PERSISTENT MULTIPLE HYPOTHESIS TRACKING FOR WIDE AREA MOTION IMAGERY

Raphael Spraul, Christine Hartung, Tobias Schuchert

Fraunhofer IOSB, Fraunhoferstraße 1, 76131 Karlsruhe, Germany

e-mail: raphael.spraul@iosb.fraunhofer.de; christine.hartung@hs-kl.de; tobias.schuchert@iosb.fraunhofer.de

ABSTRACT

Wide area motion imagery (WAMI) acquired by an airborne sensor enables continuous monitoring of large urban areas. Reliable vehicle tracking in this imagery remains challenging due to low frame rate and small object size. Many approaches solely rely on motion detections provided by frame differencing or background subtraction. Recent approaches for persistent tracking, i.e. tracking vehicles even if they become stationary, compensate for missing motion detections by combining a detection-based tracker with a second tracker based on appearance or local context. We propose a novel single tracker framework based on multiple hypothesis tracking (MHT) that enables persistent tracking in WAMI data by recovering missing motion detections with a classifier-based detector, thus avoiding the additional complexity introduced by combining two trackers. We adapt the MHT approach to the specific context of WAMI tracking by integrating an appearance-based similarity measure, vehicle-collision tests, and clutter handling. An evaluation on a region of interest in the WPAFB 2009 dataset shows state-of-the-art performance.

***Index Terms*—** multi-target tracking, wide area aerial surveillance, wide area motion imagery

1. INTRODUCTION

In recent years, wide area motion imagery (WAMI) has become increasingly interesting for wide area aerial surveillance (WAAS) systems. WAMI sensors with several square kilometers of ground coverage allow the detection and tracking of thousands of vehicles at the same time. Due to the large image size (~ 100 megapixels) the data is typically collected at $1 - 2\,\mathrm{Hz}$ and in grayscale. Small object size (approximately 10×20 pixel), weak appearance of vehicles, and vehicle travel distances of up to 100 pixel from frame to frame make tracking in WAMI particularly challenging.

Persistent tracking aims at continuously tracking vehicles even if they stop at intersections or in traffic jams, and therefore requires the handling of initially moving targets that can become stationary for an extended period of time. In such situations, the common approach of solely relying on detections that are based on object motion is not feasible anymore. In order to compensate for long-term missing motion detections, current approaches to persistent WAMI tracking combine a detection-based tracker with a second tracker that can rely on appearance [1] and local context information [2].

In this work, we present an alternative approach that enables persistent WAMI tracking with a single tracker framework based on multiple hypothesis tracking (MHT), without the need for combining two different trackers. Instead, we search for missing detections with a sliding window classifier if track velocity becomes very small and return the resulting classifier-based detection as input to the MHT tracker. The MHT tracker then processes input detections independently of whether they were generated by motion detection or with the classifier, thus providing a systematic solution to the data association problem using all available detections. The main advantage of the proposed single tracker approach is reduced complexity, as additional steps required by frameworks with two trackers, such as establishing correspondence between targets from each tracker, are eliminated.

We propose several extensions to adapt the MHT-based framework to the specific context of WAMI tracking: (a) integration of an appearance-based similarity measure to help solving the data association problem, (b) collision tests to suppress implausible vehicle maneuvers, and (c) clutter handling to remove tracks originating from false detections.

We evaluate the tracking framework on a region of interest (ROI) of the WPAFB 2009 dataset [3] proposed in [1], yielding state of the art performance. We also evaluate the influence of the individual components of the framework on the overall tracking performance.

2. RELATED WORK

Corresponding to the increasing interest in WAAS, the computer vision community has been investigating different approaches to multi-target tracking on WAMI data in the last years [1, 2, 4–10]. Many trackers rely on motion detections acquired using background subtraction or frame differencing and do not consider stationary objects. Data association is performed using the Hungarian algorithm [4, 8], multiple hypothesis tracking [6], or by an object-centric association method allowing sharing of detections among tracks [7].

