Robust Detection of Moving Vehicles in Wide Area Motion Imagery

Michael Teutsch* and Michael Grinberg

Fraunhofer IOSB, Karlsruhe, Germany
michael.grinberg@iosb.fraunhofer.de

Abstract

Multiple object tracking in Wide Area Motion Imagery (WAMI) data is usually based on initial detections coming from background subtraction or frame differencing. However, these methods are prone to produce split and merged detections. Appearance based vehicle detection can be an alternative but is not well-suited for WAMI data since classifier models are of weak discriminative power for vehicles in top view at low resolution. We introduce a moving vehicle detection algorithm that combines 2-frame differencing with a vehicle appearance model to improve object detection. Our main contributions are (1) integration of robust vehicle detection with split/merge handling and (2) estimation of assignment likelihoods between object hypotheses in consecutive frames using an appearance based similarity measure. Without using any prior knowledge, we achieve state-of-the-art detection rates and produce tracklets that considerably simplify the data association problem for multiple object tracking.

1. Introduction

The term Wide Area Motion Imagery (WAMI) denotes video data that is acquired by moving airborne cameras at a low frame-rate of about 1 Hz and a high ground coverage of several square kilometers per image. Such kind of data can be used to solve wide area surveillance tasks such as traffic monitoring, detection of abnormal behavior, or border security. However, automatically analyzing WAMI data is ambitious for numerous reasons such as camera motion, nonexistent color channels, or the large amount of data with up to hundreds of megapixels per image.

The mentioned surveillance tasks share the need for accurate tracking of moving objects in the scene. This is challenging due to the large displacement of objects between two consecutive images as a result of the low frame rate and due to large number of objects with several hundreds of vehicles per image in urban traffic scenarios. Existing tracking algorithms are based on detections coming from background subtraction or frame differencing [38, 28, 29]. As illustrated in Fig. 1, these detection methods are prone to produce both false positive (FP) detections caused by image misalignment, parallax effects, or split detections and false negative (FN) detections that occur due to weak contrast, occlusions, or merged detections. In current literature, these shortcomings of the detection methods are directly transferred to the tracking module and handled there implicitly, e.g., by introducing traffic models [29].

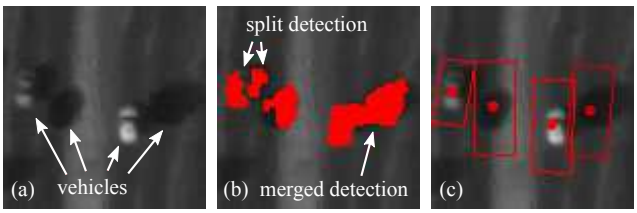

Figure 1. Motivation: (a) image region of 100 × 100 pixels with four vehicles, (b) object detection using background subtraction [28], and (c) the proposed object detection approach.

In this paper, we introduce a robust vehicle detection approach for WAMI data. Robustness is difficult to achieve as each vehicle covers only about 10 × 20 pixels in the image and thus appearance based classifier models are expected to be of weak discriminative power. In order to still apply this weak vehicle model, we avoid using context knowledge as proposed in [31, 20] but instead build up on a 2-frame differencing approach. We only assume that each moving vehicle gives us at least two blobs in the difference image. The displacement and the size of these blobs is used to reduce the search space for sliding window based vehicle detection. Our contributions are (1) integration of robust vehicle detection using sliding windows, (2) explicit handling of merged detections using k-means, (3) implicit avoidance of split and ghost detections that are typical for 2-frame differencing, and (4) estimation of assignment likelihoods between object hypotheses in consecutive frames using an appearance based similarity measure.

*M. Teutsch was with the Fraunhofer IOSB, Karlsruhe, Germany. He now works for Airbus Defence and Space Optronics, Oberkochen, Germany, e-mail: michael.teutsch@airbusds-optronics.com.

These assignment likelihoods provide association proposals that can significantly simplify the data association problem for multiple object tracking. For our evaluations, we use the WPAFB 2009 dataset provided by the U.S. Air Force Research Lab (AFRL) [1] that comes with around 18,000 fully ground-truthed vehicle tracks.

Several authors already proposed to improve multiple object tracking in WAMI data by integration of appearance information [3, 24, 5, 27, 23]. However, appearance is used to support tracks of slow, stopping or, occluded vehicles and thus the tracks are assumed to exist already. In this paper, we use appearance based vehicle detection to initialize tracks and to simplify the data association problem.

The remainder of this paper is organized as follows: literature is reviewed in Section 2. The proposed object detection approach is presented in Section 3. Experimental results are given in Section 4. We conclude in Section 5.

2. Related Work

In the last few years, several authors proposed processing chains that are able to cope with the challenges of WAMI data and some methods became common. Compensation for camera motion is achieved by assuming that the scene can be approximated by a ground plane and estimating the parameters of a global camera motion model (homography) [17]. Therefore, sparsely distributed local image features such as Harris corners [16] are detected in two consecutive images and descriptors are used to find matching pairs of features to calculate frame-to-frame homographies [25, 38, 28, 29]. These global motion models are then used to align consecutive images.

2.1. Moving Object Detection by Segmentation

Background subtraction [25, 28, 26, 27] and frame differencing [38, 18, 29, 5] are the most popular approaches for detecting moving objects in WAMI data. Actually, both methods rather are segmentation approaches: either two or three aligned images are used to calculate a difference image in which bright pixels represent changes between the images. These changes are assumed to come from the displacement of moving vehicles. Bright pixels are thresholded and clustered generating blobs. Morphological operations are applied to refine these blobs and small blobs are removed as they are assumed to be the result of noise in the difference image. Each detected object is then represented by the centroid of its related blob.

