will describe later in this section. Also, they reminded us of the requirements for practical MOT applications that indicated real-time speed especially. Thus, in this paper, we propose an online multi-object tracking method to resolve the practical tracking problems which are false positives and false negatives (missed) detections caused by occlusions and similarity between backgrounds and interested objects (persons) in video data system. Our main contributions are described as follows:

1) To apply the GMPHD filter in video data system, we extended the conventional GMPHD filter based tracking process with a hierarchical data associations (HDA) strategy. Also, we revised the equations of the GMPHD filter as a new cost function for HDA. HDA consists of detection-to-track associations (D2TA) and track-to-track associations (T2TA). The cost matrix of each association stage is solved by using the Hungarian method [60] with linear complexity $O(n^3)$ (assignment problem). These D2TA and T2TA can recover lost tracks caused by missed detections. Furthermore, we did not use image information except bounding boxes with detection confidence scores in HDA because the usage of image information (visual features) involves additional learning steps and makes it hard to achieve real-time speed.

2) To handle occlusions in video-based tracking systems, we devised "tracking merging" and "occlusion group energy minimization (OGEM)" which complements each other. "Tracking merging" relieves false positive tracks and "OGEM" recovers false "track merging" by using the occluded objects' group energy minimization. "Tracking merging" runs in tracking-level so it is different from detection-level merging such as non-maximum-suppression. To measure overlapping ratios between occluded objects, we devised a new metric named sum-of-intersection-over-each-area (SIOA) and use this metric instead of the extensively used intersection-over-union (IOU). For "OGEM", we devised a new energy function to find the optimal energy state having in a group of occluded objects. "Tracking merging" and "OGEM" follow D2TA and T2TA, respectively. We name both techniques as occlusion group management (OGM).

3) Consequently, we proposed an online multi-object tracking method with GMPHD filter and occlusion group management (GMPHD-OGM). In regards to optimization techniques, the first and second contributions locally optimize tracking processes which minimizes the association cost matrix and the occlusion group energy. We evaluated the proposed tracking method on the MOT15 [5] and MOT17 [6] benchmarks. The ablation study on the training set showed that our method is more robust than the given baseline method. The qualitative and quantitative evaluation results also showed that GMPHD-OGM efficiently handled the defined tracking problems caused by occlusion. Moreover, the proposed method achieved competitive tracking performance against state-of-the-art online MOT algorithms in terms of CLEAR MOT metrics [56] and the metrics defined in [57], especially in value of "tracking accuracy (MOTA) versus speed (FPS)".

The related works are described in Section II. In Sections III and IV, we elaborate the GMPHD filter based tracking method with HDA and OGM. In Section V, our method is evaluated against baseline method and state-of-the-art methods on the popular benchmarks MOT15 [5] and MOT17 [6]. We conclude this paper with proposal for future work in Section VI. Some preliminary results of this work were presented in Song and Jeon [16] and Song *et al.* [17].

II. RELATED WORKS

Our proposed tracking method is influenced from PHD filter based online multi-object tracking, and grouping approaches (topology and relative motion analysis).

A. PHD FILTERING THEORY

The PHD filter [1]–[3] was originally designed to deal with radar/sonar data based multi-object tracking (MOT) systems. Mahler [1] proposed a recursive Bayes filter equations for the PHD filter which optimizes MOT processes in radar/sonar systems with a random-finite set (RFS) of states and observations. Following this PHD filtering theory, Vo *et al.* [2] proposed a sequential Monte Carlo (SMC) implementation of the PHD filter by using particle filtering and clustering, named as the SMCPHD filter. In [3], Vo *et al.* [3]

implemented the governing equations by using the Gaussian mixture model as a closed-form recursion, named as the Gaussian mixture probability hypothesis density (GMPHD) filter. The SMCPHD filter is the first implementation of the PHD filter but requires relatively large amount of computation in the clustering processes for particle filtering, so the GMPHD filter has been more widely exploited for extensions and applications. The Gaussian mixture cardinalized PHD (GMCPHD) filter [62] is one of representative extensions of the GMPHD recursion and jointly propagates the posterior intensity and the posterior cardinality distribution, while the GMPHD filter propagates only the posterior intensity of objects. In their numerical experiments, even if the GMCPHD filter showed the better performance in terms of tracking accuracy and the cardinality (the number of objects) estimation, those improvements involved about three times slower speed than the GMPHD filter. Besides, in the scenario that the number of objects was changed drastically, tracking accuracy and speed became worse than that of the GMPHD filter. Most recently, an implementation of the generalized labeled multi-Bernoulli (GLMB) [63] filter was presented. Unlike the PHD and GMCPHD recursions [2], [3], [62], the GLMB filter does not need data associations by using a multi-Bernoulli RFS approximation, so it contains much more complex equations than other RFS based multi-object Bayes filters. In addition, the GMCPHD filter and the GLMB filter calculate the posterior cardinality distribution that theoretically assumes up to infinity but constrains to a finite value, e.g., 100 or 200 for a practical implementation. Especially in scenarios with a small number of objects, both filters are inefficient compared to the simplest implementation [3]. Thus, for the practical purpose (online approach with real-time speed), we employed the GMPHD filter as a base model among many RFS-based online Bayes filters.

B. ONLINE MOT USING THE PHD FILTER IN VIDEO DATA

As demand increases on online and real-time tracker in video-based tracking system [13], the GMPHD filter have been an emerging tracking model, recently. However, in the original domains the tracking algorithm should estimate true tracks (states) from a lot of detections (observations) as shown in Figure 1(a). While the radar/sonar sensors received massive number of false positives but rarely missed any observations, visual object detectors on the other hand generate much less false positives and also more missed detections as shown in Figure 1(b). That is because many additional techniques have been exploited in video-based tracking. Song and Jeon [16] extended the GMPHD filter based tracking with a two-stage hierarchical data association strategy to recover fragmented and lost tracks. They defined the affinity in the track-to-track association stage by using the tracks' linear motion and color histogram appearance. This approach is an intuitive implementation of the GMPHD filter to handle tracking problems, but it cannot correct the false associations already made in the detection-to-track association. Sanchez-Matilla and Cavallaro [23] proposed a detection confidence based data association schemes with the PHD filter. Strong (high confidence) detections initiate and propagate tracks but is weak (low confidence) detections only propagate existing tracks. This strategy works well when the detection results are reliable. However, the tracking performance is dependent on the detection performance, and is especially weak to long-term missed detections. There have been more intensive solutions [18], [40], [44] using appearance learning or motion modeling. Kutschbach *et al.* [44] added the kernelized correlation filters (KCF) [55] for online appearance update to overcome occlusion in the naive GMPHD filtering process. They demonstrated a robust online appearance learning to re-find the IDs of the lost tracks. However, the updating of appearance information of every object in every frame is a process that requires heavy computing resources. Fu *et al.* [18] added an adaptive gating technique and an online group-structured dictionary (appearance) learning strategy into the GMPHD filter. They made the GMPHD filter into a sophisticated tracking process that is fit for video-based MOT. Sanchez-Matilla *et al.* [40] proposed a global motion model based tracking by using long short-term memory models. Some methods [39], [43], [45] proposed fusion models to complement the false detections. Kutschbach *et al.* [45] designed a fusion model of blob detector [66] and head detector [67] in the GMPHD filter based tracking framework. Fu *et al.* [39] utilized full-body detector [9] and body-part detector [65] in their tracking-by-detection method. Baisa and Wallace [43] proposed another type of fusion model which tracks different types of multiple objects (persons and cars) simultaneously by using the object detector and classifier such as FRCNN [10] in parallel processes. These MOT methods with the detector fusion models and appearance learning increased tracking accuracies but inevitably reduced processing speeds, even if the GMPHD filter provided a fast online MOT framework. Besides the extensions of the GMPHD filter, in the rare case that other RFS based tracking theory was employed, Kim *et al.* [20] extended the GLMB filter [63] for MOT applications in video data. They designed a hybrid multi-object measurement likelihood using appearance learning for missed detections and groupings for occlusions. They devised a graceful extension of the GLMB filter based online MOT for video data, it still had the constraint on the number of objects like the original GLMB. Although the GMCPHD filter and the GLMB filter not only showed better accuracies in the preliminary researches [62], [63] than the GMPHD filter, since both filters involved the constrained assumption on the cardinality up to 100 or 200 and additional mathematical techniques such as divergence estimation for implementations. Grouping approach e.g., using relative motion and topological models, have already been exploited in [24] and [22]. The key difference between their methods and ours is that [24] and [22] consider the relations between all objects in a scene while we only consider topological information in the group of occluded objects. Grouping the occluded objects can exclude unnecessary associations and focus on the solving of sub-problems, and also it reduces computing time. Thus, we employed the GMPHD filter as a base model and extended the baseline method by using our proposed hierarchical data associations and occlusion group management scheme for a highly practical applications (online and real-time) without using image information except bounding boxes.

III. PROPOSED ONLINE MULTI-OBJECT TRACKING METHOD

In this section, we briefly introduce the general tracking process of the Gaussian mixture probability hypothesis density (GMPHD) filter in Subsection III-A. In III-B, we will address how to extend the GMPHD filter with a hierarchical data association strategy in video-based online MOT systems.

A. THE GMPHD FILTER

The Gaussian mixture model (GMM) of the GMPHD filter contains means, covariances, and weights which are propagated at every time step as follows: *Initialization*, *Prediction*, *Update*, and *Pruning*. We employ this basic process of the GMPHD filter but revise it to fit the video-based MOT system. The states and observations in the GMPHD filter are represented by

$$X_k = \{x_k^1, \ldots, x_k^{I_k}\}, \tag{1}$$

$$Z_k = \{z_k^1, \ldots, z_k^{J_k}\}, \tag{2}$$

where X_k and I_k denote a set of objects' states and the number of them at time k, respectively. A state vector $\boldsymbol{x_k}$ is composed of (c_x, c_y, v_x, v_y), where c_x, c_y are the center coordinates of the bounding box, and v_x and v_y are the velocities of x- and y-directions of the object, respectively. Likewise, Z_k and J_k denote a set of observations (detection responses) and the number of them at time k, respectively. An observation z_k is represented with (c_x, c_y). Equations (3) and (4) describe the basic notations of state and observation.

$$x_k^i = \{c_{x,k}, c_{y,k}, v_{x,k}, v_{y,k}\}^T, \tag{3}$$

$$z_k^j = \{c_{x,k}, c_{y,k}\}^T. \tag{4}$$

The four steps of the tracking process of the GM-PHD filter are: *Initialization*, *Prediction*, *Update*, and *Pruning* as follows.

Initialization:

$$v_0(x) = \sum_{i=1}^{I_0} w_0^i \mathcal{N}(x; m_0^i, P_0^i), \tag{5}$$

where the GMM is initialized by the initial observations from the detection responses. Besides, when an observation fails to find the association pair, i.e., failed to update the object state, the observation will initialize a new Gaussian model (a new state). The Gaussian probability function $\mathcal{N}$ represents tracking objects with weight w, mean vector $\boldsymbol{m}$, object state vector $\boldsymbol{x}$, and covariance matrix P. In this step, we set the initial velocities of the mean vector to zeros. Each weight is set to the normalized confidence value of the corresponding detection response.

Prediction:

$$v_{k-1}(x) = \sum_{i=1}^{I_{k-1}} w_{k-1}^i \mathcal{N}(x; m_{k-1}^i, P_{k-1}^i), \tag{6}$$

$$m_{k|k-1}^i = F m_{k-1}^i, \tag{7}$$

$$P_{k|k-1}^i = Q + F P_{k-1}^i (F)^T, \tag{8}$$

where we assume that the GMM representing the objects' states was initialized or active at the previous frame $k-1$ in (6). In (7) and (8), we can predict the state at time k using the Kalman filtering based on the state at time $k-1$. F is the state transition matrix and Q is the process noise covariance matrix where both F and Q are constants in our tracker. In (7), $m_{k|k-1}^i$ is derived by using the velocity of m_{k-1}^i. Covariance $P_{k|k-1}^i$ is also predicted by the Kalman filtering in (8).

