Robust Online Multi-Object Tracking based on Tracklet Confidence and Online Discriminative Appearance Learning

Seung-Hwan Bae and Kuk-Jin Yoon
Computer Vision Laboratory, GIST, Korea
{bshwan, kjyoon}@gist.ac.kr

Abstract

Online multi-object tracking aims at producing complete tracks of multiple objects using the information accumulated up to the present moment. It still remains a difficult problem in complex scenes, because of frequent occlusion by clutter or other objects, similar appearances of different objects, and other factors. In this paper, we propose a robust online multi-object tracking method that can handle these difficulties effectively.

We first propose the tracklet confidence using the detectability and continuity of a tracklet, and formulate a multi-object tracking problem based on the tracklet confidence. The multi-object tracking problem is then solved by associating tracklets in different ways according to their confidence values. Based on this strategy, tracklets sequentially grow with online-provided detections, and fragmented tracklets are linked up with others without any iterative and expensive associations. Here, for reliable association between tracklets and detections, we also propose a novel online learning method using an incremental linear discriminant analysis for discriminating the appearances of objects. By exploiting the proposed learning method, tracklet association can be successfully achieved even under severe occlusion. Experiments with challenging public datasets show distinct performance improvement over other batch and online tracking methods.

1. Introduction

The goal of multi-object tracking is to estimate the states of multiple objects while conserving their identifications under appearance and motion variations with time. In a complex scene, this problem is especially challenging due to frequent occlusion by clutter or other objects, similar appearances of different objects, and so on.

Recently, tracking-by-detection methods have shown impressive performance improvement thanks to the development of object detectors [7, 18] that provide reliable detections even in crowded scenes. The tracking-by-detection methods generally build long trajectories of objects by associating detections provided by detectors. They can be roughly categorized into batch and online methods.

Batch methods [2, 6, 14, 20, 21] usually utilize the detections of whole frames together to link fragmented trajectories (*i.e*. tracklets) due to occlusion. Given a set of detections for whole frames, short tracklets are generated by linking the detections, and the tracklets are globally associated to build longer tracklet. Thus, the global association is very crucial in this approach, and many methods [2, 6, 21] for global association have been proposed. However, the performance of the batch methods is still limited under long-term occlusion because of the difficulty in distinguishing different objects. Moreover, they usually require the detections for an entire sequence beforehand, and also require huge computation due to the iterative associations for generating globally optimized tracks. It is thus difficult to apply the batch methods to real-time applications.

On the other hand, online methods [5, 15, 16, 17, 18] can be applied to real-time applications because they sequentially build trajectories based on frame-by-frame association using online information up to the present frame. However, because it is more difficult to handle inaccurate (or even absent) detections of occluded objects in this approach, online methods tend to produce fragmented trajectories and to drift under occlusion.

In this paper, we propose a robust online multi-object tracking method in consideration of the aforementioned limitations of previous methods. The proposed method is based on (1) tracklet confidence to handle track fragments due to occlusion or unreliable detections and (2) online discriminative appearance learning to handle similar appearances of different objects in tracklet association.

To handle frequent occlusion by clutter or other objects, we first propose tracklet confidence based on the detectability and continuity of a tracklet. We then formulate the multi-object tracking problem based on the tracklet confidence and solve it by associating tracklets in different ways according to their confidence values: reliable tracklets having

high confidence are locally associated with online-provided detections, whereas fragmented tracklets having low confidence are globally associated with other tracklets and detections. Based on this strategy, tracklets can sequentially grow with online-provided detections and fragmented tracklets can be linked with others without any iterative and expensive associations.

As described above, the core steps of the proposed method are the local and global associations. In both steps, appearance modeling is crucial for associating tracklets and detections of the same object while distinguishing different objects. To this end, we also propose a novel online discriminative appearance learning taking into consideration two main issues in multi-object tracking: (1) online learning to update appearance models according to ongoing tracking results, and (2) online training sample collection for discriminating appearances of multiple tracked objects. Most previous tracking methods with online appearance learning focus on only one of these issues. [3, 5, 11] devise online learning methods, but their sample collection strategies aim at distinguishing an object from the background rather than other objects. On the other hand, [14, 20, 21] collect training samples for distinguishing different objects, but discriminative appearance models are learned in a batch manner: once training samples are collected from tracklets after low-level association, the models are simultaneously learned with all collected samples. Unlike these previous works, the proposed online learning method is designed in consideration of two issues together to learn discriminative appearance models using an incremental linear discriminant analysis (ILDA). This allows us to distinguish each object and also incrementally update learned appearance models with online tracking results. To the best of our knowledge, there has been no explicit use of ILDA to learn discriminative appearance models for multi-object tracking. By exploiting the proposed appearance learning, tracklet association can be successfully performed even under occlusion.

To sum up, the main contributions of this paper can be summarized as follows: (i) proposition of a tracklet confidence for evaluating tracklet's reliability, and two-step association using the tracklet confidence for building optimal tracklets, (ii) proposition of an online learning method for discriminating different objects and adapting learned appearances with ongoing tracking results, and (iii) proposition of a practical whole online tracking structure by effectively combining our methods, as given in Fig. 2.

2. Related Works

Some previous works related to online multi-object tracking and online appearance learning, the focus of this paper, are introduced in this section.

Given detections from a detector at each frame, an online tracking approach locally associates detections frame-by-frame to build trajectories. To associate detections frame-by-frame, [18] associates object hypotheses with detections by evaluating their affinities for appearances, positions, and sizes. [17] employs online-trained classifiers to find hypotheses of occluded objects. [5] uses a confidence density map by combining outputs of a detector and online-learned classifiers for robust association. [16] uses a part-based model to correctly associate detections under partial occlusion. However, these local association-based-tracking methods tend to produce short fragmented trajectories and to drift under occlusion because they only use the information in two consecutive frames.

