Detection and tracking of humans from an airborne platform

Adam W.M. van Eekeren, Judith Dijk, Gertjan Burghouts

TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands

ABSTRACT

Airborne platforms are recording large amounts of video data. Extracting the events which are needed to see is a time-demanding task for analysts. The reason for this is that the sensors record hours of video data in which only a fraction of the footage contains events of interest. For the analyst, it is hard to retrieve such events from the large amounts of video data by hand. A way to extract information more automatically from the data is to detect all humans within the scene. This can be done in a real-time scenario (both on-board as on the ground station) for strategic and tactical purposes and in an offline scenario where the information is analyzed after recording to acquire intelligence (e.g. a daily life pattern). In this paper, we evaluate three different methods for object detection from a moving airborne platform. The first one is a static person detection algorithm. The main advantage of this method is that it can be used on single frames, and therefor does not depend on the stabilization of the platform. The main disadvantage of this method is that the number of pixels needed for the detection is pretty large. The second method is based on detection of motion-in-motion. Here the background is stabilized, and clusters of pixels that move with respect to this stabilized background are detected as moving object. The main advantage is that all moving objects are detected, the main disadvantage is that it heavily depends on the quality of the stabilization. The third method combines both previous detection methods.

The detections are tracked using a histogram-based tracker, so that missed detections can be filled in and a trajectory of all objects can be determined. We demonstrate the tracking performance using the three different detections methods on the publicly available UCF-ARG aerial dataset. The performance is evaluated for two human actions (running and digging) and varying object sizes. It is shown that a combined detection approach (static person detection and motion-in-motion detection) gives better tracking results for both human actions than using one of the detectors alone. Furthermore it can be concluded that the minimal height of humans must be 20 pixels to guarantee a good tracking performance.

Keywords: person detection, tracking, airborne platforms, image processing, motion-in-motion.

1. INTRODUCTION

Because airborne platforms are recording large amounts of video data, extracting the events which are needed to see is a time-demanding task for analysts. Especially because only a fraction of the video footage contains events of interest. To speed up this manual procedure, a first step is to automatically detect and track all humans within the video data[1]. This can be done in a real-time scenario (both on-board as on the ground station), such as following a specific target, and in an offline scenario where the information is analyzed after recording to acquire intelligence (e.g. a daily life pattern). Based on the found tracks also further automatic analysis can be done such as action recognition[2,3], behavior recognition and threat detection. To assist the analyst the detections and tracks have to be of good quality. For example the detected pixels must mainly cover the object of interest, the number of false detections has to be minimized, the tracks have to cover enough frames and the tracks have preferably no ID switches.

In literature a wide variety of trackers can be found. A recent survey has been performed by Smeulders et al.[4]. From this survey it becomes clear that most trackers are evaluated only on a small number of videos and that their performance is good in only a subset of conditions (illumination changes, clutter, etc.). In this paper a histogram-based matching tracker is used which has some similarities with the well-known mean shift tracker as described by Comaniciu [5]. In this paper three different methods for object detection are used in combination with this tracker. The first one is a static person detection algorithm; the fastest pedestrian detector in the west (FPDW)[6]. The main advantage of this method is that it can be used on single frames, and therefor does not depend on the stabilization of the platform or how a person is moving. The main disadvantage of the method is that the number of pixels needed for the detection is pretty large. The second

Electro-Optical and Infrared Systems: Technology and Applications XI, edited by
David A. Huckridge, Reinhard Ebert, Proc. of SPIE Vol. 9249, 92490S · © 2014
SPIE · CCC code: 0277-786X/14/$18 · doi: 10.1117/12.2067568

method is based on detection of motion-in-motion. Here the background is stabilized, and clusters of pixels that move with respect to this stabilized background are detected as moving object. The main advantage is that all moving objects are detected, the main disadvantage is that it heavily depends on the camera motion and the quality of the stabilization. The third method combines both previous detection methods.

