Robust Multi-object Tracking for Wide Area Motion Imagery

Noor M. AL-Shakarji†‡, Filiz Bunyak*, Guna Seetharaman†, Kannappan Palaniappan*

* Department of Computer Science, University of Missouri-Columbia, MO 65211, USA
† Advanced Computing Concepts, U.S. Naval Research Laboratory, Washington, DC 20375, USA
‡ Department of Computer Science, University of Technology, Baghdad, Iraq

Email: *nmahyd@mail.missouri.edu, †{bunyak, palaniappank}@missouri.edu, ‡Gunasekaran.Seetharaman@rl.af.mil

Abstract—Multi-object tracking implemented on airborne wide area motion imagery (WAMI) is still challenging problem in computer vision applications. Extremely camera motion, low frame rate, rapid appearance changes, and occlusion by different objects are the most challenges. Data association, link detected object in the current frame with the existing tracked objects, is the most challenging part for multi-object tracking algorithms. The ambiguity of data association increases in WAMI datasets because objects in the scenes suffer form the lack of rich feature descriptions beside the closeness to each other, and inaccurate object movement displacement. In this paper, detection-based multi-object tracking system that uses a two-step data association scheme to ensure high tracking accuracy and continuity. The first step ensures having reliable short-term tracklets using only spatial information. The second step links tracklets globally and reduces matching hypotheses using discriminative features and tracklets history. Our proposed tracker tested on wide area imagery ABQ dataset [1]. MOTChallange [2] evaluation metrics have been used to evaluate the performance compared to some multi-object-tracking baselines for IWTS42018 [3] and VisDrone2018 [4] challenges. Our tracker shows promising results compared to those trackers.

Index Terms—Tracking, wide area motion imagery, multi-object tracking, tracking-by-detection, data association.

I. INTRODUCTION

Single object tracking [5]–[8] and multi-object tracking [9]–[12] are the two main categories of visual tracking. Wide variety of computer vision applications including video summarization, animals and cells behavior understanding, sport video analysis, and visual surveillance rely on visual object tracking. Visual object tracking is the process of localizing objects of interest over time. Real-world conditions such as background clutter, shadow and illumination changes, partial or full occlusions still make visual tracking challenging for researchers. This work focuses on multi-object tracking (MOT). Multi-object tracking aims to locate multiple objects in a scene, maintain their identities in time, and form motion trajectories for further analysis. Since the rapid progress in object detection techniques [13]–[15], tracking-by-detection [9], [16], [17] becomes popular among multi-object tracking approaches and relies heavily on improving data association approaches [18], [19] to increase the accuracy and reduce the complexity. The data association process can be performed either online [20]–[23] by exploiting only the information gathered form the past frames, or offline (batch-mode) [24], [25] by exploiting information from both past and future frames.

Some applications (e.g. online surveillance, navigation, autonomous driving etc.) require online object tracking. However, it may fail when miss detection or long occlusion of targets exists causing short trajectories (tracklets). Offline (batch mode) tends to manage this problem by optimizing multiple trajectories globally. For example, i) by considering additional cues from the entire sequence to have more complete trajectories (i.e. appearance, motion) [26], [27]. The cues adversely affect computational cost of tracking. For that, it needs to be chosen carefully to ensure efficiency and reliability. ii) By creating network flow graphs [25], [28] which is pairwise edges between observations. Although it has shown promising accuracy in multi-target tracking, it ignores kinematic constraints, since it needs more than two nodes to represent motion i.e velocity. iii) Others like [29], [30] focusing on motion information to link the tracklets especially when the appearance of the objects is ambiguous. often motion-based methods produce many identity switches.
and fragments because they fail to re-identify objects when abrupt camera motions or shot changes occur. iv) Or by using iterative hierarchical methods to link tracklets \[24, 31\]. Hierarchical methods progressively refine the targets’ association hypotheses and link them to longer tracklets which lead to efficiency in association with better accuracy.

In this paper, SCTrack (semantic color correlation tracker), a tracking-by-detection approach using a two-level data association was proposed. The first level, called local association, generates reliable short-term tracklets through consecutive frames using spatial distance for matching. The second level, called global association, links the short-term tracklets using a greedy global association scheme, where information from the whole sequence is used. This level combines spatial and temporal cues with object appearance features to ensure time efficiency while preserving tracking accuracy. The proposed pipeline was implemented on ABQ wide area motion imagery dataset \[1\] and compared with some baseline trackers whose codes are available for distribution online. Our proposed pipeline outperforms the baseline trackers. Figure 1 illustrates the proposed levels data association pipeline. The paper is organized as follows. Section 2 describes the details of the proposed pipeline including detection and tracks initialization, local and global data association. Section 3 presents the experimental results followed by the conclusion.

