Systematic Evaluation of Moving Object Detection Methods for Wide Area Motion Imagery

Lars Sommer

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
lars.sommer@kit.edu

Technical Report IES-2015-12

Abstract: Wide area motion imagery (WAMI) facilitates the surveillance of several tens of square kilometers while using only one airborne sensor platform. Typical applications such as automatic behavior recognition, scene understanding, or traffic monitoring depend on precise multiple object tracking. Therefore, moving object detection is generally used as initial step. However, reliable moving object detection for WAMI is challenging as imprecise image alignment, low object resolution and a large number of moving objects lead to split, merged, and missing detections. In the context of this report, a detailed overview of existing methods for moving object detection proposed for WAMI is given. Ten existing methods as well as a novel combination of short-term background subtraction and suppression of image alignment errors by pixel neighborhood consideration are systematically evaluated on the WPAFB 2009 dataset that contains more than 160,000 ground truth detections. Parameters that contribute most to the performance of each method, the influence of related pre-processing steps as well as the impact of varying traffic density and scenery on the performance are discussed.

1 Introduction

In recent years, wide area motion imagery (WAMI) has been attracting an increased amount of research attention as WAMI enables large area surveillance while using only one airborne sensor platform. The sensor is comprised of a matrix of multiple cameras. Images of neighboring cameras with partially overlapping field of view are stitched to form the image with large ground coverage.

The stitched images are typically collected at 1-2 Hz due to the large volume (up to 100 megapixels). As WAMI data like the publicly available WPAFB 2009 dataset [U.S09] cover several tens of square kilometers and can contain thousands of moving objects per frame, applications such as driver behavior analysis or traffic monitoring are facilitated at large scale. These applications generally depend on multiple object tracking and consequently on object detections that are used at different stages in the tracking algorithm such as track initialization or object-to-track association [BTX+14]. Object detections are obtained by object segmentation approaches based frame differencing or background subtraction.

However, moving object detection in WAMI is very challenging: The moving objects are typically in the order of 10x20 pixels due to the low spatial resolution. Thus, detection approaches based on appearance features and machine learning are unreliable in WAMI so far [PM14]. The object detection is further complicated by weak contrast between object and background, shadows and occlusions that can lead to missed detections. Although image alignment is applied for camera motion compensation, residual errors of the alignment process as well as parallax effects can result in false positive detections. Additional challenges are sudden changes in camera gain and seam artifacts due to image stitching. Seam artifacts are caused by radiometric changes across different sensors and can produce false positive detections as sweeping seams can cause bands of large difference in the difference image [KGS13]. All these challenges can affect the performance of moving object detection. Nevertheless, there exists no systematic evaluation of moving object detection in WAMI so far even though missed detections emerge the need for track linking or false positive detections can cause the initialization of false positive tracks.

In this report, several moving object detection methods that are presented in WAMI literature are summarized and extended by a novel combination of short-term background subtraction and suppression of image alignment. In total, eleven methods are systematically evaluated on four image regions of the WPAFB 2009 dataset that comprises 1,025 frames. The different image regions contain more than 160,000 ground truth detections and offering different challenges such as traffic density and varying scenery.

2 Object Detection Methods

Object classification methods that are applied on WAMI data can be distinguished into frame differencing and background subtraction. As frame differencing requires less frames, residual errors of the alignment process are reduced compared

to background subtraction [SS13]. A drawback of frame differencing compared to background subtraction is the sensitivity to detect slow moving objects whose positions partially overlap in consecutive frames and consequently can lead to missed detections. Further approaches that are widely used in aerial videos such as methods based on optical flow vectors and appearance feature based methods are not applicable [APK14] or unreliable [PM14], respectively. Thus, only frame differencing and background subtraction approaches that are presented in WAMI literature are summarized and discussed in the following subsections.