In general, noise in the difference image is a big problem. It can occur due to image misalignment, unhandled parallax effects, sudden changes in camera gain, and moving mosaic seams. This can result in a large number of FP detections. Several methods were proposed for reducing the noise: only 10 images are used to learn short-term median background models [28] instead of standard Gaussian Mixture Models (GMM) [25]. In order to suppress noise coming from parallax effects, pixel neighborhoods can be considered for difference image calculation [26] or the background gradients can be subtracted from the difference image [28]. However, misalignment is still a problem when using background models. So, the influence of misalignment is minimized by frame differencing: Saleemi and Shah [29] propose to directly subtract two consecutive images without any background model but as already mentioned the explicit handling of ghosting is difficult. Xiao *et al*. [38] introduce the subtraction of three consecutive images to avoid ghosting. This method recently became popular [18, 5]. Artifacts originating from moving mosaic seams and changes in camera gain can be reduced by applying a box filter to the difference image before objects are segmented [18]. However, frame differencing is prone to produce split detections especially for slowly moving objects.

2.2. Vehicle Detection

In this subsection, we focus on object detection methods based on sliding windows. Their applicability was already demonstrated for several tasks such as face detection [37], human detection [10], and on-road vehicle detection [33]. Although the application of sliding windows is an *exhaustive search*, several optimization methods were proposed in order to achieve real-time capabilities [6].

Some authors already proposed sliding window based vehicle detection in aerial videos [8, 36, 34] and even in wide area aerial videos [21, 31, 20]. All methods, however, have in common that learning a classifier model for vehicle appearance at low resolution in top view images leads to a model of weak discriminative power. Even for popular descriptor/classifier combinations such as Histograms of Oriented Gradients (HOG) + Support Vector Machines (SVM) [8, 36] and Haar features + AdaBoost [13, 34], the classifier only learns a rectangular shape. Considering color information can be very helpful [3, 19] but is not possible as the WPAFB dataset only provides monochromatic images. Furthermore, not only the scale but also the orientation of vehicles can vary in top view videos. This is different compared to human detection where usually at least head and torso are assumed to be in upright position.

In order to avoid a large number of FP detections, the search space of the sliding window has to be reduced. The spatial search area can be limited to areas of independent motion which has been proposed only for videos with high frame rate so far [8, 34]. Furthermore, context knowledge about the road network can be involved either by using a Geographic Information System (GIS) [38] or automatic road detection using vehicle tracks [31]. One limitation of road context is that vehicles might be missed, if they do not use the main roads. The orientation search space can be

reduced using the motion direction of individual objects in case of high frame rate videos [35] or by deriving dominant motion directions from detection locations over a certain period of time (temporal context) [20]. The latter one, however, can only be applied to busy streets where stable detection statistics can be determined. Finally, the scale search space can be limited. With known Ground Sampling Distance (GSD), the scale of the scene can be normalized and only few different vehicle sizes in a range between standard cars and large trucks need to be considered [34].

3. Vehicle Detection

A novel approach for robust vehicle detection using 2-frame differencing with an integrated vehicle appearance model is presented in this section. However, before vehicles are detected, images are aligned in order to compensate the image sequence for camera motion. Therefore, we use SIFT-like features to detect corresponding Harris corners in consecutive images to estimate frame-to-frame homographies as global camera motion model. RANSAC is applied to reject improper feature correspondences. Images are then warped using a projective transformation.

3.1. 2-Frame Differencing

In order to effectively reduce the search space for subsequent vehicle detection, we apply 2-frame differencing [29] since noise in the difference image coming from image misalignment and parallax effects can be minimized compared to 3-frame differencing or background subtraction. Inspired by preprocessing approaches for change detection, we apply histogram matching (also: histogram specification) [14] to reduce the global noise coming from rapid changes in camera gain and local mean gray-value normalization [30] to handle local noise that occurs due to moving mosaic seams. Please note that histogram matching is not applied to the large stitched images but to the cropped Areas of Interest (AOI) that we consider in our experiments in Section 4. In order to decreases the influence of noise that originates from misalignment and parallax effects, we consider pixel neighborhoods for difference image calculation as described by Saur *et al*. [30] (Eq. 4-6). Finally, we perform postprocessing by calculating the mean gradient magnitudes of both images and subtracting them from the difference image [29]. The difference image is then binarized by applying quantile based thresholding [32] instead of Otsu as proposed in [29]. Both thresholding techniques are adaptive, but we achieved better results by using the quantile based method.

Split detections and ghosting are the main problems of 2-frame differencing. Ghosting means occurrence of phantom detections (ghosts) due to wrong assignment of blobs from the difference image to the original images. This is possible since each moving object produces two blobs in the difference image (one at its old and one at its new location), but we do not know which location is old and which is new, i.e. which image contributed to which blob. This can be handled explicitly by using heuristics [29]. In Section 3.3, however, we handle both problems implicitly.

3.2. Vehicle Model

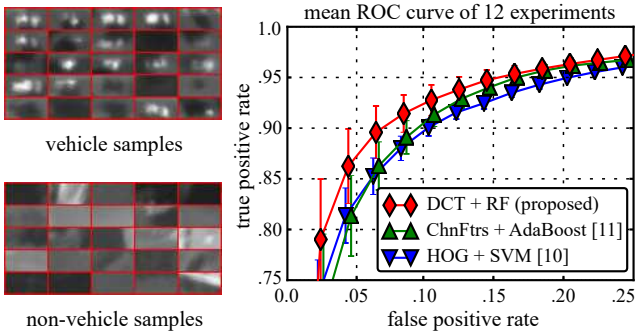

Figure 2. Samples from the classifier training data set and the resulting mean ROC curves for 12 experiments.

