Multi-Object Tracking in Airborne Video Imagery based on Compressive Tracking Detection Responses
Chen, Ting; Sahli, Hichem; Zhang, Yanning; Yang, Tao

Published in:
International Conference on Advances in Mobile Computing and Multimedia

Publication date:
2015

Document Version:
Final published version

Citation for published version (APA):
Multi-Object Tracking in Airborne Video Imagery based on Compressive Tracking Detection Responses

Ting Chen 1.School of Computer Science and Engineering, Northwestern Polytechnical University, Xi’an, PR China
2.Dept. of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Brussels, Belgium

Hichem Sahli 3.Interuniversity Microelectronics Center (IMEC), Leuven, Belgium
hsahl@vub.ac.be

Yanning Zhang, Tao Yang 1.School of Computer Science and Engineering, Northwestern Polytechnical University, Xi’an, PR China

ABSTRACT

Multi-object tracking (MOT) in airborne video is a challenging problem due to the uncertain airborne vehicle motion as well as mounted camera vibrations. Most approaches addressing tracking in such type of scenario, use data association based on motion detection responses. Such approaches fail tracking objects with low speed or static ones. To alleviate the motion detection failures, in this paper we propose a multi-object tracking system based on combining motion-detection and Compressive Tracking detection responses.

In this work, as in [1], the multi-object tracking problem is solved by associating tracklets according to their confidence values. For reliable association between tracklets and detections, we propose using Compressive Tracking (CT) as a mean to detect objects when motion-detection fails. By exploiting the compressive tracking, which allows discriminating the appearances of objects, tracklet association can be successfully achieved even when objects undertake stop-and-go motion as well as when they are partially occluded. Experiments with challenging airborne video datasets show significant tracking improvement compared to existing state-of-art methods.

Categories and Subject Descriptors
I.4.8 [Computing Methodologies]: Image Processing and Computer Vision-Scene Analysis-Tracking

General Terms
Algorithms

Keywords
Tracklet Confidence, Compressive model, Tracking

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

Copyright 2015 ACM 978-1-4503-3493-8/15/12 ...$15.00.
http://dx.doi.org/10.1145/2837126.2843846.

1. INTRODUCTION

The goal of multi-object tracking (MOT) in airborne video is to estimate the states of multiple objects while conserving their identifications under appearance and motion variations with time. This is challenging since one should consider the uncertain motion of the airborne vehicle as well as the vibration of the camera [2]. In airborne videos and aerial surveillance, due to the targets small size and their various appearance, most of the proposed MOT approaches, adopted motion-based detection and association to recover the trajectories of all targets [3, 4]. Motion-based detection rely on a background model [4, 5, 6], where objects with different motion model than the background can be detected. Such motion-based detection approaches fail detecting static objects or objects with low speed and hence not useful for long-term tracking. In the other hand, approaches using appearance models for object detection [7, 8, 1], are not suitable for airborne videos. This is mainly due changes in appearances of the tracked objects, the dynamic change of the background, the uncertain motion of the airborne vehicle as well as camera vibrations.

In this work, the multi-object tracking problem is solved by associating tracklets according to their confidence values. For reliable association between tracklets and detections, we propose using Compressive Tracking (CT) as a mean to detect objects when motion-detection fails. Compressive tracking [9, 10, 11] achieved favorably high efficiency and accuracy in several object tracking scenarios, which make it a good candidate for multi-object tracking in airborne videos. We follow [12, 1] to formulate the MOT problem by using locally and globally tracklet association based on the tracklet confidence. The tracklets with the high confidence are locally associated to the motion-detection responses, whereas tracklets with low confidence are globally associated to other tracklets using CT-based tracking results. Different from [12, 1], we adopt a compressive tracker to obtain extra observations to link the fragmented low confidence tracklets in a global fashion.

Our primary contribution in this work, is a multi-object tracker for airborne videos, that uses CT to alleviate the problems due to failures in motion-detection. We adopt the two phases tracking approach introduced in [1] and incorporate the information from CT for the tracklets association.
2. TRACKING FRAMEWORK

2.1 System Overview

We denote an object $i$ appears at frame $t$ by using a binary function as $\phi_i^t = 1$, otherwise $\phi_i^t = 0$. Let $x_t^i = [p_t^i, s_t^i, v_t^i]$, the state of the object $i$ at frame $t$, where $p_t^i, s_t^i, v_t^i$ are the location, size, velocity, respectively. Then, denote the tracklet $T^i$ of the object $i$ as a set of states up to frame $t$ as $T^i = \{x_k^i | 0 \leq k \leq t\}$, where $t$ is the start and end frame number of the tracklet. In addition, denote all the object states appeared at frame $t$ as $T_t = \{x_{n_t}^1, \cdots, x_{n_t}^{n_t}\}$, where $n_t$ is the total number of objects, and let $T_{t-1} = \{T_1^1, \cdots, T_{t-1}^{n_t}\}$ be a set of tracklets of all objects up to frame $t$.