2.1 Frame Differencing

Moving object detection methods based on frame differencing can be classified into two-frame and three-frame differencing. Two-frame differencing calculates the pixel-wise intensity difference between two consecutive frames by:

$$D(x,y) = |I_t(x,y) - \hat{I}_{t-1}(x,y)|,$$

where $D(x,y)$ is the intensity value difference at pixel (x,y) and I_t and $\hat{I}_{t-1}$ denote the intensity values of frame t and the aligned frame $t-1$. The difference image for three-frame differencing is given by the minimum of the differnece image between frame t and $t-1$ and the difference image between frame t and $t+1$:

$$D(x,y) = \min(|I_t(x,y) - \hat{I}_{t-1}(x,y)|, |I_t(x,y) - \hat{I}_{t+1}(x,y)|)$$

As two-frame differencing requires only two consecutive frames, the residual errors of the alignment process are minimal. However, each moving object produces two motion blobs in the difference image. One blob represents the object position in the current frame and an additional one represents its position in the previous frame. Saleemi and Shah [SS13] applied two-frame differencing on WAMI data and proposed to handle this so called *ghosting effect* by rejecting blobs with smaller mean gradient magnitude and intensity standard deviation in the current frame compared to the previous frame. Xiao et al. [XCSH10] applied instead three-frame differencing to avoid multiple blobs for each moving object. Additional residual errors caused by the alignment process can be suppressed by using the minimum differences of each pixel in small neighborhoods as proposed by Pollard and Antone [PA12]. Keck et al. [KGS13] extended three-frame differencing by applying a box filter to the difference image to reduce false positive detections caused by seam artifacts.

2.2 Background Subtraction

In general, moving object detection based on background subtraction is performed by calculating the difference image $D(x, y)$ between an image I_t and its corresponding background model I_{BG}:

$$D(x, y) = \min(|I_t(x, y) - I_{BG}(x, y)|)$$

A straightforward method to acquire a background model is to calculate the pixel-wise intensity median of consecutive frames. The number of frames used for background modeling applied on WAMI data range from 8 [LLB+13] to 16 [PDM11] and is thus clearly higher than the number of frames required for frame differencing. Incorporating background gradient information can be used to suppress noise in the difference image caused by parallax effects or residual errors due to the alignment process. Reilly et al. [RIS10] proposed to subtract the background gradient magnitudes from the difference images whereas Liang et al. [LLB+13] modified this approach by replacing the subtraction with an additional threshold operation. Pixels that corresponding background magnitude exceeds a given threshold are expected as noise and set to 0 in the difference image.

Calculating the pixel-wise intensity mean of consecutive frames is not considered as this approach requires four times the number of frames than median background modeling for comparable results [RIS10]. Kent et al. [KMP+12] proposed an alternative mean background approach. Instead of calculating the pixel-wise intensity mean of consecutive frames, Kent et al. [KMP+12] proposed to calculate the running mean and the standard deviation with a recursive filter. Pixels considered as moving are detected by comparing the difference between the intensity value I_t and the local mean μ to a local threshold which is given by the standard deviation multiplied with a set scaling factor.

More sophisticated approaches such as Gaussian mixture models are inapplicable for object detection in WAMI due to the high number of required frames [RIS10] as well as the sensitivity to illumination changes [SS13], parallax and registration drift [PA12]. Pollard and Antone [PA12] replaced the traditional GMM with an Interval Gaussian Mixture Model (IGMM). Each pixel is described as an interval limited by a minimum value μ_{min} and maximum value μ_{max}, instead of modeling each pixel as a mixture of Gaussians. The interval boundaries for each pixel are continuously updated by incorporating the minimum and maximum intensity values in a small neighborhood around the pixel in the current frame. Pixels that deviate more than a single global standard deviation value σ from this interval are considered as pixels belonging to a moving object.

A static background model based on an inpainting algorithm is proposed by Aeschliman et al. [APK14] Therefore, pixels assigned as objects by an initial difference image between the current and the previous image as well as pixels that correspond to objects in the previous frame are replaced based on directional and smoothness constraints to complete the background model.