II. MULTI-OBJECT TRACKING

Muli-object tracking is the process of locating and correctly associating moving objects in consecutive video frames. In this paper, a tracking-by-detection framework is proposed. Below, the main steps involving the proposed tracking is summarize:

A. Detection and Tracks Initialization

ABQ aerial wide area motion imagery dataset \[1\] for downtown Albuquerque, NM was used. The ground-truth was generated manually for 139 cars within a 200 frames subset of the ABQ dataset. The ground-truth bounding boxes are used as input to our pipeline for fair comparison with the other baseline methods and ensure eliminating the bias caused by detection failures. When a new track is formed, four groups of information corresponding to the new track are initialized and recorded for further use: (1) detected object’s location, width, height, and appearance; (2) counters for age, start frame, end frame, visible frames, and invisible frames; (3) Kalman filter parameters \(K_{i}, F_{i} = \{x_{i}, P_{i}\}\) where \(x_{i}\) represents state estimate for the object \(i\) at frame \(t\) and \(P_{i}\) represents associated covariance matrix; and (4) object velocity records. Kalman filter with a constant velocity model is used to predict the new positions of the tracked objects using their past trajectories. This information will be updated in each frame to ensure updating the status of each object.

B. Short-Term Local Association

Short-term local data association is the first step in our pipeline. The Ground-truth \((GT)^t = \{d_1, d_2, ..., d_N\}\) are assigned to the previously tracked objects \(T^{t-1} = \{T_1, T_2, ..., T_M\}\) . Where \(N\) is the number of the detected objects at frame \(t\) and \(M\) is the number of tracked objects at frame \(t-1\). Spatial distance is used for assigning GT bounding boxes with the trajectories between consecutive frames. Detected objects are assigned to existing tracks by minimizing a cost matrix using Munkres Hungarian algorithm \[18\]. Elements of the cost matrix are computed as:

\[
C(i, j) = \log \|d_i(x, y), T_j(x, y)\|_2
\]

where \(d_i(x, y)\) and \(T_j(x, y)\) are the centroids of the detected objects and predicted tracks respectively. One among four-status (new, extended, lost, and inactive track) will assigned for each detection according to the association matrix result. Figure 2 illustrates more details about each status. The short-term association process considers only the information from consecutive frames, it can not recover from temporary missing detection or association problems such as occlusion, false detection, matching ambiguities etc. Further stages of our pipeline will be used to improve the performance in these cases.

C. Global Data Association

Early terminated short tracks can be generated form various problems during object detection or data association stages. Global data association is used to link tracklets to generate longer tracks. Global association is an expensive process because it optimizes all possible hypotheses rather than only those on consecutive frames. Global association at tracklet level is used to ensure reduction in computation cost and preserving complex assignment cases.

Spatial distance, tracklet start and end times, motion directions, and appearance models are used to filter out infeasible matches between tracklets. The information required in this step was stored since the objects were initiated and kept updating during associated processing. In this work, for target appearance description, we adopt the appearance model described in \[11\]. This model combines shape and texture properties described by HoG (Histogram of Gradients) \[32\] descriptor with object color attributes described by CN (color name) \[33\] histograms and compared with our novel color correlation cost matrix. Mean square error (MSE) is used to compute the distance between two HoG descriptors. And earth mover distance (EMD) \[34\] is used to compare the distance between two color name histograms. For color information, the histogram of color information adapted from \[11\] is recorded during track initialization. Earth mover distance (EMD) is used to compare between two color histograms for more robustness to shadow and illumination changes.

Refinement process filters out the infeasible hypotheses to reduce assignment space for the tracklet candidates. For a given tracklet \(k\), at most one tracklet that belongs to the set of candidates will be assigned to \(k\) if the set is not empty after refinement process. Refinement process starts by check some conditions met with \(k\) to be refined from candidate set:
1) Refine all tracklets in candidate set that are initialized on frame borders (tracklets entering the scene).
2) Refine all tracklets that are belonging to candidate set that are born before the death of tracklet $k$.
3) Refine all tracklets belonging to candidate set that start far from the last position of tracklet $k$.
4) Refine all tracklets in candidate set that average directional motion different form $k$ direction.
5) Refine all tracklets in candidate set whose appearance distances from tracklet $k$ is larger that predefined threshold.

