IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 28, 2018, accepted October 16, 2018, date of publication November 9, 2018, date of current version December 3, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2879535*

# Learning Discriminative Appearance Models for Online Multi-Object Tracking With Appearance Discriminability Measures

**SEONG-HO LEE, MYUNG-YUN KIM, AND SEUNG-HWAN BAE, (Member, IEEE)**
Department of Computer Science and Engineering, Incheon National University, Incheon 22012, South Korea

Corresponding author: Seung-Hwan Bae (shbae@inu.ac.kr)

This work was supported in part by Incheon National University (International Cooperative) Research Grant in 2017 and in part by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant NRF-2018R1C1B6003785.

**ABSTRACT** A data association, which links detections in consecutive frames, is a key issue in multi-object tracking (MOT). For robust data association in a complex scene, a common approach is to learn object appearance models for handling appearance variations of tracked objects and improving the discriminability between objects. However, learning appearances of multiple objects during tracking is still a challenging problem due to the frequent sample contamination by occlusions and low feature discriminability by similar appearances between objects. In this paper, in order to learn each object appearance, we propose a discriminative online appearance learning using a partial least square (PLS) method. In the proposed appearance learning, we first present online sampling mining to collect training samples from tracking results. Then, we consecutively learn PLS-based subspaces during tracking and discriminate object appearances by projecting object features onto the learned spaces. Since frequent appearance updates for all tracked objects increase the tracking complexity significantly, we propose measures to evaluate the discriminability of learned object appearances and update only the appearances with low discriminability. We apply the proposed appearance learning for online MOT and compare other appearance learning methods. In addition, we evaluate the performance of our MOT method on public MOT benchmark challenge datasets and show the competitive performance compared to other state-of-the-art batch and online tracking methods.

**INDEX TERMS** Multi-object tracking, appearance discriminability measures, online appearance learning, partial least square analysis, data association, surveillance system.

## I. INTRODUCTION

MULTI-OBJECT Tracking (MOT) has been an important research area and applied for many applications [1] such as surveillance system and autonomous vehicles. Even though the substantial performance improvement has been achieved during the last years due to the recent advances in deep learning, the MOT results of the recent trackers are still behind those produced by human annotation. Many tracking failures are frequently occurred due to inaccurate detections and associations. In general, recent MOT methods are based on the tracking-by-detection approach, which builds trajectories by associating (*or* linking) detections. They can be categorized into batch and online methods according to the association manner.

Batch methods [2]–[6] build long trajectories by associating detections of whole frames using iterative associations. Even though they show high tracking accuracy even in complex scenes, applying these methods for real-time and casual applications is not appropriate since detections in whole or future frames are needed. On the other hand, online methods [1], [7]–[12] build trajectories sequentially using frame-by-frame association up to the present frame. Therefore, they can be applied for real-time applications, but their performance is lower than batch methods.

Since both methods construct trajectories by using (temporally) global and/or local associations, identifying each tracklet (*i.e.* linked detections in short frames) at frame is an important process. To this end, appearance models for tracklets are learned and then used for discriminating tracklets. Recently, due to the advances of deep learning, some deep learning-based appearance models [1], [11] have been developed and showed the impressive results. However, they are

not suitable for applications with limited resources. In addition, a large training dataset and many training process are required to apply them for MOT.

In this paper, we propose a discriminative online appearance learning to distinguish appearances of tracked objects and update the learned object appearances. The proposed learning method consists of (1) *appearance discriminability measures* to determine how much a learned appearance model can discriminate other object appearances, (2) *online sample mining* to collect training samples from the associated tracklets, (3) *online appearance learning* to generate and update feature subspaces (*i.e.* weights) for tracklets with the collected samples. More concretely, for learning a subspace, we use the partial least square (PLS) method [13] since it can learn more discriminative subspaces with labeled dataset than principle component analysis (PCA). Furthermore, we present measures to evaluate the appearance model discriminability, and update a PLS subspace when an object appearance model is evaluated to have low discriminability power. As a result, we can maintain the appearance model discriminability, and reduce the appearance learning complexity by preventing unnecessary updates for PLS subspaces with high discriminability.

We apply the proposed appearance learning method for the confidence-based data association [1], and achieve the better performance than state-of-the-art MOT methods on public available MOT benchmark datasets. We also prove the effectiveness and the benefits of the proposed methods though extensive evaluation.

## II. RELATED WORKS

In recent years, tracking methods based on tracking-by-detection can be divided into batch and online methods in terms of the data association manner. Batch tracking methods [2]–[6] build trajectories using detections of whole frames together. Even though they provide the better results than online methods in the most cases, it is hard to apply for real-time applications since they perform iterative global associations for detections of all the frames. In contrast, online tracking methods [1], [7]–[10], [14] consider detections of past and current frames only when associating tracklets. Therefore, they can be suitable for real-time applications. However, they tend to yield identity switches and track fragments by long-term occlusions since future frame information is not used.

In online tracking, a robust data association is required in order to prevent identity switches and track fragments. For the robust data association, many affinity models such as appearance, motion, and shape affinity models are exploited in many works [1], [5], [6], [8], [9], [14]. Among them, the appearance model is crucial since a visual feature is a cue to discriminate objects in many cases.

To design appearance models for MOT, a lot of appearance learning methods have been developed. Due to simplicity, some hand-crafted features (*e.g.* color histograms and histogram of gradient(HoG) [15], [16]) are applied for MOT. For instance, Huang *et al.* [17], Li *et al.* [18], and Xing *et al.* [19] extract histograms for each tracklet and then evaluate affinity between tracklets using correlation coefficient, $x^2$ distance and Bhattacharyya coefficient. Bae and Yoon [14] and Hu *et al.* [20] exploit subspace learning such as incremental linear discriminant analysis and log-euclidean riemannian subspace learning in order to reduce feature dimensionality and improve feature discriminability. Chu *et al.* [11] and Chen *et al.* [21] learn discriminative appearance models using convolutional neural networks as deep appearance models for learning more rich representation. Bae and Yoon [1], Yoon *et al.* [22], Tang *et al.* [23], and Leal-Taixé *et al.* [24] exploit the Siamese network [25] to calculate the affinity between an object pair from the network output directly. Recently, Son *et al.* [26] use a quadruplet network, an improved version of triplet network, for the same purpose. Even though exploiting deep learning improves appearance discriminability, many training samples and costly GPUs are usually required.

In general, appearance learning methods for MOT can be divided into global and object-specific appearance learning methods. The global appearance learning methods [1], [14], [27], [28] discriminate appearances of all tracked objects with an appearance model (*e.g.* a ILDA matrix [14] and a CNN model [1]). Because they learn one model only during tracking, these methods usually require less training samples and predict the affinity score faster than latter one. However, the association accuracy is lower than the object-specific learning since only one model is used when evaluating the affinity. On the other hand, the object-specific appearance learning methods [29]–[31] train an appearance model for each object in general. Therefore, they produce more accurate affinity score, but learning and inference complexity increases in proportion to the number of tracked objects.

To resolve the limitations of the object specific appearance learning, in this paper we propose an effective object-specific appearance learning based on appearance discriminability measures and subspace learning. We argue that the proposed method can reduce the learning complexity efficiently while maintaining the discriminabilty power of each appearance model due to following reasons: (1) By using the appearance discriminability measures, a few appearance models with low discriminability can be updated only during tracking. (2) By using our sample mining and the PLS method, each appearance model can be learned discriminatively with small training samples.

## III. ONLINE MULTI-OBJECT TRACKING FRAMEWORK

In this work, tracking multiple objects is based on the confidence-based data association [1]: a tracklet with high confidence is locally associated with online provided detections to grow the tracklet sequentially, but a tracklet with low confidence is globally associated with other tracklets to link fragmented tracklets. For more accurate local and global association, we combine our discriminative online appearance learning with the confidence-based data association.