The performance of tracking using the three different object detection methods is demonstrated on aerial data: the publicly available UCF-ARG aerial dataset[7]. The tracking performance is evaluated for two different human actions and varying object sizes. This evaluation gives an impression of what is possible / what are the limitations when performing tracking on aerial data.

The setup of the paper is as follows. Section 2 describes the tracking method used. In Section 3 the experimental setup is described in which the performance of the tracker is tested. The results of the experiments are presented in Section 4. Finally, conclusions will be drawn in Section 5.

2. METHOD DESCRIPTION

In this section the method is described that is used for the detection and tracking of moving objects and persons in aerial data. A flowchart of this method is depicted in Figure 1.

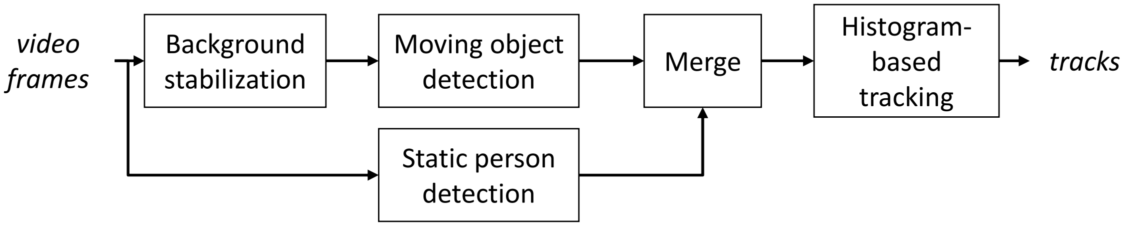


Figure 1. Flowchart of method to detect and track moving objects and persons in aerial data.

2.1 Background stabilization

The first step of the proposed method consists of frame-to-frame registration to stabilize the background of the scene. This is necessary because an aerial platform is most of the time moving, even when the camera is stabilized with hardware. The stabilization procedure consists of first calculating the optical flow between subsequent frames on a predefined grid, followed by a RANSAC[8] iteration schema to estimate an affine transformation. The assumption here is that the scene is approximately flat.

2.2 Moving object detection

After stabilizing the background, moving objects are detected by calculating the difference between registered frames. This is done by warping one or more previous frames to the actual frame and calculating the mean difference. In the experiments typically the 10th and 5th previous frame are warped to the actual frame. The resulting mean difference frame is thresholded with a fixed threshold followed by some morphological operations to remove small detections due to noise.

2.3 Static person detection

Parallel to the detection of moving objects, persons are detected using the detector proposed by Dollar et al[6]. This Fastest Pedestrian Detector of the West (FPDW) is originally trained for detecting pedestrians and it is well known for its state-of-the-art and fast performance. We selected this static person detector because a lot of aerial datasets are captured with a slightly slanted viewing angle in which persons are visible from head to toe.

2.4 Merge detections

In this step the motion and person detections are merged. Duplicate detections are removed if the detections are overlapping well. If one of the duplicate detections is fired by the person detector with high confidence, this detection is kept and the other is removed. If both duplicate detections are fired by the same detector then the smallest detection is kept if the overlap is really good, otherwise the largest detection is kept.

2.5 Histogram-based tracking

The tracking is based on histogram backprojection with a (model) histogram that is initialized with the first detection. Detections are associated with existing tracks based on the amount of overlap they have. Tracks can be lost if no association can be found, but picked up within a certain number of frames (typical 25 frames in our experiments). In the meantime, the new track position is found by histogram backprojection. The histogram and size of the track is updated in case of an association. Also, the tracks can be split if a track has multiple associations or merged if multiple tracks have a common association. After the tracking, the tracks can also be filtered based on criteria such as the minimal track length.

3. EXPERIMENTS

3.1 Data description

For the experiments the aerial UCF-ARG dataset[7] is selected, which consists of ten different human actions ranging from walking and running to digging and throwing. For each action 48 short videos (960 x 540 at 29 fps) are captured where one person (out of 12 different persons) is performing the action in a non-complex scene (parking lot surrounded by grass). To evaluate the performance of tracking for different kind of human actions, two different actions are selected for the experiments: RUNNING and DIGGING. Note that the mean video duration for RUNNING is 3 seconds and for DIGGING it is 11 seconds. A few snapshots of the visual data are shown in Figure 2 and Figure 3.