As in [13], for object detection, we adopt a motion-based detection approach using the motion history, applied on 5 frames. Let, $d_i^j$ the $j$-th detected observation at frame $t$, we denote by $D_t = \{d_i^j | 1 \leq j \leq n_d\}$ all the observations collected at frame $t$, $n_d$ being the number of observations, and $D_{t-1} = \{d_1^1, \cdots, d_{n_d}^t\}$ the set of all observation up to frame $t$. All the observations associated to object $i$ up to frame $t$ are referred to as $d_i = \{d_i^1, \cdots, d_i^t\}$.

We follow [12, 1] to formulate the MOT problem as a two phases optimization problem. The main steps of our algorithm are summarized in Algorithm 1.

During the first phase, referred to as Local Phase, the tracklets from frame $t-1$ are first locally associated with the motion-based detections of frame $t$ using the affinity model, $A(T^i, d_i^t)$ of [1], and then their confidence [1], $conf(T^i)$, is updated as follows:

\[
conf(T^i) = \frac{1}{L} \sum_{z \in [z_1^1, \cdots, z_{n_t}^t]} A(T^i, d_i^t) \times \max \left(1 \quad \beta \cdot \log \left(\frac{(L - w_l)}{L}\right), 0\right) \tag{1}
\]

where $L$ is the length of the tracklet $T^i$, defined as $L = |T^i|$, $A(T^i, d_i^t)$ is the affinity score calculated using the affinity model of [1], $\beta$ is a control parameter relying on the performance of the motion-based detection and $\beta = 2$ in our experiments, and $w_l = (t_0^i - t_i^t + 1 - L)$. The association procedure is summarized in Section 2.2. Since the tracklet confidence lies in $[0, 1]$, we consider a tracklet as a reliable tracklet with high confidence, if $conf(T^i) > 0.5$; otherwise it is considered as a fragmented tracklet with low confidence. We denote by $T^{(h)}$ the tracklets with high confidence, and by $T^{(l)}$ the tracklets with low confidence.

The local association between the tracklets and motion detections allows us to progressively grow locally optimal tracklets. However, it happens that the objects being tracked are not detected due to unreliable motion detection, small objects, static objects, objects evolving at low speed, or occluded objects. When a detection of the object is not available, the confidence of a tracklet is decreased. To solve this problem, we associate to each tracked object, $x'$, a Compressive Tracker (CT) [9], denoted as $CT_i(x')$. For each CT tracker we maintain a Bayes classifier $H(z')$, and the current tracker location, $f^i$.

During the second phase, i.e, the global association, tracklets with low confidence, $T^{(l)}$ are globally associated with other tracklets and detections. Different from [1, 12], we make use of the Compressive Tracking (CT) results ($f^i$) into the global association to solve the motion-based detection failure problem. In this phase we consider the following associations:

- Event A: $T^{(h)}$ is associated with $T^{(h)}$.
- Event B: $T^{(l)}$ is terminated.
- Event C: $T^{(l)}$ is associated to an optimal compressive tracker location $f^i$.
- Event D: $T^{(l)}$ is associated with a motion detection $d_i^t$ not associated to a high confidence tracklet.

This phase is described in Section 2.3. For both association phases (local and global) we determine the tracklet-detection pairs and the tracklet-tracklet pairs using the Hungarian algorithm [14] such that a total association cost is maximized. This will be explained in the following sections.

2.2 Local Association of Tracklets

During the local association phase pairwise association is performed to associate motion detection responses $d_i^t$ with tracklets high confidence $T^{(h)}$ [1]. A score matrix $S_{n_t \times n_d}$ defined as:

\[
S_{n_t \times n_d} = [s_{ij}]_{n_t \times n_d},
\]

\[
s_{ij} = -\ln(L\{A(T^{(h)}, d_i^t)\}, d_i^t \in D_t), \tag{2}
\]

where $n_t$ and $n_d$ is the number of the high confidence tracklets and motion-based detections at frame $t$, respectively.

Algorithm 1 The two-phases tracking framework.

