Multi-object tracking via tracklet confidence-aided relative motion analysis

Han-Mu Park
Se-Hoon Park
Kuk-Jin Yoon
Multi-object tracking via tracklet confidence-aided relative motion analysis

Han-Mu Park, Se-Hoon Park, and Kuk-Jin Yoon*
Gwangju Institute of Science and Technology, Computer Vision Lab, Gwangju, Republic of Korea

Abstract. Applications for tracking multiple objects in an image sequence are frequently challenged by various uncertainties, such as occlusion, misdetection, and abrupt camera motion. In practical environments, these uncertainties may occur simultaneously and without pattern so that they must be jointly considered to achieve reliable tracking. We propose a two-step online multi-object tracking framework that incorporates a confidence-aided relative motion network (RMN) to jointly consider various difficulties. Because of the framework’s two-step data association process and the similarity function using RMNs, the proposed method achieves robust performance in the presence of most kinds of uncertainties. In our experiments, the proposed method exhibits a very robust and efficient performance compared with other state-of-the-art algorithms. © 2017 SPIE and IS&T [DOI: 10.1117/1.JEI.26.5.050501]

Keywords: tracking; multiple object tracking; relative motion; tracklet confidence.

Paper 170403L received May 18, 2017; accepted for publication Sep. 11, 2017; published online Oct. 4, 2017.

1 Introduction

Tracking multiple objects in an image sequence is important for numerous computer vision applications.1,2 Due to advances in object detectors, tracking-by-detection approaches have recently received extensive attention.3,4 The approaches are generally formulated as problems of finding correspondences between trajectories of multiple objects and newly detected object candidates. Detection-based approaches have shown impressive performance, providing long and accurate trajectories of objects. However, the approaches still have challenging issues, because detection algorithms can fail to find objects for various reasons, such as occlusion and clutter as shown in Fig. 1. The detection failures negatively impact performance, particularly in the case of online tracking algorithms,1-6 which construct trajectories using only information up to a current frame.

Bae and Yoon3 proposed a confidence-based, online multi-object tracking (MOT) algorithm using a two-step association framework. First, the framework associates high-confidence trajectories and detection candidates. Then, the trajectories that are not associated in the first step are connected to the confident trajectories based on its confidence values. By preferentially connecting the relatively high-confidence trajectories, the algorithm can construct long and reliable object trajectories. However, the algorithm has relatively weak performance in the presence of severe occlusion and abrupt camera motions, because of limitations of its similarity function and inaccuracy in detection.

In this letter, we propose a unified MOT framework to address these kinds of uncertainties in detection-based algorithms. First, we propose an improved two-step framework that associates trajectories with candidates based on tracklet confidence and furthermore augments detection candidates by using single-object tracking. By adopting the detection augmentation step, the proposed framework improves the accuracy and reliability of trajectories in the presence of occlusion and misdetection. Second, we propose a similarity measure that considers relative motions among objects. By incorporating the relative motion concept into the similarity function, the proposed algorithm can track objects using trajectories of neighboring objects even when they are occluded.

2 Problem Formulation

2.1 Multi-Object Tracking Problem

Suppose an object \( x_i \) exists at frame \( t \) of a given image sequence, where \( i \) is the object’s index. The state of \( x_i \) can be defined as \([u_i^t, v_i^t, \dot{u}_i^t, \dot{v}_i^t, w_i^t, h_i^t]^{\top}\), where \((u_i^t, v_i^t)\) and \((\dot{u}_i^t, \dot{v}_i^t)\) represent the position and velocity of object \( i \), respectively, the values \( w_i^t \) and \( h_i^t \) denote width and height of the bounding box of object \( i \), respectively. Then, an object trajectory (tracklet) \( T_i \) up to frame \( t \) can be defined as follows:

\[
T_i = \{x_i^k | 1 \leq k \leq t, t_i \leq t \leq t_i + 1\}, \quad T_i \in T, \tag{1}
\]

where \( o_i^k \) is a binary value that indicates whether the object \( i \) appears in the frame \( k \). Also, \( t_i \) and \( t_i \) represent the start and end frame of the tracklet \( T_i \), respectively. Similarly, a detected unknown object at frame \( t \) can be denoted by \( x_i^t \in Z_t \). We can then formulate the MOT problem as follows:

\[
\hat{T}_i = \arg \max_{T_i} P(T_i|Z_t). \tag{2}
\]

As stated in Bae and Yoon’s paper,3 the problem is difficult to solve in practice because of the enormous number of possible correspondence combinations between \( T_i \) and \( Z_t \). We will address this issue by reducing the possible solution space based on tracklet confidence.