On the other hand, some methods also have been proposed for discriminative appearance models for multi-object tracking. For example, pre-defined appearance models using color and other feature histograms have been proposed in [18, 19]. However, they do not deal with appearance changes of tracked objects. To update appearance models, [3, 5, 11, 17] employ target specific appearance models with online learning such as ensemble learning [3] and online boosting [11]. However, their appearance models are trained for distinguishing an object from the background, rather than from other objects. To learn appearance models for discriminating different objects, [14, 20, 21] collect positive samples from the same tracklets and negative samples from other tracklets after low-level associations, and the models are simultaneously learned using standard AdaBoost [14, 21] or MIL instance learning [20] methods. However, these learning methods are not appropriate for updating learned appearance models online because the appearance models are learned in a batch manner.

3. Online Tracking with Tracklet Confidence

If an object i appears at frame t, we denote it by using a binary function as $v^i(t) = 1$. Otherwise, $v^i(t) = 0$. When $v^i(t) = 1$, the state of the object i is represented as $\mathbf{x}_t^i = \left(\mathbf{p}_t^i, \mathbf{s}_t^i, \mathbf{v}_t^i\right)$, where $\mathbf{p}_t^i$, $\mathbf{s}_t^i$, and $\mathbf{v}_t^i$ are the position, size, and velocity, respectively. We then define a tracklet T^i of the object i as a set of states up to frame t, and denote it as $T^i = \{\mathbf{x}_k^i | v^i(k) = 1, 1 \leq t_s^i \leq k \leq t_e^i \leq t\}$, where t_s^i and t_e^i are the time stamps of the start- and end-frame of the tracklet. In addition, a set of tracklets of all objects up to frame t is denoted as $\mathbb{T}_{1:t}$. Similarly, we denote the detection of the object i at frame t as $\mathbf{z}_t^i$, and a set of all detections up to frame t as $\mathbb{Z}_{1:t}$. The online multi-object tracking problem can then be formulated to find the optimal $\mathbb{T}_{1:t}$ by maximizing the posterior probability for a given $\mathbb{Z}_{1:t}$ as

$$\hat{\mathbb{T}}_{1:t}^{\text{MAP}} = \underset{\mathbb{T}_{1:t}}{\text{argmax}}\, p\left(\mathbb{T}_{1:t} | \mathbb{Z}_{1:t}\right). \tag{1}$$

Note that directly solving Eq. (1) is not feasible in practice because the possible combinations of $\mathbb{T}_{1:t}$ and $\mathbb{Z}_{1:t}$ is innumerable. Therefore, we reformulate the problem

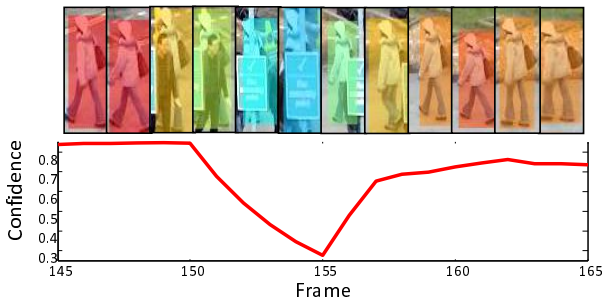


Figure 1. Tracklet confidence variation of an object (in the PETS-L1 sequence) under occlusion. Under occlusion, the confidence decreases, but it then gradually increases by association.

with the tracklet confidence and propose a practical solution based on the reformulation.

3.1. Tracklet Confidence

Tracklet confidence can be intuitively interpreted as how well the constructed tracklet matches the real trajectory of the object. We determine a reliable tracklet with high confidence based on the following requirements:

• Length: a short tracklet tends to be unreliable. A long tracklet is more likely to be a correct tracklet of an object.

• Occlusion: a severely occluded tracklet by other tracklets is not appropriate as a reliable tracklet.

• Affinity: a high affinity between a tracklet and an associated detection indicates that the tracklet is reliable.

The tracklet confidence $conf\left(T^i\right)$ can then be modeled based on the above requirements as

$$conf\left(T^i\right) \models \left(\frac{1}{L}\sum\nolimits_{k\in[t_s^i,t_e^i],v^i(k)=1}\Lambda\left(T^i,\mathbf{z}_k^i\right)\right)\times \max\left(\left(1+\beta\cdot\log\left((L-w)/L\right)\right),\ 0\right), \tag{2}$$

where L is the cardinality of T^i (*i.e.* the length of a tracklet), given as $L=|T^i|$, and w is the number of frames in which the object i is missing due to occlusion by other objects or unreliable detection, and is given as $w=t_e^i-t_s^i+1-L$. The first term in Eq. (2) is the average affinity score between the tracklet and associated observations (*i.e.* detection): a high affinity score increases the confidence. Here, the affinity can be defined by using several cues. We define the affinity in Sec. 5.1. The second term in Eq. (2) is also computed with L and w together, and decreases for short or heavily occluded tracklets. β is a control parameter relying on the performance of a detector. When a detector shows high accuracy, β should be set to a large value. The first and the second terms are all closely related to the detectability and the continuity of a tracklet. Figure 1 shows the confidence variation of an object under occlusion.

