DGLR-2006-038

MULTIPLE OBJECTS TRACKING IN AN UAV CAMERA SEQUENCE

N. Frietsch, O. Meister, C. Schlaile, J. Wendel, and G. F. Trommer
Institute of Systems Optimization
University of Karlsruhe (TH)
Kaiserstr. 12, 76128 Karlsruhe

ABSTRACT

This paper focuses on the detection and tracking of multiple moving objects in a camera sequence, which is provided by a small UAV in a hover-and-stare scenario. Two different algorithms for the segmentation of the moving objects are described. The first approach bases on the compensation of the camera motion followed by image subtraction and appropriate filtering. Consecutive images are assumed to be related by homographies. The second algorithm needs no stabilization: a dense optical flow field between two consecutive frames allows the identification of moving areas by processing the outliers of a RANSAC estimation of the homography. Furthermore, a strategy for tracking these objects over time is described including a Kalman filter and a heuristic rule set for reducing false alarms. Finally, experimental results on in-flight images are presented and the performances of the developed algorithms are compared. Both algorithms are able to detect and track moving objects. The algorithm without preceding stabilization needs distinct object characteristics for a good performance. The results of the method with camera motion compensation showed that this algorithm is more robust against false alarms.

1. INTRODUCTION

The research activity in the field of unmanned aerial vehicles (UAV) has increased constantly over the last few years. Besides large UAVs which can carry significant payloads, small UAVs are of interest due to their cost efficiency and their ease of deployment. These UAVs are used for a wide range of applications like surveillance and reconnaissance missions, therefore the ability to hover is often desirable. An onboard digital camera acquires important information about the observed situation. The used UAV is an electrically powered four rotor helicopter (shown in Figure 1). Due to the small size and weight of the aircraft, the onboard processing power is limited and any image processing has to take place on the ground station. The image data is transmitted in flight via a radio downlink.

In order to increase the situation awareness of the operator, an automated processing of the received video stream is advantageous. The first step of most monitoring or surveillance applications is the segmentation of the objects that move differently from the background of the scene. Starting from this partial result, the objects can be tracked, classified and their behavior can be interpreted.

A lot of research has been done in the field of moving objects detection. Starting from the processing of video streams acquired by stationary cameras, a number of algorithms for moving camera scenes have been recently proposed. Motion detection in a static scene can be done for example by simply subtracting the background. Processing the image stream of a moving camera is much more challenging. There exist two basic approaches: The first one includes the stabilization of the image stream by compensating the relative motion between camera and the background of the scene followed by a method to extract the independently moving parts. The registration is often done by applying a geometric transformation like an affine transformation [1] [2] or a homography [3]. The independently moving objects are then detected for example by residual optical flow [1] or background subtraction [2] [3].

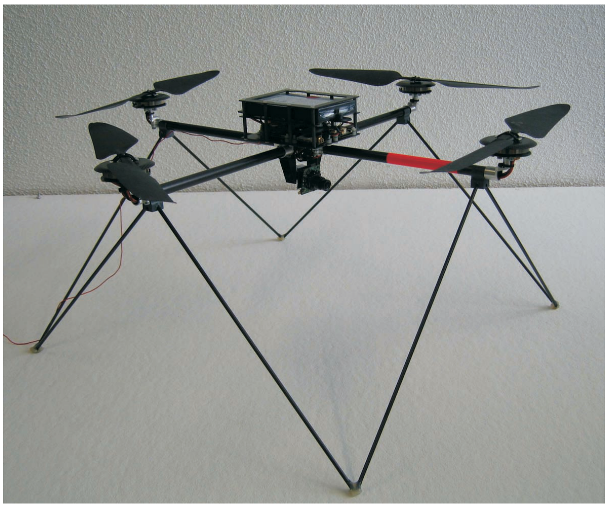


Fig. 1: The UAV with the camera used for acquiring image data. Take-off weight is below 1 kg, diameter is 52 cm.

The second approach for processing the image stream of a moving camera includes no preceding stabilization. Different strategies have been derived. In [4], an algorithm is proposed that recovers the scene structure, the trajectories of the moving objects and the camera motion at once. It is assumed that moving objects can be approximated by single points and that they move linearly with constant velocity. Another procedure is described in [5]. Moving objects are detected by processing the outliers detected during computation of the homography. The necessary point correspondences are found by focusing on whole clusters of features. A specific object is finally identified by searching for known features of the object of interest.