2.3 Proposed Method

The combination of median background modeling and neighborhood consideration is expected to be a powerful approach that has not been reported yet. Preliminary experiments indicated good recall values in case of median background subtraction even for sequences with slow moving objects whose positions partially overlap between consecutive frames. However, median background modeling causes a high number of false positive detections due to parallax effects and the image alignment process. Neighborhood consideration seems to be an appropriate alternative for incorporating background gradient information to suppress false positive detections as noise caused by parallax effects as well as image alignment are in the order of a few pixels. Thus, the intensity value difference $D(x, y)$ between the current frame I_t and the corresponding median background model I_{BG} is given by the minimum difference between pixel (x, y) in the current frame and all pixels (x_i, y_j) in a given neighborhood N of the background model:

$$D(x, y) = \min_{i,j}(|I_t(x, y) - I_{BG}(x_i, y_j)|)$$

3 Experimental Results

In total, eleven object detection methods are considered for the evaluation. An overview of these methods is listed in Table 3.1. The performance of the selected methods is evaluated on four image regions of the WPAFB 2009 dataset [U.S09]. The image regions are selected with regard to the image regions evaluated by Basharat et al. [BTX+14] and Keck et al. [KGS13]. The WPAFB 2009 dataset comprises 1,025 frames with annotated GT. In the context of this report, stopping and parked objects are removed from the GT in order to determine the correct number of missing detections. The four image regions shown in Fig. 3.1 consist of 2,278×2,278 pixels and represent different challenges such as traffic density and varying scenery. The performance of each method is evaluated by means of

$$precision = \frac{TP}{FP + TP}$$

and

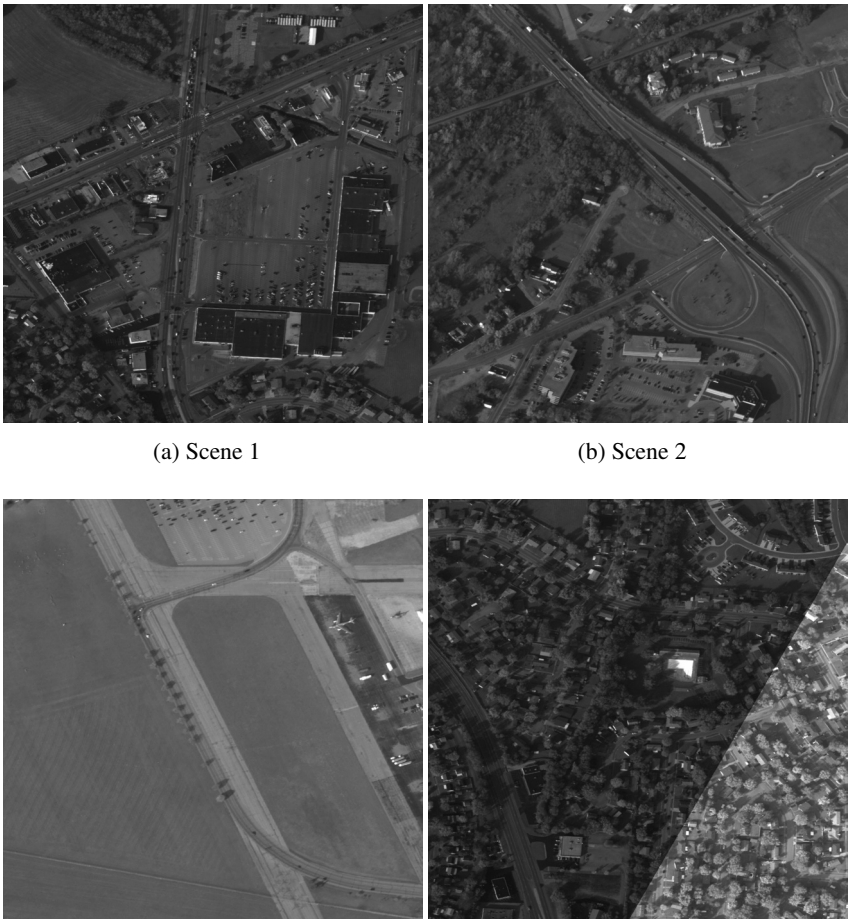
$$recall = \frac{TP}{TP + FN},$$

where TP, FP and FN are the number of true positive, false positive and false negative detections. In order to be consistent with the literature, the centroid of each blob is considered as a detection. Thus, each detection is represented by a point. Detections with annotated GT within a radius of 20 pixels are defined as TP otherwise as FP. GT objects without associated detection are defined as FN. The distance is set to 20 pixel, since GT annotations can differ from the center of the object. Furthermore, the blob centroid is often shifted from the annotated GT position due to appendant shadows.