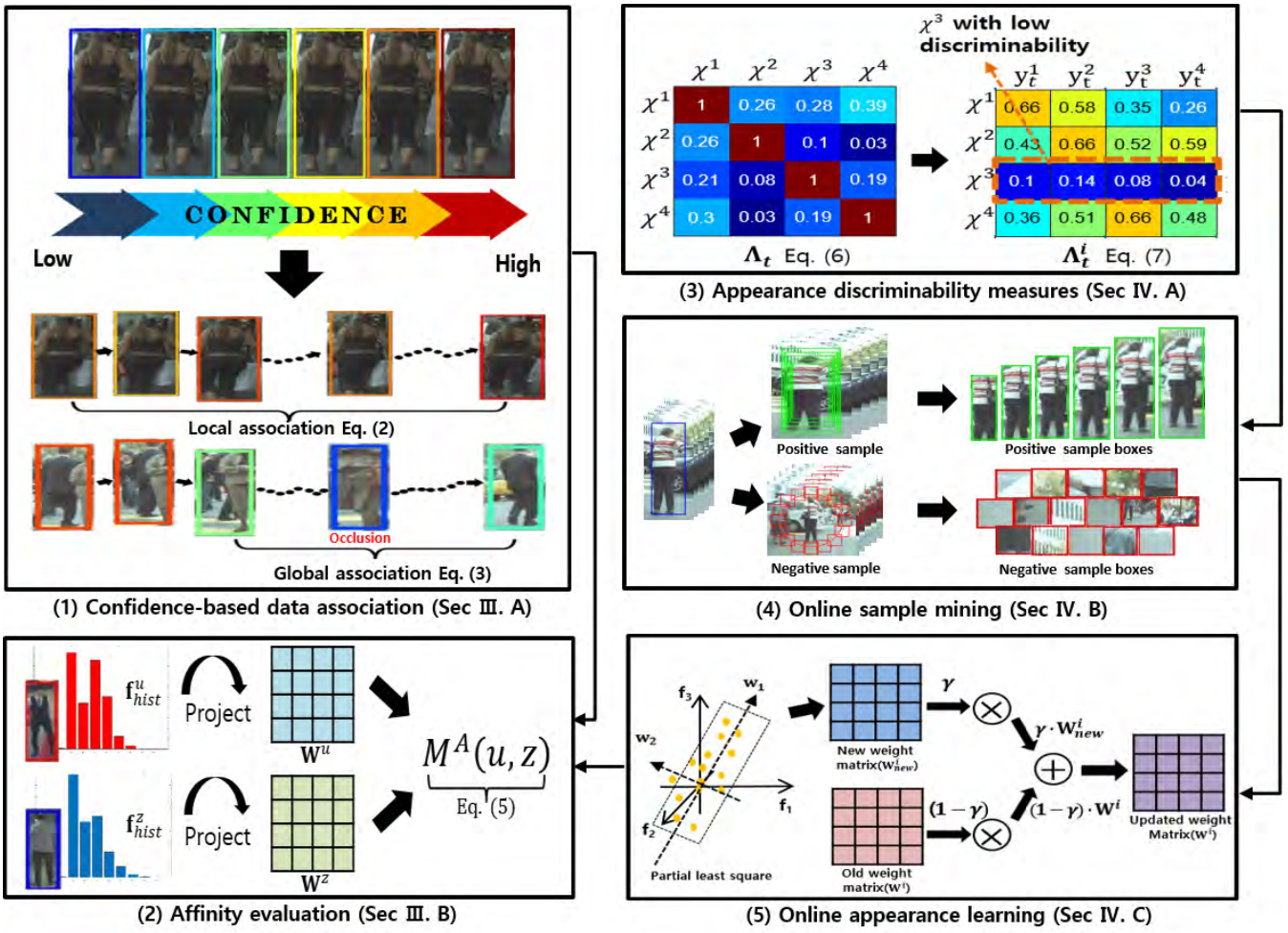


**FIGURE 1.** The overall framework consists of confidence-based data association and discriminative online appearance learning. We divide tracklets to be locally or globally associated according to their confidence. In Fig. 1(1), the confidence-based data association is depicted. Figure 1(2) shows projected features on the PLS weight matrices and appearance affinity evaluation with the projected feature. In order to learn a discriminative appearance model for each object, we propose appearance learning methods shown in Fig. 1(3), 1(4), and 1(5). Figure 1(3) shows appearance discriminability measures computed by Eq. (6) and Eq. (7). For a tracklet with low discriminability, online sample mining and online appearance learning are performed to update its appearance model with collected samples during tracking as shown in Fig. 1(4) and 1(5), respectively.

The overall framework of incorporating both methods is shown in Fig. 1. In the next section, we present details of each method for the framework.

### A. CONFIDENCE BASED DATA ASSOCIATION

Given detections from a trained detector, a detection at frame $t$ is represented as $\mathbf{d}_t = [d_x, d_y, d_w, d_h]$, where $d_x$, $d_y$, $d_w$ and $d_h$ are x and y positions, width, and height, respectively. We then define a tracklet $\chi^i$ as a set of associated detections up to frame $t$ as $\chi^i = \{\mathbf{d}_k^i | 1 \le t_s^i \le k \le t_e^i \le t\}$, where $t_s^i$ and $t_e^i$ are the time stamps of the start- and end-frame of the tracklet. Then, an online multi-object tracking problem can be considered to find a detection $\mathbf{d}_t^i$ which can be associated with a tracklet $\chi^i$ at each frame $t$.

To determine $\mathbf{d}_t^i$ at each frame, we exploit the confidence-based data association. To this end, we first evaluate the confidence of a tracklet $conf(\chi^i)$ in consideration of the length and continuity of a tracklet and the affinity with an associated detection as follows:

$$conf(\chi^i) \models \left(\frac{1}{L}\sum\nolimits_{k\in[t_e^i,t_s^i],v^i(k)=1} M\left(\chi^i, \mathbf{d}_k^i\right)\right) \times \left(1-exp^{-\beta\cdot\sqrt{(L-\lambda)}}\right), \quad (1)$$

where $v^i(t)$ is a binary function for representing 'existness' of $\mathbf{d}_k^i$. If an associated detection for object $i$ exists at frame $t$, $v^i(k) = 1$. Otherwise, $v^i(t) = 0$. $L$ is the length of a tracklet $\chi^i$ as $L = |\chi^i|$, and $\lambda$ is the number of frames in which the object $i$ is missing due to occlusion by other objects or unreliable detection as $\lambda = t_e^i - t_s^i + 1 - L$. $\beta$ is a control parameter relying on the performance of a detector. When a detector shows high accuracy, $\beta$ should be set to a large value ($\beta$ is set to 1.2 as done in [1]). The average affinity $M\left(\chi^i, \mathbf{d}_k^i\right)$ between the tracklet and detection is computed by Eq. (4).

Once the confidence scores of tracklets are computed by Eq. (1), we perform local and global association adaptively

according to their confidence. A tracklet with high confidence $\chi^{i(hi)}$ is considered as a reliable tracklet, and is locally associated with a detection in order to grow it progressively. When $h$ tracklets with high confidence and a detection set $D_t = \{\mathbf{d}_t^j\}_{j=1}^m$ are given at frame $t$, we compute a local association score matrix $S$ as

$$S = [s_{ij}]_{h\times m}, \quad s_{ij} = -M(\chi^{i(hi)}, \mathbf{d}_t^j), \ \mathbf{d}_t^j \in D_t \tag{2}$$

However, a tracklet with low confidence $\chi^{i(lo)}$ is considered as a fragmented trajectory by occlusions. To link fragmented tracklets into one, we associate $\chi^{i(lo)}$ with $\chi^{i(hi)}$ or a detection $\mathbf{y}_t^j$ not associated with any $\chi^{i(hi)}$ in the local association. Assume that there exist $\eta$ non-associated detections ($\eta \leq m$), and $h$ and $l$ tracklets with high and low confidence, respectively. Then, we perform global association by considering following events:

- Event A: $\chi^{i(lo)}$ is associated with $\chi^{j(hi)}$,
- Event B: $\chi^{i(lo)}$ is terminated,
- Event C: $\chi^{i(lo)}$ is associated with $\mathbf{y}_t^j$.

We then define a global association score matrix $G$ for all the events as follows:

$$G_{(l+\eta)\times(h+l)} = \begin{bmatrix} A_{l\times h} & B_{l\times l} \\ -\theta_{\eta\times h} & C_{\eta\times l} \end{bmatrix}, \tag{3}$$

Here, $A = [a_{ij}]$ represents the event A, where $a_{ij} = -M(\chi^{i(lo)}, \chi^{j(hi)})$ is the association cost computed by the affinity between them using Eq. (4). $B = \text{diag}[b_1, \ldots, b_l]$ models the event B, where $b_i = -(1 - conf(\chi^{i(lo)}))$ is the cost to terminate $\chi^{i(lo)}$, and $C = [c_{ij}]$ represents the event C, where $c_{ij} = -M(\chi^{i(lo)}, \mathbf{y}_t^j)$ is the association cost computed by Eq. (4). A threshold $\theta = 0.5$ is employed to select reliable association pairs having high affinity scores.

Once the association score matrices ($S$ or $G$) are computed, we determine optimal matching pairs in each matrix using the Hungarian algorithm [32] such that the total affinity score in the matrix is maximized. Then, detections of the associated pairs are linked each other in a sequential manner, and confidences of all existing tracklets are updated by Eq. (1).