A classifier model must be learned before sliding window based vehicle detection can be applied. Therefore, we use the Ground Truth (GT) to extract four training and test datasets for classification at different regions in the WPAFB dataset. These regions do not overlap with the regions used for the experiments in Section 4. In the original GT, each track is given by a set of points, where each point represents the vehicle's position in one frame. We use the track motion direction of the GT to rotate vehicle samples in horizontal direction, semi-automatically generate a bounding box for each sample with an appropriate size and scale this box to 32×16 pixels as visualized in Fig. 2. The aim of rotating the vehicle samples is to learn a classifier model of higher discriminative power compared to unrotated samples since less background is included in the training data. A similar idea has been described by Shi *et al*. [31].

We propose to use a descriptor that is based on Discrete Cosine Transform (DCT) [12] and a Random Forest (RF) classifier that is robust to bad training samples due to the bagging learning strategy [7]. Those bad samples can be heavily blurred objects or vehicles without recognizable shape due to shadows. The RF gives us probabilities for object existence as confidence values. In our evaluation, we compare the proposed model with the popular models Histograms of Oriented Gradients (HOG) + Support Vector Machine (SVM) [10] and Integral Channel Features (ChnFtrs) + AdaBoost classifier [11]. Each of the four datasets contains between 100 and 500 vehicle and non-vehicle samples, respectively. In order to evaluate the three different descriptor/classifier combinations, each classifier is trained on each dataset (i.e. four models for each descriptor/classifier) and evaluated with the other three datasets (i.e. 12 experiments for each descriptor/classifier). We add 2,000 non-

vehicles to each evaluation dataset verify the robustness of the classifier model. The results are visualized as a mean Receiver Operating Characteristic (ROC) curve including the standard deviation. The proposed DCT + RF model outperforms the other methods especially for parametrization that generates low FP rates which is important for the sliding window as we expect much more non-vehicles than vehicles. The final vehicle model is trained using samples from all four datasets together.

3.3. Vehicle Detection and Tracklet Hypotheses

In this subsection, we consider the special case of single vehicle detection. The general case for multiple vehicles including split and merge handling is presented in Section 3.4. The main assumption for our proposed vehicle detection approach is that each moving object produces at least two blobs in the difference image when 2-frame differencing is applied. This is the case for fast and slowly moving vehicles if the contrast compared to the background is strong enough. We then assume that each blob contains exactly one vehicle or no vehicle at all. These assumptions are necessary since naïvely applying a sliding window in all possible scales and orientations would lead to a large number of FP detections at structures that look similar to vehicles such as road markings, curbs, or buildings. However, 2-frame differencing is sufficient to derive enough information so that we can achieve robust vehicle detection without using context knowledge such as dominant track directions [31] or road networks [20].

The proposed approach is illustrated in Fig. 3. For two consecutive images I_t and I_{t+1} at time t and $t+1$, 2-frame differencing produces two blobs (white color). Initially, we do not know which image contributes to which blob (ghosting) and whether there is a vehicle inside any blob or not (potential FP). However, in order to prove object existence and achieve robustness, we not only aim to detect the object in one image but also to find the correctly matching detections between two frames. Furthermore, as soon as the correct match is found it can be interpreted as a tracklet hypothesis between the two frames. A sliding window (green box) is then used around both blob locations in each image in order to locate the vehicle in both images. The orientation of the vehicle is assumed to correspond to the motion direction, so we orient the sliding window according to the line connecting both blobs (red line). The search is restricted to two search spaces s_1 and s_2 (black boxes). They are determined from the blob position and size. In each image multiple detection hypotheses d_t^i and d_{t+1}^j appear. Bright red color indicates a high classifier confidence. This confidence is given by probabilities $P(d_t^i)$ and $P(d_{t+1}^j)$ of the RF classifier where a high probability represents a high certainty that there is a vehicle at the current sliding window position.

Now, we aim to find the best matching detections between I_t and I_{t+1} by using the similarity probability $P(S|d^i, d^j)$ for two detection hypotheses d^i and d^j. Therefore, we calculate an appearance descriptor adopted from face recognition [2]: we subdivide each detection window in non overlapping blocks of 8×8 pixels and extract histograms of uniform Local Binary Patterns (LBP) and local variance (VAR) [22] in each block (each histogram has 59 bins). In order to avoid histogram sparsity, LBP and VAR are calculated with three different radii $R \in \{0.5, 1.0, 2.0\}$. Furthermore, we calculate the gray-value histogram of the entire window (256 bins). These histograms are concatenated to a descriptor of size 1,200 and normalized. In this way, we can capture all available appearance information: local texture (LBP), local variance (VAR), and brightness (histogram). The Hellinger distance $H(d^i, d^j)$ [4] can be interpreted as dissimilarity probability and is used to calculate the similarity probability $P(S|d^i, d^j) = 1.0 - H(d^i, d^j)$ between two detection hypotheses d^i and d^j. So, the confidence of each tracklet hypothesis $\theta_{t,t+1}^{i \to j}$ can be calculated by the likelihood

$$p(\theta_{t,t+1}^{i \to j}) = P(S|d_t^i, d_{t+1}^j) \cdot P(d_t^i) \cdot P(d_{t+1}^j) \qquad (1)$$

where detection i and j do not come from the same search space s. Tracklet $\theta_{t,t+1}$ that represents the best matching detection in Fig. 3 is then given by

$$\theta_{t,t+1}^* = \arg\max_{i,j}(p(\theta_{t,t+1}^{i \to j})), \qquad (2)$$

where $*$ indicates the best match for this single vehicle detection problem.

Up to now, we did not discuss image rescaling which is important to detect vehicles of different size. Rescaling is introduced by calculating the best tracklet for each scale and keeping the tracklet with highest confidence. Inspired by [34] we use three different scales for (1) standard cars, (2) large cars and small trucks, and (3) trucks and busses.