Update:

$$v_{k|k}(x) = \sum_{i=1}^{I_{k|k}} w_k^i(z) \mathcal{N}(x; m_{k|k}^i, P_{k|k}^i), \tag{9}$$

$$q_k^i(z) = \mathcal{N}(z; H m_{k|k-1}^i, R + H P_{k|k-1}^i (H)^T), \tag{10}$$

$$w_k^i(z) = \frac{w_{k|k-1}^i q_k^i(z)}{\sum_{l=1}^{I_{k|k-1}} w_{k|k-1}^l q_k^l(z)}, \tag{11}$$

$$m_{k|k}^i(z) = m_{k|k-1}^i + K_k^i (z - H m_{k|k-1}^i), \tag{12}$$

$$P_{k|k}^i = [I - K_k{}^i H] P_{k|k-1}^i, \tag{13}$$

$$K_k^i = P_{k|k-1}^i (H)^T (H P_{k|k-1}^i (H)^T + R)^{-1}. \tag{14}$$

Update step is to update a state x represented by a Gaussian model with a mean vector m and a covariance matrix P. An optimal z is found by data association which will be presented in III-B in detail. In the perspective of tracker, the update step follows after the data association. After finding the set of z, the GMM is updated from (6) to (9). R denotes the observation noise covariance. H denotes the observation matrix to transit a state vector to an observation vector. Both R and H are constants in our tracker.

Pruning:

$$\tilde{X}_k = \{m_k^i : w_k^i \geq \theta_w, i = 1, \ldots, I_k\}, \tag{15}$$

$$\tilde{W}_k = \{w_k^i : m_k^i \in \tilde{X}_k, i = 1, \ldots, I_k\}, \tag{16}$$

$$\tilde{W}_k = \{w_{\tilde{k},1}, \ldots, \tilde{w}_{k,\tilde{I}_k}\}, \tilde{I}_k = |\tilde{W}_k|, \tag{17}$$

$$w_k^i = \frac{\tilde{w}_k^i}{\sum_{l=1}^{\tilde{I}_k} \tilde{w}_k^l}. \tag{18}$$

$$X_k = \tilde{X}_k. \tag{19}$$

Pruning step handles the false positive tracks caused by the false positive detections. The states with the weights under threshold θ_w are pruned as in (15). We experimentally set θ_w to 0.1 and the weights of the surviving states are normalized as shown in (18).

B. HIERARCHICAL DATA ASSOCIATION

Video-based tracking systems have inherent problems as shown in Figure 1(b). Generally, when objects are not

Algorithm 1 Proposed Online MOT Algorithm

```
                                   ▷ $k$ : the current frame number
                             ▷ $X_{k-1}$ : a set of states at time $k-1$
                             ▷ $Z_k$ : a set of observations at time $k$
                                ▷ $\sigma_m$ : threshold for track merging
                    ▷ $\tau_{T2T}$ : the minimum track length for T2TA
                  ▷ $\theta_{T2T}$ : the maximum frame interval for T2TA
         ▷ $T^{live}$ : a {key:id,value:tracklet} set of live tracklets
         ▷ $T^{lost}$ : a {key:id,value:tracklet} set of lost tracklets
procedure GMPHD_OGM($k$,$X_{k-1}$,$Z_k$,$\sigma_m$,$\tau_{T2T}$,$\theta_{T2T}$,$T^{live}$,$T^{lost}$)
    $l = |X_{k-1}|$;                         // the number of states
    $m = |Z_k|$;                             // the number of observations
    $G_{k-1}, G_k$;                          // a set of occlusion groups at time k-1 and k.
    if $k = 1$ or $l = 0$ then
        Initialize states $X'_k$ with $Z_k$;
        $G_{k-1} = G_k$;
        $X_k = MERGE(X'_k, \sigma_m, G_k)$;
        return $X_k$;
    end if
  /* 1. Detection-to-Track Association (D2TA) */
    $C_{D2T}[1 \dots l][1 \dots m]$;         // for cost matrix
    $P_{D2T}[1 \dots l]$;                    // for pairing observations' indices
    /*predict states $X_{k-1}$ to be $X_{k|k-1}$*/
    for $i = 1$ to $l$ do
        $X_{k|k-1}[i] = PREDICT(X_{k-1}[i])$;
    end for
    /*calculate the GMPHD filter cost matrix $C_{D2T}$*/
    for $i = 1$ to $l$ do
        for $j = 1$ to $m$ do
            $C_{D2T}[i][j] = COST_{D2T}(X_{k|k-1}[i], Z_k[j])$;
        end for
    end for
    /*find min-cost pairs by the Hungarian method*/
    $P_{D2T} = HungrianMethod(C_{D2T})$;
    /*update and birth states*/
    /*update $X_{k|k-1}$ with the min-costly observations*/
    for $i = 1$ to $l$ do
        $X'_k[i] = UPDATE(X_{k|k-1}[i], Z_k[P_{D2T}[i]])$;
    end for
    /*prune $X_{k|k-1}$ with the weight under 0.1*/
    for $i = 1$ to $l$ do
        $X'_k[i] = PRUNE(X_{k|k-1}[i])$;
    end for
    for $j = 1$ to $m$ do
        if $Z_k[j]$ is not assigned to update any state then
            Initialize newly birth state $x$ with $Z_k[j]$;
            $X'_k = X'_k \cup \{x\}$;
        end if
    end for
  /* 2. Merge States and Find Occlusion Groups */
    $G_{k-1} = G_k$;
    $X_k = MERGE(X'_k, \sigma_m, G_k)$;
    /*manage tracklet pool after D2TA and MERGE*/
    for $i = 1$ to $|X_k|$ do
        if $X_k[i]$ is active then
            update $T^{live}[X_k[i].id]$ with $X_k[i]$;
            delete $T^{lost}[X_k[i].id]$;
        else
            update $T^{lost}[X_k[i].id]$ with $X_k[i]$;
            delete $T^{live}[X_k[i].id]$;
        end if
    end for
  /* 3. Track-to-Track Association (T2TA) */
    $t_1 = |T^{lost}|$;                      // the number of lost tracklets
    $t_2 = |T^{live}|$;                      // the number of live tracklets
    $C_{T2T}[1 \dots t_1][1 \dots t_2]$;     // for cost matrix
    $P_{T2T}[1 \dots t_1]$;                  // for pairing observations' indices
    /*calculate the GMPHD filter cost matrix $C_{T2T}$*/
    for $i = 1$ to $t_1$ do
        for $j = 1$ to $t_2$ do
            $C_{T2T}[i][j] = COST_{T2T}(T^{lost}[i], T^{live}[j], \tau_{T2T}, \theta_{T2T})$;
        end for
    end for
    /*find min-cost pairs by the Hungarian method*/
    $P_{T2T} = HungrianMethod(C_{T2T})$;
    /*update tracklets and manage tracklet pool after T2TA*/
    for $i = 1$ to $l$ do
        $X'_k[i] = UPDATE(X_{k|k-1}[i], Z_k[P_{D2T}])$;
    end for
    for $i = 1$ to $|X_k|$ do
        if $X_k[i]$ is active then
            update $T^{live}[X_k[i].id]$ with $X_k[i]$;
            delete $T^{lost}[X_k[i].id]$;
        else
            update $T^{lost}[X_k[i].id]$ with $X_k[i]$;
            delete $T^{live}[X_k[i].id]$;
        end if
    end for
  /* 4. Occlusion Group Energy Minimization (OGEM) */
    if $k > 1$ and $|G_{K-1}| > 0$ then
        $OGEM(k, G_{k-1}, X_k)$;
        /*manage tracklet pool after OGEM*/
        for $i = 1$ to $|X_k|$ do
            if $X_k[i]$ is active then
                update $T^{live}[X_k[i].id]$ with $X_k[i]$;
                delete $T^{lost}[X_k[i].id]$;
            else
                update $T^{lost}[X_k[i].id]$ with $X_k[i]$;
                delete $T^{live}[X_k[i].id]$;
            end if
        end for
    end if
    return $X_k$;                            // return final states $X_k$
end procedure
```

detected, the objects' IDs are frequently changed and the tracks are fragmented if only detection-to-track association is employed. To prevent these problems, we take advantage of the hierarchical data association (HDA) strategy that has been widely used in many online multi-object tracking methods [14], [16], [17], [23], [26]. Thus, in this paper, we propose a simple HDA scheme with just two stages, a detection-to-track (D2T) association stage, followed by a and track-to-track (T2T) association stage. We implement the both association methods with the two individual GMPHD filtering processed as shown in 2. In each stage, the tracking process follows the GMPHD filter as given in Section III-A but utilizes two different observations and states. With them, we derive a cost function based on the weight w_k from (11) as follows:

$$Cost(x^i_{k|k-1}, z^j_k) = -\ln w^i_k(z^j_k), \tag{20}$$

where w^i_k indicate the weight value, assuming that observation z^j_k updates state $x^i_{k|k-1}$. We use $-\ln w^i_k(z^j_k)$ as a cost between $x^i_{k|k-1}$ and z^j_k. Then, cost matrix C can be built from every pair of values in the state set $X_{k|k-1}$ and observation set Z_k as follows:

$$C[i, j] = Cost(X_{k|k-1}[i], Z_k[j]). \tag{21}$$

After the cost matrix C is built, the Hungarian method [60] is used to solve it. Then, the optimal pairs between observations and states are found, and consequently state $x_{k|k-1}$ is updated to x_k in D2T and T2T associations. In III-Band III-B, we introduce the definition of observations and states in each association stage with more detailed usage of the cost function.

Stage 1. Detection-to-Track Association (D2TA). Observation set Z_k is filled with detection responses at time k. We assume that state set X_{k-1} already exists from time $k-1$, and then $X_{k|k-1}$ is predicted by using the Kalman filtering as shown in (6)-(8). Thus, the cost matrix C_{D2T} is easily calculated with these sets $X_{k|k-1}$ and Z_k.

Stage 2. Detection-to-Track Association (D2TA). A simple temporal analysis of tracklet is introduced. A tracklet means a fragment of the track, and becomes a calculation unit. Before T2TA, all tracklets are categorized into two types, according to success or failure of tracking at the present time k as follows:

$$T^{lost}_k = \{\tau^{lost}_{1,k}, \dots, \tau^{lost}_{i,k}\}, \tag{22}$$

$$\tau^{lost}_{i,k} = \{a^i_s, .., a^i_t\}, \quad 0 \le s < t < k, \tag{23}$$

$$T^{live}_k = \{\tau^{live}_{1,k}, \dots, \tau^{live}_{j,k}\}, \tag{24}$$

$$\tau^{live}_{j,k} = \{a^j_s, .., a^j_t\}, \quad 0 \le s < t,\ t = k, \tag{25}$$

$$a^i_s = \{c_{x,s}, c_{y,s}, v_{x,s}, v_{y,s}\}^T, \tag{26}$$

$$a^i_t = \{c_{x,t}, c_{y,t}, v_{x,t}, v_{y,t}\}^T, \tag{27}$$

where T indicates a set of tracks, and "*live*" indicates that tracking succeeds at time k. "*lost*" indicates that tracking fails at time k. The two attributes are not compatible then $T^{lost}_k \cup T^{live}_k = T^{all}_k$ and $T^{lost}_k \cap T^{live}_k = \phi$ are satisfied. Then, for the T2TA, observation set Z_k is filled with the first (oldest) elements a^j_ss of "*live*" tracklets. However, the state set $X_{k|k-1}$ is not filled with the last (most recent) elements a^i_ts of "*lost*" tracklets. One prediction step is needed as follows:

$$x^i_t = \{c_{x,t}, c_{y,t}, \frac{c_{x,t} - c_{x,s}}{t-s}, \frac{c_{y,t} - c_{y,s}}{t-s}\}^T, \tag{28}$$

$$x^i_{k|k-1} = F^{T2T} x^i_t, \tag{29}$$

$$F^{T2T} = \begin{pmatrix} 1 & 0 & d_f & 0 \\ 0 & 1 & 0 & d_f \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{30}$$

$$d_f(i, j) = \text{frame difference between } a^i_t \text{ and } a^j_s. \tag{31}$$

In (30) $\frac{c_{x,t}-c_{x,s}}{t-s}$ and $\frac{c_{y,t}-c_{y,s}}{t-s}$ are the averaged velocities in the directions of the x-axis and y-axis, respectively. The velocities are calculated by subtracting the center position of the first object state a^i_s from that of the last state a^i_t, and dividing it by the frame difference $t - s$ which is equivalent to the length of "*lost*" tracklet $\tau^{lost}_{i,k}$. D2TA has the identical time interval "*1*" between states and observations in transition matrix F, whereas in T2TA, each cost of matrix C_{T2T} has different time interval (frame difference) between states and observations. Variable d_f depends on which state of "*lost*" tracklet and observation of "*live*" tracklet are paired. (29) means the prediction process of state with linear motion analysis. Finally, the cost matrix C_{T2T} is filled by (29) and the oldest element a^j_s of live tracklet $\tau^{live}_{j,k}$.