III. EXPERIMENTAL RESULTS

We have tested and evaluated our multi-object tracking SCTrack tracker on 200 frames of ABQ dataset [1]. The goal is to track the moving cars in the scene. We used the ground-truth bounding boxes for 139 cars as a detection to our pipeline in order to have fair comparison result with the other baseline trackers. We have evaluated the performance of our tracker using the multi-object tracking challenge development kit [35]. The evaluation process includes the following metrics described in [2]: mostly track (MT), mostly lost (ML), partially track (PT), identity switches (IDs), track fragmentation (FM), false positives (FP), false negatives (FN), multi-object tracking precision (MOTP), and multi-object tracking accuracy (MOTA). We also use (RANK) metric which is the average ranking for the all evaluation metrics. We compared our tracker with competitive approaches whose code are accessible and are considered as baseline for IWTS42018 [3] and VisDrone2018 [4] challenges. Table I shows the comparison results with GOG [28], CMOT [23], IHTLS [29] trackers. The accuracy of multi-object tracking result is best reflected by (MOTA) and (RANK) metrics between the compared trackers. Our SCTrack tracker outperforms other comparable trackers. SCTTrack tracker takes into account spatial, appearance, and kinematic information to ensure reliable trajectory linking. While GOG [28] basically uses a pairwise edges between network flow graphs to link trajectories results in ignoring kinematic constraints between observations. CMOT [23] is an online tracker that does not exploit the visual information form future frames which is important for increase performance under unpredictable objects displacement and camera motion. IHTLS [29] focuses on motion information to link the tracklets that leads to many identity switches especially with WAMI datasets since the motion is quite challenging. Figure 3 shows the multi-object tracking results in different $t$ time.

IV. CONCLUSIONS

Detection-based multi-object tracking system that uses a two-step cascaded data association scheme to ensure time efficiency by reducing the hypotheses through the steps. While preserving tracking accuracy by having a discriminative object appearance models. Low cost local association operates at object level on consecutive frames relying only on spatial distance, and a robust tracklets linking step using discriminative object appearance models are used. Experiments on ABQ dataset shows promising results against baseline trackers. We are in the process of extending our appearance model for
Fig. 3. Our multi-object tracking results in different t time.

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Rank ↓</th>
<th>Rell ↑</th>
<th>Prcn ↑</th>
<th>FAR ↓</th>
<th>CT</th>
<th>ML ↑</th>
<th>FT ↓</th>
<th>FB ↑</th>
<th>FN ↓</th>
<th>IDs ↓</th>
<th>FM ↑</th>
<th>MOTA ↑</th>
<th>MOTP ↑</th>
<th>MOTA ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC Track</td>
<td>1.2</td>
<td>100.0</td>
<td>98.6</td>
<td>0.57</td>
<td>139</td>
<td>139</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>133</td>
<td>0.57</td>
</tr>
<tr>
<td>GOG [28]</td>
<td>1.9</td>
<td>97.8</td>
<td>100</td>
<td>0.01</td>
<td>139</td>
<td>135</td>
<td>4</td>
<td>0</td>
<td>182</td>
<td>205</td>
<td>39</td>
<td>95.3</td>
<td>100</td>
<td>97.8</td>
</tr>
<tr>
<td>CMOT [23]</td>
<td>3.2</td>
<td>32.1</td>
<td>52.1</td>
<td>19.98</td>
<td>139</td>
<td>49</td>
<td>15</td>
<td>75</td>
<td>3997</td>
<td>-2.5</td>
<td>3983</td>
<td>798</td>
<td>48.8</td>
<td></td>
</tr>
<tr>
<td>IHTLX [29]</td>
<td>3.3</td>
<td>49.1</td>
<td>48.8</td>
<td>21.47</td>
<td>139</td>
<td>52</td>
<td>12</td>
<td>75</td>
<td>4293</td>
<td>4237</td>
<td>768</td>
<td>-11.7</td>
<td>99.7</td>
<td></td>
</tr>
</tbody>
</table>

TABLE I
Tracker performance comparison using three baseline tracker for IWTS4201 and VisDrone2018 challenges.

Further improved results. And having more clues for tracking the occluded objects.

REFERENCES
[25] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-