The proposed method is able to implicitly solve the ghosting problem, achieve robust vehicle detection under consideration of different scales and orientations, and provide tracklet hypotheses that can simplify the data association problem of multiple object tracking. Considering all detection hypotheses coming from s_1 and s_2 in I_t and I_{t+1} is necessary to find the best match even in dense traffic, where there may be vehicles in each search space but we want to find the only correct match.

3.4. Split and Merge Handling

As we usually do not have such simple cases as shown in Fig. 3, we aim to achieve robust vehicle detection in presence of many vehicles and even under split and merge conditions that are likely to occur when 2-frame differencing is applied. An example with four vehicles is visualized in Fig. 4 (leftmost image). There are split detections

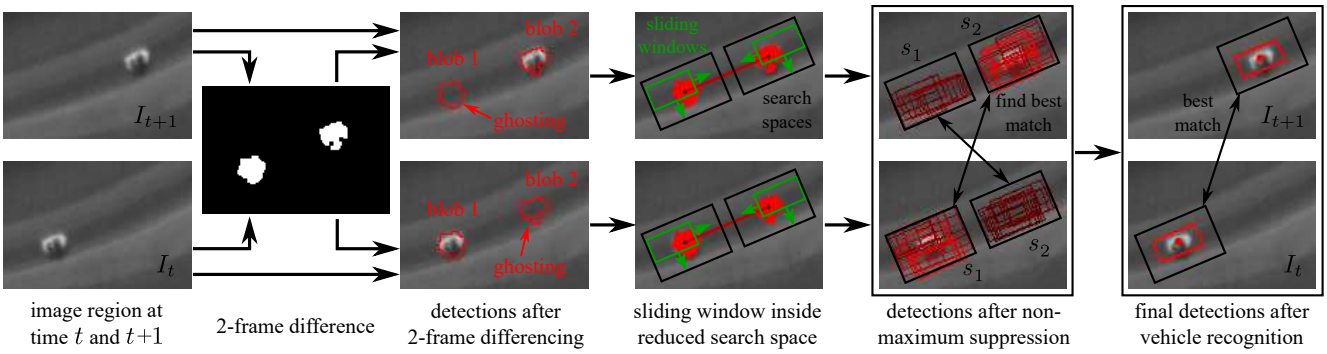

Figure 3. Vehicle detection: for image I_t and I_{t+1}, the 2-frame difference produces two blobs. Position, size, and displacement of the blobs are used to determine two search spaces (black boxes) for a sliding window (green boxes). The best matching detections between I_t and I_{t+1} are found by using a similarity measure and their centers are used as start and end point in order to generate a tracklet hypothesis.

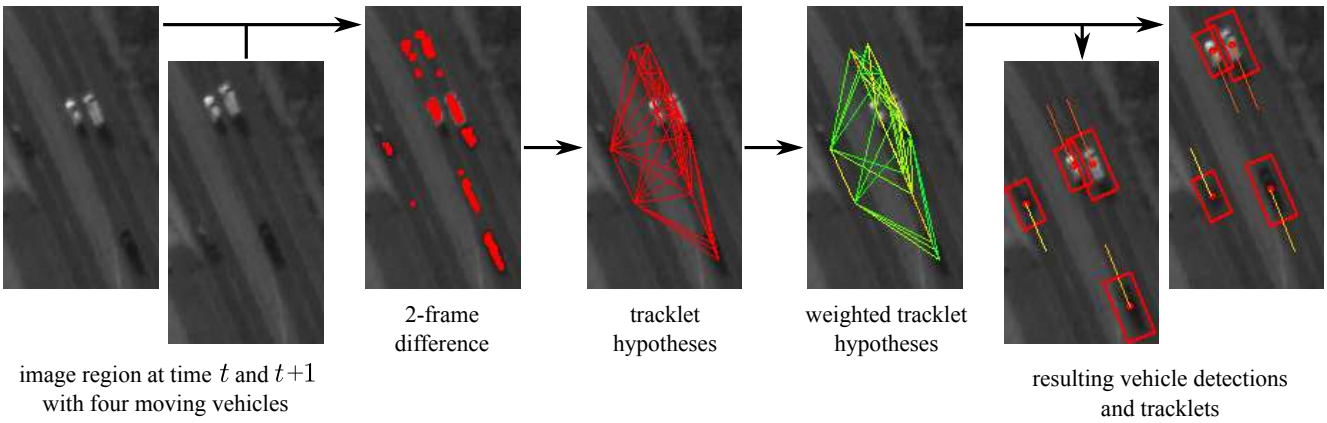

Figure 4. Multiple vehicle detection is achieved by interconnecting all blobs with a distance less than 100 pixels and by analyzing all interconnected blobs as described in Fig. 3. The tracklet likelihoods are visualized in colors (red = high, green = low likelihood). Weak tracklets are rejected. The tracklets can be used for multiple object tracking. By solving a local optimization problem we generate the resulting vehicle detections that are used in our experiments in Section 4.

and very small blobs due to weak contrast. Hence, rejecting small blobs can cause FN detections. So, we interconnect all blobs whose centroids have a Euclidian distance of 100 pixels (i.e. a velocity around 90 km/h or 56 mph) or less as shown in Fig. 4 (middle image). This results in a graph-like representation where each connection (or edge) illustrates a tracklet hypothesis. The likelihood for each tracklet hypothesis is calculated by using the approach presented in Fig. 3. In Fig. 4, these likelihood values are visualized in colors: red and orange indicate a high likelihood while green represents a low likelihood. If there are different vehicles inside the search spaces (potential mismatch), low tracklet likelihoods are expected either due to weak appearance similarity or due to low classifier confidence resulting from the incorrect orientation of the sliding window. In Fig. 4, we see another advantage of using oriented sliding windows for vehicle detection: if we would use bounding rectangles without rotating them as it was done in [29], one of the two bright vehicles that drive close to each other would probably be considered as a duplicate or split detection because of the strong overlap of the bounding rectangles.