Input:
- The tracked objects $T_{t-1} = \{x_{1,t-1}^1, \cdots, x_{n_t-1}^{n_t-1}\}$,
- The tracked objects locations obtained by compressive tracking $C_{t-1} = \{d_{1,t-1}^1, \cdots, d_{n_t,t-1}^{n_t-1}\}$, and compressive tracking classifiers $H_{t-1} = \{H(z_{1,t-1}^1), \cdots, H(z_{n_t,t-1}^{n_t-1})\}$ at frame $t-1$,
- Motion-based detections $D_t = \{d_1^1, \cdots, d_{n_d}^t\}$ at frame $t$.

Local Phase: For each tracked object
1: if $x_{1,t-1}^{n_t}$ can be observed by $d_i^t$ then
5: else
9: else
13: end if
17: end if

Global Phase:
2: Update the $x_i^t$ and $conf(T^i)$ by associated $d_i^t$.
3: Get the compressive tracking result $f^i$ by Eq.(4).
4: Initialize/Update the classifier $H(z_i^t)$.
5: Update/Terminate the classifier $H(z_i^t)$.

Output: The object set $T_t = \{x_1^t, \cdots, x_{n_t}^t\}$ with their compressive tracker attributes $C_t = \{d_1^t, \cdots, d_{n_d}^t\}$ and $H_t = \{H(z_1^t), \cdots, H(z_{n_d}^t)\}$. 

$\text{conf}(T^i)$ is the motion-based detections of frame $T^i$. The main steps of our algorithm are summarized in Algorithm 1.
The Hungarian algorithm [14] is used to determine tracklet-detection pairs with the maximized total affinity in $S_{n_b \times n_d}$. The detection $d_i^t$ is associated with $T_i^{(b)}$ when the association cost of a tracklet-detection pair is less than a pre-defined threshold, $-\ln(\theta)$, the position and velocity of a tracklet $T_i^{(b)}$ is updated with the associated $d_i^t$, and the size of the object is updated by averaging the associated detections of recent past frames [1]. The confidence of the tracklet $\text{conf}(T_i^{(b)})$ is updated by using $d_i^t$ in Eq(1).

### 2.3 Global Association of Tracklets

As mentioned above, this phase allows associating globally low confidence tracklets either to high confidence tracklets, or to motion detection responses, $d_i^t$, not associated to a high confidence. Moreover, to alleviate the problem of motion detection failure, and hence missing associations, we make use of the CT trackers to associate low confidence tracklets to other high confidence tracklets.

Let, $\{q_i\}_{i=1}^{n_q}$, the motion-based detection responses that have not been associated to any high confidence tracklets during the previous local association. We denote by $C_i = \{l_i^1, \ldots, l_i^{n_c}\}$ the current locations of the objects as determined by the $n_c$ CT trackers. The following association events are considered:

- **Event A**: $T_i^{(b)}$ is associated with $T_i^{(a)}$,
- **Event B**: $T_i^{(b)}$ is terminated,
- **Event C**: $T_i^{(b)}$ is associated to an optimal compressive tracker location $l_i^j$,
- **Event D**: $T_i^{(b)}$ is associated with a motion detection $d_i^t$ not associated to a high confidence tracklet.

The cost matrix for the above events is defined as follows:

$$
G_{(n_j+n_c+n_q)\times(n_b+n_d)} = \begin{bmatrix}
A_{n_j\times n_b} & B_{n_j\times n_d} \\
-\log(\theta)_{n_j\times n_b} & C_{n_j\times n_d} \\
-\log(\theta)_{n_d\times n_b} & D_{n_d\times n_j}
\end{bmatrix}
$$

where $A = [a_{ij}]$ represents $T_i^{(i)}$ is associated with $T_i^{(b)}$, and $a_{ij} = -\ln(A(T_i^{(i)}, T_i^{(b)}))$ is the association cost calculated by the affinity model of Bae et al. [1]. $B = \text{diag}[b_1, \ldots, b_{n_c}]$ models $T_i^{(b)}$ is terminated and $b_i = -\ln(1-\text{conf}(T_i^{(b)}))$ is the cost to terminate $T_i^{(i)}$. $C = [c_{ij}]$ represents $T_i^{(b)}$ is associated with $l_i^j$, where $c_{ij} = -\ln(A(T_i^{(b)}, l_i^j))$. $D = [d_{ij}]$ represents $T_i^{(b)}$ is associated with $q_i^j$, where $d_{ij} = -\ln(A(T_i^{(b)}, q_i^j))$.

The optimal association pairs, which minimize the global association cost in $G$, are determined using the Hungarian algorithm [14], and the tracklets and their confidence values are updated with the results.