The pseudo-code in Algorithm 1 includes the procedures presented in this section. *Initialization*, *Prediction*, Cost-minimization, *Update*, and *Pruning* in D2TA correspond to each of line 5-10, 13-15, 16-21, 22-24, and 25-27 in Algorithm 1. Tracklet-categorization, Cost-minimization, *Update* in T2TA correspond to line 36-44, 49-54, and 55-66 in Algorithm 1, respectively.

IV. OCCLUSION GROUP MANAGEMENT SCHEME

In Section III, we pointed out that the proposed online multi-object tracking method is based on the GMPHD filtering theory with a two-stage hierarchical data association. However, the tracking results from that method still has uncertainties, even if we effectively extend the conventional GMPHD filter to be suitable for video-based tracking. To improve the tracking performance, we focused on handling occlusions which can generate false positive detections. The false positive detections inevitably generate false positive tracks as shown in Figure 3 and the second row (D2TA) of Figure 5. To resolve the tracking problems, we design a new occlusion group management (OGM) scheme. OGM consists of "*Track Merging*" and "*Occlusion Group Energy Minimization (OGEM)*" routines which execute just after D2TA and T2TA, respectively. Figure 2 shows the tracking pipeline with those two components of OGM. Not only is our OGM technique able to decrease false positive tracking results , it is also able to prevent occluded tracks from falsely performing

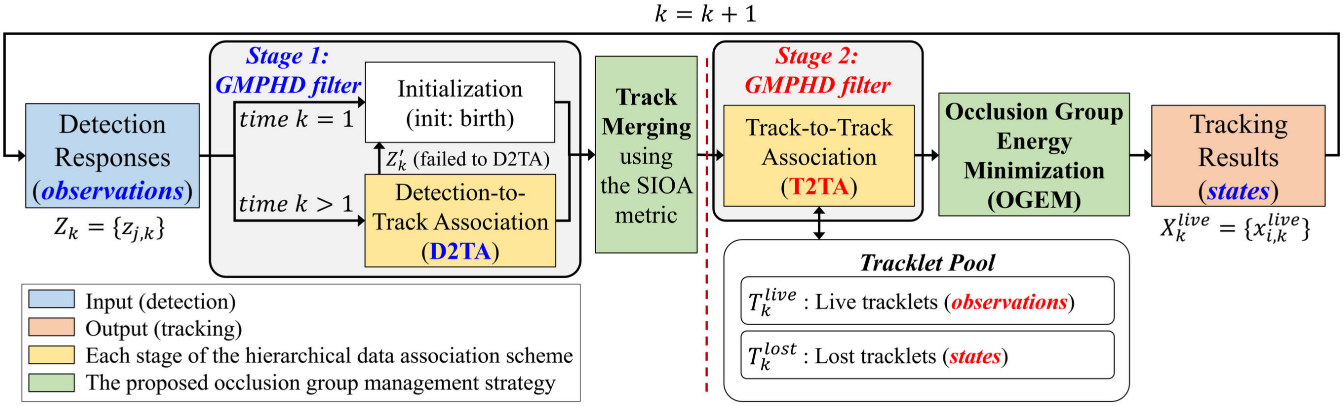

$k = k + 1$
Detection Responses (*observations*)
$Z_k = \{z_{j,k}\}$
Stage 1: GMPHD filter
time $k = 1$
time $k > 1$
Initialization (init: birth)
Z'_k (failed to D2TA)
Detection-to-Track Association (D2TA)
Track Merging using the SIOA metric
Stage 2: GMPHD filter
Track-to-Track Association (T2TA)
Occlusion Group Energy Minimization (OGEM)
Tracking Results (*states*)
$X_k^{live} = \{x_{i,k}^{live}\}$
Tracklet Pool
T_k^{live} : Live tracklets (*observations*)
T_k^{lost} : Lost tracklets (*states*)
Input (detection)
Output (tracking)
Each stage of the hierarchical data association scheme
The proposed occlusion group management strategy

FIGURE 2. Flow chart of the proposed online multi-object tracking method. The red dotted line divides the proposed hierarchical data association into two stages (stage 1 and 2). Each stage has an individual GMPHD filtering based tracking process. The states and observations of each stage are marked as blue and red, i.e., D2TA and T2TA, respectively. The meanings and usages of those states and observations are described in Section III-B.

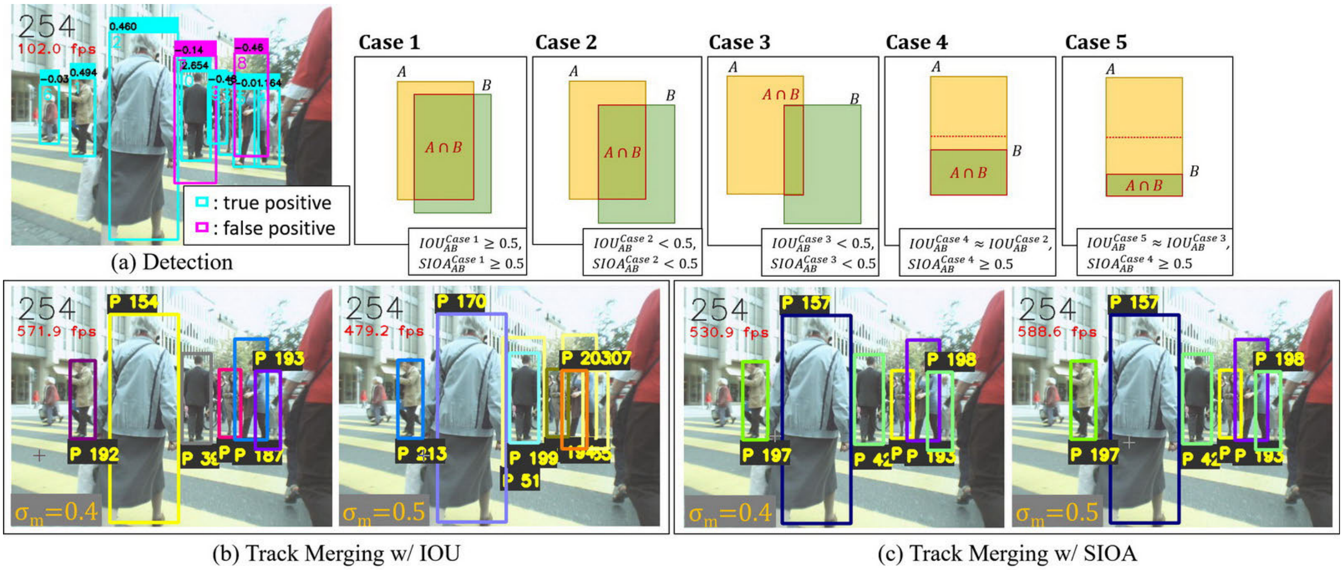

254
102.0 fps
: true positive
: false positive
(a) Detection
Case 1
Case 2
Case 3
Case 4
Case 5
A
B
$A \cap B$
$IOU_{AB}^{Case\,1} \geq 0.5$, $SIOA_{AB}^{Case\,1} \geq 0.5$
$IOU_{AB}^{Case\,2} < 0.5$, $SIOA_{AB}^{Case\,2} < 0.5$
$IOU_{AB}^{Case\,3} < 0.5$, $SIOA_{AB}^{Case\,3} < 0.5$
$IOU_{AB}^{Case\,4} \approx IOU_{AB}^{Case\,2}$, $SIOA_{AB}^{Case\,4} \geq 0.5$
$IOU_{AB}^{Case\,5} \approx IOU_{AB}^{Case\,3}$, $SIOA_{AB}^{Case\,4} \geq 0.5$
571.9 fps
479.2 fps
530.9 fps
588.6 fps
σ_m=0.4
σ_m=0.5
(b) Track Merging w/ IOU
(c) Track Merging w/ SIOA

FIGURE 3. Case study about "Track Merging" and the qualitative results on MOT17-05-DPM training sequence at frame 254. The overlapping ratios between the occluded objects can be measured with the *IOU* and *SIOA* metrics. When the sizes of A and B differ largely (more than twice size) such as case 4 and 5, IOU does not distinguish between not only Case 2 and 4 but also Case 3 and 5. For the same detection results with true and false positives, under the different merging thresholds σ_m values 0.4 and 0.5, SIOA is less sensitive to merge size variant false positive bounding boxes than IOU.

"track merging". The effectiveness of the proposed OGM method is discussed in our experimental results in Section V in more details.

A. TRACK MERGING

Merging neighboring objects when the distances between them is under a threshold is proposed in [3] already. However, it uses only point-to-point distances without considering regional information e.g., overlapping ratio between visual objects (bounding boxes). To measure the overlapping ratio, the intersection-over-union (IOU) metric, which was originally designed to measure mAP in object detection research fields [58], [59], is used. However, the IOU metric is good for refining the detection bounding boxes but is not flexible enough to measure overlapping ratios for merging the objects. Figure 3 explains that reason by a case study. The case study mainly assumes that the number of detection responses (observations) is larger than the number of real objects in occlusions. When the observations most likely include false positive detections, the object states paired those observations are also most likely the false positive states. So, to handle and consider the characteristics of those observations with the false positive errors, we propose a new metric named sum-of-intersection-over-each-area (SIOA). The IOU and SIOA metrics are formulated as follows:

$$IOU_{AB} = \frac{area(A) \cap area(B)}{area(A) \cup area(B)}, \tag{32}$$

$$SIOA_{AB} = (\frac{area(A) \cap area(B)}{area(A)} + \frac{area(A) \cap area(B)}{area(B)})/2, \tag{33}$$

Algorithm 2 Track Merging using the SIOA Metric

```
                                        ▷ $X_k$ : a set of states at time $k$
                                        ▷ $\sigma_m$ : threshold for merging
▷ $G_k$ : a set {key:id,value:states} of occlusion groups at time k
function Merge($X_k$,$\sigma_m$,$G_k$)
    $l = |X_k|$;                      // $l$ : the number of states $X_k$
    Let $M[1 \dots l][1 \dots l]$ be the array set to all $false$;
    /* measure occlusion ratio between all states by using the SIOA metric */
    for $i = 1$ to $l$ do
        for $j = i + 1$ to $l$ do
            $r_{occ} = SIOA_{X_k[i],X_k[j]}$;// SIOA occlusion ratio.
            if $r_{occ} > \sigma_m$ then
                $M[i][j] = true$;      // check to be merged
                $M[j][i] = true$;      // double check
            else if $r_{occ} \le \sigma_m$ and $r_{occ} > 0$ then
                $id_i = X_k[i].id$, $id_j = X_k[j].id$;
                if $id_i < id_j$ then
                    $G_k[id_i] = G_k[id_i] \cup \{X_k[i], X_k[j]\}$;
                else
                    $G_k[id_j] = G_k[id_j] \cup \{X_k[i], X_k[j]\}$;
                end if
            end if
        end for
    end for
    /* merge the states where SIOA value > $\sigma_m$ */
    for $i = 1$ to $l$ do
        for $j = i + 1$ to $l$ do
            if $M[i][j] = true$ then
                if $X_k[i].id < X_k[j].id$ then
                    $X_k[i] = 0.9 * X_k[i] + 0.1 * X_k[j]$;
                    Deactivate state $X_k[j]$;
                else
                    $X_k[j] = 0.9 * X_k[j] + 0.1 * X_k[i]$;
                    Deactivate state $X_k[i]$;
                end if
            end if
        end for
    end for
    return $X_k$;
end function
```

where A and B indicate two different objects. *area* represent a bounding box $(x, y, width, height)$. Algorithm 2 describes the proposed track merging method. Track merging with the SIOA metric follows after the D2T association as presented in Subsection III-B and Figure 2.