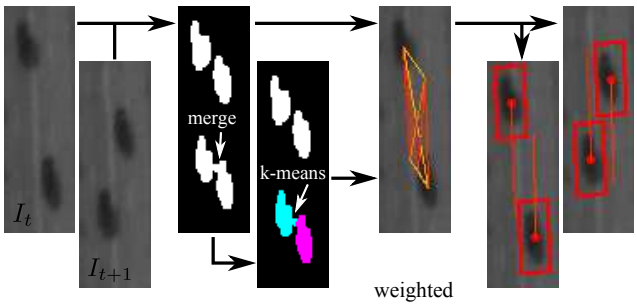

Figure 5. Merges are handled by splitting large blobs using k-means. New tracklet hypotheses that originate from this split are added to the graph and processed as described in Fig. 4.

In general, split detections can be handled by applying a non-maximum suppression (NMS) to detect all tracklets that overlap or cross each other and reject the ones with

lower likelihood. Merge handling is visualized in Fig. 5. Blobs inside the difference image that exceed a certain diameter (we use 25 pixels which is about 2.5 times the expected width of a standard car) are split using k-means with $k = 2$. In Fig. 5 this leads to generation of two new blobs that are highlighted in cyan and magenta color. The tracklet hypotheses originating from these two blobs are added to the graph.

A large vehicle will cause undesired merge handling. In such cases we expect the classifier to detect this vehicle multiple times at a similar position so that split handling can correct the erroneous merge assumption.

3.5. Extracting Detections for Evaluation

The resulting tracklets (weighted tracklet hypotheses) are transferred to a multiple object tracking algorithm that takes sets of interconnected tracklets as an input instead of pure detections. In WAMI data with a low framerate of 1 or 2 Hz this considerably simplifies the data association problem, which is particularly complex for scenes with several hundreds of moving vehicles. Precalculated tracklet likelihoods additionally facilitate solving of the data association problem. The tracking algorithm, however, is not discussed in this paper. In order to measure the performance of the vehicle detection and to compare it with state-of-the-art detection methods, we thus need to derive detections from the sets of interconnected tracklets. To do so we seek for a subset of tracklets that have high likelihoods and may co-exist w.r.t certain constraints (i.e. no crossing tracklets and no detection sharing by two tracklets). This is done by solving a local optimization problem for each local graph of interconnected tracklets. As a result, a set of detections can be derived for each frame as shown in Fig. 4 (rightmost image).

Interconnecting detections as described in Section 3.4 and shown in Fig. 4 may lead to large local graphs in areas with dense traffic. Solving the optimization problem by considering all possible solutions may thus become quite inefficient. In order to accelerate this step, we use a randomized greedy algorithm that is described in detail in the appendix of this paper. This randomized greedy algorithm is able to find a good solution even for large graphs of connected tracklets where we cannot consider all possible solutions. As mentioned above, it is only used for generating detections that can be evaluated in Section 4 and does not belong to the proposed vehicle detection approach.

4. Experimental Results

For our experiments, we use subregions of the WPAFB 2009 dataset that provides ground truth for detections and tracks. In the *test* sequence that lasts for approximately 7 minutes, we crop the AOIs 34, 40, and 41 as proposed by Basharat *et al*. [5]. The most important test cases are covered: (1) low/intermediate traffic density with homogeneous background in AOI 34, (2) dense traffic with slowly moving and stopping vehicles at an intersection in AOI 40, and (3) low traffic density in an urban area with textured background (trees, buildings) in AOI 41. We do not consider persistent tracking [24, 27] but only focus on moving objects. Hence, we removed stopping or parked vehicles from the GT. There are 459 tracks with 27,240 individual detections in AOI 34, 696 tracks with 41,943 detections in AOI 40, and 266 tracks with 12,426 detections in AOI 41.

We compare our proposed vehicle detection method to state-of-the-art frame differencing and background subtraction algorithms. 2-frame differencing with explicit avoidance of ghosting [29] is the baseline approach. Furthermore, we compare with 3-frame differencing [38], 3-frame differencing with additional box filter applied to the difference image [18], 10-frame median background subtraction with background gradient suppression [28], and background subtraction with an Interval GMM that considers pixel neighborhoods [26]. In order to provide a fair comparison, we apply exactly the same image alignment approach, pre-processing methods (histogram matching and local mean gray-value normalization), and thresholding technique as described in Section 3. We avoid comparing our method to vehicle detection approaches that use context knowledge [31, 20]. Although these methods derive their knowledge directly from the tracks, the detection performance highly depends on the tracking algorithm and the traffic density.

There are two main parameters: the minimum blob size and the threshold that is used to binarize the difference images. We fix the blob size after optimization and vary the binarization threshold to generate precision-recall curves. Typical values for the minimum blob size are 30 pixels for 2-frame differencing, 60 for 3-frame differencing, and 70 for background subtraction. 2-frame differencing needs this low minimum blob size since slowly moving vehicles produce only small blobs. We consider a detection as true positive (TP) if the point-to-point distance to the next GT object is less than 20 pixels (i.e. 5 m). This large distance is necessary since shadows can severely pull away detections from the GT point that is located at the object center. Each GT object is assigned only once, so all additional detections nearby are counted as FP detections. Consequently, for a merged detection we consider one detection as TP and obtain FNs for all remaining GT objects within the merge region.