In this paper, two different algorithms are described. The first one uses image stabilization. The homography is chosen as geometric transform between consecutive frames. Moving areas are segmented by image subtraction. But in contrast to [3], there is no background frame created, that is subtracted from each newly received frame. In each step, we warp the new frame $\mathbf{F_{i+1}}$ onto the current frame $\mathbf{F_i}$ and subtract them directly. After an appropriate filtering, the detected moving objects are identified and only their coordinates are warped onto a reference frame $\mathbf{F_0}$ followed by a tracking operation using a Kalman filter. This procedure has the advantage that an inaccurately estimated homography matrix only affects the particular detection step. Otherwise the reference background image and therefore the following detection process is distorted. For better interpretability, the whole images can be warped onto a mosaic frame, where the tracked objects can be marked.
The second algorithm needs no compensation of the relative camera motion. A dense field of reliable point correspondences is calculated. The homography and resulting outliers are determined with a RANSAC algorithm [6]. These outliers comprise three different types: wrongly matched feature pairs, points where the background is not planar enough and points belonging to independently moving objects. After merging these points into areas by region growing, the same tracking algorithm is used as described above.

This paper is organized as follows. In the next section, the motion detection algorithm with preceding stabilization of the images is described, followed by the detection algorithm without stabilization in Section 3. In Section 4, the reduction of false alarms and the tracking of the extracted moving objects is discussed as well as a suitable visualization of the results for a human operator. Furthermore, the performance of the algorithms when processing real in-flight image sequences are compared in Section 5 and finally conclusions are drawn.

2. MOTION DETECTION WITH PRECEDING STABILIZATION

On the ground station, a preprocessing of the raw video-stream has to take place. For de-interlacing,

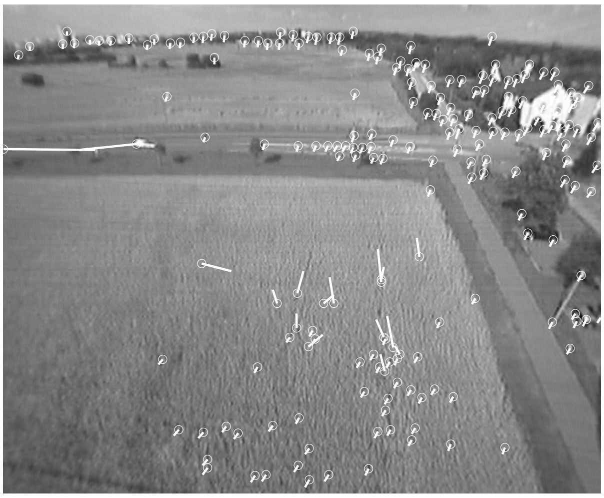


Fig. 2: One frame of a preprocessed UAV-camera sequence with tracked feature points.

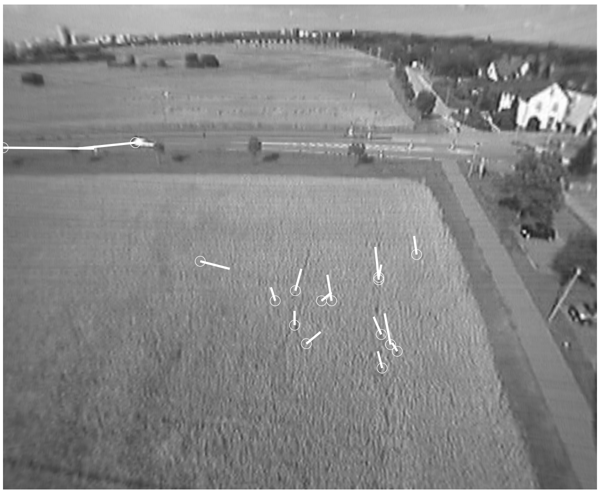


Fig. 3: The outliers: wrongly matched feature pairs and features belonging to a moving car.

a simple procedure works good enough. Only the fields containing the odd rows of the images are taken and interpolated to full resolution. This increases the interpretability for the human operator. Finally the radial distortion introduced by the wide angle camera needs to be corrected before the images can be processed.