B. OCCLUSION GROUP ENERGY MINIMIZATION

The occlusion group energy minimization method is devised to prevent the true objects which are occluded to others from false merging. In other words, track merging may merge occluded objects with correct number of observations into the states with less number of real objects. That causes tracking errors such as false negatives and fragmented tracks.

Algorithm 3 Occlusion Group Energy Minimization

```
                                        ▷ $k$ : the current frame number
                         ▷ $G_{k-1}$ : a set of occlusion groups at time $k-1$
                                        ▷ $X_k$ : a set of states at time $k$
function OGEM($k$,$G_{k-1}$,$X_k$)
    $l = |G_{k-1}|$;     // $l$ : the number of the groups $G_{k-1}$.
    $n = |X_k|$;         // $n$ : the number of the states $X_k$.
    /*build the GMMs for all occlusion groups at time k-1*/
    /*a GMM is used for the defined energy function in (34)*/
    for $i = 1$ to $l$ do
        $p_i = {}^{|G_{k-1}[i]|}P_2$       //the number of topological vectors.
        $GMM[1 \dots p_i]$;      //the Gaussian mixture for a group.
        for $j = 1$ to $p_i$ do //iterate topologies in a group.
            Initialize a Gaussian mixture $GMM[j]$ with
            the mean vectors $m$ having topological info and
            the covariance matrix $R$ as defined in (34)
        end for
        $E_{min} = DBL\_MAX$;          //variable for the min-energy.
        $h_{min} = 0$;       //index to the optimal hypothesis.
        for $h = 1$ to $|H|$ do        //iterate topological hypotheses.
            if $E(h) < E_{min}$ then        //find the optimal hypothesis.
                $h_{min} = h$;
                $E_{min} = E(h)$;
            end if
        end for
        Update $G_{k-1}[i]$ with $h_{min}$;
        for $g$ in $G_{k-1}[i]$ do       //iterate group $G_{k-1}[i]$.
            Find the state $x$ with $g.id$ in $X_k$;
            if $x$ is in $X_k$ then
                $X_k[g.id] = g$;
            else
                $X_k = X_k \cup \{g\}$;
            end if
        end for
    end for
    return $X_k$;
end function
```

Thus, we propose a new energy minimization model to prevent false merging, named "*Occlusion Group Energy Minimization (OGEM)*". Each group of occluded objects has an energy function represented by a Gaussian mixture model (GMM) as follows:

$$E(h) = -\ln \sum \mathcal{N}(t|\mu, R), \tag{34}$$

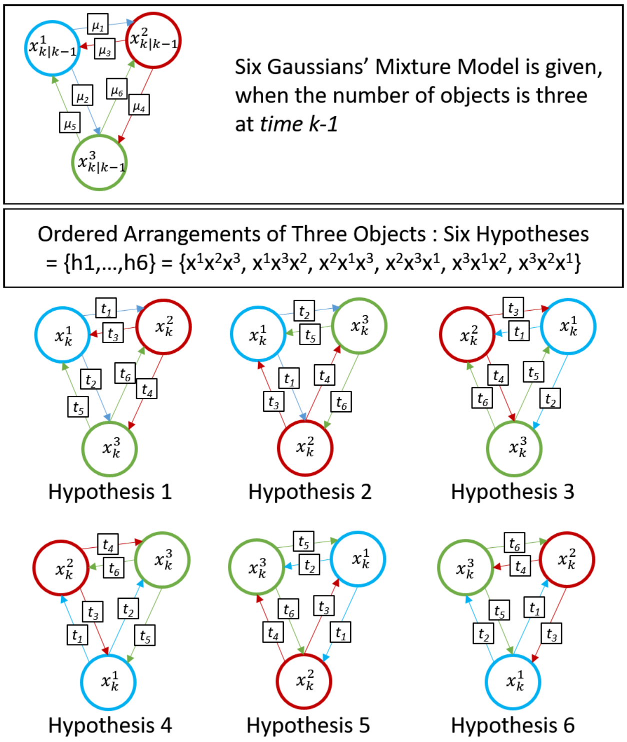

Six Gaussians' Mixture Model is given, when the number of objects is three at *time k-1*
Ordered Arrangements of Three Objects : Six Hypotheses = {h1,...,h6} = {x¹x²x³, x¹x³x², x²x¹x³, x²x³x¹, x³x¹x², x³x²x¹}
Hypothesis 1
Hypothesis 2
Hypothesis 3
Hypothesis 4
Hypothesis 5
Hypothesis 6

FIGURE 4. Illustration of the proposed occlusion group energy minimization represented by the Gaussian mixture model. Six hypotheses exist in the case of three occluded objects.

where h, t, μ, and R indicate hypothesis, topological vector, mean vector, and Gaussian covariance matrix (noise), respectively. A Gaussian probability function $\mathcal{N}$, i.e., component of the GMM, indicates a topological position vector between two objects in an occlusion group, which is given at time $k-1$. The Gaussian function has a mean vector m which denotes the topological position, i.e., relative position between the predicted center positions at time k of the objects in the group. Those objects are denoted by $x_{k|k-1}$ and the notation $k|k-1$ indicates the prediction at time k from position and velocity at time $k-1$. If there are three occluded objects in a group, six hypotheses exists as shown in Figure 4. One hypothesis is a set of six topological vectors $\{t_1, t_2, t_3, t_4, t_5, t_6\}$. For example, μ_1 is calculated by $x^2_{k|k-1} - x^1_{k|k-1}$ and t_1 is calculated by $x^2_k - x^1_k$. In the case that an object state x^d_k becomes "*lost*" (inactive) by falsely performing "track merging" in occlusion, we build a new hypothesis using a $x^d_{k|k-1}$as a dummy. Then the dummy added hypothesis recovers the false merged object. If there are n occluded objects in a group, $n(n-1)$ hypotheses exists with the condition $1 < n < 4$. Then with these topological models, we can find an optimal one among all hypotheses making the group cost minimal.

While track merging runs after D2TA, the proposed occlusion group energy minimization follows Track-to-Track Association (T2TA) as described in Figure 2. Figure 5 includes some examples to illustrate that the proposed group energy minimization complements track merging step. The tracking process from detection responses to tracking results at frame 42 shows it.

In summary, both "*Track Merging*" and "*Occlusion Group Energy Minimization*" assume occlusion situations, and the GMPHD filtering is adopted as the base algorithm. The pseudo-code of the proposed occlusion group management scheme are described in Algorithms 2 and 3. Also, both methods correspond to lines 8, 35 and 67-78 in Algorithm 1. From now on we will use GMPHD-OGM as the abbreviation for the proposed method, which is the online multi-object tracking algorithm with the GMPHD filter and occlusion group management.

V. EXPERIMENTS

In this section, we present the development environment used in this study including parameter settings, and also discuss evaluation results of the GMPHD-OGM tracker which include an ablation study with baselines and comparisons to state-of-the-art MOT methods. The GMPHD-OGM tracker is implemented in Visual C++ with OpenCV3.4.1 and boost1.61.0 libraries, and without any GPU-accelerated libraries such as CUDA. All experiments are conducted on Windows 10 with an Intel i7-7700K CPU @ 4.20GHz and DDR4 32.0GB RAM. The source code of the GMPHD-OGM tracker is available at https://github.com/SonginCV/GMPHD-OGM.

A. PARAMETER SETTING

Our proposed tracking method involves several parameter settings. The key parameters are summarized in Table 1. Parameter σ_m indicates the threshold for "Track Merging" which is set to 0.5 in terms of the SIOA metric. 0.5 is set not only empirically but also by considering the occlusion cases between the detected object's bounding boxes that are of different sizes as shown in Case 4 and 5 of Figure 3. τ_{T2T} and θ_{T2T} are related to track-to-track association (T2TA) of the hierarchical data association, whose parameters are selected adaptively, scene-by-scene. The optimal values of τ_{T2T} and θ_{T2T} are gained from the ablation study presented in Figure 8. We use the optimal parameter settings obtained from the study in both the training and test sequences.

The GMPHD filtering process has a set of static parameters. The matrices $F, Q, P, R,$ and H are used in *Prediction* and *Update*. Also, θ_w is used in *Pruning*. Based on experimentation, we set the parameters for the GMPHD filter's tracking process as follows:

$$F = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad Q = \frac{1}{2}\begin{pmatrix} 5^2 & 0 & 0 & 0 \\ 0 & 10^2 & 0 & 0 \\ 0 & 0 & 5^2 & 0 \\ 0 & 0 & 0 & 10^2 \end{pmatrix},$$

$$P = \begin{pmatrix} 5^2 & 0 & 0 & 0 \\ 0 & 10^2 & 0 & 0 \\ 0 & 0 & 5^2 & 0 \\ 0 & 0 & 0 & 10^2 \end{pmatrix}, \quad R = \begin{pmatrix} 5^2 & 0 \\ 0 & 10^2 \end{pmatrix},$$