The results are shown in Fig. 6. We significantly improve both precision and recall compared to the baseline approach [29]. The reason is that we reduce the minimum blob size to only 5 pixels (reduces FNs) but reject most of the emerging FPs at the same time by using the vehicle detector. We clearly outperform the other methods with respect to precision. However, the rather high FN rate of the baseline

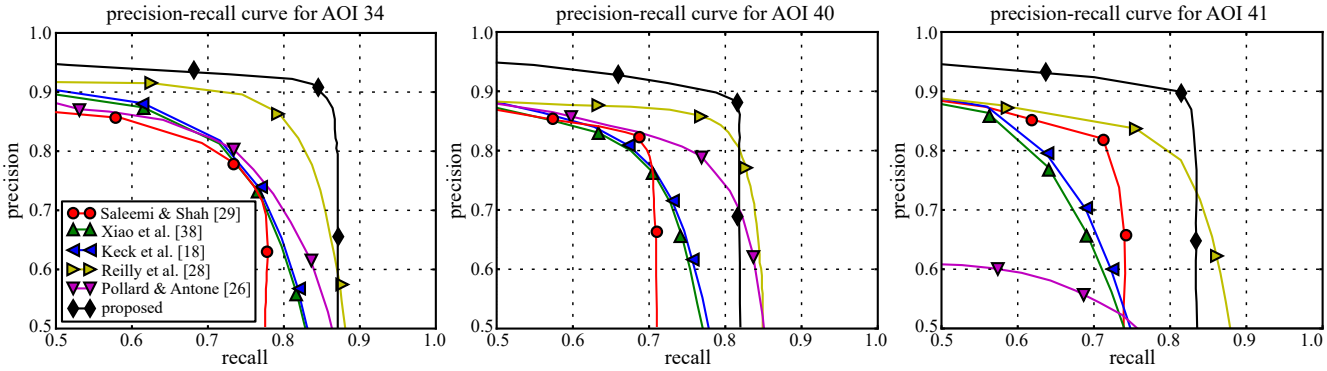

Figure 6. Detection results given as precision-recall curves for AOI 34, AOI 40, and AOI 41. The baseline approach [29] is significantly improved by the proposed method. Other approaches taken from the literature are outperformed with respect to precision.

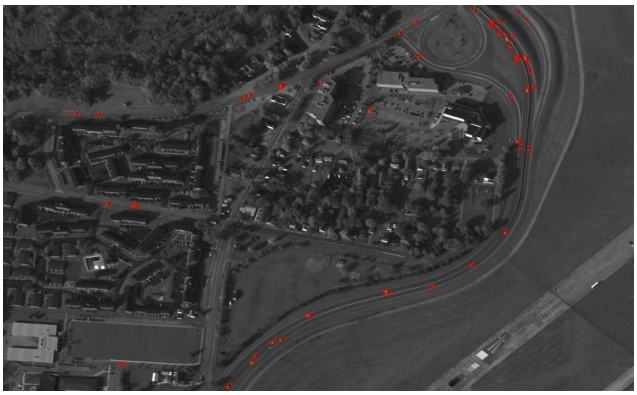

Figure 7. Example detections (red boxes) and ground truth (green dots) in AOI 34.

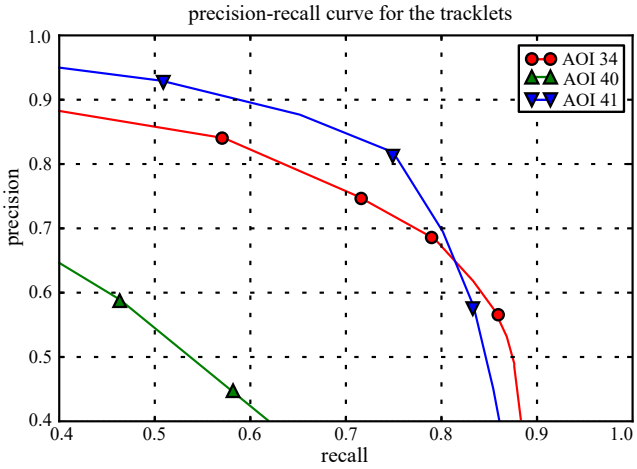

Figure 8. Precision and recall of the resulting tracklets without optimization.

method cannot be handled by our approach and limits our recall. Compared to the second best method [28] and with a fixed recall of 0.8, our approach produces about 2,000 less FPs in AOI 34, 4,000 in AOI 40, and 750 in AOI 41. The mean point-to-point distance of all TPs to the GT is similar with 5.45 pixels compared to 5.75 pixels. In Fig. 7, some example detection results (red boxes) and GT objects (green dots) for AOI 34 are shown.

The precision-recall curve for the tracklets is visualized in Fig. 8. We used the GT tracks to determine GT tracklets between pairs of frames. The local optimization was not applied here since tracking algorithms are supposed to use the entire set of tracklets in combination with a longer temporal sliding window [29, 9] to achieve a better solution for the association problem. Promising precision-recall curves are achieved for AOI 34 and AOI 41. However, numerous ambiguities in AOI 40 demonstrate that accurate tracking is still challenging in scenes with dense traffic. In order to show the suitability of our detection method for multiple object tracking, we connected tracklets with overlapping detections (red boxes) using a greedy algorithm and plot all tracks with a length of 50 frames or more in Fig. 9. Each track is visualized in a different color so that it is possible to distinguish between different object IDs. Connecting the tracklets to tracks based on the detection overlap is obvious since each moving object gets two identical detections per timestep t: one from the tracklet between frame $t-1$ and t and one from the tracklet between frame t and $t+1$. So, we can expect a strong overlap between the two detections and can use this overlap to connect the tracklets. Even without using a motion model or track linking, we can generate a large number of correct and complete tracks. Furthermore, we can see that tracklets are connected accurately even during non-linear vehicle motion such as U-turns. The reason is that vehicle appearance does not change much between two consecutive frames and, thus, vehicles are still correctly re-identified. All these observations indicate that the data association problem for multiple object tracking can be significantly simplified using the proposed vehicle tracklets.