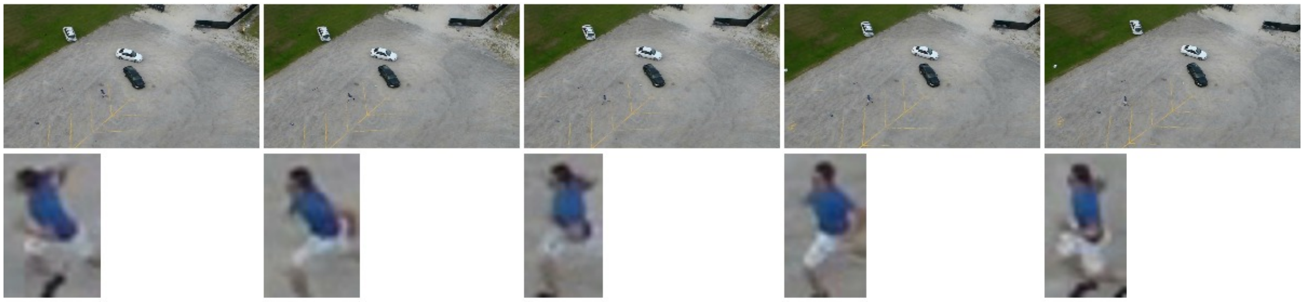


Figure 2. Snapshots showing the human action RUNNING from the UCF-ARG dataset. Upper row: original frames, lower row: zoom in on person.

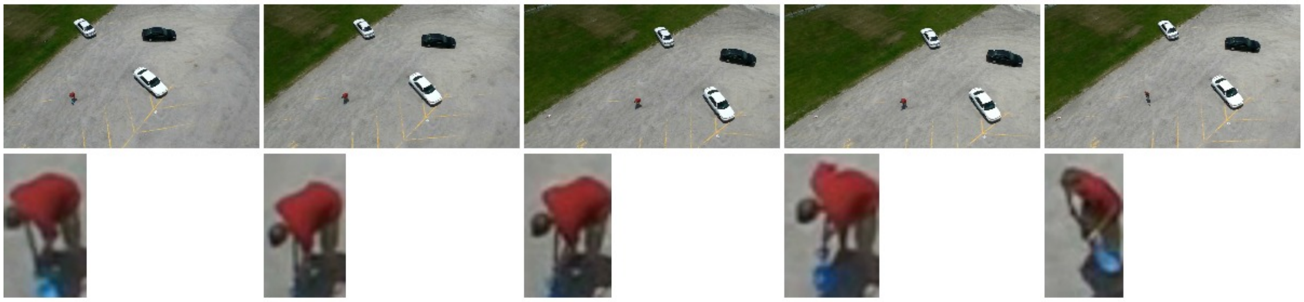


Figure 3. Snapshots showing the human action DIGGING from the UCF-ARG dataset. Upper row: original frames, lower row: zoom in on person.

To test the relation of the tracking performance with the resolution of the data, the original video frames are also scaled with a factor of 0.7 and 0.5. In the original video frames the persons are ranging from 30 to 55 pixels in length. All original video frames are manually annotated by drawing a bounding box around the person which performed the specified action.

3.2 Modes of method

This subsection describes the different modes for which the tracking method is tested. The first mode is when only a static person detection algorithm (FPDW detector[6]) is used before tracking. In this mode the threshold on the detection confidence is varied. The second mode is when only moving object detection is performed before tracking. In this mode the moving object detection threshold is varied. The third mode is when both moving object detection and person detection are performed before tracking. In this mode the person detection threshold is used which gives the highest True Positive Rate (TPR), while the moving object detection threshold is varied.