Source	Object Detection Method
Saleemi [SS13]	2-frame Differencing + Ghost Handling
Xiao [XCSH10]	3-frame Differencing
Keck [KGS13]	3-frame Differencing + Box Filter
Pollard [PA12]	3-frame Differencing + Neighborhood
Pollard [PA12]	Interval Gaussian Mixture Model
Shi [SLBH12]	Median Background
Reilly [RIS10]	Median Background + Gradient Magnitude Suppression
Liang [LLB$^+$13]	Median Background + Gradient Magnitude Thresholding
Kent [KMP$^+$12]	Mean Background + Local Thresholding
Aeschliman [APK14]	Inpaint
Proposed	Median Background + Neighborhood

Table 3.1: Evaluated methods for moving object detection.

Prior to moving object detection, the camera motion is compensated by image alignment. After image alignment global histogram matching (HM) [GW02] is used to adjust camera gain and illumination variation, followed by local Gaussian mean filtering (MF) [SKS14] to reduce seam artifacts. Fig. 3.2 shows the impact of histogram matching and Gaussian mean filtering exemplarily for two-frame differencing by means of difference images. Large intensity differences are markedly reduced by HM (Fig. 3.2(b)) compared to no HM (Fig. 3.2(a)). However, the left image region still exhibits large differences in intensity. The reason for these differences is intensity discontinuities in the images due to stitching. These so called seam artifacts are suppressed by additional MF as depicted in Fig. 3.2(c). The impact of HM and MF on the performance is illustrated in

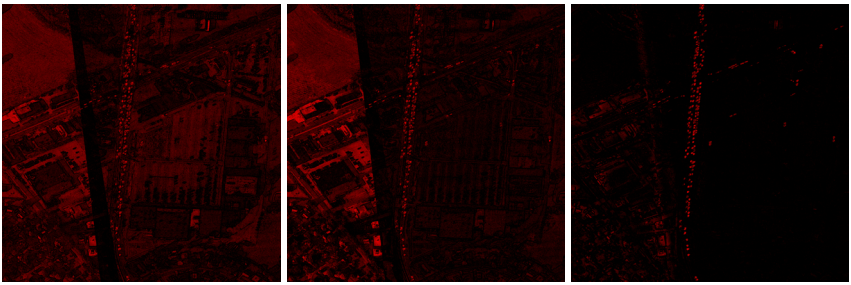


(a) Scene 1 (b) Scene 2

(c) Scene 3 (d) Scene 4

Figure 3.1: Image sections offering different challenges such as traffic density and varying scenery used for evaluation.

Fig. 3.3(a). As thresholding is used to distinguish pixels into objects and non-objects, it is expected that the threshold value has the highest impact on the performance. Thus, the shown precision-recall curves are generated by varying the



(a) no HM/MF (b) HM (c) HM+MF

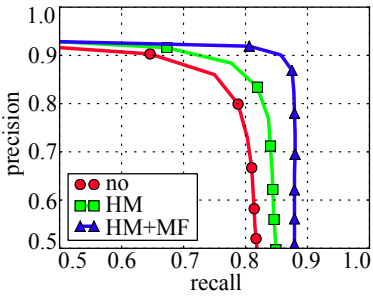
Figure 3.2: Difference images for two-frame differencing without global histogram matching (HM) and local Gaussian mean filtering (MF) (a), with HM (b) and with HM and MF (c).

threshold value. The performance without HM and MF is considerably increased by applying HM whereas HM is outperformed by additional MF. Similar results are obtained for the other object detection methods as well for all scenes.

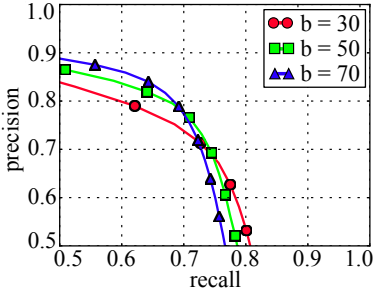
The performance of each object detection method is influenced by several parameters. The influence of relevant parameters is separately evaluated and optimized for each object detection method with regard to precision and recall. Thus, precision-recall curves are generated by varying the threshold value. In the following the parameters that contributed most to the performance are discussed. The corresponding precision-recall curves are given in Fig. 3.3(b)-3.3(f) exemplarily by means of scene 1.