5. Conclusions and Outlook

In this paper, we integrated a novel robust vehicle detection approach to extend and improve 2-frame differenc-

Figure 9. Example tracks for AOI 34 generated by simple greedy connection of the tracklets.

ing in WAMI data. This approach not only outperforms other state-of-the-art detection methods w.r.t. precision but also provides weighted tracklet hypotheses in consecutive frames that can considerably simplify data association for multiple object tracking. This simplification is of interest as the frame rate in WAMI data is usually only between 1 and 2 Hz and thus the data association problem can become very complex in scenes with several hundreds of moving vehicles. One drawback of our approach is the large number of FNs that occur during 2-frame differencing and cannot be handled with the proposed approach. This number could be reduced by introducing persistent tracking [27] in a way that appearance information is combined with the motion model. In this manner, the appearance model can be used to detect stopping vehicles, too.

Appendix: Randomized Greedy Algorithm

The idea of the proposed vehicle detection and tracklet generation approach is to provide vehicle tracklets $\theta^{*i}_{t,t+1}$ between two consecutive frames t and $t+1$. These tracklets are arranged in local sets (or graphs) $\mathfrak{G}^g_{t,t+1}$ (with $g \in \{1,\dots,G\}$) of interconnected tracklets as shown in Fig. 4 (4th and 5th images). A tracking algorithm can now interconnect these local sets of tracklets over several frames and use cost functions to determine longer tracklets or tracks, for example. We, however, need to derive detections from the tracklets in order to compare our proposed approach to other moving object detection methods taken from the literature. So, in each of these local sets $\mathfrak{G}^g_{t,t+1}$ we need to find the subset $\Theta^g_{t,t+1}$ of strongest tracklets that may co-exist w.r.t. certain plausibility constraints (e.g., no crossing tracklets, no detection sharing by two tracklets, etc.) and reject all others as visualized in the rightmost image of Fig. 4. This is done by solving a local optimization problem for each set $\mathfrak{G}^g_{t,t+1} = \{\theta^{*1}_{t,t+1},\dots,\theta^{*N}_{t,t+1}\}$. We use a randomized greedy algorithm that generates C solutions $\Theta^{g_c}_{t,t+1} = \{\theta^{*1}_{t,t+1},\dots,\theta^{*M}_{t,t+1}\}$ with $M < N$ for each $\mathfrak{G}^g_{t,t+1}$ and chooses the best $\Theta^g_{t,t+1}$ w.r.t. maximum likelihood. The $\Theta^{g_c}_{t,t+1}$ can be considered as joint association events. In each iteration c, the greedy algorithm picks tracklets with highest likelihoods $\theta^{*max}_{t,t+1} = \arg\max_i p(\theta^{*i}_{t,t+1})$ with $\theta^{*i}_{t,t+1} \in \mathfrak{G}^g_{t,t+1}$. If tracklets $\theta^{*j}_{t,t+1}$ exist that cross or share blobs with $\theta^{*max}_{t,t+1}$, they are either collected in subset $\mathfrak{T}$ if $|p(\theta^{*j}_{t,t+1}) - p(\theta^{*max}_{t,t+1})| < \epsilon$ or rejected otherwise (non-maximum suppression). Only one tracklet is picked randomly from the subset $\mathfrak{T}$ and added to solution $\Theta^{g_c}_{t,t+1}$. The other tracklets in $\mathfrak{T}$ are rejected. The likelihood of each solution $\Theta^{g_c}_{t,t+1}$ is then calculated by

$$p(\Theta^{g_c}_{t,t+1}) = \prod_{i=1}^{M} p(\theta^{*i}_{t,t+1}) \quad \text{with} \quad \theta^{*i}_{t,t+1} \in \Theta^{g_c}_{t,t+1}. \tag{3}$$

Picking tracklets from a set $\mathfrak{G}^g_{t,t+1}$ in this way may leave one or even multiple detections without an association. Such leftover detections are interpreted as a result of either a FP detection or a missing detection (FN detection). FP detections can occur due to parallax effects while typical FN detections emerge when vehicles are occluded or leave the image. The first case is modeled by rejecting the respective detection with a fixed FP likelihood (pseudo-tracklet $\theta^{*FP}_{t,t+1}$). The second case is modeled by introducing a pseudo-correspondence with a fixed FN likelihood (pseudo-tracklet $\theta^{*FN}_{t,t+1}$). The likelihoods of the pseudo-tracklets are also incorporated into the product in Eq. 3. This approach is inspired by Grinberg *et al.* [15]. In general, both likelihood values may be chosen using typical FP and FN rates, however we choose a fixed value of $p(\theta^{*FP}_{t,t+1}) = p(\theta^{*FN}_{t,t+1}) = 0.4$.

After having defined multiple joint association events $\Theta^{g_c}_{t,t+1}$ for each set $\mathfrak{G}^g_{t,t+1}$ we pick the best solution

$$\Theta^g_{t,t+1} = \arg\max_c p(\Theta^{g_c}_{t,t+1}). \tag{4}$$

The resulting overall set of tracklets $\Theta_{t,t+1}$ in each frame tuple $(t, t+1)$ is then given by

$$\Theta_{t,t+1} = \bigcup_{g=1}^{G} \Theta^g_{t,t+1}. \tag{5}$$

From this set of tracklets, we take the detections d^i_t (see Section 3.3) and use them for the evaluation in Section 4.

Acknowledgments

Michael Teutsch likes to thank Dr. Mubarak Shah, Dr. Imran Saleemi, and Dr. Haroon Idrees for helpful advice and productive discussions during his research visit at the Center for Research in Computer Vision (CRCV) at the University of Central Florida (UCF). Furthermore, he thanks the Karlsruhe House of Young Scientists (KHYS) for funding this research visit.