### 2.4 Compressive Tracking Model

In this section, we give a short overview of Compressive Tracking (CT) and how it is used in our framework to estimate the current tracker location, the interested reader is referred to [9, 10, 11] for more details. In CT a high-dimensional feature being a series of rectangle features which have different locations and sizes within the region of interest are used. This large set of features is then compressively sensed using a very sparse measurement matrix which satisfies the restricted isometry property. This can then be efficiently computed for real-time tracking. By taking the tracking as the online classification task of the object and background, the naive Bayes classifier is used to model the real-time tracking process. For each frame, some positive samples near the target and some negative samples farther away from the object are cropped to update the classifier parameter. To predict the next target location, samples are drawn around the current target location and sent to the classifier. The sample with the maximal classification score are selected as the most probable target location $l_t$ [9].

In our framework, the most probable target location $l_t$ in frame $t$ is given by:

$$
l_t^* = \arg \max_{l_t^* \in \mathbb{R}^d, l_t \in \mathbb{R}^d} \left( H(z_t^*) \cdot w^k \right)
$$

with $z_t^* = \{z_t^1, z_t^2, \ldots, z_t^n \}$, are the samples around location $l_t-1$ (candidate samples), and the weights $w^k$ is set to $w_k = 1$ in case object $i$ is not associated to any motion detection response, and defined as in Eq(5) if object $i$ is associated to the motion detection response $d_i^t$.

$$
w^k = \frac{||l(z_t^k) - d_i^t||_2}{\sum_{k \in \mathbb{R}^d} ||l(z_t^k) - d_i^t||_2}$$

### 3. EXPERIMENT

We evaluate our approach on DARPA VIVID dataset [15]. The popular evaluation metrics defined in [16] are used for performance comparison: the number of objects in ground-truth (GT), the ratio of mostly tracked trajectories (MT), the ratio of mostly lost trajectories (ML), the number of fragments (Frag) and ID switches (IDS).

We compared our approach to [1] and [3]. All of the approaches adopted the same motion-based detection configuration, and a window size of 5 frames was defined to remove the unreliable shorter tracklet. For both [1] and [3] we used the publicly available MATLAB codes provided by the authors and we selected the p1 system in [1] in our experiment. We set one random matrix for each high confidence tracklet, and the dimension of each random matrix have been randomly (uniform) set between 40 and 100. Similar to [9, 10, 11], the search radius to detect the object location is set to $\gamma = 20$ and about 1100 samples are generated. The algorithm has been implemented in MATLAB on PC with an Intel Core 3.40 GHz CPU with 4GB RAM.

As shown in Figure 1(a), the green boxes are the motion-based detection, and the red boxes in #404 and #451 are the compressive appearance detection for each high confidence tracklets when motion-based detection failed detecting the targets. With the help of the compressive appearance detection for data association, our algorithm increase the MT ratio and reduce the IDs switches ratio.

A quantitative comparison between our approach and other algorithm is given in Table 1. Both [1] and [3] perform well with the good motion-based detections, while poor performs with failed detection. As expected, our algorithm improves performance for most metrics, such as the noticeably reduced ML and Frags rates and increased MT rate against the [1] and [3]. The tracking results are shown in Figure 1(b).

The main limitation of the proposed approach is the possible high computation time which is proportional to the number...
Table 1: Performance comparison.

<table>
<thead>
<tr>
<th>Method</th>
<th>MT(%)</th>
<th>ML(%)</th>
<th>Frag</th>
<th>IDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>74.7</td>
<td>10.5</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>[3]</td>
<td>68.4</td>
<td>16.6</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Our</td>
<td>80.6</td>
<td>6.2</td>
<td>7</td>
<td>16</td>
</tr>
</tbody>
</table>

of high confidence tracklets, and hence number of compressive tracking classifier. For the DARPA VIVID dataset [15], where the number of objects is less than 10, the average run-time is 0.38 sec/frame.

4. CONCLUSIONS

We proposed a multi-object tracking method based on tracklet confidence with compressive appearance model to handle the motion-based detection failure. The online multi-object tracking problem is formulated based on the tracklet confidence: the tracklets with the high confidence are associated with motion-detection responses, whereas tracklets with low confidence are linked to the compressive tracking responses and other tracklets. Experiments with challenging public datasets show the improved tracking performance compared to existing state-of-art methods.

5. ACKNOWLEDGMENTS

This work is supported by the CSC-VUB scholarship (No. 201406290121), the National Natural Science Foundation of China (No. 61272288, No. 61502364, No. 61231016, No. 61303123), Northwestern Polytechnical University (NPU) New AoXiang Star (No.G2015KY0301), Fundamental Research Funds for the Central Universities (No.3102015AX007), NPU New People and Direction (No.13GH014604) and NPU AoXiang Star (No.12GH0311).

6. REFERENCES