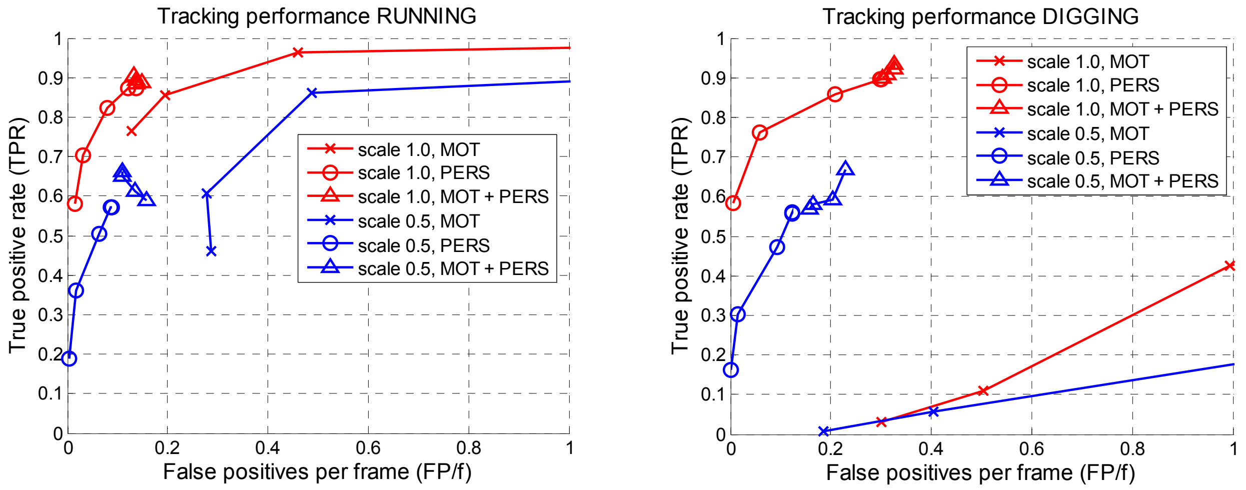
3.3 Evaluation criteria

The tracking method is evaluated on the following criteria:

1. **ROC curve**: this curve shows the relation between the True Positive Rate (TPR) versus the number of False Positives per frame (FP/f) for different algorithmic settings. Such a curve gives an insight in the quality of the tracking.
2. **Track length**: the length of all found tracks (including the lost frames) is analyzed. How longer the length of a track, the more information this track can provide to an analyst.
3. **ID switches**: each track has a unique ID number. If this number changes (e.g. due to temporal detection loss or a false merge with a nearby track) it is counted as an ID switch. Ideally no ID switches occur.