References

[1] AFRL. Wright-Patterson Air Force Base (WPAFB) dataset. https://www.sdms.afrl.af.mil/index.php?collection=wpafb2009, 2009. 2

[2] T. Ahonen, A. Hadid, and M. Pietikäinen. Face Description with Local Binary Patterns: Application to Face Recognition. *PAMI*, 28(12):2037–2041, Dec. 2006. 4

[3] S. Ali, V. Reilly, and M. Shah. Motion and Appearance Contexts for Tracking and Re-Acquiring Targets in Aerial Videos. In *CVPR*, 2007. 2

[4] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012. 4

[5] A. Basharat, M. 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. 2, 6

[6] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*, 2012. 2

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct. 2001. 3

[8] X. Cao, C. Wu, J. Lan, P. Yan, and X. Li. Vehicle Detection and Motion Analysis in Low-Altitude Airborne Video Under Urban Environment. *IEEE CSVT*, 21(10):1522–1533, Oct. 2011. 2

[9] B. Chen and G. Medioni. Motion Propagation Detection Association for Multi-target Tracking in Wide Area Aerial Surveillance. In *AVSS*, 2015. 7

[10] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005. 2, 3

[11] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral Channel Features. In *BMVC*, 2009. 3

[12] H. K. Ekenel and R. Stiefelhagen. Local appearance based face recognition using discrete cosine transform. In *EUSIPCO*, 2005. 3

[13] A. Gaszczak, T. P. Breckon, and J. Han. Real-time people and vehicle detection from UAV imagery. In *Proc. of SPIE Vol. 7878*, 2011. 2

[14] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*, pages 94–102. Prentice Hall, 2nd edition, 2002. 3

[15] M. Grinberg, F. Ohr, and J. Beyerer. Feature-based probabilistic data association (FBPDA) for visual multi-target detection and tracking under occlusions and split and merge effects. In *IEEE ITSC*, 2009. 8

[16] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of the Fourth Alvey Vision Conf.*, 1988. 2

[17] R. Hartley and A. Zisserman. *Multiple-View Geometry in Computer Vision*. Cambridge University Press, Mar. 2004. 2

[18] M. A. Keck, L. Galup, and C. Stauffer. Real-time tracking of low-resolution vehicles for wide-area persistent surveillance. In *WACV*, 2013. 2, 6

[19] A. Khembavi, D. Harwood, and L. Davis. Vehicle Detection Using Partial Least Squares. *PAMI*, 6(33):1250–1265, Apr. 2011. 2

[20] P. Liang, H. Ling, E. Blasch, G. Seetharaman, D. Shen, and G. Chen. Vehicle detection in wide area aerial surveillance using temporal context. In *FUSION*, 2013. 1, 2, 3, 4, 6

[21] P. Liang, G. Teodoro, H. Ling, E. Blasch, G. Chen, and L. Bai. Multiple Kernel Learning for Vehicle Detection in Wide Area Motion Imagery. In *FUSION*, 2012. 2

[22] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *PAMI*, 24(7):971–987, July 2002. 4

[23] Y. Pang, X. Shi, B. Jia, E. Blasch, C. Sheaff, K. Pham, G. Chen, and H. Ling. Multiway histogram intersection for multi-target tracking. In *FUSION*, 2015. 2

[24] R. Pelapur, S. Candemir, F. Bunyak, M. Poostchi, G. Seetharaman, and K. Palaniappan. Persistent Target Tracking Using Likelihood Fusion in Wide-Area and Full Motion Video Sequences. In *FUSION*, 2012. 2, 6

[25] A. A. G. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu. Multi-Object Tracking Through Simultaneous Long Occlusions and Split-Merge Conditions. In *CVPR*, 2006. 2

[26] T. Pollard and M. Antone. Detecting and Tracking All Moving Objects in Wide-Area Aerial Video. In *CVPRW*, 2012. 2, 6

[27] J. Prokaj and G. Medioni. Persistent Tracking for Wide Area Aerial Surveillance. In *CVPR*, 2014. 2, 6, 8

[28] V. Reilly, H. Idrees, and M. Shah. Detection and Tracking of Large Number of Targets in Wide Area Surveillance. In *ECCV*, 2010. 1, 2, 6, 7

[29] I. Saleemi and M. Shah. Multiframe Many-Many Point Correspondence for Vehicle Tracking in High Density Wide Area Aerial Videos. *IJCV*, 104(2):198–219, Sept. 2013. 1, 2, 3, 5, 6, 7

[30] G. Saur, W. Krüger, and A. Schumann. Extended image differencing for change detection in UAV video mosaics. In *Proceedings of SPIE Vol. 9026*, 2014. 3

[31] X. Shi, H. Ling, E. Blasch, and W. Hu. Context-Driven Moving Vehicle Detection in Wide Area Motion Imagery. In *ICPR*, 2012. 1, 2, 3, 4, 6

[32] F. Shih. *Image Processing and Pattern Recognition: Fundamentals and Techniques*. Wiley, 2010. 3

[33] S. Sivaraman and M. M. Trivedi. Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis. *IEEE Transactions on ITS*, 14(4):1773–1795, Dec. 2013. 2

[34] M. Teutsch and W. Krüger. Robust and fast detection of moving vehicles in aerial videos using sliding windows. In *CVPRW*, 2015. 2, 3, 4

[35] M. Teutsch, W. Krüger, and J. Beyerer. Evaluation of Object Segmentation to Improve Moving Vehicle Detection in Aerial Videos. In *AVSS*, 2014. 3

[36] S. Türmer, F. Kurz, P. Reinartz, and U. Stilla. Airborne Vehicle Detection in Dense Urban Areas Using HoG Features and Disparity Maps. *IEEE JSTARS*, 6(6):2327–2337, Dec. 2013. 2

[37] P. Viola and M. Jones. Robust Real-time Face Detection. *IJCV*, 57(2):137–154, 2004. 2

[38] J. Xiao, H. Cheng, H. Sawhney, and F. Han. Vehicle Detection and Tracking in Wide Field-of-View Aerial Video. In *CVPR*, 2010. 1, 2, 6