3.2. Formulation with Tracklet Confidence

To effectively solve the online multi-object tracking problem, we reformulate the online multi-object problem Eq. (1) by using the tracklet confidence as

$$\hat{\mathbb{T}}_{1:t}^{\text{MAP}} = \underset{\mathbb{T}_{1:t}}{\text{argmax}} \iint \underbrace{p\left(\mathbb{T}_{1:t}^{(lo)}|\mathbb{T}_{1:t}^{(hi)},\mathbb{Z}_{1:t}\right)}_{Global\ phase}\underbrace{p\left(\mathbb{T}_{1:t}^{(hi)}|\mathbb{Z}_{1:t}\right)}_{Local\ phase} d\mathbb{T}_{1:t}^{(hi)}d\mathbb{T}_{1:t}^{(lo)}. \tag{3}$$

Here, $\mathbb{T}_{1:t}^{(hi)}$ and $\mathbb{T}_{1:t}^{(lo)}$ represent a set of tracklets with high confidence and a set of tracklets with low confidence, respectively. As shown in Eq. (3), the problem is solved in two phases: tracklets with high confidence are locally associated with online-provided detections, while tracklets with low confidence, which are more likely to be fragmented, are globally associated with other tracklets and detections. To be more concrete, the tracklets with high confidence are first considered to be locally associated with detections because more reliable detections originate from them rather than from tracklets with low confidence. The local association between the tracklets and detections allows us to progressively grow locally optimal tracklets with online provided-detections. The object being tracked, however, is frequently not detected due to occlusion or unreliable detectors in complex scenes. When a detection of the object is not available, the confidence of a tracklet is decreased. Therefore, we can consider the tracklets with low confidence as fragmented tracklets, and globally associate them with other tracklets and detections to link them. The overall framework of the proposed method is shown in Fig. 2. Here, since the tracklet confidence lies in $[0,1]$, we consider a tracklet as a reliable tracklet with high confidence when $conf(T^i)>0.5$; otherwise it is considered as the fragmented tracklet with low-confidence. In our experiment, the tracking performance, however, is not significantly affected by this threshold.

3.3. Local Association of Tracklets

In the local association, tracklets with high confidence, $T^{i(hi)}$, are sequentially grown with a set of detections at frame t, $\mathbb{Z}_t$. Pairwise association is performed to associate detection responses with tracklets. When h tracklets with high confidence and n detections exist at frame t, a score matrix $S_{h\times n}$ can be defined as

$$S=[s_{ij}]_{h\times n},\ s_{ij}=-\log\left(\Lambda(T^{i(hi)},\mathbf{z}_t^j)\right),\ \mathbf{z}_t^j\in\mathbb{Z}_t, \tag{4}$$

where the affinity $\Lambda(T^{i(hi)},\mathbf{z}_t^j)$ is computed by Eq. (9). We then determine tracklet-detection pairs using the Hungarian algorithm [1] such that the total affinity in $S_{h\times n}$ is maximized. When the association cost of a tracklet-detection pair is less than a pre-defined threshold, $-\log(\theta)$, $\mathbf{z}_t^j$ is associated with $T^{i(hi)}$. For the tracklet $T^{i(hi)}$ having associated detection $\mathbf{z}_t^j$, the following procedure is performed:

(i) The position and velocity of a tracklet are updated with the associated $\mathbf{z}_t^j$. The size of the object is also updated

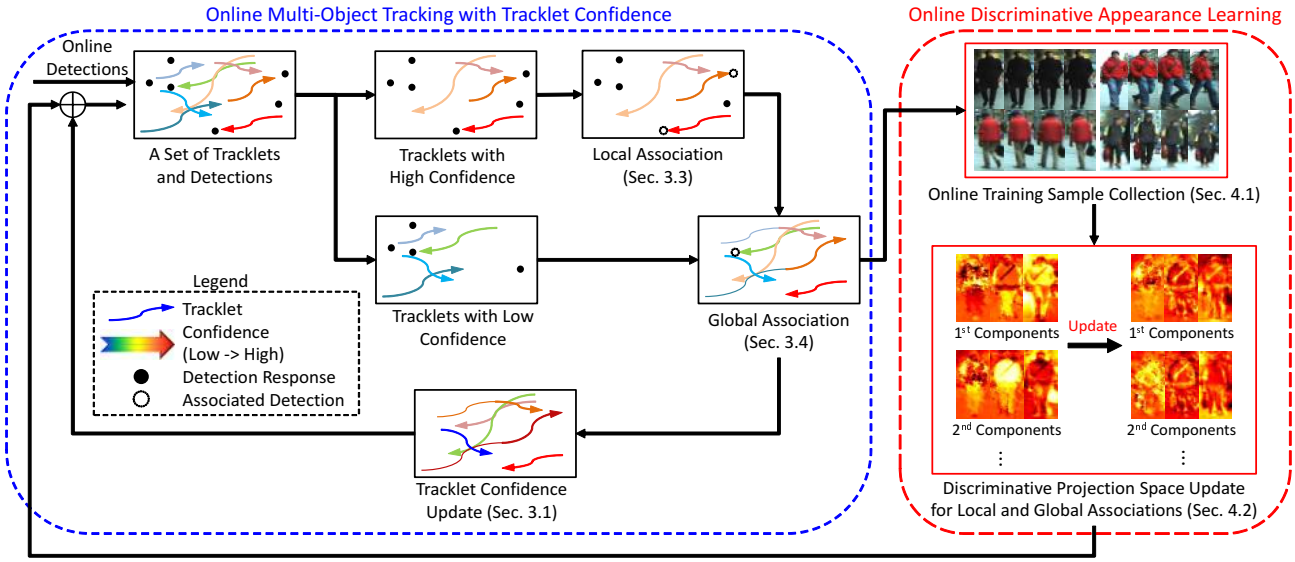


Figure 2. Proposed framework for robust online multi-object tracking. Colors of tracklets indicate their confidence values.

by averaging the associated detections of recent past frames.