2.1. Compensation of the Background Motion

For compensating the relative motion between camera and background of the scene, it is required that most pixels in the images belong to the static background. The image sequence is supposed to be taken in a hover-and-stare scenario, therefore the translational component of the camera-motion is small. Under the further assumption, that the observed scene is sufficiently far away and therefore well described by a plane, the geometrical relation between two consecutive frames can be described by a homography.

The homography matrix $\mathbf{H}$ associates two subsequent frames $\mathbf{F_1}$ and $\mathbf{F_2}$ by the equation

$$\vec{x}_2 \sim \mathbf{H} \cdot \vec{x}_1, \tag{1}$$

with the homogeneous coordinates $\vec{x}_1$ and $\vec{x}_2$. To determine the entries of the matrix, exact coordinates of corresponding points have to be calculated. In the first frame, features are selected and tracked with subpixel accuracy in the second frame by using an iterative Lucas-Kanade-algorithm [7] on different image resolution levels. This tracking algorithm works best, if the selected features have a neighborhood with high intensity variances [8]. With a Harris-corner-detector, points are selected and the strongest corners are kept for registration. The calculation of the homography with its eight degrees of freedom requires at least four point correspondences. For robustness reasons, around 200 points are selected in the first frame.

In opposite to other described algorithms, the features found in the first frame are not tracked over the whole image sequence. Instead, the best features for tracking are newly selected in each image.

Before calculating the homography, the points are normalized using isotropic scaling [9]. In addition, outliers in the form of mismatched feature pairs or points belonging to moving objects in the scene need to be eliminated, as the homography is calculated by minimizing a least-square criterion. The robust RANSAC algorithm is used for this task. It iteratively finds a subset of the feature pairs resulting in a homography matrix so that most points fit well.

Figure 2 shows an image with the tracked feature points marked by circles and their displacements by solid lines. In Figure 3, the features are plotted that are classified as outliers by the RANSAC algorithm.

With the calculated homography matrix the two images can be warped onto each other to compensate the camera motion. This is done for every two subsequent images.

2.2. Motion segmentation

Having calculated the homography concatenating two subsequent frames, the first frame is warped onto the second frame. The two images are subtracted and the resulting borders from the transformation are cropped. For smoothing the differential image, a median filter is considered an appropriate filter method (see Figures 4 and 5).

In the next step, a threshold is applied to obtain a black and white image in which white parts refer to areas that do not move consistently with the background (see Figure 6). The applied threshold is quite difficult to choose. Some methods model the histogram of an image by a sum of different probability density functions and choose accordingly a threshold to segment the image. As the obtained differential image is very noisy and the background pixels dominate the scene, no modeling gave reasonable results for the analyzed image sequences. The manually chosen threshold was suited for a whole image sequence acquired under similar lighting conditions.

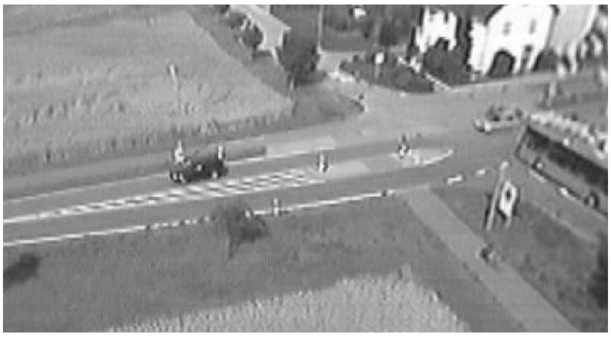


Fig. 4: Part of a gray value image, one of the images used for subtraction.

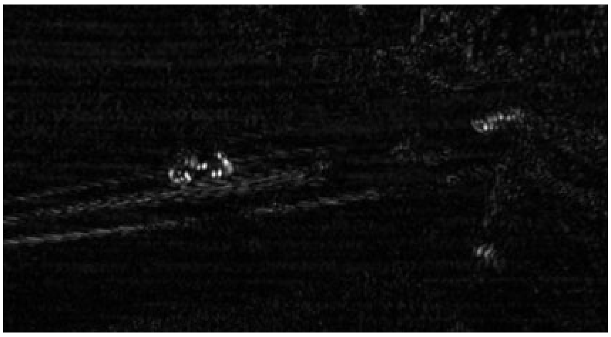


Fig. 5: Median filtered differential image with increased contrast.