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad \theta_w = 0.1,$$

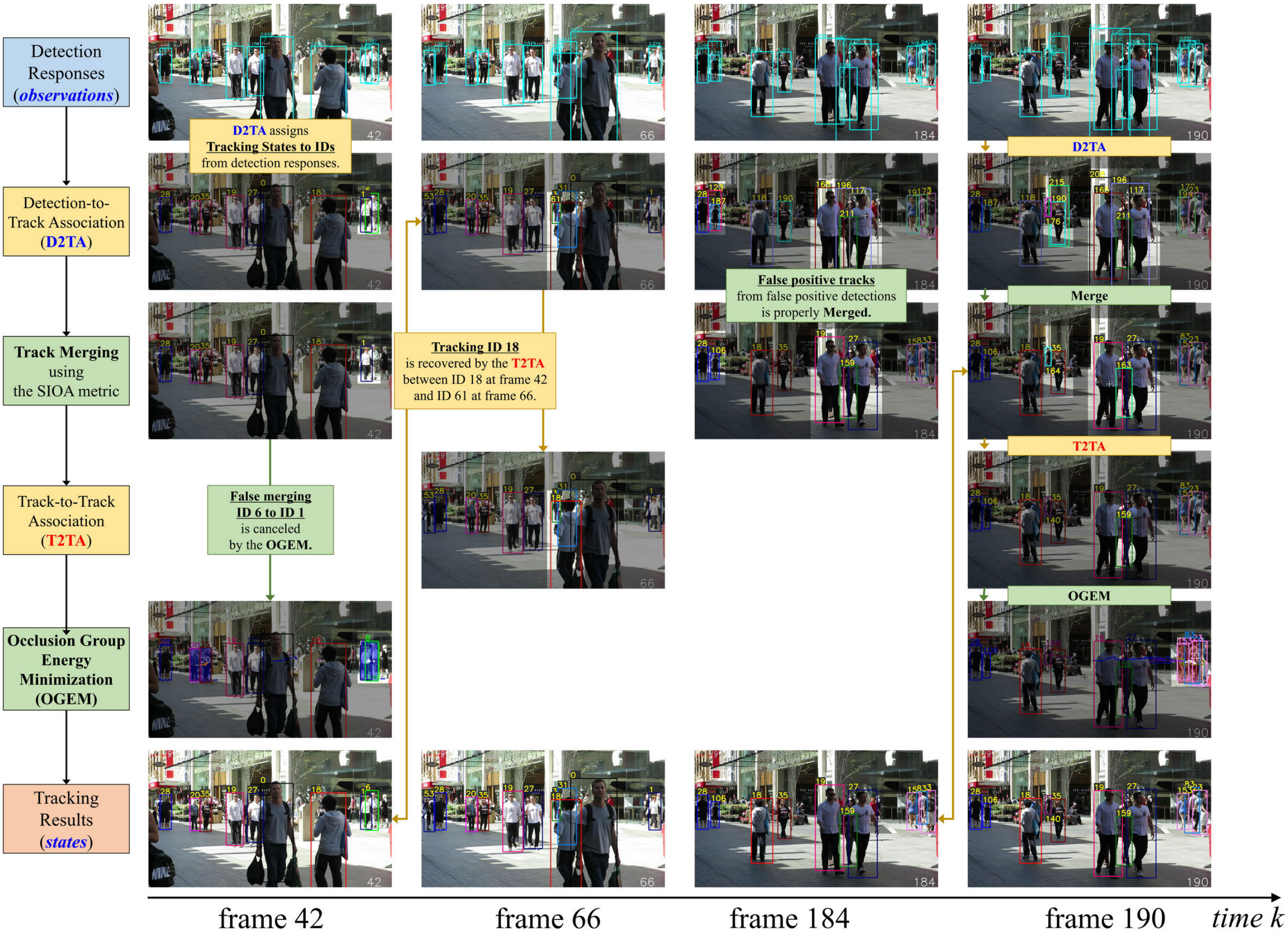

Detection Responses (observations)
Detection-to-Track Association (D2TA)
Track Merging using the SIOA metric
Track-to-Track Association (T2TA)
Occlusion Group Energy Minimization (OGEM)
Tracking Results (states)
D2TA assigns Tracking States to IDs from detection responses.
False merging ID 6 to ID 1 is canceled by the OGEM.
Tracking ID 18 is recovered by the T2TA between ID 18 at frame 42 and ID 61 at frame 66.
False positive tracks from false positive detections is properly Merged.
D2TA
Merge
T2TA
OGEM
frame 42
frame 66
frame 184
frame 190
time k

FIGURE 5. Illustration of the proposed multi-object tracking process with the qualitative results on MOT17-08-DPM test sequence. The whole process consists of four components which are D2TA, Merge, T2TA, and OGEM. Qualitative tracking results at frame 42, 66, 184, and 190 present that all components are complementary to each other for handling tracking problems.

TABLE 1. Parameter settings for "Track Merging" and track-to-track association (T2TA).

Symbol	Description	Value
σ_m	threshold for track merging.	0.5
τ_{T2T}	the minimum track length for T2TA	$1, 2, 3$
θ_{T2T}	the maximum frame interval for T2TA	5 to 100

B. EVALUATION RESULTS

In this section, we evaluate the proposed method against state-of-the-art online ([14]–[16], [18]–[26], [38]–[45]) and offline ([27]–[37], [46]–[52]) MOT methods in terms of the comprehensive MOT metrics [56], [57]. The metrics gracefully measure multi-object tracking performance from the detailed perspectives such as multi-object tracking accuracy (MOTA), multi-object tracking precision (MOTP), mostly tracked objects (MT), mostly lost objects (ML), total number of false positives (FP), total number of false negatives (FN, i.e., missed tracks), total number of identity switches (IDS), total number of times that a trajectory is fragmented (Frag), and processing speed (FPS, i.e., frames per second). Among these metrics, MOTA is normally proposed as a key metric, because it considers three error sources including FP, FN, and IDS, comprehensively. More details of the metrics are described in Table 3. The evaluation results contain not only the tracking results on the MOT15 and MOT17 test datasets but also an ablation study on the MOT15 training dataset.

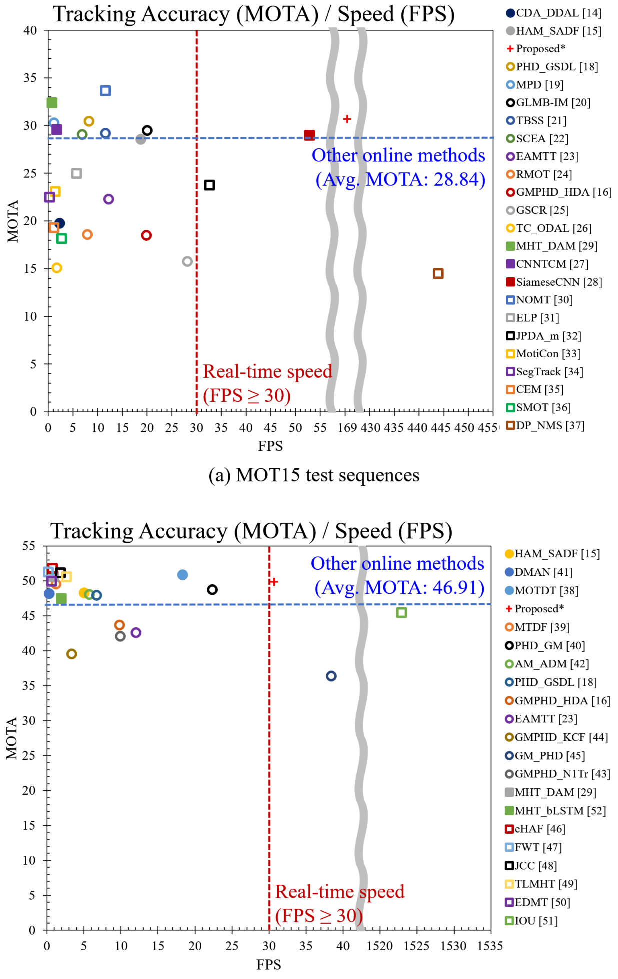
In the ablation study, we employ one baseline and two proposed methods, to find optimal parameters settings and to prove the effectiveness of the proposed method. The baseline method is the GMPHD filter based tracker with hierarchical data association (HDA) but without the occlusion group management (OGM). The first and second proposed methods are the GMPHD filter based tracker with the HDA and OGM by using the IOU metric and the SIOA metric for measuring occlusion ratio. We name these three methods as GMPHD-HDA, $p1$:GMPHD-OGM (w/ IOU), and $p2$:GMPHD-OGM (w/ SIOA). The evaluation results are shown in Table 4 and Figure 8. The scene-by-scene optimal parameter settings of those three methods are obtained from the results of the ablation study as shown in Figure 8. The same τ_{T2T} and θ_{T2T}

TABLE 2. Datasets specifications. The average object density (the average number of objects in a frame) of MOT17-test is three times that of MOT15-test.

Dataset	CAM	Sequence	FPS	Resolution	Density
MOT15-test	Static	ADL-Rundle-3	30	1920x1080	16.3
		AVG-TownCentre	2.5	1920x1080	15.9
		KITTI-16	10	1224x370	8.1
		PETS09-S2L2	7	768x576	22.1
		TUD-Crossing	25	640x480	5.5
		Venice-1	30	1920x1080	10.1
	Moving	ADL-Rundle-1	30	1920x1080	18.6
		ETH-Crossing	14	640x480	4.6
		ETH-Jelmoli	14	640x480	5.8
		ETH-Linthescher	14	640x480	7.5
		KITTI-19	10	1238x374	5.0
Average					10.6
MOT17-test	Static	MOT17-01	30	1920x1080	14.3
		MOT17-03	30	1920x1080	69.8
		MOT17-08	30	1920x1080	33.8
	Moving	MOT17-06	14	640x480	9.9
		MOT17-07	30	1920x1080	33.8
		MOT17-12	30	1920x1080	9.6
		MOT17-14	25	1920x1080	24.6
Average					31.8

settings are applied to the whole training sequences with the range {1, 2, 3} and {5, 10, 20, 30, 50, 70, 100}, respectively. $p1$ is an improvement over the GMPHD-HDA in terms of the upper bound of tracking accuracy (the maximum MOTA). $p2$, on the other hand, shows that the upper bound and lower bound of tracking accuracy increases more than $p1$. Besides, with θ_{T2T} over 20, the maximum and minimum values of MOTA increase. Figure 3 shows the comparison results of "Track Merging" between using the IOU metric and the SIOA metric when the detection results contain a lot of false positives. Also, through the case study on occlusion, we observe that the IOU metric cannot consider size-variant detections with false positives and too sensitive to be used for merging as shown in Figure 3. On the other hand, the SIOA metric can cope with the detections of varying sizes. Table 4 provides the quantitative results on the MOT15 training dataset with the merging threshold σ_m values. $p1$ and $p2$ show the improved MOTA, ML, FP, IDS, Speed compared to GMPHD-HDA. $p1$ and $p2$ achieved the best MOTA scores with the different merging thresholds σ_m which are 0.3 and 0.5, respectively. Comparing $p1$ and $p2$, $p1$ has more sensitive results depending on the σ_m value but also $p2$ shows the more overall improvements with MT, ML. The ablation study proves that $p2$ is not only the best in MOTA performance overall but it is also more robust and less sensitive in parameters than the baseline method and $p1$. Thus, we selected $p2$:GMPHD-OGM (w/ SIOA) as our final tracking model.

Figure 5 shows some qualitative results of our tracking method in stages of the overall process. Detection results (observations) initialize tracking objects (states). Subsequently, the states are associated with the proper observations by the detection-to-track association (D2TA) using the GMPHD filtering process, frame-by-frame. Then "Track Merging" merges false positive tracks having the SIOA value $\geq$ the merging threshold $\sigma_m = 0.5$, which are most likely generated from false positive detections. If objects are occluded ($0 < SIOA < \sigma_m$), their IDs can be switched or changed. The track-to-track association (T2TA) can recover their IDs. The occlusion group energy minimization (OGEM) process can recover false merging, i.e., the case that "Track Merging" merges true tracks, by calculating the probability hypothesis of the Gaussian mixture model and optimizing the energy of the groups of occluded objects, as described in Subsection IV-B.



(a) MOT15 test sequences

(b) MOT17 test sequences

FIGURE 6. Comparisons of tracking accuracy against speed with the state-of-the-art methods on the (a) MOT15 and (b) MOT17 test sequences. We assumed that at least 30 FPS is required for real-time speed.

Table 5 and 6 show the quantitative results on MOT15 and MOT17 test dataset, respectively. Those benchmark datasets have two crucial differences. First, MOT15 provides the detection results of ACF [8] and MOT17 provides three types of detections: DPM [9], FRCNN [10], and SDP [11]. Since ACF and DPM exploit hand-crafted features learning and models, they show relatively poor performance compared to the DNN based detectors FRCNN and SDP. In other words,

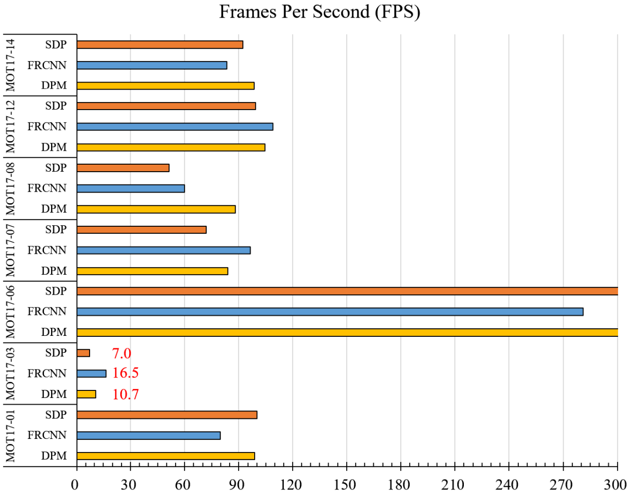


FIGURE 7. Speed comparison of the proposed tracking method on MOT17 test dataset which provides three types of detection results for each scene, including DPM [9], FRCNN [10], and SDP [11].