### B. AFFINITY EVALUATION

Since the local Eq. (2) and global Eq. (3) association score matrices are evaluated with affinities between objects, for the more accurate affinity evaluation we use several models when describing a tracklet. Here, $\chi^i$ is represented as $\{A, S, Q\}$, where $A$, $S$ and $Q$ are appearance, shape and motion models, respectively. Then, an overall affinity between a pair of two objects can be defined with those models as

$$M(u, z) = M^A(u, z) \cdot M^S(u, z) \cdot M^Q(u, z), \tag{4}$$

where $u$ and $z$ can be a tracklet or a detection. Each affinity is computed as follows:

$$M^A(u, z) = \max\left(\cos\left(\mathbf{f}_{proj}^u, \mathbf{f}_{proj}^z\right), 0\right),$$

$$M^S(u, z) = \exp\left(-\left\{\frac{\hat{d}_h^u - \hat{d}_h^z}{\hat{d}_h^u + \hat{d}_h^z} + \frac{\hat{d}_w^u - \hat{d}_w^z}{\hat{d}_w^u + \hat{d}_w^z}\right\}\right),$$

$$M^Q(u, z) = \mathcal{N}\left(\hat{\mathbf{d}}_{tail}^u + \mathbf{v}_F^u\Phi; \hat{\mathbf{d}}_{head}^z, O^F\right) \times \mathcal{N}\left(\hat{\mathbf{d}}_{head}^z + \mathbf{v}_B^z\Phi; \hat{\mathbf{d}}_{tail}^u, O^B\right), \tag{5}$$

where $\hat{\mathbf{d}}$ means updated states with Kalman filtering [33]. The shape affinity $M^S(u, z)$ is calculated with their updated width and height. The motion affinity $M^Q(u, z)$ is calculated with $u$ tail (*i.e.* the last updated position) and $z$ head (*i.e.* the first updated position) with time gap $\Phi$. The forward velocity $\mathbf{v}_F^u$ is calculated from the head to tail of $u$, but the backward velocity $\mathbf{v}_B^z$ is calculated from the tail to the head of $z$. We assume that the difference between the predicted position computed with the velocity and the updated position follows Gaussian distribution. For the appearance affinity $M^A(u, z)$, we use the proposed appearance model. We first extract a color histogram $\mathbf{f}_{hist}^u$ for each tracklet from an image, and produce a compact and discriminative feature $\mathbf{f}_{proj}^u$ by projecting $\mathbf{f}_{hist}^u$ onto the learned PLS subspace $W^u$ (*i.e.* $\mathbf{f}_{proj}^u = W^u\mathbf{f}_{hist}^u$) from Eq. (10). We then evaluate $M^A(u, z)$ by computing a cosine similarity between projected features $\mathbf{f}_{proj}^u$ and $\mathbf{f}_{proj}^z$. In the next section, we provide the details of training $W$ during online MOT.

## IV. DISCRIMINATIVE APPEARANCE MODEL

As discussed, an appearance model is key for the local and global association. Using an appearance model with low discriminability decreases the overall association accuracy since it yields high appearance affinity scores for different tracklets. Recently, Bae and Yoon [1] and Chu *et al.* [11] exploit deep learning methods such as a convolutional network in order to learn a more robust appearance model. They also show that using the deep appearance models with rich representation shows the better MOT performance than using the shallow appearance models [14], [20].

Nevertheless, we believe that shallow appearance models could still achieve the comparable performance through efficient learning compared to deep appearance models.[1] In addition, a shallow appearance model with much fewer parameter is more suitable for online MOT because it usually dose not require a large training set and has the lower complexity of learning a model. Therefore, we propose a shallow appearance model using PLS and use it for online MOT.

In particular, to increase appearance discriminability further, we learn and update a object-specific appearance model for each object using the online sample mining and appearance learning. However, the frequent appearance updates for all objects increase the learning complexity significantly. Therefore, we present measures to evaluate appearance discriminability power and update only the appearance with low discriminability.

### A. APPEARANCE DISCRIMINABILITY MEASURES

To evaluate the discriminability of the learned appearance models for whole tracklets, we define an overall appearance

[1] We provide the comparison with deep learning methods in Table 1.

discriminability measure at frame $t$ as $\Lambda_t$. From Eq. (5) we can evaluate an appearance affinity $M^A(\chi^i, \chi^j)$ between two tracklets $\chi^i$ and $\chi^j$ with the cosine distance of their projected features $\mathbf{f}^i_{proj}$ and $\mathbf{f}^j_{proj}$. Suppose that there exist $n$ tracklets consisting of $h$ high confidence and $l$ low confidence tracklets at frame $t$. Then, we can compute $\Lambda_t$ with the appearance affinity between other different tracklets as

$$\Lambda_t = 1 - \frac{1}{n^2 - n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} M^A(\chi^i, \chi^j) - \sum_{i=1}^{n} M^A(\chi^i, \chi^i)\right), \tag{6}$$

where the second term means the average appearance affinity between different tracklets. Thus, $\Lambda_t$ lies in [0, 1]. When $\Lambda_t \geq TR_1$, we consider learned appearance models still maintain high discriminability to distinguish each other. Otherwise, they are considered to have low discriminability and needed to be updated with collected samples during online tracking.

However, because updating all the models is computationally expensive, we first find which tracklet's appearance model is old and then update the corresponding one only. For achieving this, we present $\Lambda^i_t$ to measure discriminability of $i$-th tracklet's appearance model as

$$\Lambda^i_t = 1 - \frac{1}{\eta}\sum_{j=1}^{\eta}\left(M^A(\chi^i, \mathbf{y}^j_t)\right), \tag{7}$$

where $\mathbf{y}^j_t$ is non-associated detections with $\chi^i$, and $\eta$ is the number of the non-associated detections. $M^A(\chi^i, \mathbf{y}^j_t)$ is also computed by Eq. (5). Therefore, $\Lambda^i_t$ measures directly how much the appearance model can discriminate detections originated from other objects or cluttered background. In our experiment, we set $TR_1$ and $TR_2$ to 0.25 and 0.2, respectively. If $\Lambda^i_t < TR_2$, we update the appearance model of the $i$-th tracklet as provided in Sec. IV-C.

### B. ONLINE SAMPLE MINING

For each object $i$, we denote positive $Z^{i,+}_t$ and negative sample $Z^{i,-}_t$ sets as

$$\begin{aligned} Z^{i,+}_t &= \left\{\left(\mathbf{f}^k_{hist}, +1\right)\right\}_{k=1}^{g^+}, \\ Z^{i,-}_t &= \left\{\left(\mathbf{f}^k_{hist}, -1\right)\right\}_{k=1}^{g^-}, \end{aligned} \tag{8}$$

where $g^+$ and $g^-$ are the number of positive and negative samples. $\mathbf{f}^k_{hist}$ is a color histogram feature with dimension $\varrho$ extracted from positive $D^{i,+}_t$ and negative $D^{i,-}_t$ box sets.

#### 1) POSITIVE SAMPLE BOX

Given a bounding box $\mathbf{d}^i = [d_x, d_y, d_w, d_h]$, we represent a rescaled box as $\mathbf{d}^i_{res} = [d_x, d_y, d_w \cdot \psi, d_h \cdot \psi]$, where $\psi$ is a scale factor. We initially set to $\psi = 0.7$ and increase $\psi$ with the interval 0.1 until an overlap ratio $\alpha_{over}$ for an intersection region over an union region between $\mathbf{d}^i$ and $\mathbf{d}^i_{res}$ is below to 0.75. We generate a positive box set $D^{i,+}_t = \left\{\mathbf{d}^i, \mathbf{d}^{i,k}_{res}\right\}_{k=1}^{g^+-1}$ with the original and rescaled boxes, where $\mathbf{d}^{i,k}_{res}$ has $\alpha_{over} \geq 0.75$ for $\mathbf{d}^i$.

#### 2) NEGATIVE SAMPLE BOX

To improve the appearance discriminability between an object and other objects nearby (*or* scene clutter), we collect negative sample boxes around the object. Given an object bounding box $\mathbf{d}^i$, we define a negative sample box as $\mathbf{d}^i_{neg} = \left[d_x + \beta\cos(\omega), d_y + \beta\sin(\omega), d_w/\zeta_w, d_h/\zeta_h\right]$. Here, $\beta = \frac{\rho\sqrt{d_w^2 + d_h^2}}{2}$ and $\omega = \frac{2\pi k}{g^-}$. $k \in \{1, \ldots, g^-\}$ is a negative sample index. In our experiment, we set $\rho$, $\zeta_w$ and $\zeta_h$ to 1.2, 2 and 4, respectively. As a result, a negative sample set $D^{i,-}_t = \left\{d^{i,k}_{neg}\right\}_{k=1}^{g^-}$ is generated by collecting $d^{i,k}_{neg}$ with different $k$. Figure 2 demonstrates positive and negative sample boxes around an object.