To remove the occurring noise pattern and to merge the different small blobs belonging to one object, morphological filtering is applied. First, an opening of the image with a small rectangle of size (2×2)-pixels as structuring element is used to eliminate the noise. Therefore the size of the moving objects needs to be at least (2×2)-pixels. Next, the image is filtered with a closing operation to leave every object in the majority of cases as a single blob (see Figure 7).

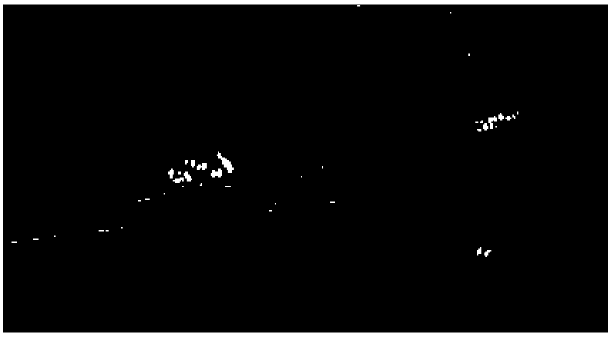


Fig. 6: Differential image after applying a threshold.

The next step is the sorting of the detected blobs and the tracking of the moving objects over time,

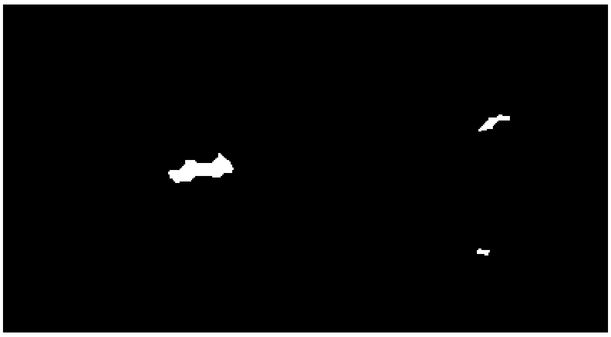


Fig. 7: Result of applying the opening and closing filters.

described in Section 4.

3. MOTION DETECTION WITHOUT PRECEDING STABILIZATION

The process of the detection of moving objects is complicated by the independent 3D-motion of the camera. In this section, our approach to motion detection without preceding image stabilization is described. The applied preprocessing of the image stream is described in Section 2.

3.1. Theoretical consideration

At first, a scene without independently moving objects is considered. If the scene is sufficiently planar, the projection of the relative 3D motion field between camera and scene onto the image plane – the so called 2D motion field [10] – gives a largely smooth 2D vector field. An independently moving object in the planar scene introduces now discontinuities in this 2D-motion field. Either, it moves in the direction of the relative motion of the background, then the absolute value is larger at this spot, or the direction of the motion is different, then there is a discontinuity in the angle component that can be detected.

The only field that can be extracted from the image data is the optical flow field. As a matter of fact, the 2D-motion field is only well approximated by the optical flow at places where the intensity variances are high. Additional assumptions are necessary to calculate an optical flow vector in every pixel. The examination of the smoothness of this dense vector field and the extraction of discontinuities does not necessarily reflect the discontinuities of the 2D-motion field.

Our approach is to calculate the displacements only in distinct points, where the spatial variances are high. The tracking gives then reliable results for the optical flow vector. The resulting vector field is not dense enough to extract discontinuities. Under the assumption that most of the pixels belong to the static and planar background of the scene, it is possible to estimate a homography as in the preceding algorithm that characterizes the relative background motion. Outliers are points that were wrongly matched or places where the assumption that the background is planar does not hold. Finally, these outliers can belong to objects that move independently from the background.

3.2. Extraction of moving Objects

The first step is to calculate a relatively dense optical flow field. The actual image is dividing into small regions. In each region, the best features for tracking are selected and their displacement in the next frame is obtained by the Lucas-Kanade-Algorithm. To increase the robustness of the detection process, the associated features in the second image are tracked back to the first frame. The resulting coordinates are compared to the original ones and wrongly matched pairs are sorted out. In Figure 8, part of a processed image is shown and in Figure 9 the dense field of strong features is plotted.