While the above trackers only track moving objects, recent work focusses on enabling persistent tracking, i.e. track-

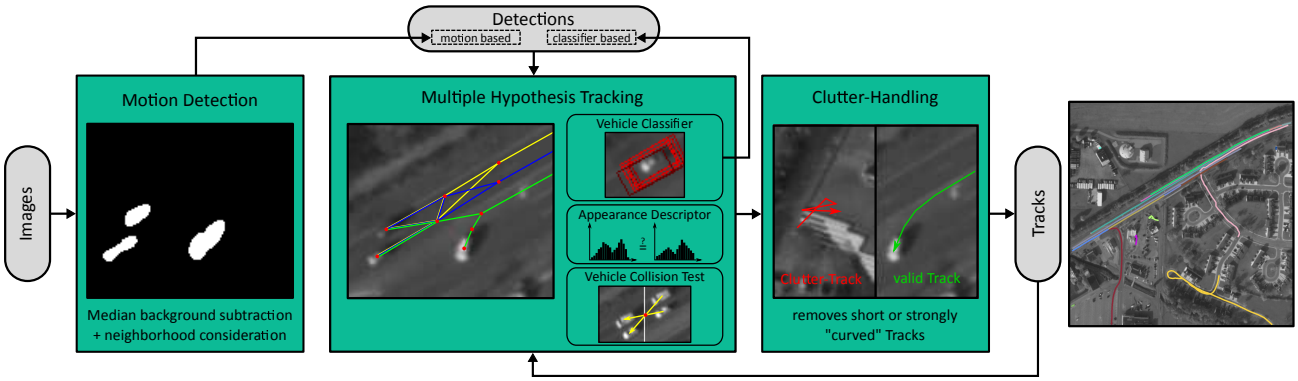

Fig. 1. Proposed framework for persistent WAMI tracking: Motion-based vehicle detections are generated via background subtraction and used as input for the MHT module. A special appearance descriptor and collision tests support the data association process. Missing motion detections caused by slow or stopping vehicles trigger the generation of classifier-based detections to facilitate persistent tracking. Clutter handling deletes invalid tracks to reduce the number of false alarms.

ing targets also if they become stationary. To handle the resulting long-term missing motion detections, Prokaj *et al.* [1] extend the capabilities of a detection-based tracker by combining it with a regression tracker that does not rely on detections. Chen *et al.* [2] couple a detection-based tracker with a local context tracker. To the best of our knowledge, MHT has not yet been extended for persistent tracking in WAMI data.

The alternative to using two trackers, i.e. recovering missing motion detections by means of an appearance-based detector and providing them directly to the detection-based tracker, remains challenging since only weak appearance information is available in WAMI data. Teutsch *et al.* [11], however, demonstrate that a sliding window classifier can efficiently support robust vehicle detection in WAMI data after classifier search space restriction. Furthermore, in [12] a systematic evaluation of appearance descriptors applicable to WAMI data is presented, with a combined descriptor of Local Binary Patterns and local variance showing best results. Neither the classifier-based detection approach presented in [11] nor the appearance descriptor proposed in [12] have been adopted for persistent WAMI tracking yet.

3. TRACKING FRAMEWORK

Fig. 1 illustrates the proposed *detect-before-track* framework for persistent WAMI tracking. The individual components of the framework are described in detail below.

3.1. Motion-based Vehicle Detection

For moving vehicle detection we use a median background subtraction approach which yielded best results in a survey on WAMI detection methods by Sommer *et al.* [13]: At each time step k a median background image I_{BG} is created from the six previous frames. To suppress errors due to imprecise image alignment and parallax effects a small neigborhood is considered for calculating the difference image D_k:

$$D_k(\mathbf{x}) = \min_{\Delta\mathbf{x}\in N} |I_k(\mathbf{x}) - I_{BG}(\mathbf{x}+\Delta\mathbf{x})|. \quad (1)$$

We choose N as a 7×7 neighborhood around the current pixel. To find the moving objects, D_k is binarized with a quantile threshold T_q [14]. After morphological operations, object blobs are obtained by connected component labeling. Blobs below a minimum blob size are rejected. The centroids of valid blobs are considered as detections (measurements).

3.2. Multiple Hypothesis Tracking

For data association between tracks and available vehicle detections we use MHT [15, 16] and follow the "track-oriented" MHT approach [17, 18]. Each vehicle gets its own track tree, where branches represent different possible vehicle trajectories, called track hypotheses. To estimate the track motion, we use a Kalman filter with a vehicle motion model with constant velocity and turn rate. The main idea in MHT is to delay data association until enough information is collected to resolve ambiguities. The gathered information is stored in a track score for each track hypothesis that can be calculated recursively for the curent time step k:

$$S(k) = S(k-1) + \Delta S(k). \quad (2)$$

Rather than relying only on motion information, e.g. how close a measurement is to the predicted position, we also integrate a special appearance descriptor to calculate the similarity between two vehicle positions (see Sec. 3.3). As Kim *et al.* in [19], we therefore split the track score update $\Delta S(k)$ in a motion $\Delta S_{mot}(k)$ and an appearance based part $\Delta S_{app}(k)$:

$$\Delta S(k) = \Delta S_{mot}(k) + \Delta S_{app}(k). \quad (3)$$

To keep the number of track hypotheses small, we use a tree depth of 2 and do not initialize a new track tree for every

measurement. Instead only measurements that are not used for updating a track hypothesis or that lie close to the image borders produce new track trees. To find the optimal global hypothesis in the current frame a multi-dimensional assignment problem has to be solved. The optimal global hypothesis is the best (in terms of maximum sum of track scores) set of non-conflicting track hypotheses. Similary to [20] we formulate this task as a Maximum Weighted Independent Set Problem and use the same algorithms as [19] for solving it.

3.3. Appearance Descriptor

For calculating the similarity probability between a track hypothesis appearance in image I_{k-1} and in I_k we adapt a descriptor presented in [12]: Around the previous and the current hypothesis position we generate a 32×16 pixel large bounding box which is oriented in vehicle motion direction. We subdivide each box in non overlapping blocks and extract histograms of uniform Local Binary Patterns (LBP) and local variance (VAR) [21] in each block. In order to avoid histogram sparsity, LBP and VAR are calculated with three different radii. Furthermore a gray-value histogram of the entire box is calculated. These histograms are concatenated to a descriptor that captures all available appearance information like local texture (LBP), local variance (VAR) and brightness (histogram). To calculate the appearance score update for the track hypothesis that uses the measurement i at time $k-1$ and the measurement j at time k the Hellinger distance is needed. The Hellinger distance $H(d^i_{k-1}, d^j_k)$ can be interpreted as dissimilarity probability and is used to calculate the similarity probability $P(S|d^i_{k-1}, d^j_k) = 1 - H(d^i_{k-1}, d^j_k)$ between the descriptor d^i_{k-1} at the old track hypothesis position and a descriptor d^j_k at the current track hypothesis position. Thus we calculate the appearance score update as follows:

$$\Delta S^{i \to j}_{app}(k) = \ln \frac{P(S|d^i_{k-1}, d^j_k)}{P(S|d^j_k, b^j_k)}, \tag{4}$$

where b^j_k is the descriptor at the position of the j^{th} measurement in the background image. According to that, our $\Delta S_{app}(k)$ is a ratio between the similarity to the previous vehicle appearance and the similarity to the background. If the background similarity is high, it is more likely that the assigned measurement is a false detection and vice versa.

3.4. Vehicle Collision Test

MHT produces many track hypotheses especially in scenes with dense traffic. To assist the algorithm in finding an optimal global hypothesis we included a vehicle collision test to check if different track hypotheses exclude each other. Fig. 2 illustrates the collision test with an example. The track tree for car I spawns two track hypotheses, 1 and 2. The track tree for car II holds the hypotheses 3 and 4. The optimal global hypothesis cannot contain the intersecting track hypotheses 2 and 3 at the same time, because this would represent a vehicle collision. If such collisions are detected, the information on conflicting hypotheses is passed on to the tracker by inserting additional edges in the MHT conflict graph ([19]).

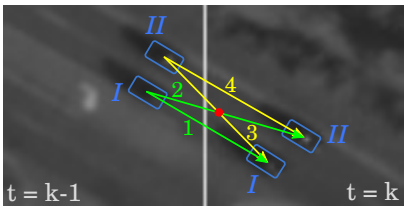

Fig. 2. Vehicle collision test: Arrows represent track hypotheses. 2 and 3 are in conflict because they represent a collision.

3.5. Classifier-based Detections and Persistent Tracking

Persistent Tracking cannot be done by a *detect-before-track* framework only relying on motion-based detections. Slow or stopped vehicles will not be detected by the median background subtraction approach, because they become part of the background image. To enable persistent tracking we therefore search for the missing detections with a vehicle classifier, if the track velocity becomes very small. We can effectively reduce the search space, because the position and orientation of the vehicle are approximately known from the Kalman filter state estimate. If the classifier finds valid vehicle detections, they are returned as input to the MHT-tracker, which then processes all input detections independently of whether they were generated by motion detection or with the classifier.