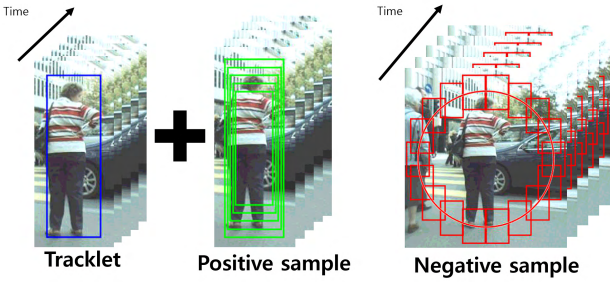


**FIGURE 2. Positive and negative sample boxes generated by online sample mining Sec. IV-B. By rescaling tracklet boxes, positive sample boxes (green boxes) are collected. Negative sample boxes (red boxes) are collected around the object.**

### C. ONLINE APPEARANCE LEARNING

In many MOT methods, online appearance learning is required since the learned models are outdated by appearance changes or occlusions. A main issue of online appearance learning in MOT is to reduce the learning complexity while keeping appearance discriminability. For reducing the complexity, we determine the right time to update an appearance model using the appearance discriminability measures Eq. (6) and Eq. (7) instead of updating it per frame.[2] In addition, we exploit the PLS method to discriminate object appearances since it produces more discriminative subspaces than PCA [34]. Let us denote a sample set of the $i$-th tracklet collected from $t - \Delta + 1$ to $t$ frames as $Z^i_{t-\Delta+1|t}$, where $Z^i_{t-\Delta+1|t}$ consists of $Z^{i,+}_{t-\Delta+1|t}$ and $Z^{i,-}_{t-\Delta+1|t}$ as defined in Eq. (8). Using the NIPALS algorithm [13], we learn a new PLS weight vector $\mathbf{w}$ with dimension $\varrho$ at each iteration as follows:

$$\begin{aligned} \mathbf{w} &= \frac{F^T\mathbf{e}}{\mathbf{e}^T\mathbf{e}}, \quad \mathbf{w} = \frac{\mathbf{w}}{\|\mathbf{w}\|}, \\ \mathbf{r} &= F\mathbf{w}, \quad p = \frac{\mathbf{o}^T\mathbf{r}}{\mathbf{r}^T\mathbf{r}}, \ \mathbf{e} = \frac{\mathbf{o}p}{\sqrt{p^Tp}}, \end{aligned} \tag{9}$$

[2]We provide the comparison of updating models with appearance discriminability measures and updating them per frame in Table 4.

where $F = \{\mathbf{f}^1_{hist}, \mathbf{f}^2_{hist}, \ldots, \mathbf{f}^g_{hist}\}$ is the appearance feature matrix with dimension $g \times \varrho$ consisting of $g$ histogram features with dimension $\varrho$ in $Z^i_{t-\Delta+1|t}$. $\mathbf{r}$, $\mathbf{o}$ and $\mathbf{e}$ are $g$-dimensional feature score, label, and label score vectors, respectively. $p$ is a label loading value. By learning $\mathbf{w}$ for $\tau$ iterations, we can produce a PLS weight matrix $W = \{\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^\tau\}^T$.

Now, we can learn a weight matrix $W^i$ for the $i$-th tracklet with $Z^i_{t-\Delta+1|t}$ using Eq. (9). To update $W^i$, we generate a new weight matrix $W^i_{new}$ with a sample set $Z^i_{t-\Delta+1|t}$ and merge it with the learned $W^i$ as follows:

$$W^i \longleftarrow \gamma W^i_{new} + (1-\gamma)W^i, \tag{10}$$

where the updated $W^i$ is used for generating a PLS feature $\mathbf{f}^i_{proj} = W^i\mathbf{f}^i_{hist}$ and $\mathbf{f}^i_{proj}$ is used for affinity evaluation in Eq. (5). Since the dimension ($\tau = 40$) of $\mathbf{f}^i_{proj}$ is lower than the dimension ($\varrho = 144$) of $\mathbf{f}^i_{hist}$, we can improve the association speed by using $\mathbf{f}^i_{proj}$. $\gamma$ is the hyper parameter for balancing the weights of $W^i_{new}$ and $W^i$. Therefore, $\gamma$ is between 0 and 1. When $\gamma = 1$, $W^i$ is updated with $W^i_{new}$ only. In our case, we set $\gamma$ to 0.5 (refer to Sec. V-B for more details of $\gamma$).

All the procedures of the proposed appearance learning are summarized in Algorithm1.

**Algorithm 1** The Proposed Discriminative Online Appearance Learning

**Input** : Weight matrices $\{W^i\}^n_{i=1}$ for $n$ tracklets.
**Output**: Updated weight matrices $\{W^i\}^n_{i=1}$

```
//Evaluating overall appearance discriminability
Compute Λ_t using Eq. (6);
if Λ_t < TR_1 then
    for i ← 1 to n do
        //Evaluating appearance discriminability of each tracklet
        Compute Λ^i_t using Eq. (7);
        if Λ^i_t < TR_2 then
            // Updating old W^i
            Generate Z^i_{t−Δ+1|t};
            Learn W^i_new with Z^i_{t−Δ+1|t} by Eq. (9);
            Update W^i using Eq. (10);
        end
    end
end
```

### D. DISCUSSION OF APPEARANCE DISCRIMINABILITY MEASURES

It is possible to skip the overall appearance discriminability evaluation $\Lambda_t$ in Eq. (6) and try to find an appearance model with low discriminability via only an appearance discriminability measure for each tracklet $\Lambda^i_t$ in Eq. (7). However, although some appearance models show low appearance discriminability, association pairs can be determined well due to other appearance models with high discriminability. Therefore, in many cases it is unnecessary to update an appearance model with low discriminability immediately if the other appearance models still maintain high discriminability. In order to prevent frequent appearance updates, we therefore exploit both measures $\Lambda_t$ and $\Lambda^i_t$ together. As a result, we can boost the tracking speed by reducing the number of appearance update. This is proven in Table 4.

## V. EXPERIMENTS

### A. MOT SYSTEM IMPLEMENTATION

#### 1) DATASET

To evaluate our method, we use the 2016 and 2017 multiple object tracking (MOT16/17) challenge benchmark sets [35] for pedestrian tracking. The MOT16 dataset includes 7 training and 7 test sequences captured from moving or static cameras with different frame rates. The MOT17 dataset also includes 7 training and 7 test sequences, and 3 detection sets by applying DPM [36], Faster-RCNN [37], and SDP [38] are provided per sequence. Therefore, 21 different training and 21 different test sets are provided in the MOT17 dataset, respectively. Also, the crowded density of objects is different each other. For a fair comparison, we use only detections and ground truth provided in the MOT16/17 challenges.

#### 2) EVALUATION METRICS

We use common metrics which are also used in the MOT benchmark challenge: the multiple object tracking accuracy (MOTA↑), multiple object tracking precision (MOTP↑), the ratio of mostly tracked trajectories (MT↑), the ratio mostly lost trajectories (ML↓), the number of track fragment (FG↓), false alarms per frame (FAF↓), the number of false positives (FP↓), the number of false negative (FN↓), the number of identity switches (IDS↓) and tracker speed in frames per second (Hz↑). Here, ↑ and ↓ represent that higher and lower scores are better results, respectively.

### B. PERFORMANCE EVALUATION

#### 1) SYSTEM PARAMETERS

We have tuned all parameters from experiments and unchanged them for all the evaluation. As shown in Fig. 1, our framework consists of confidence-based multi-object tracking (Fig. 1. (1)-(2)) and discriminable online appearance learning (Fig. 1. (3)-(5)) parts. In the first part, the most of parameters are set to be identical with [1] except for $\varrho$. $\varrho$ is a dimension of histogram feature and is tuned to 144.

The appearance learning part includes the following parameters ($\psi$, $\rho$, $\zeta_w$, $\zeta_h$, $\Delta$, $\tau$, $g$, $\gamma$, $TR1$ and $TR2$). $\psi$, $\rho$, $\zeta_w$, $\zeta_h$ are parameters to collect training samples in the online sample mining method. The setting values of these parameters are provided in Sec. IV-B. In fact, changing these parameters affects the number of training samples $g$. Therefore, we provide the evaluation results for the speed and accuracy of the our tracker by changing $g$ in Fig. 4.