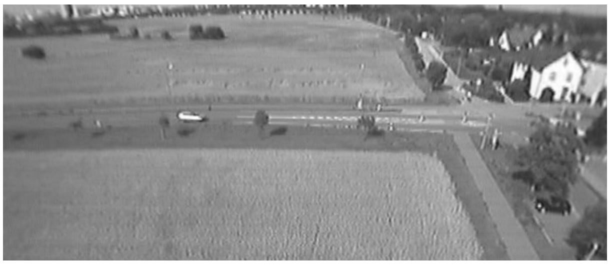


Fig. 8: Part of an image acquired with the hovering UAV.

Successful detection and tracking depends on various conditions. The first one is that moving objects need to feature corners or other trackable structures in order to be detectable by the algorithm. Furthermore, the density of the resulting vector field restricts the minimal size of the observed objects.

Based on the coordinate pairs, a homography matrix is estimated by using a RANSAC algorithm as before and outliers are sorted out. These outliers are processed. Wrongly matched feature pairs are supposed to be not time-consistent and will be sorted out by the following tracking process. At this point, different methods for merging these single

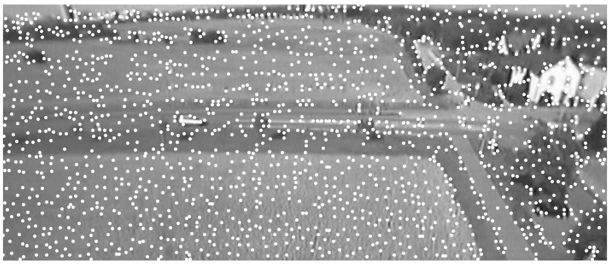


Fig. 9: Extracted feature points for calculating a relatively dense optical flow field.

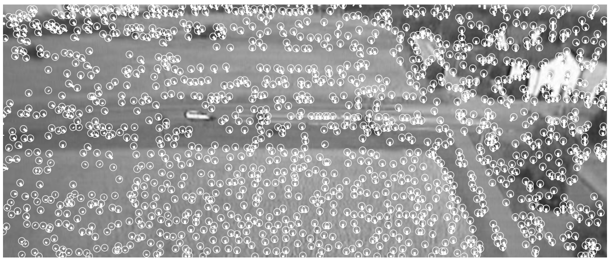


Fig. 10: Inliers describing the relative background motion

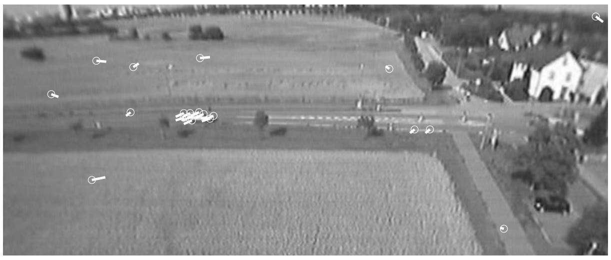


Fig. 11: Outliers including independently moving points

points into areas are possible like applying active contours around regions with more than one outlier point. If characteristics of the searched objects are known, one might use some sort of template matching, too [5].

Assuming that the objects are silhouetted against the background by their gray value, a region growing procedure is used in the following. Successively, each outlier is used as seed point for a region growing process. Starting from the current seed point a connected component is built. A pixel belongs to this area, if it is connected to the seed point by neighbour-pixels that already belong to the area and if its gray value is in a predefined range around the gray value of the seed pixel or the connected neighbour pixel respectively. If the connected area is too large, it is rejected. This happens in uniformly colored regions as for example for points belonging to the sky. The applied threshold should be selected according to the maximum expected size of the objects. The result is a black-and-white image where white regions belong to moving parts of the image. Finally, morphological filtering is used to merge holey regions. The following interpretation and tracking of the objects is described in the next section.