2.2 Problem Formulation with Tracklet Confidence

To measure the reliability of potential tracklets, we adopt the confidence measure that was proposed by tracklet confidence and online discriminative appearance learning (TC_ODAL).3 The measure evaluates each tracklet in three aspects: length, occlusion, and affinity. First, the confidence in a tracklet is likely to be proportional to its length because a long trajectory means that the object was reliably traced. Second, a tracklet containing many occluded frames is likely to be unreliable, because occluded frames give noisy information about the object. Third, a tracklet that has high affinity with associated detection candidates is likely to be reliable. Based on the observations, a tracklet confidence measure \( \text{conf}(T_i) \) is defined as follows:

\[1017-9909/2017/$25.00 \text{© } 2017 \text{ SPIE and IS&T]}

*Address all correspondence to: Kuk-Jin Yoon, E-mail: kjyoon@gist.ac.kr
Confidence update

\[
\text{conf}(T_i) \triangleq \max \left( 1 + \beta \cdot \log \left( \frac{L - w}{L} \right), 0 \right) \cdot \left( \frac{1}{L} \sum_{k \in [i']_{d(k)}^n} \Lambda(T_i, z^k_t) \right),
\]

where \( L \) is the length of the associated tracklet, \( w \) is the number of occluded frames, and \( \Lambda(T_i, z^k_t) \) is the similarity function for tracklets and detections. The first term of Eq. (3) takes into consideration the effects from occlusion and the length of the tracklet. The logarithmic term greatly reduces tracklet confidence as the number of occluded frames increases, whereas its inner term, \((L - w)/L\), denotes the ratio of non-occluded frames to the total number of frames in the tracklet. The second term indicates the affinity between the tracklet and detections, which is defined by averaging their similarity values. Based on the measure, we can reformulate the MOT problem as follows:

\[
\hat{T}_i = \arg \max_{T_i} \int \int P(T_i | T_i^{(h)}, T_i^{(l)}) P(T_i^{(h)} | T_i^{(h)}, Z_t) P(T_i^{(h)} | Z_t) dT_i^{(h)} dT_i^{(l)},
\]

where \( T_i^{(h)} \) and \( T_i^{(l)} \) are the sets of tracklets that have high and low confidences, respectively. The first term indicates the probability of generating \( T_i^{(h)} \) from \( T_i^{(h)} \) and \( T_i^{(l)} \). The second term indicates the probability of generating \( T_i^{(l)} \), which is a subordinate to \( P(T_i^{(h)} | Z_t) \). The last term indicates the probability of generating \( T_i^{(h)} \), which can be obtained by computing the similarity between tracklets and detections.

3 Proposed Multi-Object Tracking Framework

To solve the reformulated MOT problem, we propose a two-step association scheme by generalizing TC_ODAL. Our framework consists of a local and a global association process as shown in Fig. 2. During the local association step, the probability \( P(T_i^{(h)} | Z_t) \) is estimated by computing the similarity between detections and high-confidence tracklets. Then, low-confidence tracklets are associated with high-confidence tracklets to connect fragmented tracklets during the global association step.

In contrast with the framework of TC_ODAL, we additionally incorporate a single-object tracking step using a correlation filter and a similarity measure based on a relative motion network (RMN). By incorporating these steps, the proposed framework can handle uncertainties caused by several factors, such as misdetection or abrupt camera motion.