4. RESULTS

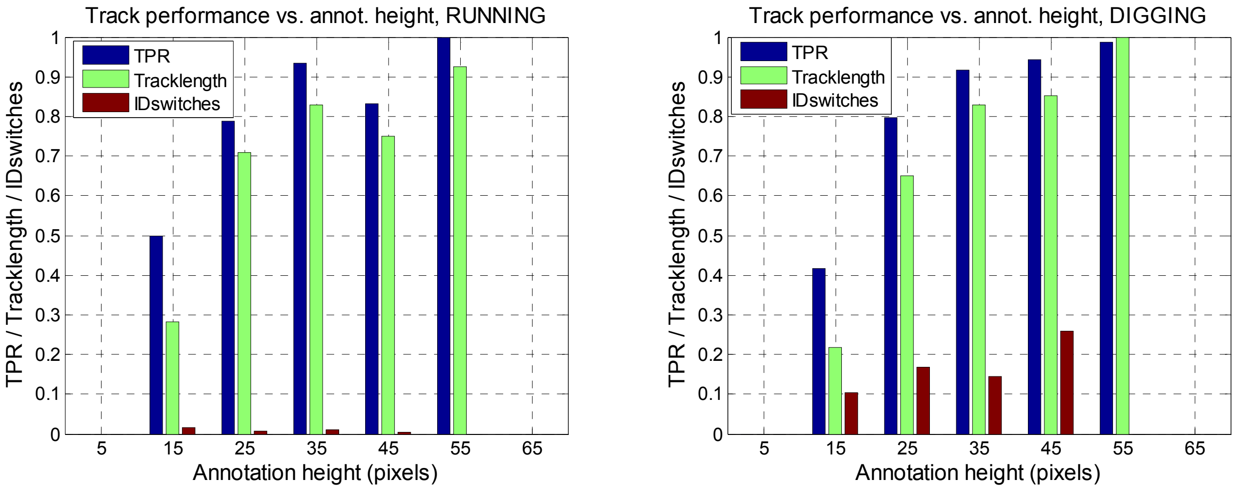
As said in the previous section the tracker is tested on visual data containing one of two different human activities: RUNNING and DIGGING. The main reason for choosing these activities is that the first one is very dynamic while the second one is more static. In Figure 4 ROC curves are depicted for tracking both activities in the original data and in the 0.5 rescaled data for the three different detection modes. Furthermore it must be noted that the 'filtering' step in the tracking uses the following criteria: 1) minimum track length of 30 frames and 2) at least 5% of the tracked frames must have the label 'person' if such a label is available. The effect of filtering is mainly a reduction of the number of false positives, while only a slightly decrease of the true positive rate takes place.



(a) (b)

Figure 4. ROC curves for the tracking performance on aerial data containing two human activities: RUNNING and DIGGING. The tracking is done on the original data and on the 0.5 rescaled data in three different detection modes: 1) moving object detection (MOT), 2) static person detection (PERS) and 3) combined detection (MOT + PERS).

The results in Figure 4 show that the tracking performance is improved by using a combined detection (MOT+PERS) instead of only moving object detection (MOT) or only person detection (PERS). For RUNNING moving object detection works well, but by using the combined detection the FP/f can be reduced significantly, while only a small decrease of the TPR takes place. For DIGGING the performance of moving object detection is really low, while the static person detection still has reasonable performance. The combined detector gives better results than the person detector with slightly more FP/f. The results show furthermore that the tracking performance for both activities decreases significantly when lower resolution data is available.



(a) (b)

Figure 5. Tracking performance versus annotation height for tracking using the combined detection mode of RUNNING (a) and DIGGING (b). Blue bars indicate the mean True Positive Rate (TPR) for each bin. The green bars indicate the mean ratio of the measured tracklength and the maximum possible tracklength. The red bars indicate the mean number of track ID switches within a video divided by ten (to fit in the figure). The bin boundaries are multiples of ten, so 10, 20,…

In Figure 5 an overview is presented of a few performance measures (True Positive Rate, Tracklength and number of ID switches) for tracking using the combined detection mode (MOT + PERS) for different heights of annotated persons. For both activities it is clear that the TPR and tracklength decrease for smaller annotation height. This means that when persons are smaller in the field-of-view of a camera they are harder to track. A lower limit on good tracking of persons seems to be 20 pixels in height. Below this size the TPR and tracklength decrease significantly.

Furthermore it can be noted that the number of track ID switches for DIGGING is larger than for RUNNING. This might be explained by the fact that DIGGING gives local motion detection on the body instead of full motion detection which is the case for RUNNING and therefore might give multiple tracks on one person which results in ID switches.

Some snapshots in Figure 6 and Figure 7 show good and bad tracking performance for RUNNING. The bad performance is explained by the low contrast with the background.