**TABLE 1. Performance comparison with other MOT systems on the 2016 and 2017 MOT challenge benchmark. The results are sorted according to the setting and MOTA score. (More results can be found in the 2016 and 2017 MOTChallenge website.)**

| Benchmark | Method | Setting | Learning | Detector | MOTA ↑ | MOTP ↑ | FAF↓ | MT ↑ | ML ↓ | FP ↓ | FN ↓ | IDS ↓ | FG ↓ | Hz ↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MOT16 Challenge Test Set (7 sequences) | Proposed (AM_ADM) | Online | Shallow | Public | 40.1 % | 75.4 % | 1.4 | 7.1 % | 46.2 % | 8503 | 99891 | 789 | 1736 | 5.8 |
| | STAM16 [11] | Online | Deep | Public | 46.0 % | 74.9 % | 1.2 | 14.6 % | 43.6 % | 6895 | 91117 | 473 | 1422 | 0.2 |
| | CDA_DDALv2 [1] | Online | Deep | Public | 43.9 % | 74.7 % | 1.1 | 10.7 % | 44.4 % | 6450 | 95175 | 676 | 1795 | 0.5 |
| | TBSS [39] | Online | Shallow | Public | 44.6 % | 75.2 % | 0.7 | 12.3 % | 43.9 % | 4136 | 96128 | 790 | 1419 | 3.0 |
| | oICF [40] | Online | Shallow | Public | 43.2 % | 74.3 % | 1.1 | 11.3 % | 48.5 % | 6651 | 96515 | 381 | 1404 | 0.4 |
| | EAMTT_pub [41] | Online | Shallow | Public | 38.8 % | 75.1 % | 1.4 | 7.9 % | 49.1 % | 8114 | 102452 | 965 | 1657 | 11.8 |
| | OVBT [42] | Online | Shallow | Public | 38.4 % | 75.4 % | 1.9 | 7.5 % | 47.3 % | 11517 | 99463 | 1321 | 2140 | 0.3 |
| | JCmin_MOT [43] | Online | Shallow | Public | 36.7 % | 75.9 % | 0.5 | 7.5 % | 54.4 % | 2936 | 111890 | 667 | 831 | 14.8 |
| | GMPHD_HDA [44] | Online | Shallow | Public | 30.5 % | 75.4 % | 0.9 | 4.6 % | 59.7 % | 5169 | 120970 | 539 | 731 | 13.6 |
| | QuadMOT16 [26] | Batch | Deep | Public | 44.1 % | 76.4 % | 1.1 | 14.6 % | 44.9 % | 6388 | 94775 | 745 | 1096 | 1.8 |
| | LINF1 [45] | Batch | Shallow | Public | 41.0 % | 74.8 % | 1.3 | 11.6 % | 51.3 % | 7896 | 99224 | 430 | 963 | 4.2 |
| | GMMCP [46] | Batch | Shallow | Public | 38.1 % | 75.8 % | 1.1 | 8.6 % | 50.9 % | 6607 | 105315 | 937 | 1669 | 0.5 |
| | DP_NMS [47] | Batch | Shallow | Public | 26.2 % | 76.3 % | 0.6 | 4.1 % | 67.5 % | 3689 | 130557 | 365 | 638 | 5.9 |
| MOT17 Challenge Test Set (7 sequences and 3 detection sets) | Proposed (AM_ADM17) | Online | Shallow | Public | 48.1 % | 76.7 % | 1.4 | 13.4 % | 39.7 % | 25207 | 265393 | 2217 | 5031 | 5.7 |
| | MOTDT17 [48] | Online | Deep | Public | 50.9 % | 76.6 % | 1.4 | 17.5 % | 35.7 % | 24069 | 250768 | 2474 | 5317 | 18.3 |
| | PHD_GSDL17 [10] | Online | Shallow | Public | 48.0 % | 77.2 % | 1.3 | 17.1 % | 35.6 % | 23199 | 265954 | 3998 | 8886 | 6.7 |
| | EAMTT [41] | Online | Shallow | Public | 42.6 % | 76.0 % | 1.7 | 12.7 % | 42.7 % | 30711 | 288474 | 4488 | 5720 | 1.4 |
| | GMPHD_KCF [49] | Online | Shallow | Public | 39.6 % | 74.5 % | 2.9 | 8.8 % | 43.3 % | 50903 | 284228 | 5811 | 7414 | 3.3 |
| | FWT [50] | Batch | Deep | Public | 51.3 % | 77.0 % | 1.4 | 2.14 % | 35.2 % | 24101 | 247921 | 2648 | 4279 | 0.2 |
| | jCC [3] | Batch | Deep | Public | 51.2 % | 75.9 % | 1.5 | 20.9 % | 37.0 % | 25937 | 247822 | 1802 | 2984 | 1.8 |
| | MHT_DAM [5] | Batch | Deep | Public | 50.7 % | 77.5 % | 1.3 | 20.8 % | 36.9 % | 22875 | 252889 | 2314 | 2865 | 0.9 |
| | EDMT17 [51] | Batch | Deep | Public | 50.0 % | 77.3 % | 1.8 | 21.6 % | 36.3 % | 32279 | 247297 | 2264 | 3260 | 0.6 |

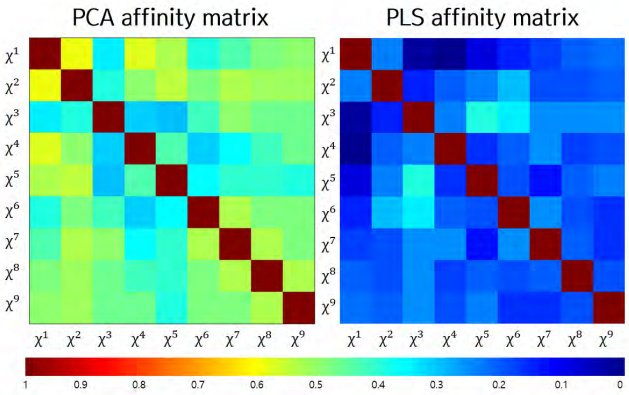


**FIGURE 3. Appearance affinity matrices using PCA and PLS. Each element of these matrices represents an appearance affinity score between tracklets.**

In addition, $\Delta$ controls the frame interval to collect training samples. More specifically, the samples for each tracklet are collected from $t - \Delta + 1$ to $t$ frames. $\tau$ is the number of PLS vectors used in Eq. (9). $\gamma$ is the hyper parameter for balancing the weights of $W^i_{new}$ and $W^i$ used in Eq. (10). $TR1$ and $TR2$ are thresholds used for measuring appearance discriminability of a tracklet as described in Sec. IV-A. Since these parameters ($\Delta$, $\tau$, $TR1$ and $TR2$) could affect performance of our system, we have investigated the sensitivity of our tracker over these parameters in Fig. 5. In addition, we evaluate the performance of our tracker for different $\gamma$ in Table 5.

Based on the evaluation results in Fig. 4-5 and Table 5, we determined the values of the hyper parameters which maximize the speed and accuracy of our tracker together. As a result, we set $g$, $\Delta$, $\tau$, $\gamma$, $TR1$ and $TR2$ to 24, 5, 40, 0.5, 0.25, and 0.2.

### 2) MOT16/17 CHALLENGE EVALUATION

We evaluated our tracking system on the MOT Benchmark website [35], and compared with other state-of-the-art tracking systems. Table 1 shows the performance of our system on the test sets in the MOT16/17 challenges. In Table 1, we divide tracking systems into online and batch, and appearance models into shallow and deep models. Table 1 shows that our system achieves the better results for several metrics than other systems. In particular, our system achieves the best MOTA rate among online tracking systems using shallow appearance models on MOT17. In MOT16, our system is superior to other online systems using shallow models. Only two online tracking systems using shallow models [39], [40] show the higher MOT scores than ours, but our system is much faster than [39], [40]. In addition, our system shows the comparable performance with the recent tracking systems [1], [11], [26], [48] using deep learning. These results indicate that the proposed appearance learning can build trajectories under occlusions by discriminating object appearances accurately.

### 3) COMPARISON OF APPEARANCE MODELS

To evaluate our appearance learning method, we have implemented several MOT systems with different appearance models:

(a1) Object-specific appearance models using PLS
(a2) A global appearance model using PLS
(a3) Object-specific appearance models using color histogram
(a4) Object-specific appearance models using PCA

Here, all the systems (a1-a4) use our sample mining and appearance discriminability measures in common, but use different features. (a1) and (a2) use the learned features using PLS. In (a1), we generate a PLS matrix $W^i$ per object and produce a PLS feature $\mathbf{f}^i_{proj}$ by projecting a histogram feature $\mathbf{f}^i_{hist}$ on $W^i$ (*i.e.* $\mathbf{f}^i_{proj} = W^i\mathbf{f}^i_{hist}$). However, in (a2), we generate a common PLS matrix $W^{global}$ with object labels, and generate $\mathbf{f}^i_{proj}$ by projecting $\mathbf{f}^i_{hist}$ on $W^{global}$ (*i.e.* $\mathbf{f}^i_{proj} = W^{global}\mathbf{f}^i_{hist}$). (a3) uses a color histogram feature $\mathbf{f}^i_{hist}$ for each object.