3.1 Local Association Step

In this step, detections are associated with high-confidence tracklets. This step models the third probability \( P(T_i^{(h)} | Z_t) \) of Eq. (4). Suppose the number of high-confidence tracklets is \( h \) and the number of detections is \( n \), then a matching score matrix \( S \), which describes the similarity between tracklets and detections, can be defined as follows:

\[
S = [s_{ij}]_{h \times n}, \quad s_{ij} = -\log(\Lambda(T_i^{(h)}, z^k_t)), \quad T_i^{(h)} \in T_i^{(h)}, \quad z^k_t \in Z_t,
\]

where \( \Lambda(T_i^{(h)}, z^k_t) \) is a similarity function based on the concept of an RMN as defined in Sec. 3.4. After computing \( S \), the Hungarian algorithm is applied to maximize the total affinity of the score matrix. Finally, the confidence of each tracklet is updated by using Eq. (3).

3.2 Global Association Step

In the global association step, low-confidence tracklets are connected with high-confidence tracklets or unmatched detections. This step models the first and second probability values, \( P(T_i^{(h)} | T_i^{(h)}, T_i^{(l)}) \) and \( P(T_i^{(l)} | T_i^{(h)}, Z_t) \), of Eq. (4). The threefold processing of low-confidence tracklets is defined. First, low-confidence tracklets are associated with high-confidence tracklets. In this case, an element of the score matrix \( A_{h \times l} = [a_{ij}] \) is defined as \( a_{ij} = -\log(\Lambda(T_i^{(l)}, y_i)) \), where \( y_i \) is an unmatched detection and \( n \) is the total number of unmatched detections. Last, some low-confidence tracklets are not associated with any other tracklets or detections, and these tracklets should be terminated. In this final case, an element of the score matrix \( C_{n \times l} = [c_t] \) is defined as \( c_t = -\log(1 - \text{conf}(T_i^{(l)})) \). Altogether, we can define a unified score matrix \( G \) to represent the cases as follows:

\[
G_{(l+n) \times (h+l)} = \begin{bmatrix} A_{h \times l} & C_{n \times l} \\ D_{n \times h} & B_{n \times l} \end{bmatrix}.
\]
where \(D_{n\times h}\) is a threshold matrix. After computing \(G\), the Hungarian algorithm\(^3\) is also applied. Finally, the confidence values of all tracklets are updated using Eq. (3).

### 3.3 Augmentation of Detected Object Candidates

To augment a set of detected object candidates, we iteratively apply the single-object tracking algorithm, which is based on a correlation filter.\(^4\) At the start of each iteration, the single-object tracking algorithm finds object candidates based on the positions of previously detected object trajectories. Since this step is a potential computational overhead for the proposed framework, we must consider trade-offs between tracking accuracy and speed. In this sense, the correlation filter provides an appropriate balance because the method is extremely fast and results in accuracy that is comparable to other state-of-the-art algorithms. Finally, the set of tracked object candidates for the local and global association steps is updated during each iteration as follows:

\[
\hat{Z}_t = \{Z_t \cup D_t\}, \quad D_t = \{d_t|\text{PSR}(d_t) > \theta_D\},
\]

where \(d_t^i\) is a tracked object from the previously determined trajectory \(T_{t-1}\), which is included in \(\hat{Z}_t\) only when the peak-to-sidelobe ratio (PSR) of \(d_t^i\) is greater than the threshold \(\theta_D\). In this way, the PSR measures the degree of appearance variation of an object in the correlation-filter-based tracking algorithm,\(^5\) and thus, it represents the reliability of the tracked prediction \(d_t^i\).

### 3.4 Similarity Function Based on Relative Motion Network

A function \(\Lambda\) that measures similarity between tracklets and detections is generally defined by considering three aspects of the trajectory: appearance, size, and motion.\(^3,6\) In this letter, we define the similarity function by using a probabilistic formulation as follows:

\[
\Lambda(z_t^i, T_t^i) \triangleq P(z_t^i|T_t^i, R_t^i) = P_a(z_t^i|T_t^i)P_s(z_t^i|T_t^i)P_m(z_t^i|T_t^i, R_t^i),
\]

\[
\approx P_a(z_t^i|x_t^i)P_s(z_t^i|x_t^i)P_m(z_t^i|x_t^i, R_t^i),
\]

where the terms \(P_a(z_t^i|T_t^i)\), \(P_s(z_t^i|T_t^i)\), and \(P_m(z_t^i|T_t^i, R_t^i)\) indicate appearance, size, and motion similarity between the \(k\)th detection and the tracklet of the \(i\)th object at frame \(t\). These probabilities are approximated by \(P_a(z_t^i|x_t^i)\), \(P_s(z_t^i|x_t^i)\), and \(P_m(z_t^i|x_t^i, R_t^i)\) based on a first-order Markov chain. The term \(R_t^i\) indicates the RMN. These terms are defined as follows.