TABLE 3. Evaluation metrics. MOTA has been mainly used for measuring tracking performance as a key metric.

Measure	Better	Perfect	Description
MOTA	↑	100%	Multiple Object Tracking Accuracy [56]. This measure combines three error sources: false positives, missed targets and identity switches.
MOTP	↑	100%	Multiple Object Tracking Precision [56]. The misalignment between the annotated and the predicted bounding boxes.
MT	↑	100%	Mostly tracked targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at least 80% of their respective life span.
ML	↓	0 %	Mostly lost targets. The ratio of ground-truth trajectories that are covered by a track hypothesis for at most 20% of their respective life span.
FP	↓	0	Total number of false positives.
FN	↓	0	Total number of false negatives (missed targets).
IDS	↓	0	Total number of identity switches. Please note that we follow the stricter definition of identity switches as described in [57].
Frag	↓	0	Total number of times a trajectory is fragmented (i.e. interrupted during tracking).
FPS	↑	∞	Processing speed (in frames per second excluding the detector) on the benchmark.

DPM generates more false positives than FRCNN and SDP do, and especially ACF misses much more objects than others do. Thus, in MOT15, state-of-the-art trackers shows wider range of MOTA distribution than that in MOT17. Among online methods, our final proposed method $p2$ with $\sigma_m = 0.5$ achieves the second best MOTA 30.7 vs. the best speed 169.5 fps in MOT15 and the second best MOTA 49.9 vs. the second best speed 30.7 fps in MOT17. The trackers with DNN [14], [38] shows the top MOTA scores in MOT15 and MOT17 but we think that our proposed method achieves competitive performance and efficiency to consider real-time application. In Figure 6(a), the proposed method is located in a spot well to the right of the graph, which indicates the effectiveness of our occlusion group based object analysis (OGM), compared to other relation analysis between all objects in the scene [22], [24]. However, in MOT17, the speed of our proposed method decreases to 30.7 fps. This speed still belongs to real-time processing time but is not outstanding compared to other online methods. The decrease in speed is caused by the second different point between MOT15 and MOT17 as described in Table 2. MOT15 includes 5,783 frames with 721 tracks, 61,440 bounding boxes, and an object density of 10.6 i.e., the average number of objects in a frame, whereas MOT17 includes 17,757 frames with 2355 tracks, 564,228 bounding boxes, and an object density of 31.8. Due to the fact that MOT17 has scenes with much higher density but also accurate detection results so tracking accuracy and processing time increase. Figure 6(b) shows the increase in number of detected points also the performances of state-of-the-art methods are saturated on the spot with MOTA around 50 and speed under 5 fps. Figure 7 explains the drastic decrease of speed. In MOT17-03, the speed is around 10 fps since many objects appear in the scene with an object density of 69.8. That makes the number of track-to-track association (T2TA) greatly increase. However, the proposed method is still comparable to state-of-the-art methods and is positioned in a spot for real-time application as shown in Figure 6(b). Besides our tracking algorithm (GMPHD-OGM), there are many PHD filter based online approaches [16],

TABLE 4. Quantitative evaluation results on MOT15 training dataset. The proposed methods namely $p1$: GMPHD-OGM (w/ IOU) and $p2$: GMPHD-OGM (w/ SIOA) are compared to one baseline method GMPHD-HDA. GMPHD-HDA employs the GMPHD filtering with hierarchical data association (HDA). GMPHD-OGM is equal to GMPHD-HDA with the proposed occlusion group management (OGM). The IOU and SIOA metrics are used for "Track Merging" in $p1$ and $p2$, respectively. The optimal values of the merging threshold σ_m are underlined and the best scores are in bold in terms of the MOTA metric.

Tracker	σ_m	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Frag↓	Speed↑
GMPHD-HDA	n/a	34.8 %	72.3 %	14.4 %	47.8 %	4,042	21,646	338	572	212.4 fps
$p1$: GMPHD-OGM (w/ IOU)	0.2	34.5 %	**72.4 %**	13.4 %	49.8 %	3,594	22,226	285	550	201.1 fps
	0.3	35.6 %	72.3 %	14.0 %	48.6 %	3,537	21,901	**278**	548	**228.4 fps**
	0.4	35.3 %	**72.4 %**	14.0 %	48.2 %	3,667	21,865	291	562	201.9 fps
	0.5	34.7 %	72.3 %	14.4 %	47.8 %	4,044	21,661	340	577	205.3 fps
$p2$: GMPHD-OGM (w/ SIOA)	0.3	34.5 %	72.3 %	14.2 %	49.2 %	**3,496**	22,336	297	**559**	216.3 fps
	0.4	35.4 %	**72.4 %**	14.6 %	48.4 %	3,556	21,930	284	**540**	216.3 fps
	0.5	**35.8 %**	72.2 %	15.0 %	47.6 %	3,569	21,758	295	545	221.0 fps
	0.6	35.5 %	72.3 %	**15.6 %**	**47.2 %**	3,702	21,724	312	556	221.9 fps
	0.7	34.7 %	72.2 %	**15.6 %**	**47.2 %**	4,159	**21,519**	368	567	202.9 fps

TABLE 5. Quantitative evaluation results on MOT15 test dataset. The proposed method is compared to state-of-the-art in terms of the MOT metrics. For each mode, i.e, online and offline, the first and the second best scores are highlighted in red and blue in terms of each metric.

Mode	Tracker	DNN	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Frag↓	Speed↑
Online	CDA_DDAL [14]	O	**32.8 %**	70.7 %	9.7 %	42.2 %	4,983	35,690	614	1,583	2.3 fps
	HAM_SADF [15]	O	28.6 %	71.1 %	10.0 %	44.0 %	7,485	35,910	460	1,038	18.7 fps
	*p*2: GMPHD-OGM (w/ SIOA)	X	30.7 %	**71.6 %**	11.5 %	**38.1 %**	6,502	35,030	1,034	1,351	**169.5 fps**
	PHD_GSDL [18]	X	30.5 %	71.2 %	7.6 %	41.2 %	6,534	35,284	879	2,208	8.2 fps
	MDP [19]	X	30.3 %	71.3 %	**14.0 %**	38.4 %	9,717	32,422	680	1,500	1.1 fps
	GLMB-IM [20]	X	29.5 %	71.0 %	13.2 %	40.1 %	10,556	**32,020**	670	1,260	20.0 fps
	TBSS [21]	X	29.2 %	71.3 %	6.8 %	43.8 %	6,068	36,779	649	1,508	11.5 fps
	SCEA [22]	X	29.1 %	71.1 %	8.9 %	47.3 %	**6,060**	36,912	604	1,182	6.8 fps
	EAMTT [23]	X	22.3 %	69.6 %	5.4 %	52.7 %	7,924	38,982	833	1,485	12.2 fps
	RMOT [24]	X	18.6 %	69.6 %	5.3 %	53.3 %	12,473	36,835	684	1,282	7.9 fps
	GMPHD_HDA [16]	X	18.5 %	70.9 %	3.9 %	55.3 %	7,864	41,766	**459**	1,266	19.8 fps
	GSCR [25]	X	15.8 %	69.4 %	1.8 %	61.0 %	7,597	43,633	514	**1,010**	28.1 fps
	TC_ODAL [26]	X	15.1 %	70.5 %	3.2 %	55.8 %	12,970	38,538	637	1,716	1.7 fps
Offline	MHT_DAM [29]	O	32.4 %	71.8 %	**16.0 %**	43.8 %	9,064	**32,060**	435	826	0.7 fps
	CNNTCM [27]	O	29.6 %	71.8 %	11.2 %	44.0 %	7,786	34,733	712	943	1.7 fps
	SiameseCNN [28]	O	29.0 %	71.2 %	8.5 %	48.4 %	**5,160**	37,798	639	1,316	52.8 fps
	NOMT [30]	X	**33.7 %**	**71.9 %**	12.2 %	44.6 %	7,762	32,547	442	823	11.5 fps
	ELP [31]	X	25.0 %	71.2 %	7.5 %	43.8 %	7,345	37,344	1,396	1,804	5.7 fps
	JPDA_m [32]	X	23.8 %	68.2 %	5.0 %	58.1 %	6,373	70,084	**365**	869	32.6 fps
	MotiCon [33]	X	23.1 %	70.9 %	4.7 %	52.0 %	10,404	35,844	1,018	1,061	1.4 fps
	SegTrack [34]	X	22.5 %	71.7 %	5.8 %	63.9 %	7,890	39,020	697	**737**	0.2 fps
	CEM [35]	X	19.3 %	70.7 %	8.5 %	46.5 %	14,180	34,591	813	1,023	1.1 fps
	SMOT [36]	X	18.2 %	71.2 %	2.8 %	54.8 %	8,780	40,310	1,148	2,132	2.7 fps
	DP_NMS [37]	X	14.5 %	70.8 %	5.0 %	**40.8 %**	13,171	34,814	4,537	3,090	**444.8 fps**

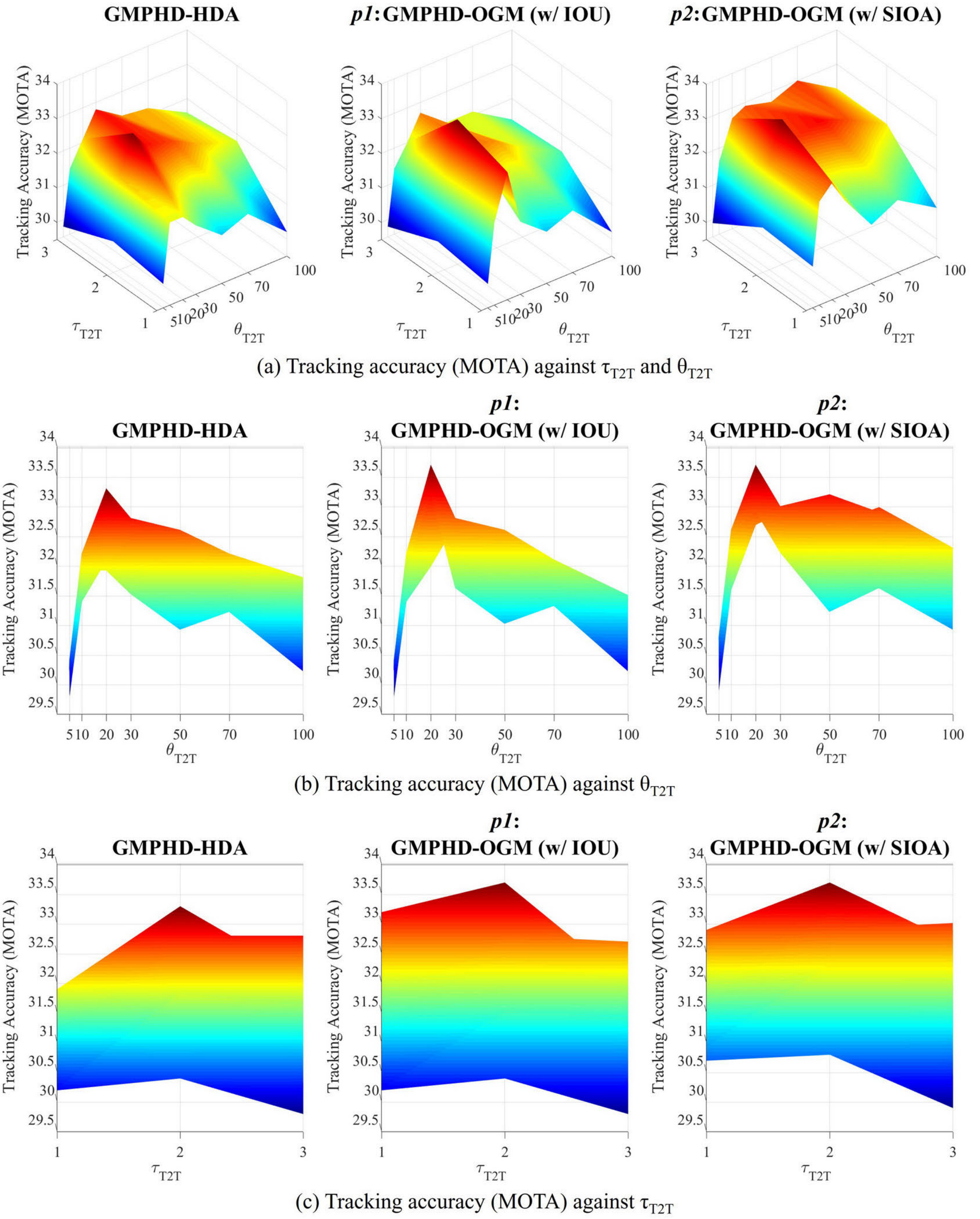
TABLE 6. Quantitative evaluation results on MOT17 test dataset. The proposed method is compared to state-of-the-art in terms of the MOT metrics. For each mode, i.e, online and offline, the first and the second best scores are highlighted in red and blue in terms of each metric. Only our method and [51] does not utilized any complex visual features except bounding boxes.