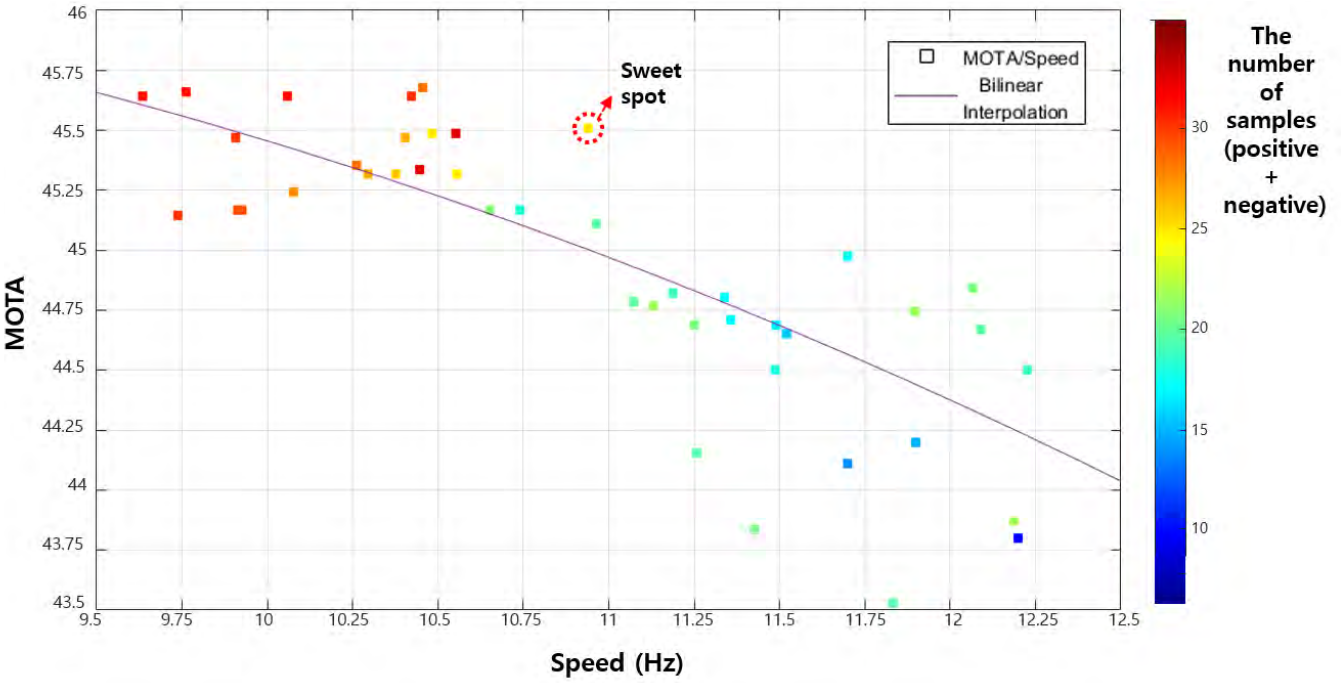


**FIGURE 4.** Speed and accuracy by using different $g$ on the MOT16-09 sequence. Marker colors are changed according to $g$. $g$ is selected by online sample mining (IV-B).

**TABLE 2.** Performance comparison with difference appearance (App.) models.

| Benchmark | Method | MOTA ↑ | MOTP ↑ | FAF ↓ | MT ↑ | ML ↓ | FP ↓ | FN ↓ | IDS ↓ | FG ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| MOT16 | (a1) with proposed app. | 34.1 % | 76.8 % | 1.5 | 9.1 % | 47.8 % | 8198 | 64231 | 359 | 914 |
| Challenge | (a2) with a global app. | 32.7 % | 76.4 % | 1.7 | 9.3 % | 47.4 % | 9233 | 64636 | 489 | 1006 |
| Training Set | (a3) with color hist. | 31.7 % | 76.4 % | 2.0 | 9.2 % | 48.2 % | 10634 | 64261 | 490 | 1051 |
| (7 sequences) | (a4) with PCA | 31.6 % | 76.5 % | 1.8 | 7.2 % | 48.9 % | 9418 | 65597 | 520 | 1026 |

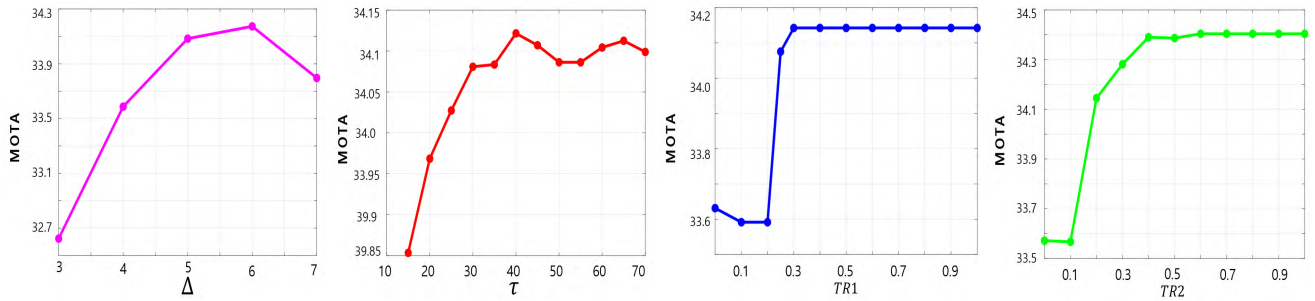


**FIGURE 5.** Sensitivity analysis for appearance learning parameters on MOT16 training set. (a)-(d) show the MOTA rates of our system when $\Delta$, $\tau$, $TR1$ and $TR2$ change. In each evaluation, other parameter values are fixed.

In (a4), we generate a PCA matrix for each object and project $\mathbf{f}^i_{hist}$ on each PCA matrix.

In Table 2, we evaluate (a1)-(a4) for several metrics on a MOT16 training set. As shown, we obtain the best results for the most metrics when using (a1). When comparing (a1) and (a2), (a1) greatly reduces FP. This means that the object-specific appearance model is more accurate to discriminate each object with background. We also know that the features learned by PLS (a1) and (a2) have more discriminability power than the color histogram feature (a3) and the feature learned by PCA (a4).

For more comparison between PLS and PCA, we compare appearance affinity matrices between different tracklets by using PLS and PCA in Fig. 3. As shown, the matrix

**TABLE 3.** The speed (Hz) of our system is computed on the MOT16/17 training sets. The average speed (Avg.) in each set is calculated by dividing the total running time into the total number of frames for whole sequences. Here, DPM, FRCNN and SDP represent detection sets obtained from different detectors as described in Sec. V-A.

| Sequence | MOT16-02 | MOT16-04 | MOT16-05 | MOT16-09 | MOT16-10 | MOT16-11 | MOT16-13 | Avg. |
|---|---|---|---|---|---|---|---|---|
| Hz ↑ | 4.88 | 4.43 | 26.98 | 10.61 | 10.7 | 11.76 | 13.23 | 8.38 |
| **Sequence** | **MOT17-02 -DPM** | **MOT17-04 -DPM** | **MOT17-05 -DPM** | **MOT17-09 -DPM** | **MOT17-10 -DPM** | **MOT17-11 -DPM** | **MOT17-13 -DPM** | **Avg.** |
| Hz ↑ | 4.61 | 4.18 | 25.75 | 10.79 | 11.15 | 12.28 | 14.56 | 8.23 |
| **Sequence** | **MOT17-02 -FRCNN** | **MOT17-04 -FRCNN** | **MOT17-05 -FRCNN** | **MOT17-09 -FRCNN** | **MOT17-10 -FRCNN** | **MOT17-11 -FRCNN** | **MOT17-13 -FRCNN** | **Avg.** |
| Hz ↑ | 4.71 | 4.12 | 25.41 | 11.37 | 9.85 | 14.48 | 9.79 | 7.98 |
| **Sequence** | **MOT17-02 -SDP** | **MOT17-04 -SDP** | **MOT17-05 -SDP** | **MOT17-09 -SDP** | **MOT17-10 -SDP** | **MOT17-11 -SDP** | **MOT17-13 -SDP** | **Avg.** |
| Hz ↑ | 3.55 | 3.16 | 21.81 | 9.29 | 9.40 | 13.34 | 10.43 | 6.60 |

learned by PLS produces more discriminative affinity scores; lower scores for the pairs of different objects, but higher scores for the pairs of the same objects. Since the appearance affinity matrix is used for the association and appearance discriminability measures, our appearance model using PLS can improve the overall tracking performance as shown in Table 2.