In addition to the threshold value, all methods are affected by the minimum blob size. The minimum blob size is the minimal object size in pixels that is expected. Thus, detections with fewer pixels are associated as false detections and are rejected. The impact of the blob size on the performance of three-frame differencing is shown in Fig. 3.3(b). More false positives due to noise are rejected for larger minimum blob sizes. Consequently, the precision is increasing with increasing minimum blob sizes. In contrast, the recall is decreasing, since more small objects or partially detected objects are discarded.

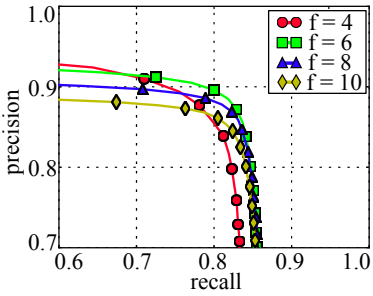
The further parameters that are discussed are only relevant for particular methods. Methods based on median background modeling are affected by the number of frames used for modeling the median background. The precision-recall curve for various number of frames is depicted in Fig. 3.3(c). The precision is increasing



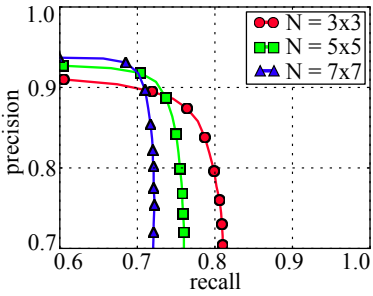
(a) Various pre-processing



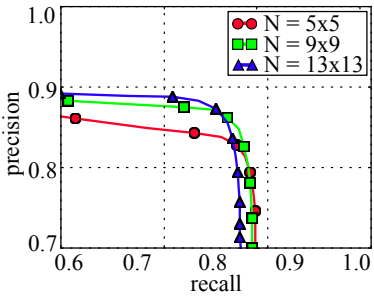
(b) Various minimum blob sizes b



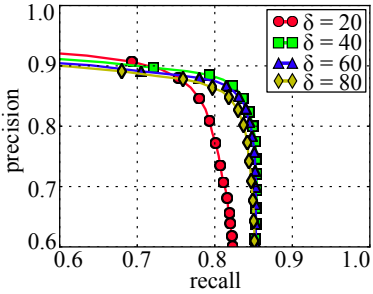
(c) Various number of frames f



(d) Various neighborhood sizes N



(e) Various neighborhood sizes N



(f) Various gradient thresholds δ

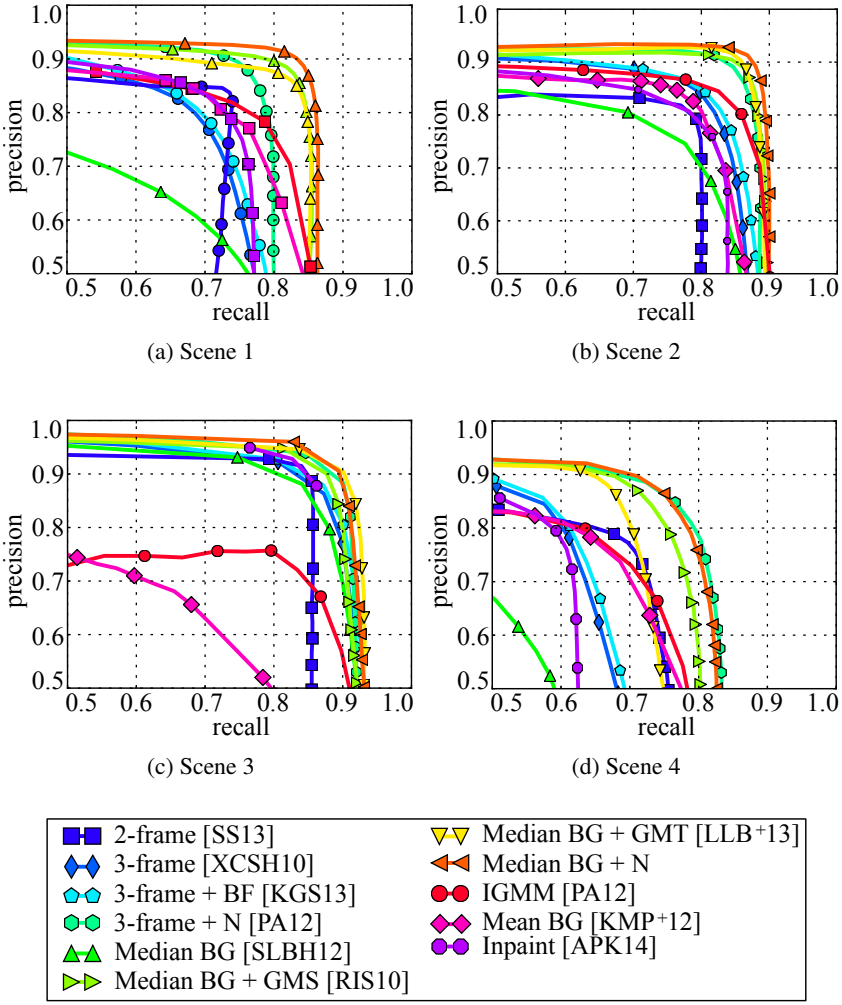
Figure 3.3: Variation of pre-processing steps (a) and optimization of parameters that contributed most to the object detection performance: (b) minimal blob size (in pixels), (c) number of frames used for median background modeling, (d)-(e) neighborhood size (in pixels) in case of 3-frame differencing and median background subtraction and (f) gradient magnitude threshold value δ in case of median background subtraction.