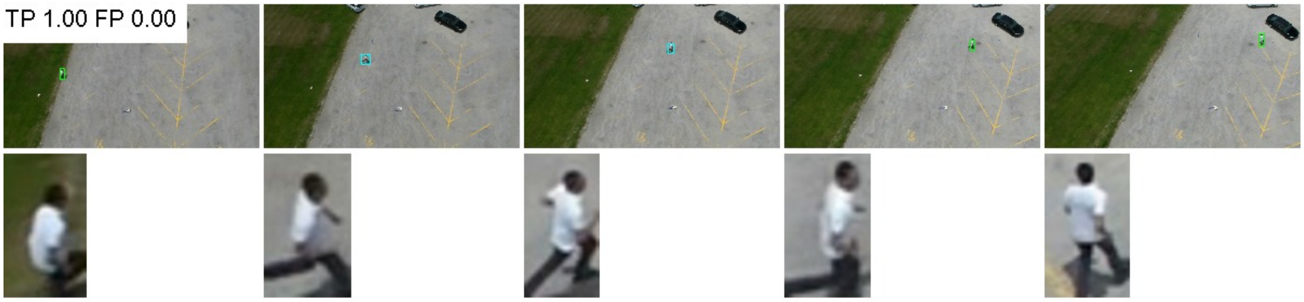


Figure 6. Snapshots showing a perfect tracking result of the human action RUNNING. Upper row: original frames with detections (green = person detection, cyan = motion detection), lower row: zoom in on person based on annotations.

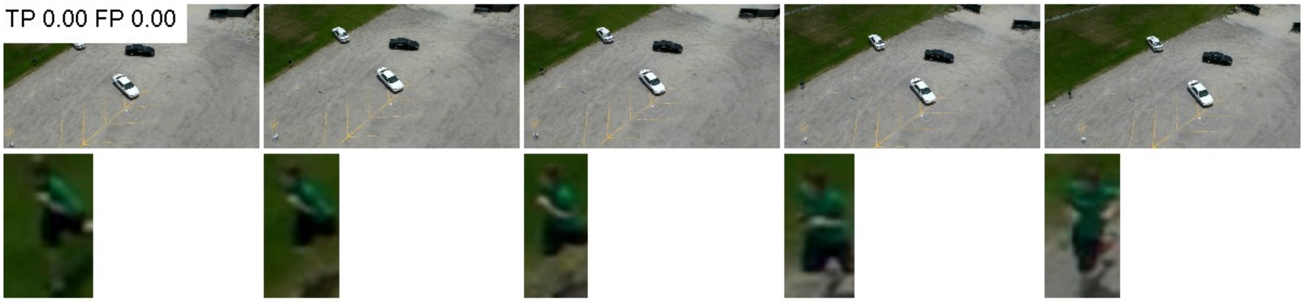


Figure 7. Snapshots showing a complete missed tracking result of the human action RUNNING. Upper row: original frames containing no detections, lower row: zoom in on person based on annotations showing low contrast.

Some examples of the tracking performance for DIGGING are shown in Figure 8 and Figure 9. The weak performance of the last example might be explained because the person is observed from the side and might therefore be harder to detect by the person detector.

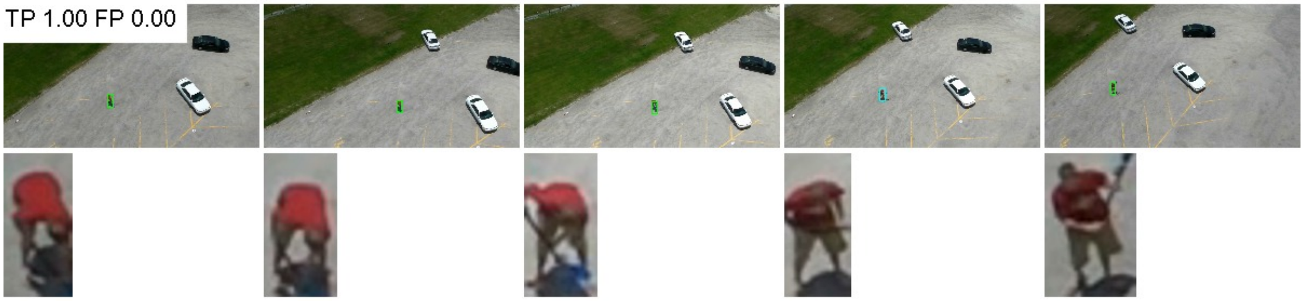


Figure 8. Snapshots showing a perfect tracking result of the human action DIGGING. Upper row: original frames with detections (green = person detection, cyan = motion detection), lower row: zoom in on person based on annotations.