**Appearance similarity:** The appearance similarity, \(P_a(z_t^i|x_t^i)\), is estimated by computing the Bhattacharyya coefficient of color histograms as follows:

\[
P_a(z_t^i|x_t^i) = \frac{\sum_{n=1}^{N_c} \sqrt{H^n(z_t^i)H^n(x_t^i)}}{N_c},
\]

where \(H^n(\cdot)\) denotes the \(n\)th element of each color histogram.

**Size similarity:** The size similarity, \(P_s(z_t^i|x_t^i)\), is estimated by comparing the heights of bounding boxes as follows:

\[
P_s(z_t^i|x_t^i) = \frac{1 - \frac{|h(T_t^i) - h(z_t^i)|}{h(T_t^i) + h(z_t^i)}}{h(T_t^i) + h(z_t^i)},
\]

where \(h(\cdot)\) denotes the height of a bounding box of an object or detection.

**Motion similarity:** The motion similarity, \(P_m(z_t^i|x_t^i, R_t^i)\), is defined by using the RMN\(^6\) to handle abrupt camera motion. Suppose a relative motion \(r_{i;j}^t\) between two objects, \(x_t^i\) and \(x_t^j\), is defined as the differences between the positions and velocities of the objects as follows:

\[
r_{i;j}^t = |x_{t}^{i} - x_{t-1}^{i}, v_{t}^{i} - v_{t-1}^{i}, a_{t}^{i} - a_{t-1}^{i}|.
\]

Then, an RMN \(R_t^i\) of the object \(x_t^i\) is defined as follows:

\[
R_t^i = \{\langle x_t^i, x_t^j \rangle | \langle x_t^i, x_t^j \rangle \in \{T_t^{(k)} \cup D_t\}\},
\]

where \(\theta_R\) is a predefined threshold for finding reliable objects. Also, \(\langle x_t^i, x_t^j \rangle\) represents the self-motion of \(x_t^j\), and \(\langle x_t^i, x_t^j \rangle\) represents the relative motion between the two objects \(x_t^i\) and \(x_t^j\).

In contrast to the original definition of RMN,\(^6\) we construct the network using only reliable objects that are obtained from high-confidence tracklets because unreliable objects can corrupt the estimation of the relative motion of \(x_t^j\). In addition, we also use the single-object tracking algorithm to construct a set of reliable tracked objects \(\hat{D}_t\), which is then incorporated into the set of neighboring objects to address the misdetection problem. By using the enhanced RMN, the state \(x_t^i\) can be estimated from the earlier state \(x_{t-1}^i\) of the previous frame as follows:

\[
x_{t}^{i;\check{\cdot}} = f(x_{t-1}^{i;\check{\cdot}}, x_{t-1}^{j;\check{\cdot}}) + w
\]

\[
= F(\left[x_{t-1}^{i;\check{\cdot}}, v_{t-1}^{i;\check{\cdot}}, a_{t-1}^{i;\check{\cdot}}\right] + r_{i;j}^{(t)} + w_{t-1}^{(t)} + h_{t-1}^{(t)}) + w,
\]

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Comparison of the proposed algorithm to the state-of-the-art methods on the MOT challenge dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td>MOTA ↑</td>
</tr>
<tr>
<td>MDP(^4)</td>
<td>30.3</td>
</tr>
<tr>
<td>SCEA(^5)</td>
<td>29.1</td>
</tr>
<tr>
<td>OMT(_{DFH})(^1)</td>
<td>21.2</td>
</tr>
<tr>
<td>RMOT(^6)</td>
<td>18.6</td>
</tr>
<tr>
<td>TC(_{ODAL})(^3)</td>
<td>15.1</td>
</tr>
<tr>
<td>CDA(_{DDL})(^2)</td>
<td>32.8</td>
</tr>
<tr>
<td>Proposed</td>
<td>31.1</td>
</tr>
</tbody>
</table>

Note: Bold, italic, and bold-italic values denote the best, the second-best, and the third-best performances in each evaluation metric, respectively.