Mode	Tracker	DNN	MOTA↑	MOTP↑	MT↑	ML↓	FP↓	FN↓	IDS↓	Frag↓	Speed↑
Online	MOTDT [38]	O	**50.9 %**	76.6 %	17.5 %	35.7 %	24,069	250,768	2,474	5,317	18.3 fps
	HAM_SADF [15]	O	48.3 %	77.2 %	17.1 %	41.7 %	20,967	269,038	**1,871**	**3,020**	5.0 fps
	DMAN [41]	O	48.2 %	75.7 %	19.3 %	38.3 %	26,218	263,608	2,194	5,378	0.3 fps
	*p*2: GMPHD-OGM (w/ SIOA)	X	49.9 %	77.0 %	**19.7 %**	38.0 %	24,024	255,277	3,125	3,540	30.7 fps
	MTDF [39]	X	49.6 %	75.5 %	18.9 %	**33.1 %**	37,124	**241,768**	5,567	9,260	1.2 fps
	PHD_GM [40]	X	48.8 %	76.7 %	19.1 %	35.2 %	26,260	257,971	4,407	6,448	22.3 fps
	AM_ADM [42]	X	48.1 %	76.7 %	13.4 %	37.7 %	25,061	265,495	2,214	5,027	5.7 fps
	PHD_GSDL [18]	X	48.0 %	77.2 %	17.1 %	35.6 %	23,199	265,954	3,998	8,886	6.7 fps
	GMPHD_HDA [16]	X	43.7 %	76.5 %	11.7 %	43.0 %	25,935	287,758	3,838	5,056	9.2 fps
	EAMTT [23]	X	42.6 %	76.0 %	12.7 %	42.7 %	20,711	288,474	4,488	5,720	12.0 fps
	GMPHD_N1Tr [43]	X	42.1 %	**77.7 %**	11.9 %	42.7 %	**18,214**	297,646	10,698	10,864	9.9 fps
	GMPHD_KCF [44]	X	39.6 %	74.5 %	8.8 %	43.3 %	50,903	284,228	5,811	7,414	3.3 fps
	GM_PHD [45]	X	36.4 %	76.2 %	4.1 %	57.3 %	23,723	330,767	4,607	11,317	**38.4 fps**
Offline	MHT_DAM [29]	O	50.7 %	77.5 %	20.8 %	36.9 %	22,875	252,889	2,314	2,865	0.9 fps
	MHT_bLSTM [52]	O	47.5 %	77.5 %	18.2 %	41.7 %	25,981	268,042	2,069	3,124	1.9 fps
	eHAF [46]	X	**51.8 %**	77.0 %	**23.4 %**	37.9 %	33,212	**236,772**	1,834	2,739	0.7 fps
	FWT [47]	X	51.3 %	77.0 %	21.4 %	**35.2 %**	24,101	247,921	2,648	4,279	0.2 fps
	JCC [48]	X	51.2 %	75.9 %	20.9 %	37.0 %	25,937	247,822	1,802	2,984	1.8 fps
	TLMHT [49]	X	50.6 %	**77.6 %**	17.6 %	43.4 %	22,213	255,030	**1,407**	**2,079**	2.6 fps
	EDMT [50]	X	50.0 %	77.3 %	21.6 %	36.3 %	32,279	247,297	2,264	3,260	0.6 fps
	IOU [51]	X	45.5 %	76.9 %	15.7 %	40.5 %	**19,993**	281,643	5,988	7,404	**1,522.9 fps**

[18], [23], [39], [40], [43]–[45] that have been proposed in the past decade. Tables 5 and 6 include them in the comparison against them, GMPHD-OGM achieves not only the best MOTA, MOTP, MT, ML, FN, and speed scores in MOT15 but also the best MOTA, MT, and Frag scores (and second best in MOTA and speed) in MOT17. Our proposed method is distinguished in terms of tracking accuracy (MOTA) vs. speed (FPS), even though we did not utilize any complex visual features except bounding boxes as described in Figure 6.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed an efficient online multi-object tracking method with a Gaussian mixture probability hypothesis density (GMPHD) filter and an occlusion group management (OGM) named GMPHD-OGM. In our proposed method, the GMPHD filter is exploited for the implementation of an online and real-time MOT method. Since the GMPHD filter is originally designed to handle MOT in radar/sonar systems, we revised the filter to fit to video data system. To resolve missed tracks problems of MOT in



(a) Tracking accuracy (MOTA) against τ_{T2T} and θ_{T2T}

(b) Tracking accuracy (MOTA) against θ_{T2T}

(c) Tracking accuracy (MOTA) against τ_{T2T}

FIGURE 8. Ablation study with one baseline method GMPHD-HDA and two proposed methods *p*1:GMPHD-OGM (with IOU) and *p*2:GMPHD-OGM (with SIOA). The final proposed method is *p*2. Three graphs indicates the MOTA scores' distributions against (a) the minimum track length for T2TA (τ_{T2T}) and the maximum frame interval for T2TA (θ_{T2T}), (b) τ_{T2T}, and (c) θ_{T2T}. *p*2 shows overall improvements in upper and lower bound of MOTA score.

video, we extended the conventional GMPHD filtering process with a hierarchical data association (HDA) strategy. Next, to solve the occlusion problems, we proposed an occlusion group management (OGM) scheme that is composed of "Track Merging" and "Occlusion Group Energy Minimization (OGEM)". "Track Merging" reduced the

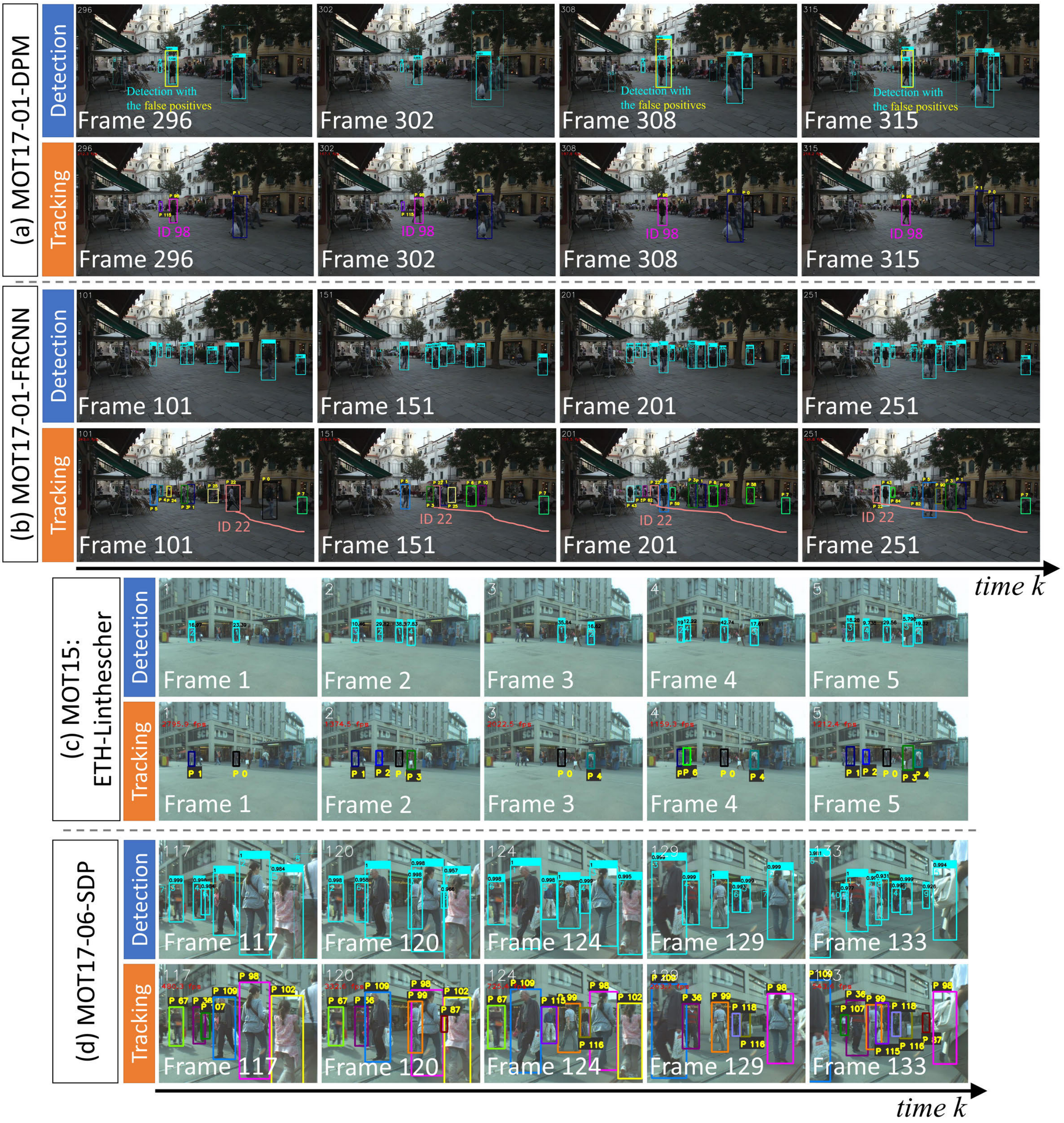


FIGURE 9. Examples of the qualitative tracking results on the four test sequences: (a) MOT17-01-DPM, (b) MOT17-01-FRCNN, (c) MOT15: ETH-Linthescher, and (d) MOT17-06-SDP. In (a), our tracker estimated a true track (ID 98) from the false positive detections by using our Merging technique. In (b), our hierarchical data association (HDA) successfully preserved ID 22 from the false negative detections resulting in occlusions while the proposed occlusion group energy minimization (OGEM) prevented the false Merging in occlusions. (c) and (d) share the same image sequence but have two different detections which are ACF [8] and SDP [11], respectively. In (c), despite false negative detections, HDA recovered ID 1 and ID 3 when the objects were detected again. On the other hand, in (d) SDP detected most of pedestrians even in crowd situation (occlusions), OGEM prevented false Merging and HDA recovered the reappearing objects ID 26, ID 99, and ID 107 after the occlusions.

number of false positives by merging them, while the OGEM prevented false ‘‘Track Merging’’ between true tracks. Also instead of using the IOU metric, we designed a new metric named sum-of-intersection-over-each-area (SIOA) to measure the occlusion ratio between visual objects. The effectiveness of our tracker (GMPHD-OGM) was verified by the ablation study and also by the evaluation results on the MOT15 [5] and MOT17 [6] benchmarks. The ablation study shows that the GMPHD-OGM (w/ SIOA) is more efficient at solving the defined problems than other methods such as GMPHD-HDA and GMPHD-OGM (w/ IOU). The GMPHD-OGM (w/ SIOA) achieves the best MOTA scores in the MOT15 and MOT17 datasets, respectively, in comparison with other PHD filter based online trackers [16], [18], [23], [39], [40], [43]–[45] but also our proposed method shows the competitive value in ''MOTA versus speed'' against state-of-the-art online MOT methods. As a future work, we will extend the proposed tracking model in terms of scalability and modularity considering not only when the number of objects is over a hundred but also various environments such as aerial views from drones, underwater scenes, and weather conditions (rain, fog, and snow), simultaneously achieving state-of-the-art level tracking accuracy and real-time speed. We expect that our tracker will be extended for a universal online and real-time MOT framework.