We use the same sliding window classifier as Teutsch *et al.* (see [11] for details), which consists of a Discrete Cosine Transform [23] descriptor and a Random Forest [24] classifier. The classifier is trained with scaled vehicle and non-vehicle samples extracted from the WPAFB 2009 dataset.

3.6. Clutter Handling

To reduce the number of false alarms the tracking framework validates finished tracks and deletes clutter tracks. Clutter tracks are non-vehicle tracks originating from false detections, which often appear at high buildings due to parallax effects or at image seams. In our framework, we delete tracks with a life time equal or shorter than two frames and tracks with a very short covered total distance. For remaining tracks with a duration lower than six frames, a total curvature is calculated by summing up the angles between the line segements of the track path. If P is the track path as a polygonal line with nodes $a_1, ..., a_n$ and α_i the angle between the vectors $a_{i+1} - a_i$ and $a_{i+2} - a_{i+1}$, then the total curvature κ is calculated as

$$\kappa(P) = \sum_{i=1}^{n-2} \alpha_i. \tag{5}$$

Method	precision	recall	f-score	N-MODA	S/T	B/T	MOTA
Chen *et al.* [2] (DBT+LCT)	**0.990**	0.606	0.752	0.600	**0.015**	**0.317**	0.599
Chen *et al.* [10] (DBT)	0.987	0.550	0.706	0.543	0.200	0.500	0.540
Prokaj *et al.* [1] (DBT+RGR)	0.960	0.539	0.690	0.516	0.237	1.022	0.512
Prokaj *et al.* [22] (DBT)	0.985	0.504	0.667	0.497	0.249	1.515	0.493
Reilly *et al.* [4] (DBT)	0.940	0.573	0.712	0.536	0.851	1.293	0.522
Proposed	0.932	**0.657**	**0.770**	**0.609**	0.373	1.005	**0.602**

Table 1. Tracking results in comparison to the literature. Best results are highlighted. DBT: detection-based trackers, DBT+LCT: DBT combined with local-context tracker, DBT+RGR: DBT combined with regression tracker.

Assuming that valid vehicle tracks have smooth trajectories with small total curvature, we reject tracks with total curvatures greater than $\frac{\pi}{3}$ as invalid.

4. EXPERIMENTAL RESULTS

4.1. Setup and Metrics

For evaluating our tracking approach and comparing it to other trackers, we use an ROI of size 1408×1408 pixels of the WPAFB 2009 dataset. To ensure comparability of results, we use the same sequence and evaluation code also used in [1, 2], provided to us by the authors of [1]. The chosen sequence comprises 1025 frames and 410 ground truth (GT) tracks. We quantitatively evaluate detection and tracking performance with the common metrics recall, precision, f-score, N-MODA and MOTA [25]. S/T and B/T denote the number of ID-switches and breaks per GT-track, respectively.

4.2. Results

Table 1 shows the quantitative results in comparison to the literature, as published in [2]. Compared to other trackers, the proposed tracker yields highest f-score and N-MODA and therefore best overall detection performance. We operate our tracker at lower precision and higher recall (i.e. we suppress fewer false positives in order to lose fewer true positives). Regarding ID-switches, some other trackers show better performance. However, the proposed approach still yields best MOTA, which is a combined measure for all tracking errors (false negatives, false positives, and ID-switches), and hence best overall tracking performance of the compared trackers.

Table 2 shows the results for different test configurations T_i of the proposed tracking framework after deactivating individual components, thus demonstrating their influence on the total tracking performance. Best performance is achieved by combining all components (test T_1). Comparing tests T_2, T_3, and T_5 to test T_1 shows the contributions of appearance descriptor, vehicle collision tests, and clutter handling, respectively. These components could be easily integrated into other tracking frameworks to improve performance. Comparing T_4 to T_1 shows that activating the generation of classifier-based detections for persistent tracking improves the MOTA value from 0.543 to 0.602 (10.9%). In comparison, combining a detection-based tracker with a local context tracker (see DBT+LCT in Table 1) improves MOTA from 0.540 to 0.599 (10.9%), while combination with a regression tracker (see DBT+RGR in Table 1) improves MOTA from 0.493 to 0.512 (3.9%). Therefore, classifier-based detection generation yields equal or superior improvements and qualifies as a valid alternative to combining two trackers for persistent WAMI tracking.

The average computation time of our unoptimized code for tracking was 1.25 seconds per image on a desktop with 3.33 GHz CPU and 4 GB memory (see [2] for computation time of other approaches).