(ii) $conf(T^{i(hi)})$ is updated using $\mathbf{z}_t^j$ by Eq. (2).

Here, it is possible to skip this local association and try to solve the problem via only the global association which is described in the next section. However, in this case, much more computation is required and the performance is also degraded. This is because the local association greatly reduces the ambiguity in the global association as well as the computation cost. This is proven in Sec. 5.

3.4. Global Association of Tracklets

In the global association, tracklets with low confidence $T^{i(lo)}$, which are more likely to be fragmented, are globally associated with other tracklets and detections. Suppose that there exist h and l tracklets with high and low confidence, respectively. Since association events are mutually exclusive, we only consider n detections, $\mathbb{Y}_t = \{\mathbf{y}_t^j\}_{j=1}^n \subseteq \mathbb{Z}_t$ in associating $T^{i(lo)}$, where $\mathbb{Y}_t$ is a set of detections not associated with any $T^{i(hi)}$ in the local association. The following association events are then considered:

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

The cost matrix for all events is defined as follows:

$$G_{(l+n)\times(h+l)} = \begin{bmatrix} A_{l\times h} & B_{l\times l} \\ -\log(\theta)_{n\times h} & C_{n\times l} \end{bmatrix}, \quad (5)$$

Here, $A = [a_{ij}]$ represents event A, where $a_{ij} = -\log(\Lambda(T^{i(lo)}, T^{j(hi)}))$ is the association cost calculated by the affinity between them using Eq. (9). $B = \text{diag}\,[b_1, ..., b_l]$ models event B, where $b_i = -\log(1 - conf(T^{i(lo)}))$ is the cost to terminate $T^{i(lo)}$, and $C = [c_{ij}]$ represents event C, where $c_{ij} = -\log(\Lambda(T^{i(lo)}, \mathbf{y}_t^j))$ is the association cost obtained using Eq. (9). The same threshold θ used in the local association is also employed to select a reliable association pair having a high affinity score. Once the cost matrix is computed, the optimal association pairs, which minimize the global association cost in G, are determined using the Hungarian algorithm [1], and the tracklets and their confidence values are updated with the results.

4. Discriminative Appearance Learning

As mentioned, the appearance modeling is very important in both the local and global association for associating tracklets and detections of the same object while distinguishing different objects. In this section, we present our approach, which considers the two main issues described in Sec. 1, to learn discriminative appearance models. In the proposed learning method, online training samples are collected from tracked objects, and a discriminative projection space is updated with the collected samples using ILDA. By projecting the appearance models of tracklets into the discriminative projection space, we make the appearances of tracklets more discriminative.

It is emphasized that, as an online learning method to update discriminative appearances with new samples, we use the ILDA method, while other methods [5, 17] employ the online boosting method [11]. The main reason for using ILDA is that appearances of multiple objects can be distinguished with a single updated LDA projection matrix, whereas specific classifiers of the objects [5, 17] are required in the boosting method. We can thereby significantly reduce the computational complexity in appearance learning. A further benefit of using ILDA lies in its ability

Figure 3. Training samples from the trackets with high confidence (red) and low confidence (blue).

to memorize the discriminative information for a long time. This makes it possible to accurately identify objects even under significant pose and appearance changes and long-term occlusion.

4.1. Training Sample Collection

At each frame, we collect N image patches with different locations and scales around the refined locations of tracklets for discriminating different tracklets. Since images patches from the tracklets with low confidence are likely to be polluted by occlusion as shown in Fig. 3, we only extract image patches from the tracklets with high confidence.

For each image patch (*i.e.*, sample), we create a low-level feature $\mathbf{f}_l$ by concatenating templates extracted from HSV color channel images. A set $\mathbb{B} = \{(\mathbf{f}_l, y_l)\}_{l=1}^{N}$ consisting of features $\mathbf{f}_l$ and the labels of tracklets (*i.e.* ID) y_l is then constructed. In our experiments, the patch is resized to 96x32, and the dimension of the feature is 9216. Since the feature dimension is very high, directly exploiting the high-dimensional feature to distinguish each object is not effective. Therefore, we project the high-dimensional feature onto a low-dimensional subspace using ILDA.

4.2. Online-Learning Algorithm

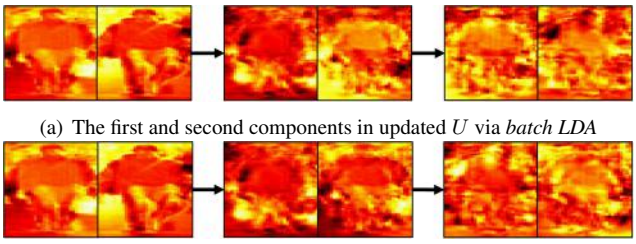
In the batch LDA, a projection matrix U is constructed by maximizing class separability of the given training set

$$\hat{U} = \underset{U}{\operatorname{argmax}} \frac{\left|U^{\mathrm{T}} S_B U\right|}{\left|U^{\mathrm{T}} S_T U\right|}, \tag{6}$$

where the between-class scatter matrix S_B and the total scatter matrix S_T are calculated by

$$\begin{aligned} S_B &= \textstyle\sum_{i=1}^{C} n_i \left(\mathbf{m}_i - \mu\right)\left(\mathbf{m}_i - \mu\right)^{\mathrm{T}}, \\ S_T &= \textstyle\sum_{l} \left(\mathbf{f}_l - \mu\right)\left(\mathbf{f}_l - \mu\right)^{\mathrm{T}}, \end{aligned} \tag{7}$$

where C is the total number of classes (*i.e.* number of tracklets), n_i is the sample number of class (*i.e.* tracklet) i, $\mathbf{m}_i$ is the mean feature vector of class i, and μ is the global mean feature. The problem defined by Eq. (6) can be solved by computing an eigenvector matrix of $S_T^{-1} S_B$. However, it is necessary to incrementally update the existing projection matrix with updated samples because not all training samples are available in online multi-object tracking. Although a number of ILDA methods have been proposed, we employ