APPENDIX. QUALITATIVE RESULTS IN THE TEST DATA

In appendix, we present a tracking diagram based on the test sequences of MOT15 and MOT17 datasets as shown in Figure 9 which includes the qualitative results based on the four representative test sequences to show the effectiveness of our proposed tracker.

REFERENCES

[1] R. P. S. Mahler, ''Multitarget Bayes filtering via first-order multitarget moments,'' *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[2] B.-N. Vo, S. Singh, and A. Doucet, ''Sequential Monte Carlo implementation of the PHD filter for multi-target tracking,'' in *Proc. 6th Int. Conf. Inf. Fusion (FUSION)*, Jul. 2003, pp. 792–799.

[3] B. N. Vo and W. K. Ma, ''The Gaussian mixture probability hypothesis density filter,'' *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, Nov. 2006.

[4] B.-T. Vo, ''Random finite sets in multi-object filtering,'' Ph.D. dissertation, School of Elect., Electron. Comput. Eng., Univ. Western Australia, Perth, WA, Australia, 2008.

[5] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, ''MOTChallenge 2015: Towards a benchmark for multi-target tracking,'' Apr. 2015, *arXiv:1504.01942*. [Online]. Available: https://arxiv.org/abs/1504.01942

[6] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, ''MOT16: A benchmark for multi-object tracking,'' May 2016, *arXiv:1603.00831*. [Online]. Available: https://arxiv.org/abs/1603.00831

[7] A. Geiger, P. Lenz, and R. Urtasun, ''Are we ready for autonomous driving? The KITTI vision benchmark suite,'' in *Proc. Int. Conf. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3354–3361.

[8] P. Dollár, R. Appel, S. Belongie, and P. Perona, ''Fast feature pyramids for object detection,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

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

[10] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Montréal, QC, Canada, vol. 1, Dec. 2015, pp. 91–99.

[11] F. Yang, W. Choi, and Y. Lin, ''Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2129–2137.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[13] S. Murray, ''Real-time multiple object tracking—A study on the importance of speed,'' M.S. thesis, School Comput. Sci. Commun., Nat. Inst. Inform., Tokyo, Japan, 2017.

[14] S.-H. Bae and K.-J. Yoon, ''Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 595–610, Mar. 2018.

[15] Y.-C. Yoon, A. Boragule, Y.-M. Song, K. Yoon, and M. Jeon, ''Online multi-object tracking with historical appearance matching and scene adaptive detection filtering,'' in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[16] Y. Song and M. Jeon, ''Online multiple object tracking with the hierarchically adopted GM-PHD filter using motion and appearance,'' in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Oct. 2016, pp. 1–4.

[17] Y.-M. Song, Y.-C. Yoon, K. Yoon, and M. Jeon, ''Online and real-time tracking with the GM-PHD filter using group management and relative motion analysis,'' in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[18] Z. Fu, P. Feng, F. Angelini, J. Chambers, and S. M. Naqvi, ''Particle PHD filter based multiple human tracking using online group-structured dictionary learning,'' *IEEE Access*, vol. 6, pp. 14764–14778, 2018.

[19] Y. Xiang, A. Alahi, and S. Savarese, ''Learning to track: Online multi-object tracking by decision making,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4705–4713.

[20] D. Y. Kim, B.-N. Vo, B.-T. Vo, and M. Jeon, ''A labeled random finite set online multi-object tracker for video data,'' *Pattern Recognit.*, vol. 90, pp. 377–389, Jun. 2019.

[21] X. Zhou, P. Jiang, Z. Wei, H. Dong, and F. Wang, ''Online multi-object tracking with structural invariance constraint,'' in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2018, pp. 1–13.

[22] J. Hong Yoon, C.-R. Lee, M.-H. Yang, and K.-J. Yoon, ''Online multi-object tracking via structural constraint event aggregation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1392–1400.

[23] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, ''Online multi-target tracking with strong and weak detections,'' in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Oct. 2016, pp. 84–99.

[24] J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon, ''Bayesian multi-object tracking using motion context from multiple objects,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2015, pp. 33–40.

[25] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle, ''Online multi-person tracking based on global sparse collaborative representations,'' in *Proc. IEEE Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2414–2418.

[26] S.-H. Bae and K.-J. Yoon, ''Robust online multi-Object tracking based on tracklet confidence and online discriminative appearance learning,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1218–1225.

[27] B. Wang, L. Wang, B. Shuai, Z. Zuo, T. Liu, K. L. Chan, and G. Wang, ''Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 386–393.

[28] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler, ''Learning by tracking: Siamese CNN for robust target association,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 33–40.

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

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

[31] N. McLaughlin, J. M. Del Rincon, and P. Miller, ''Enhancing linear programming with motion modeling for multi-target tracking,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2015, pp. 71–77.

[32] S. H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid, ''Joint probabilistic data association revisited,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3047–3055.

[33] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, ''Learning an image-based motion context for multiple people tracking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3542–3549.

[34] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid, ''Joint tracking and segmentation of multiple targets,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5397–5406.

[35] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 58–72, Jan. 2014.

[36] C. Dicle, O. I. Camps, and M. Sznaier, "The way they move: Tracking multiple targets with similar appearance," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2304–2311.

[37] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2011, pp. 1201–1208.

[38] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[39] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2277–2291, Sep. 2019.

[40] R. Sanchez-Matilla and A. Cavallaro, "A predictor of moving objects for first-person vision," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2189–2193.

[41] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Feb. 2019, pp. 366–382.

[42] S.-H. Lee, M.-Y. Kim, and S.-H. Bae, "Learning discriminative appearance models for online multi-object tracking with appearance discriminability measures," *IEEE Access*, vol. 6, pp. 67316–67328, 2018.

[43] N. L. Baisa and A. Wallace, "Development of a n-type GM-PHD filter for multiple target, multiple type visual tracking," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 257–271, Feb. 2019.

[44] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora, "Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug./Sep. 2017, pp. 1–5.

[45] V. Eiselein, D. Arp, M. Pätzold, and T. Sikora, "Real-time multi-human tracking using a probability hypothesis density filter and multiple detectors," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2012, pp. 325–330.

[46] H. Sheng, Y. Zhang, J. Chen, Z. Xiong, and J. Zhang, "Heterogeneous association graph fusion for target association in multiple object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3269–3280, Nov. 2019.

[47] R. Henschel, L. Leal-Taixé, D. Cremers, and B. Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1428–1437.

[48] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation & multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[49] H. Sheng, J. Chen, Y. Zhang, W. Ke, Z. Xiong, and J. Yu, "Iterative multiple hypothesis tracking with tracklet-level association," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[50] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, "Enhancing detection model for multiple hypothesis tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 18–27.

[51] E. Bochinski, V. Eiselein, and T. Sikora, "High-speed tracking-by-detection without using image information," in *Proc. IEEE Int. Workshop Traffic Street Surveill. Saf. Secur. (AVSS)*, Sep. 2017, pp. 1–6.

[52] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear LSTM," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 200–215.

[53] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 539–546.

[54] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3239–3248.

[55] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, Mar. 2015.

[56] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *J. Image Video Process.*, vol. 2008, pp. 1–10, Feb. 2008.

[57] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: HybridBoosted multi-target tracker for crowded scene," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 2953–2960.

[58] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[59] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[60] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, Nov. 1987.

[61] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, D, J. Basic Eng.*, vol. 82, pp. 35–45, 1960.

[62] B. T. Vo, B. N. Vo, and A. Cantoni, "Analytic implementations of the cardinalized probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3553–3567, Jul. 2007.

[63] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3460–3475, Jul. 2013.

[64] G. Brooker, "Detection of targets in noise," in *Sensors and Signals*. Sydney, NSW, Australia: Univ. of Sydney, 2006, ch. 10, pp. 283–302.

[65] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 34–55.

[66] R. H. Evangelio, T. Senst, and T. Sikora, "Detection of static objects for the task of video surveillance," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2011, pp. 534–540.

[67] M. Pätzold, R. H. Evangelio, and T. Sikora, "Counting people in crowded environments by fusion of shape and motion information," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal Based Surveill. (PETS Workshop)*, Aug./Sep. 2010, pp. 157–164.

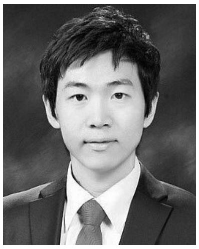
[68] Y. Xia and J. Sattar, "Visual diver recognition for underwater human-robot collaboration," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2019, pp. 6839–6845.

[69] J. Sattar and G. Dudek, "Where is your dive buddy: Tracking humans underwater using spatio-temporal features," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (ICRA)*, Oct./Nov. 2007, pp. 3654–3659.

[70] N. Dong, Z. Jia, J. Shao, Z. Li, F. Liu, J. Zhao, and P.-Y. Peng, "Adaptive object detection and visibility improvement in foggy image," *J. Multimedia*, vol. 6, no. 1, pp. 14–21, Feb. 2011.

[71] C. Sakaridis, D. Dai, S. Hecker, and L. Van Gool, "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 687–704.

YOUNG-MIN SONG received the B.S. degree in computer science and engineering from Chungnam National University, Daejeon, South Korea, in 2013, and the M.S. degree information and communications from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2015, where he is currently pursuing the Ph.D. degree in electrical engineering and computer science. His research interests include multiobject tracking and data fusion.



KWANGJIN YOON received the B.S. degree in computer science from Sahmyook University, Seoul, South Korea, in 2009, the M.S. degree in image engineering from Chung-Ang University, Seoul, in 2011, and the Ph.D. degree in electrical engineering and computer science from the Gwangju Institute of Science and Technology, in 2019. His research interests include multiobject tracking, computer vision, and deep learning.

YOUNG-CHUL YOON received the B.S. degree in electronics and communications engineering from Kwangwoon University, Seoul, South Korea, and the M.S. degree in electrical engineering and computer science from the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2019. He is currently working as a Computer Vision Researcher with LG Electronics. His research interests are multiobject tracking and deep learning.

KIN CHOONG YOW received the B.Eng. degree in electrical engineering from the National University of Singapore, in 1993, and the Ph.D. degree in information engineering from the Cambridge University, U.K., in 1998. From 1998 to 2012, he was an Assistant and then Associate Professor from the School of Computer Engineering, Nanyang Technological University, Singapore. In 2012, he was a Professor with the Cloud Computing Centre, Shenzhen Institute of Advanced Technology, China. From 2013 to 2018, he was an Associate Professor with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, South Korea. In 2018, he joined the University of Regina, Regina, Saskatchewan, Canada, where he is an Associate Professor with the Faculty of Engineering and Applied Science. His current research interests include computer vision, artificial intelligence, and machine learning.

MOONGU JEON received the B.S. degree in architectural engineering from Korea University, Seoul, South Korea, in 1988, and the M.S. and Ph.D. degrees in computer science and scientific computation from the University of Minnesota, Minneapolis, MN, USA, in 1999 and 2001, respectively. As a Postgraduate Researcher, he worked on optimal control problems with the University of California at Santa Barbara, Santa Barbara, CA, USA, from 2001 to 2003, and then moved to the National Research Council of Canada, where he worked on the sparse representation of high-dimensional data and the image processing, until 2005. In 2005, he joined the Gwangju Institute of Science and Technology, Gwangju, South Korea, where he is currently a Full Professor with the School of Electrical Engineering and Computer Science. His current research interests are in machine learning, computer vision, and artificial intelligence.