#### 4) SPEED AND ACCURACY

We analyze the speed and accuracy of our tracking system with the proposed appearance learning. As mentioned, the performance is affected by the number of training samples ($g = g^+ + g^-$) used for learning a PLS weight matrix Eq. (9). To find out the relationship between them, we evaluate our system by changing $g$ on the MOT16-09 sequence. For evaluating the accuracy, we use the MOTA metric which shows overall tracking performance. As shown in Fig. 4, using more training samples can improve the accuracy but reduce the speed. When using $g = 24$ with $g^+ = 8$ and $g^- = 16$, we achieve the best result (sweet spot) to trade-off the speed and accuracy. Here, the sweet spot point $g$ maximizes the summation for the weighted MOTA and Hz scores. We set the weights to 0.7 and 0.3 to increase the tracking accuracy more. Therefore, for all the experiment, we set $g$ to 24.

In addition, Table 3 provides the details for the speed of our system on the MOT16/17 training set. In average, it performs in 8.38 Hz and 7.53 Hz on the MOT16/17 training sets, respectively. The tracking speed is different according to the object crowded density in a sequence. In particular, ours shows the real-time speed (26.98 Hz) in the less crowded scene (MOT16-05).

For the more comparison, we compare the speed of our system with other systems on the MOT16/17 test sets. As shown in Table 1, our system has the competitive speed compared to other systems. On the MOT16 challenge, our system is faster than other state-of-the-art tracking systems [1], [11], [26], [39], [40]. On the MOT17 challenge, our system is also faster than [3], [5], [41], [49]–[51]. We emphasize that our system achieves the average speed of the 5.8 Hz and 5.7 Hz without using GPU (*e.g.* MOTDT17 [48] ) and parallel programming. Therefore, there is a still room to improve the speed by applying the techniques.

#### 5) SENSITIVITY ANALYSIS

We also analyze the sensitivity of our system over $\Delta$, $\tau$, $TR1$ and $TR2$. In order to evaluate sensitivity, we extensively evaluate the overall multi-object tracking accuracy (MOTA) of our system by changing the values of those parameters on the MOT16 training set consisting of 7 sequences. Figure 5 shows the sensitivity evaluation results. As shown, the MOTA scores of our system are not changed much according to the parameter values. This indicates that our system is not sensitive to the hyper parameters. From the experimental results, we set $\Delta$ and $\tau$ to 5 and 40 in consideration of tracking accuracy and speed together. In addition, we tune $TR1$ and $TR2$ to be 0.25 and 0.2, respectively. Although these parameter values could not provide the best MOTA scores shown in Fig. 5(c)-5(d), we set the values to boost the tracking speed. As a result, compared to the speed of our system using $TR1 = 1$ and $TR2 = 1$, the speed increases approximately 7.3 times.

#### 6) EVALUATION OF APPEARANCE DISCRIMINABILITY MEASURES

To evaluate the effectiveness of the proposed appearance discriminability measures $\Lambda_t$ and $\Lambda_t^i$ in Eq. (6) and Eq. (7), we compare tracking performance by applying the measures $\Lambda_t$ and $\Lambda_t^i$ in a different way. We implement the following systems:

(b1) system with all $\Lambda_t$ and $\Lambda_t^i$
(b2) system without $\Lambda_t$
(b3) system without $\Lambda_t$ and $\Lambda_t^i$

From the evaluation results of (b1)-(b3) shown in Table 4, we can see the effect of the proposed appearance discriminability measures. When comparing the system

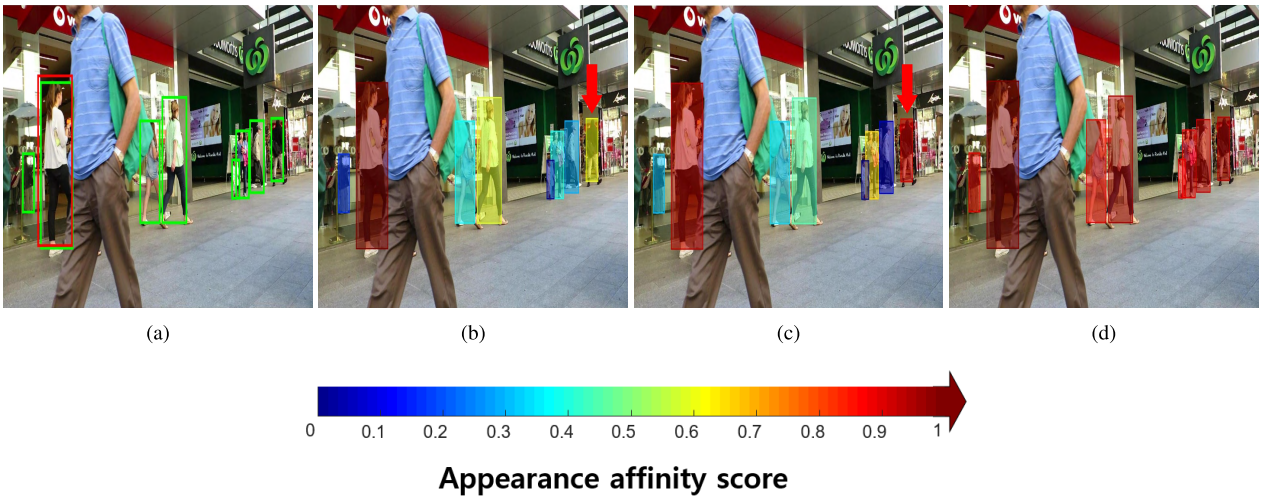


**FIGURE 6.** (a) A tracklet and detections are marked with red and green boxes, respectively. (b)-(d) Appearance affinity scores between a tracklet and other detections are computed by using the PLS, PCA, color histogram, respectively.


**FIGURE 7.** Tracking results by using our discriminative appearance learning method under occlusions. Each column shows a tracked object before/during/after the occlusion, respectively. On the sequence (MOT16-02) captured in a static camera, we also depict the trajectory of the tracked object. (a) MOT16-02 frame 452. (b) MOT16-02 frame 457. (c) MOT16-02 frame 464. (d) MOT16-05 frame 254. (e) MOT16-05 frame 259. (f) MOT16-05 frame 266. (g) MOT16-11 frame 756. (h) MOT16-11 frame 769. (i)MOT16-11 frame 782.

(b1) and (b2), exploiting the overall appearance discriminability $\Lambda_t$ allows us to improve the speed by reducing the number (#) of appearance model update. Compared to (b3), which updates all appearance models per frame, (b1) is faster about 7.3 times. In terms of the accuracy, (b3) shows the slightly higher MOTA due to the lower FP score than (b1). However, when considering the accuracy and speed together, we confirm that using the both $\Lambda_t$ and $\Lambda_t^i$ is more effective.

**FIGURE 8.** (a)-(n) Tracking results using the proposed appearance model on the 2016 MOTChallenge dataset.

**TABLE 4.** Evaluation of the proposed appearance discriminability measures $\Lambda_t$ and $\Lambda_t^i$.

| Benchmark | Method | MOTA ↑ | MOTP ↑ | FAF↓ | MT ↑ | ML ↓ | FP ↓ | FN ↓ | IDS ↓ | FG ↓ | Hz ↑ | Update # |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MOT16 | (b1) with all (proposed) | 34.1 % | 76.8 % | 1.5 | 9.1 % | 47.8 % | 8198 | 64231 | 359 | 914 | 8.38 | 670 |
| Training Set | (b2) without $\Lambda_t$ | 34.1 % | 76.8 % | 1.5 | 9.1 % | 47.8 % | 8207 | 64231 | 359 | 914 | 7.95 | 1499 |
| (7 sequences) | (b3) without $\Lambda_t$ and $\Lambda_t^i$ | 34.4 % | 76.8 % | 1.5 | 8.5 % | 47.8 % | 7998 | 64024 | 362 | 918 | 1.15 | 46024 |

### 7) ONLINE LEARNING EVALUATION

We evaluate the effect of our online appearance learning by changing the hyper parameter $\gamma$ used for balancing old and new weight matrices in Eq. (10):

(c1) Updating $W^i$ with old and new models ($\gamma = 0.5$)
(c2) Fixing $W^i$ with the learned model initially ($\gamma = 0$)
(c3) Updating $W^i$ with a new model only ($\gamma = 1$)

In Table 5, the comparison results of (c1)-(c3) are shown. The proposed (c1) achieves the best MOTA, MT, FN, IDS, and FG scores. In particular, (c1) greatly reduces the IDS and FG. When comparing (c2) and (c3), both methods show almost similar performance for the most metrics. These results verify the effect of our online learning which updates appearance models with old and new models.