---

\(^{1}\) OMT\(_{DFH}\) is an abbreviation for Online Multi-target Tracking with Dynamic Fusion Hierarchical

\(^{2}\) CDA\(_{DDL}\) is the abbreviation for Category Detection and Association with Decoupled Learning

\(^{3}\) TC\(_{ODAL}\) is the abbreviation for Track Clustering with Online Detection and Association

\(^{4}\) MDP is the abbreviation for Markov Decision Process

\(^{5}\) SCEA is the abbreviation for State-Space Correlation Filter

\(^{6}\) RMOT is the abbreviation for Real-Time Multi-object Tracking

---
where $\mathbf{F}$ denotes a state transition matrix based on the constant velocity model and $\mathbf{w}$ is the white Gaussian noise model. Based on this RMN model, we define the motion similarity as follows:

$$P_m(z_i^t | x_i^t, \mathcal{R}_i^t) \approx P_m(z_i^t | \langle x_i^t, x_i^t \rangle^*),$$

$$\langle x_i^t, x_i^t \rangle^* = \arg\max_{\langle x_i^t, x_i^t \rangle \in \mathcal{R}_i} P(z_i^t | \langle x_i^t, x_i^t \rangle),$$

(14)

where $\mathcal{R}_i$ is a set of all RMNs $\{\mathcal{R}_i = \bigcup_{j=1}^{N_{obj}} \mathcal{R}_j \}$ and $\langle x_i^t, x_i^t \rangle^*$ is an object pair that mostly contributes to the relative motion. The contribution of each object pair to the relative motion is determined by computing $P(z_i^t | \langle x_i^t, x_i^t \rangle)$, which is defined by using the Pascal score\footnote{9} between bounding boxes of a detected object and a reprojected object, as shown here

$$P(z_i^t | \langle x_i^t, x_i^t \rangle) = \frac{\text{area}(\mathcal{B}(z_i^t) \cap \mathcal{B}(f(x_{t-1}^j, x_{t-1}^j)))}{\text{area}(\mathcal{B}(z_i^t) \cup \mathcal{B}(f(x_{t-1}^j, x_{t-1}^j)))},$$

(15)

where $\mathcal{B}(\cdot)$ denotes a bounding box.

### 4 Experimental Results

In the experiment, we used the MOT challenge benchmark dataset\footnote{10} that consists of several image sequences obtained from a static or dynamic camera. We compared the proposed method with other state-of-the-art online MOT algorithms, specifically TC_ODAL\footnote{3} MDP\footnote{4} SCA\footnote{5} OMT_DFH\footnote{6} RMOT\footnote{7} CDA_DDALpb.\footnote{8} For evaluation, we adopted seven metrics: the multiple object tracking accuracy (MOTA), the relative motions between objects as well as tracking of individual objects to track multiple moving objects in a robust way. The relative motion information can help to track objects by inferring its location from neighboring objects even if an object is severely occluded by others. This effect can be observed from the experimental result. The proposed algorithm achieved the best performances in terms of IDS and FG, even when a simple appearance model is used.

### 5 Conclusion

In this letter, we have proposed a unified MOT framework to tackle the difficulties of detection-based tracking algorithms. The proposed framework consists of a two-step association process to generate long and accurate object trajectories. First, tracklets and detected object candidates are associated to construct high-confidence trajectories. Then, the confident trajectories are associated with low-confidence trajectories. We also proposed a similarity function that incorporates the concept of an RMN. In our experiments, the proposed algorithm exhibited more robust performance than current state-of-the-art algorithms.

### Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (No. NRF-2015R1A2A1A1005455).

### References