(a) The first and second components in updated U via *batch LDA*

(b) The first and second components in updated U via *incremental LDA*

Figure 4. Updated bases on ETHMS-Bahnhof: Incrementally learned bases are almost identical to those of batch LDA.

the ILDA method using sufficient spanning sets developed by [13] due to the basis of its effectiveness. The procedure is as follows.[1] Given eigenspace models of S_T and S_B, $\{\mu_d, M_d, P_d, \Sigma_d, \}_{d=1,2}$ and $\{\mu_d, M_d, Q_d, \Delta_d\}_{d=1,2}$, where μ_d and M_d are the mean vector and the total number of samples in dataset d, $P_d(Q_d)$ and $\Sigma_d(\Delta_d)$ are eigenvector and eigenvalue matrices of $S_{T,d}(S_{B,d})$. The combined scatter matrices $S_{T,3} \backsimeq P_3 \Sigma_3 P_3^T$ and $S_{B,3} \backsimeq Q_3 \Delta_3 Q_3^T$, and their eigenspace models can then be obtained using the algorithm [12]. However, using sufficient spanning sets Φ and Ψ computed by QR decomposition[2] for eigenvector matrices P_3 and Q_3, the eigenproblems of the combined scatterer matrices can be efficiently solved as

$$\begin{aligned} S_{T,3} &= P_3 \Sigma_3 P_3^T \Rightarrow \Phi^T S_{T,3} \Phi = R_T \Sigma_3 R_T^{\mathrm{T}} \\ S_{B,3} &= Q_3 \Delta_3 Q_3^T \Rightarrow \Psi^T S_{B,3} \Psi = R_B \Delta_3 R_B^T. \end{aligned} \tag{8}$$

By computing the eigendecomposition of the r.h.s., eigenvalues, Σ_3 and Δ_3, and eigenvectors, R_T and R_B, are yielded. After removing non-significant components in R_T and R_B according to Σ_3 and Δ_3, minimal sufficient spanning sets of the combined eigenvectors are obtained as $P_3 = \Phi R_T$ and $Q_3 = \Psi R_B$. The detailed pseudo code of the ILDA algorithm is given in the supplementary material. Figure 4 shows the updated projection matrices using batch LDA and ILDA, proving the accuracy of ILDA.

To verify the effectiveness of the ILDA method, we compare the performance of ILDA and online boosting methods with the ETHMS dataset [10] as shown in Fig. 5[3]. We can see that the ILDA method is much more effective than the boosting method in terms of computation cost and identification accuracy. In our evaluation, the testing time is similar, but the training of ILDA is much faster, because C classifiers for C objects are trained in boosting as done in [5, 17], while only one LDA matrix is updated in ILDA.

[1] More detailed description can be found in the supplementary material.

[2] Φ and Ψ are orthonormal matrices spanning the combined scatter matrices, *e.g.* $P_3 = \Phi R = h([P_1, P_2, \mu_1 - \mu_2])R$, where h is an orthonormalization function and R is a rotation matrix.

[3] We use a uniqueness score $(r_i / r_{close}, i \neq close)$ where r_i is the affinity score between the object i current appearance and its online-learned appearance, and r_{close} is the highest score among affinity scores between current appearances of other objects and the learned appearance of the object i. A higher uniqueness score reflects better performance.

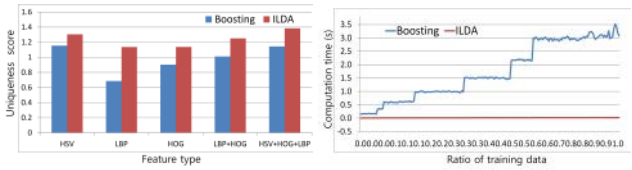


Figure 5. Performance comparison between incremental LDA and online boosting methods.

5. Experiments

5.1. Implementation

We have implemented the proposed online tracking system using MATLAB. Our code is available at https://cvl.gist.ac.kr/.

Affinity model: Although a tracklet T^i can be described with several cues in different ways, we describe T^i with three elements $\{A^i, S^i, M^i\}$, where A^i, S^i and M^i represent appearance, shape, and motion models, respectively. An affinity measure to determine how well two tracklets (or a tracklet and a detection) are matched is then defined as

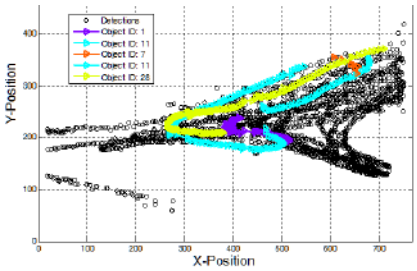
$$\Lambda\left(X,Y\right) = \Lambda^A\left(X,Y\right)\Lambda^S\left(X,Y\right)\Lambda^M\left(X,Y\right), \quad (9)$$

where X and Y can be tracklets or detections. The affinity score is computed based on affinities of appearance, shape, and motion models as follows:

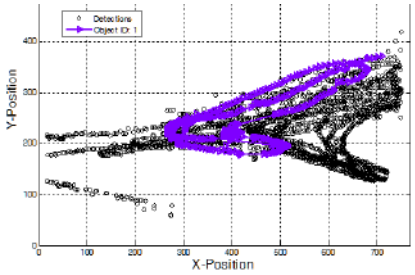
$$\begin{aligned}\Lambda^A\left(X,Y\right) &= \frac{\left(U^T f(X)\cdot U^T f(Y)\right)}{\|U^T f(X)\|\|U^T f(Y)\|},\\ \Lambda^S\left(X,Y\right) &= \exp\left(-\left\{\frac{h_X-h_Y}{h_X+h_Y}+\frac{w_X-w_Y}{w_X+w_Y}\right\}\right), \quad (10)\\ \Lambda^M\left(X,Y\right) &= \mathcal{N}\left(\mathbf{p}_X^{tail}+\mathbf{v}_X^F\Theta;\mathbf{p}_Y^{head},O^F\right)\\ &\quad\times\mathcal{N}\left(\mathbf{p}_Y^{head}+\mathbf{v}_Y^B\Theta;\mathbf{p}_X^{tail},O^B\right).\end{aligned}$$

For appearance models, $f(X)$ and $f(Y)$, the appearance affinity $\Lambda^A\left(X,Y\right)$ is evaluated on the online-learned discriminative projection matrix U using ILDA as described in Sec. 4. The shape affinity $\Lambda^S\left(X,Y\right)$ is calculated with the height h and width w of objects. $\Lambda^M\left(X,Y\right)$ is the motion affinity between X tail (*i.e.* the last refined position) and Y head (*i.e.* the first refined position) with the frame gap Θ. The forward velocity $\mathbf{v}_X^F$ is evaluated from the head to the tail of X, while the backward velocity $\mathbf{v}_Y^B$ is evaluated from the tail to the head of Y. The difference between the predicted position computed with the velocity and the refined position is assumed to follow a Gaussian distribution. Note that only the forward motion is used when evaluating affinity between a tracklet and a detection.

Dataset and detection: For the performance evaluation, we use the following datasets: CAVIAR [8], VS-PETS 2009 (PETS) [9], and ETH Mobile scene (ETHMS) [10]. Although the CAVIAR dataset consists of 26 video sequences, only 20 sequences were used, as listed in [14], to ensure a fair comparison with other methods. In the PETS dataset, tracking sequences S2.L1 and S2.L2 were used. In the ETHMS dataset, the SUNNY DAY and BAHNHOF sequences of street scenes taken by a moving camera



(a) Fragmented trajectories using system (p2)



(b) A single long trajectory using system (p4)

Figure 6. Detections (black circles) and estimated trajectories of a single object (color lines) for PETS-L1 sequence over 520 frames.

were selected. For VS-PETS 2009, ETHMS, and CAVIAR datasets, we used the public available detections given by [2], [14], and [21], respectively.

System parameters: All parameters have been found experimentally, and remained unchanged for all datasets. From an extensive evaluation, we find that most parameters do not affect the overall performance of our system much. In the affinity model in Eq. (10), all parameters (*i.e.* positions, sizes and velocities) are automatically determined by tracking results except for O^F and O^B, which were set to $\text{diag}[30^2\ 75^2]$. The same threshold $\theta = 0.4$ is used for the local and global association.

5.2. Performance Evaluation

Evaluation metrics: We use the common CLEAR MOT [4] consisting of multiple metrics. The multiple object tracking precision (MOTP ↑) evaluates the intersection area over the union area of bounding boxes. The multiple object tracking accuracy (MOTA ↑) calculates the accuracy composed of false negatives (F.Neg. ↓), false positives (F.Pos. ↓), and identity switching (IDS ↓). In addition, the metrics used in [14, 21] are computed: the number of trajectories in the ground truth (GT), the ratio of mostly tracked trajectories (MT ↑), the ratio of mostly lost trajectories (ML ↓), the ratio of partially tracked trajectories (PT), *i.e.*, $1 - PT - ML$, and the number of track fragments (FG ↓). Here, ↑ represents that higher scores indicate better results, and ↓ denotes that lower scores indicate better results.

Evaluation: In Table. 1, a quantitative comparison between our systems and other tracking systems is given. The implemented systems (p1)-(p4) are described as follows:

(a) Tracking results without online learned appearance models

(b) Tracking results with online learned appearance models

Figure 7. Tracking results: IDs (7,14) are swapped by occlusions in the top rows, but IDs (7,9) are correctly kept in the bottom rows.

(p1) System *without online-learned appearance models*[4];
(p2) System *without global association*;
(p3) System *without local association*;
(p4) System *with all proposed methods*.

From the evaluation results of (p1)-(p4), we can see the effect of each part in the proposed method. As expected, the proposed system (p4) improves performance for most metrics, but our other systems (p1)-(p3) are still comparable with other systems. In particular, our system (p4) noticeably reduces the ML and FG rates and increases MT rate against the system (p2). Figure 6 supports this analysis: with the global association, longer trajectories are built by linking fragmented trajectories. Furthermore, by comparing our systems (p1) and (p4), we can see that the proposed learning method allows us reduction of IDS and FG. Figure 7 shows that IDs of objects are correctly maintained using our learning method. More results are shown in Fig. 8.

Overall, our system achieves better performance in terms of MOTP and MOTA. Although we could not find the MOTP and MOTA for the ETHMS and CAVIAR datasets in other studies, other metrics show the robustness of our system. When compared to other online tracking systems [5, 15, 18, 19], our system (p4) provides far superior performance. Compared to batch tracking systems [2, 14, 20, 21], our system (p4) is still competitive. Notably, this improvement is achieved without using future frames and without employing multiple (color, shape, and/or texture) features (*c.f.* [5, 14, 20, 21]). In addition, we achieve the best performance in terms of MT and ML. This implies that our system can robustly construct trajectories under challenging conditions. Our system, however, produces slightly more IDS and FG than other batch systems [14, 21] in return.