**TABLE 5.** Performance evaluation of the proposed online appearance learning by changing the hyper parameter $\gamma$.

| Benchmark | Method | MOTA ↑ | MOTP ↑ | FAF↓ | MT ↑ | ML ↓ | FP ↓ | FN ↓ | IDS ↓ | FG ↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| MOT16 | (c1) with $\gamma = 0.5$ (proposed) | 34.1 % | 76.8 % | 1.5 | 9.1 % | 47.8 % | 8198 | 64231 | 359 | 914 |
| Training Set | (c2) with $\gamma = 0$ | 32.7 % | 76.9 % | 1.5 | 7.5 % | 49.1 % | 7969 | 65836 | 481 | 1018 |
| (7 sequences) | (c3) with $\gamma = 1$ | 32.6 % | 76.9 % | 1.5 | 7.5 % | 49.1 % | 7979 | 65916 | 479 | 1016 |

#### 8) QUALITATIVE EVALUATION

Figure 6 shows appearance affinity scores between a tracklet and detections, and the scores are computed by learned appearance models with the PLS, PCA, and color histogram. In Fig. 6(a), we mark a tracklet and detections with red and green bounding boxes, respectively. In Fig. 6(b)- 6(d), we represent the affinity scores of detections with colored boxes. As shown, appearance models learned by a PLS weight matrix produce more discriminative scores (*i.e.* lower association scores) for detections from other objects than those learned by PCA and color histogram. In particular, the color histogram generates high association scores even for many other detections as shown in Fig. 6(d). When comparing PLS and PCA, PLS can still produce a lower association score for the detection (the arrow with red color) which is difficult to distinguish due to the similar clothes.

Figure 7 shows the tracking results of the proposed method under severe occlusions. Even though the objects are fully occluded by other objects as in Fig. 7(b), 7(e), and 7(h), we can correctly maintain their IDs using our discriminative appearance models after occlusions.

Figure 8 shows the tracking results using the proposed appearance model on the 2016 MOTChallenge dataset. Our system robustly tracks the most objects even though the objects are frequently occluded and their appearance and motion are changed with time.

## VI. CONCLUSION

In this paper, we have proposed a discriminative online appearance learning for multi-object tracking. We present online sample mining and online appearance learning methods to learn and update appearance models of tracked objects with incoming tracking results. To alleviate computational complexity of learning appearance models, we propose appearance discriminability measures to evaluate the appearance discriminability between all tracklets and for each tracklet. We then determine and update appearance models with low discriminability only.

We have verified the effectiveness of the proposed methods from extensive evaluation. In addition, our method achieves the improved performance over state-of-the-art tracking methods on the MOT16/17 benchmark challenges. Although we combine our appearance learning method with the confidence-based data association method in this paper, we believe that the proposed appearance learning method can be compatible for other online and batch tracking methods since it does not depend on association methods.

## ACKNOWLEDGMENT

*(Seong-Ho Lee and Myung-Yun Kim contributed equally to this work.)*

## REFERENCES

[1] 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.

[2] X. Wang *et al.*, "Greedy batch-based minimum-cost flows for tracking multiple objects," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4765–4776, Oct. 2017.

[3] M. Keuper, S. Tang, Z. Yu, B. Andres, T. Brox, and B. Schiele, "A multi-cut formulation for joint segmentation and tracking of multiple objects," *CoRR*, Jul. 2016. [Online]. Available: https://arxiv.org/abs/1607.06317

[4] S. S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 19, no. 1, pp. 5–18, Jan. 2004.

[5] 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.

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

[7] 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.*, Sep. 2012, pp. 325–330.

[8] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association with online target-specific metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1234–1241.

[9] J. Wei, M. Yang, and F. Liu, "Learning spatio-temporal information for multi-object tracking," *IEEE Access*, vol. 5, pp. 3869–3877, Mar. 2017.

[10] 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, Mar. 2018.

[11] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4836–4845.

[12] A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4225–4232.

[13] H. Wold, "11—Path models with latent variables: The NIPALS approach," in *Quantitative Sociology* (International Perspectives on Mathematical and Statistical Modeling), H. Blalock, A. Aganbegian, F. Borodkin, R. Boudon, and V. Capecchi, Eds. New York, NY, USA: Academic, 1975, pp. 307–357.

[14] 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. 2014, pp. 1218–1225.

[15] B. Benfold and I. D. Reid, "Stable multi-target tracking in real-time surveillance video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3457–3464.

[16] J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 2, pp. 283–298, Jun. 2009.

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

[18] 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.

[19] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1200–1207.
[20] W. Hu, X. Li, W. Luo, X. Zhang, S. J. Maybank, and Z. Zhang, "Single and multiple object tracking using log-Euclidean Riemannian subspace and block-division appearance model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2420–2440, Dec. 2012.
[21] L. Chen, H. Ai, C. Shang, Z. Zhuang, and B. Bai, "Online multi-object tracking with convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 645–649.
[22] Y. Yoon, A. Boragule, K. Yoon, and M. Jeon, "Online multi-object tracking with historical appearance matching and scene adaptive detection filtering," *CoRR*, May 2018. [Online]. Available: https://arxiv.org/abs/1805.10916
[23] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person reidentification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3701–3710.
[24] 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*, Jun. 2016, pp. 418–425.
[25] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, vol. 1, Jun. 2005, pp. 539–546.
[26] J. Son, M. Baek, M. Cho, and B. Han, "Multi-object tracking with quadruplet convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3786–3795.
[27] B. Yang and R. Nevatia, "An online learned CRF model for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2034–2041.
[28] C.-H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 685–692.
[29] C.-H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1217–1224.
[30] 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.
[31] X. Song, J. Cui, H. Zha, and H. Zhao, "Vision-based multiple interacting targets tracking via on-line supervised learning," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 642–655.
[32] F. Bourgeois and J.-C. Lassalle, "An extension of the munkres algorithm for the assignment problem to rectangular matrices," *Commun. ACM*, vol. 14, no. 12, pp. 802–804, Dec. 1971.
[33] J. D. Hamilton, *Time Series Analysis*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1994.
[34] H.-N. Qu, G.-Z. Li, and W.-S. Xu, "An asymmetric classifier based on partial least squares," *Pattern Recognit.*, vol. 43, no. 10, pp. 3448–3457, Oct. 2010.
[35] A. Milan, L. Leal-Taixé, I. D. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *CoRR*, Mar. 2016. [Online]. Available: https://arxiv.org/abs/1603.00831
[36] 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.
[37] 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, 2015, pp. 91–99.
[38] 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.*, Jun. 2016, pp. 2129–2137.
[39] 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.*, Sep. 2018, p. 203.
[40] H. Kieritz, S. Becker, W. Hübner, and M. Arens, "Online multi-person tracking using integral channel features," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2016, pp. 122–130.
[41] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro, "Online multi-target tracking with strong and weak detections," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 84–99.
[42] Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horaud, "Tracking multiple persons based on a variational Bayesian model," in *Proc. ECCV Workshops*, Amsterdam, The Netherlands, vol. 9914, Oct. 2016, pp. 52–67.
[43] A. Boragule and M. Jeon, "Joint cost minimization for multi-object tracking," in *Proc. IEEE Conf. Adv. Video Signal Based Surveill.)*, Aug. 2017, pp. 1–6.
[44] 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.
[45] L. Fagot-Bouquet, R. Audigier, Y. Dhome, and F. Lerasle, "Improving multi-frame data association with sparse representations for robust near-online multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 774–790.
[46] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4091–4099.
[47] 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.
[48] C. Long, A. Haizhou, Z. Zijie, and S. Chong, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," in *Proc. IEEE Conf. Multimedia Expo*, Jul. 2018, pp. 1–6.
[49] 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. IEEE Int. Workshop Traffic Street Surveill. Safety Secur. (AVSS)*, Aug. 2017, pp. 1–5.
[50] R. Henschel, L. Leal-Taixe, 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*, 2018, pp. 1428–1437.
[51] 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. 2143–2152.

**SEONG-HO LEE** is currently pursuing the B.S. degree with the Department of Computer Science and Engineering, Incheon National University, South Korea. His current research interests are multi-object tracking, object detection, deep learning, and zero-shot learning.

**MYUNG-YUN KIM** is currently pursuing the B.S. degree with the Department of Computer Science and Engineering, Incheon National University, South Korea. His current research interests are multi-object tracking, object detection, and deep learning.

**SEUNG-HWAN BAE** (M'18) received the B.S. degree in information and communication engineering from Chungbuk National University in 2009 and the M.S. and Ph.D. degrees in information and communications from the Gwangju Institute of Science and Technology in 2010 and 2015, respectively. He was a Senior Researcher with the Electronics and Telecommunications Research Institute, South Korea, from 2015 to 2017. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Incheon National University, South Korea. His research interests include multi-object tracking, object detection, deep learning, dimensionality reduction, and medical image analysis.