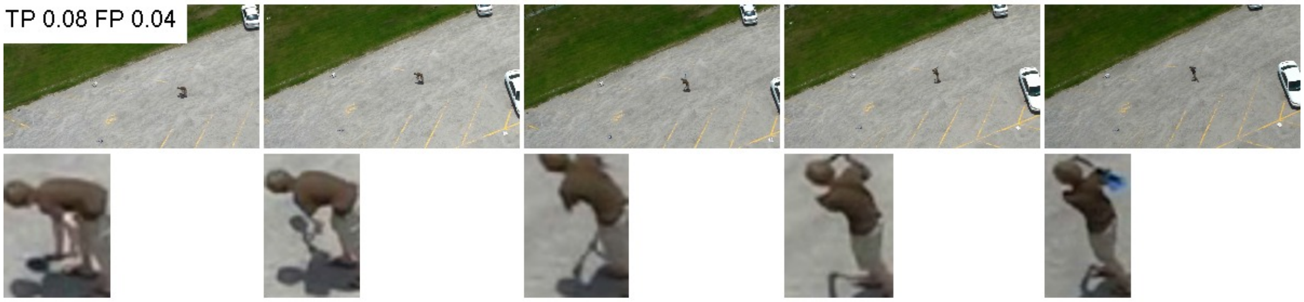


Figure 9. Snapshots showing a bad tracking result of the human action DIGGING. Upper row: original frames showing no detections lower row: zoom in on person based on annotations.

5. CONCLUSIONS

The results show that tracking of humans that perform different activities in visual aerial data using a well-known tracking approach is possible. The tracking performance (TPR and tracklength) decreases significantly when persons are smaller than 20 pixels in length. Furthermore it is shown that using a combined motion- and person detection before tracking increases the overall tracking performance; increasing the TPR on one hand and reducing the number of false detections on the other hand. In this way an analyst is best assisted with his task and not misled by false detections. Automatic tracking will significantly reduce the workload of an analyst. Also it enables further automatic analysis such as action recognition, behavior recognition and threat detection.

Although the implementation of the tracking method in this paper is not running in real-time (1 fps), after speedup it should be capable to use in a real-time scenario (both on-board as on the ground station).

REFERENCES

[1] Trinh, H., Li, J., Miyazawa, S., Moreno, J.., Pankanti, S., "Efficient UAV video event summarization," 21st Int. Conf. Pattern Recognit. ICPR, 2226–2229, IEEE (2012).

[2] Bouma, H., Hanckmann, P., Marck, J.-W., Penning, L., den Hollander, R., ten Hove, J.-M., van den Broek, S., Schutte, K.., Burghouts, G., "Automatic human action recognition in a scene from visual inputs," SPIE Def. Secur. Sens., 83880L–83880L, International Society for Optics and Photonics (2012).

[3] Burghouts, G., Schutte, K., Bouma, H.., den Hollander, R., "Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos," Mach. Vis. Appl. **25**(1), 85–98 (2014).

[4] Smeulders, A. W. M., Chu, D. M., Cucchiara, R., Calderara, S., Dehghan, A.., Shah, M., "Visual Tracking: an Experimental Survey," IEEE Trans Pattern Anal Mach Intell (2014).

[5] Comaniciu, D., Ramesh, V.., Meer, P., "Real-time tracking of non-rigid objects using mean shift," CVPR, 142–149, IEEE (2000).

[6] Dollár, P., Belongie, S.., Perona, P., "The Fastest Pedestrian Detector in the West.," BMVC **2**, 7, Citeseer (2010).

[7] Reddy, K., "UCF-ARG Data Set (http://crcv.ucf.edu/data/UCF-ARG.php)."

[8] Fischler, M. A.., Bolles, R. C., "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM **24**(6), 381–395 (1981).