Speed: Our system was implemented on a PC with a 3.07 GHZ CPU without parallel programming. The run-time relies on the number of detections. For less crowded (PETS-L1, CAVIAR) and crowded (ETHMS) scenes, the run-times are about 0.20 and 0.45 (sec/frame), excluding detection costs, and appearance learning is the most expensive part, accounting for 30% of the total computation. Here, we can reduce the run-time by about 0.05 (sec/frame) on average by performing the global association every 10 frames (the performance was degraded in return), which is performed every frame in our implementation.

[4] As an appearance model, we used the HSV color histogram with 192 bins, and computed appearance affinity using the Bhattacharyya distance.

6. Conclusion

We have proposed a robust online multi-object tracking method based on tracklet confidence and the online discriminative appearance learning. We build optimal tracklets by sequentially linking tracklets and detections using the proposed local and global association according to their confidence. Furthermore, the proposed online appearance learning allows us to discriminate multiple objects in both associations even in complex sequences. Extensive experimental results compared with those of state-of-the-art systems verity the effectiveness and robustness of our method.

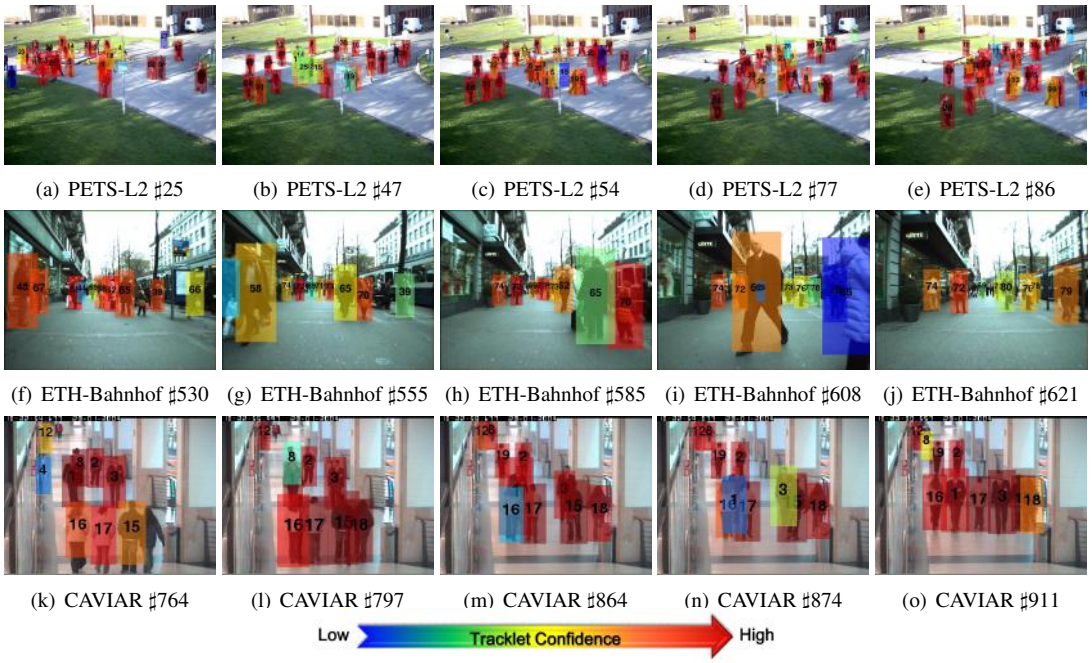
Acknowledgements: This work was supported by the Basic Science Research (2012R1A1A1010871) and Global Frontier R&D (2013M3A6A3075453) programs funded by the National Research Foundation of Korea (NRF).

References

[1] R. Ahuja, T. Magnanti, and J. Orlin. Network Flows. *Prentice Hall*, 1993. 3, 4
[2] A. Andriyenko, S. Roth, and K. Schindler. Continuous Energy Minimization for Multi-Target Tracking. *IEEE TPAMI*, vol. 35, no. 1, 2014. 1, 6, 7, 8
[3] S. Avidan. Ensemble tracking. In *CVPR*, pages 494–501, 2005. 2
[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP J. Image and Video Processing*, 2008. 6
[5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. J. V. Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE TPAMI*, 33(9):1820–1833, 2011. 1, 2, 4, 5, 7, 8
[6] W. Brendel, M. Amer, and S. Todorovic. Multiobject tracking as maximum weight independent set. In *CVPR*, pages 1273–1280, 2011. 1
[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 1
[8] CAVIAR dataset. http://homepages.inf.ed.ac.uk/rbf/caviardata1/. 6
[9] VS-PETS dataset. http://www.cvg.rdg.ac.uk/pets2009/index.html. 6
[10] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008. 5, 6
[11] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR*, pages 260–267, 2006. 2, 4
[12] P. M. Hall, A. D. Marshall, and R. R. Martin. Merging and Splitting Eigenspace Models. *IEEE TPAMI*, 22(9):1042–1049, 2000. 5
[13] T.-K. Kim, B. Stenger, J. Kittler, and R. Cipolla. Incremental linear discriminant analysis using sufficient spanning sets and its applications. *IJCV*, 91(2):216–232, 2011. 5
[14] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *CVPR*, pages 1217–1224, 2011. 1, 2, 6, 7, 8
[15] F. Poiesi, R. Mazzon, and A. Cavallaro. Multi-target tracking on confidence maps: An application to people tracking. *CVIU*, 117(10):1257–1272, 2013. 1, 7, 8

Table 1. Performance comparison. For each dataset, online tracking systems are recorded with blue color, and the best performance is marked with red color. In addition, systems evaluated with the same detections used in our system are marked with an asterisk ⋆.