Components	T_1	T_2	T_3	T_4	T_5
App. descr.	✓	✗	✓	✓	✓
Veh. coll. test	✓	✓	✗	✓	✓
Classif. det.	✓	✓	✓	✗	✓
Clutter hand.	✓	✓	✓	✓	✗
MOTA	0.602	0.565	0.582	0.543	0.289

Table 2. MOTA for various tests after deactivating individual framework components (appearance descriptor, vehicle collision tests, classifier-based detections, clutter handling).

5. CONCLUSIONS

Current approaches for persistent WAMI tracking often combine a detection-based tracker with a second tracker to compensate for missing motion detections when vehicles slow down or stop. We developed an alternative MHT framework for persistent multi-target tracking in WAMI data that recovers missing detections with a classifier and depends only on a single tracker, with the advantage of avoiding additional complexity introduced by handling two trackers. We integrated several extensions that could also be used to improve performance of other trackers, e.g. a special appearance descriptor that assesses vehicle similarities. The proposed framework achieves best overall detection (f-score) and tracking performance (MOTA) in comparison to other state-of-the-art WAMI trackers. In the future, we plan to integrate improved split-merge-handling and optimize computation time.

References

[1] J. Prokaj and G. Medioni, "Persistent tracking for wide area aerial surveillance," in *CVPR*, 2014.

[2] B. Chen and G. Medioni, "Exploring local context for multi-target tracking in wide area aerial surveillance," in *WACV*, 2017.

[3] U.S. Air Force Research Laboratory (AFRL), "WPAFB 2009 dataset," `https://www.sdms.afrl.af.mil/in-dex.php?collection=wpafb2009`.

[4] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," in *ECCV*, 2010.

[5] J. Xiao, H. Cheng, H. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *CVPR*, 2010.

[6] M. Keck, L. Galup, and C. Stauffer, "Real-time tracking of low-resolution vehicles for wide-area persistent surveillance," in *WACV*, 2013.

[7] I. Saleemi and M.k Shah, "Multiframe many–many point correspondence for vehicle tracking in high density wide area aerial videos," *International Journal of Computer Vision*, 2013.

[8] C. Aeschliman, J. Park, and A. C. Kak, "Tracking vehicles through shadows and occlusions in wide-area aerial video," *IEEE Transactions on Aerospace and Electronic Systems*, 2014.

[9] A. Basharat, M. W. Turek, Y. Xu, C. Atkins, D. Stoup, K. Fieldhouse, P. Tunison, and A. Hoogs, "Real-time multi-target tracking at 210 megapixels/second in wide area motion imagery," in *WACV*, 2014.

[10] B. Chen and G. Medioni, "Motion propagation detection association for multi-target tracking in wide area aerial surveillance," in *AVSS*, 2015.

[11] M. Teutsch and M. Grinberg, "Robust detection of moving vehicles in wide area motion imagery," in *CVPR Workshops*, 2016.

[12] M. Cormier, L. W. Sommer, and M. Teutsch, "Low resolution vehicle re-identification based on appearance features for wide area motion imagery," in *WACVW*, 2016.

[13] L. W. Sommer, M. Teutsch, T. Schuchert, and J. Beyerer, "A survey on moving object detection for wide area motion imagery," in *WACV*, 2016.

[14] F. Y. Shih, *Image Processing and Pattern Recognition: Fundamentals and Techniques*, Wiley, 2010.

[15] D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, 1979.

[16] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*, Artech House Inc, 1999.

[17] T. Kurien, "Issues in the design of practical multitarget tracking algorithms," in *Multitarget-Multisensor Tracking: Advanced Applications*, Y. Bar-Shalom, Ed. Artech House, 1990.

[18] I. J. Cox and S. L. Hingorani, "An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996.

[19] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *ICCV*, 2015.

[20] D. J. Papageorgiou and M. R. Salpukas, "The maximum weight independent set problem for data association in multiple hypothesis tracking," in *Optimization and Cooperative Control Strategies*, 2009.

[21] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[22] J. Prokaj, M. Duchaineau, and G. Medioni, "Inferring tracklets for multi-object tracking," in *CVPR Workshops*, 2011.

[23] H. K. Ekenel and R. Stiefelhagen, "Local appearance based face recognition using discrete cosine transform," in *EUSIPCO*, 2005.

[24] L. Breiman, "Random forests," *Machine Learning*, 2001.

[25] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan, "The clear 2006 evaluation," 2007.