with fewer frames as the number of false positive detections caused by parallax effects or image alignment is reduced. The recall is almost the same for 6 to 10 frames. However, even less frames result in more missed detections. Reason for this is an inadequate estimated background especially in areas with dense traffic or intersections.

False positive detections caused by parallax effects or image alignment can be suppressed by neighborhood consideration. The impact of the applied neighborhood size on the performance of three-frame differencing and median background subtraction is shown in Fig. 3.3(d) and Fig. 3.3(e), respectively. More false positive detections are suppressed with increasing neighborhood sizes. However, the recall is decreasing with increasing neighborhood sizes as more small objects or partially detected objects are suppressed as well. Practical sizes are in the range of 3×3 to 5×5 pixels in case of three-frame differencing and slightly larger in case of median background subtraction as more errors are accumulated due to the number of used frames.

Background gradient information can be used to suppress false positive detections caused by parallax effects or image alignment as well. The impact of the gradient magnitude threshold δ on the performance of median background subtraction is illustrated in Fig. 3.3(f). As described in Section 2.2 gradient magnitudes above this threshold are expected to be caused by parallax and alignment errors and set to 0. The precision increases with lower threshold values as more errors are suppressed. In contrast, the recall is almost constant for threshold values between 40 and 80, but decreases considerably for lower threshold values as more objects are suppressed.

Fig. 3.4 shows the precision-recall curves of all methods for Scene 1-4. The parameters optimized for Scene 1 are adjusted for all Scenes. Median background subtraction without suppression of errors due to parallax effects or image alignment exhibits the worst performance of all methods for Scene 1,2,4. In contrast, the best performances are achieved for methods based on background subtraction that suppress these errors. Median background subtraction with neighborhood consideration outperforms both background gradient information based methods. Three-frame differencing with neighborhood consideration exhibits worse performance for Scene 1-3, whereas the performance is slightly better for Scene 4. The reason for the better performance is that Scene 4 showing a residential area densely covered with buildings and tree is more error prone to image alignment and parallax effects that are accumulated by the number of used frames. The weaker recall for Scene 1 indicates instead that median background models are more effective to detect slow moving objects especially in dense traffic. The other frame differencing methods show markedly worse performance compared



(a) Scene 1

(b) Scene 2

(c) Scene 3

(d) Scene 4

Figure 3.4: Precision-recall curves of all object detection methods for all four image regions.

to Three-frame differencing with neighborhood consideration. The impact of the locally applied box filter to suppress seam artifacts is marginal since these artifacts are partially suppressed during the pre-processing. The performance of the

further background subtraction based methods is comparable to the frame differencing methods without neighborhood consideration except for Scene 3. The precision of IGMM for Scene 3 that is expected to be less challenging is considerably worse. The adaptive interval model and the fixed standard deviation used to segment pixels in object and non-object is not able to compensate for severe illumination changes and consequently results in a large number of false positive detections. The same difficulty is observed for the running mean approach which shows even poorer performance for this Scene.