Dataset (Sequence)	Method	MOTP	MOTA	F. Neg	F. Pos	IDS	GT	MT	PT	ML	FG
PETS (S2.L1)	Conf. Map [5]	56.30%	79.70%	—	—	—	—	—	—	—	—
	⋆Energy Min. [2]	—	80.20%	—	—	11	23	91.30%	4.35%	4.35%	6
	⋆PRIMPT [14]	—	—	—	—	1	19	78.90%	21.10%	0.00%	23
	OLMOAP [20]	—	—	—	—	0	19	89.50%	10.50%	0.00%	9
	proposed (p1 - w/o online learning)	69.39%	78.19%	1.32%	20.30%	16	23	100%	0.00%	0.00%	10
	proposed (p2 - w/o global assoc.)	69.01%	77.38%	2.18%	20.41%	10	23	100%	0.00%	0.00%	12
	proposed (p3 - w/o local assoc.)	66.39%	78.49%	1.37%	19.99%	6	23	100%	0.00%	0.00%	4
	proposed (p4 - with all)	69.59%	83.04%	1.19%	19.41%	4	23	100%	0.00%	0.00%	4
PETS (S2.L2)	⋆Energy Min. [2]	—	59.40%	—	—	99	74	37.84%	45.95%	16.22%	73
	Conf. Map [5]	51.30%	50.00%	—	—	—	—	—	—	—	—
	proposed (p1 - w/o online learning)	54.52%	67.79%	17.89%	13.77%	46	74	66.22%	32.43%	1.35%	53
	proposed (p2 - w/o global assoc.)	55.96%	67.56%	17.32%	14.59%	42	74	63.51%	35.14%	1.35%	49
	proposed (p3 - w/o local assoc.)	55.92%	69.58%	14.02%	15.72%	57	74	70.27%	28.38%	1.35%	48
	proposed (p4 - with all)	53.92%	70.12%	14.99%	14.35%	45	74	71.62%	27.03%	1.35%	52
ETHMS (Sunny& Bahnhof)	MT-TBD [15]	—	–	—	—	45	125	62.40%	29.60%	8.00%	69
	⋆PRIMPT [14]	—	—	—	—	11	125	58.40%	33.60%	8.00%	23
	⋆Online CRF [21]	—	—	—	—	11	125	68.00%	24.80%	7.20%	19
	proposed (p1 - w/o online learning)	59.66%	70.40%	25.40%	3.77%	44	126	65.87%	27.78%	6.35%	50
	proposed (p2 - w/o global assoc.)	60.04%	67.90%	28.34%	3.53%	23	126	68.25%	27.78%	3.97%	57
	proposed (p3 - w/o local assoc.)	57.47%	72.46%	22.79%	4.29%	48	126	71.43%	25.40%	3.17%	44
	proposed (p4 - with all)	64.01%	72.03%	23.64%	4.16%	18	126	73.81%	23.81%	2.38%	38
CAVIAR	Two-steps [19]	—	—	—	—	14	140	84.29%	12.14%	3.57%	24
	PRIMPT [14]	—	—	—	—	4	143	86.00%	13.30%	0.70%	17
	OLMOAP [20]	—	—	—	—	5	143	89.10%	10.20%	0.70%	11
	Part-based tracking [18]	—	—	—	—	17	140	75.70%	17.90%	6.40%	35
	proposed (p1-w/o online learning)	85.26%	81.37%	17.45%	1.23%	25	143	85.34%	13.61%	1.05%	28
	proposed (p2 - w/o global assoc.)	84.36%	80.15%	18.26%	1.59%	21	143	77.49%	14.66%	7.85%	33
	proposed (p3 - w/o local assoc.)	86.75%	83.43%	14.92%	1.20%	24	143	87.96%	12.04%	0.00%	18
	proposed (p4 - with all)	87.15%	86.52%	11.38%	1.00%	9	143	89.53%	10.47%	0.00%	8



(a) PETS-L2 ♯25 (b) PETS-L2 ♯47 (c) PETS-L2 ♯54 (d) PETS-L2 ♯77 (e) PETS-L2 ♯86

(f) ETH-Bahnhof ♯530 (g) ETH-Bahnhof ♯555 (h) ETH-Bahnhof ♯585 (i) ETH-Bahnhof ♯608 (j) ETH-Bahnhof ♯621

(k) CAVIAR ♯764 (l) CAVIAR ♯797 (m) CAVIAR ♯864 (n) CAVIAR ♯874 (o) CAVIAR ♯911

Figure 8. Tracking results for the PETS, ETHMS, and CAVIAR datasets. At each frame, tracklets with different confidence are illustrated with different color boxes. The identities of tracked objects are marked in black. (Refer to the supplementary material for more results.)

[16] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, pages 1815–1821, 2012. 1, 2

[17] X. Song, J. Cui, H. Zha, and H. Zhao. Vision-based multiple interacting targets tracking via on-line supervised learning. In *ECCV*, pages 642–655, 2008. 1, 2, 4, 5

[18] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247–266, 2007. 1, 2, 7, 8

[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 *CVPR*, pages 1200–1207, 2009. 2, 7, 8

[20] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *CVPR*, pages 1918–1925, 2012. 1, 2, 7, 8

[21] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *CVPR*, pages 2034–2041, 2012. 1, 2, 6, 7, 8