4 Conclusion

In the context of this report, eleven object detection methods were evaluated on four different challenging image regions of the WPAFB 2009 dataset. For this purpose, the impact of pre-processing steps as well as parameters contributing most to the performance was discussed. The performance can be considerably increased by applying histogram matching and local Gaussian mean filtering to adjust camera gain and illumination variation and to suppress seam artifacts. The strong impact of various parameters on the object detection performance exhibits that the adjustment of these parameters is not neglible with regard to the following applications. The best performance overall is achieved by median background subtraction with neighborhood consideration that slightly outperforms other approaches for the suppression of errors caused by imprecise image alignment and parallax effects. The fact that other methods exhibit considerably worse performance indicates the importance of the suppression of these kinds of errors. Nevertheless, the impact of optimized moving object detection on existing multiple object tracking algorithms itself has to be analyzed.

Bibliography

[APK14] Chad Aeschliman, Johnny Park, and Avinash C. Kak. Tracking Vehicles Through Shadows and Occlusions in Wide-Area Aerial Video. *IEEE Transactions on Aerospace and Electronic Systems*, 50(1):429–444, 2014.

[BTX+14] Arslan Basharat, Matt Turek, Yiliang Xu, Chuck Atkins, David Stoup, Keith Fieldhouse, Paul Tunison, and Anthony Hoogs. Real-time Multi-Target Tracking at 210 Megapixels second in Wide Area Motion Imagery. *WACV*, pages 839–846, 2014.

[GW02] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*, pages 94–102. Prentice Hall, 2nd edition, 2002.

[KGS13] Mark Keck, Luis Galup, and Chris Stauffer. Real-time Tracking of Low-resolution Vehicles for Wide-Area Persistent Surveillance. In *WACV*, 2013.

[KMP+12] Phil Kent, Simon Maskell, Oliver Payne, Sean Richardson, and Larry Scarff. Robust Background Subtraction for Automated Detection and Tracking of Targets in Wide Area Motion Imagery. *Proc. SPIE*, 8546, 2012.

[LLB+13] Pengpeng Liang, Haibin Ling, Erik Blasch, Guna Seetharaman, Dan Shen, and Genshe Chen. Vehicle Detection in Wide Area Aerial Surveillance using Temporal Context. In *FUSION*, 2013.

[PA12] Thomas Pollard and Matthew Antone. Detecting and Tracking All Moving Objects in Wide-Area Aerial Video. In *CVPRW*, 2012.

[PDM11] Jan Prokaj, Mark Duchaineau, and Gerard Medioni. Inferring Tracklets for Multi-Object Tracking. In *CVPRW*, 2011.

[PM14] Jan Prokaj and Gerard Medioni. Persistent Tracking for Wide Area Aerial Surveillance. In *CVPR*, 2014.

[RIS10] Vladimir Reilly, Haroon Idrees, and Mubarak Shah. Detection and Tracking of Large Number of Targets in Wide Area Surveillance. In *ECCV*, 2010.

[SKS14] Günter Saur, Wolfgang Krüger, and Arne Schumann. Extended image differencing for change detection in UAV video mosaics. In *Proceedings of SPIE Vol. 9026*, 2014.

[SLBH12] Xinchu Shi, Haibin Ling, Erik Blasch, and Weiming Hu. Context-Driven Moving Vehicle Detection in Wide Area Motion Imagery. In *ICPR*, 2012.

[SS13] Imram Saleemi and Mubarak Shah. Multiframe Many-Many Point Correspondence for Vehicle Tracking in High Density Wide Area Aerial Videos. *IJCV*, 104(2):198–219, 2013.

[U.S09] U.S. Air Force Research Lab. SDMS: WPAFB 2009 Dataset. `https://www.sdms.afrl.af.mil/index.php?collection=wpafb2009`, 2009.

[XCSH10] Jiangjian Xiao, Hui Cheng, Harpreet Sawhney, and Feng Han. Vehicle Detection and Tracking in Wide Field-of-View Aerial Video. In *CVPR*, 2010.