4. TRACKING AND VISUALIZATION

For tracking the moving objects, it is necessary to refer the coordinates to the same frame. The first frame of the scene is chosen as reference frame. The center-coordinates of the moving objects as well as the coordinates of the bounding boxes are transformed into the coordinate system of this frame. The tracking of the moving objects is then performed by a Kalman filter [11]. The used system model is based on motion with constant speed:

$$\text{(2)} \quad \begin{pmatrix} p_{x,k+1} \\ p_{y,k+1} \\ v_{x,k+1} \\ v_{y,k+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & \Delta T \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} p_{x,k} \\ p_{y,k} \\ v_{x,k} \\ v_{y,k} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \eta_x \\ \eta_y \end{pmatrix}$$

The state vector contains the position in pixels $(p_x, p_y)^T$ and the velocity $(v_x, v_y)^T$ in pixels per sampling time ΔT. The system noise $(\eta_x, \eta_y)^T$ is supposed to affect only the velocity. This simple assumption on the objects behavior in combination with the robust Kalman filter algorithm gives good results for most moving objects like cars or people. Temporary occlusions are bridged by the Kalman filter. The association of the newly detected objects to the tracked objects is done by choosing the smallest Mahalanobis distance between the predicted state-vector and the measurements.

Due to several reasons like optical distortion or the fact that the homography model is only an approximation, not every moving area corresponds to an object that is moving with respect to the background. In order to reduce false alarms, a set of higher-level heuristic rules is additionally implemented.

Under the assumption that a lot of those artifacts are not temporally consistent, an object is only considered reliable after it has been successfully tracked a fixed number of frames. In addition, points of the background, where the planarity assumption holds, are supposed to keep their position in the reference frame. The sum of the estimated velocity vectors over time is close to zero. This case is not considered by the Kalman filter model and needs to be eliminated by a higher-level part of the tracking algorithm.

Mosaic images – widely used in computer vision – give a panoramic view of the scene and are suited for improving the visualization for the human operator. As before, the first frame of the image sequence is chosen as reference frame. The homography $\mathbf{H}_{1,i}$ that warps the actual frame onto the first frame is given by

$$\text{(3)} \quad \mathbf{H}_{1,i} = \prod_{j=1}^{i-1} \mathbf{H}_{j,j+1}$$

Each processed frame is transformed with the corresponding homography matrix. The frames are simply superimposed. There is no interpolation necessary, because the mosaic is not further processed. The tracking objects are marked by their bounding boxes.

5. EXPERIMENTAL RESULTS

In this section, the performance of the developed algorithms is illustrated using an in-flight UAV camera sequence. The received image stream contains 25 frames per second. Different scenes were observed. In Figures 12 and 14, the results of the processing of an 200-frame sequence are shown by exemplary frames.

The first algorithm works very well under human supervision. Cars as well as a cyclist were reliably detected and tracked (see Figures 12 and 13). False alarms due to noise and inaccuracy of the homography model were almost always successfully suppressed. As the movement of the objects between two consecutively processed frames needs to be big enough to be detectable in the differential image, not every frame is processed. In the implemented system, the rate of the processed frames is fixed over an image sequence and chosen according to what type of object is tracked. Therefore the algorithm is not able to detect very slowly and fast moving objects at the same time. Pedestrians are detected and tracked also well, if enough time has passed between consequent, processed image frames.

Figures 14 and 15 show the results of the object detection with the second presented algorithm. The second approach turned out to be much more sensitive to changes of parameters than the algorithm with preceding stabilization. The classification of the tracked features belonging to moving objects as outliers worked well but the performance of the region growing for merging small blobs was dependent on the intensity of the objects. The intensity of the objects needed to be explicitly silhouetted against the gray value of the adjacent background pixels for reliable detection. The false alarm rate was much higher, too.

The performance of the tracking algorithm was reliable even if basic model assumptions were made. A problem occurred, if the spatial images of two objects touched or even overlapped for a lengthy time. They were detected as one single moving area in the differential frame and the tracking algorithm failed. After the separation, one of the Kalman filters was reinitialized. A more sophisticated model of the object is investigated at the moment to solve this problem.

6. CONCLUSION

In this paper, two different concepts for the detection of moving objects in an UAV camera sequence were presented. The first algorithm is based on the preceding compensation of the relative camera motion by estimating a geometric transformation. The identification of the moving areas happened with the help of differential frames followed by appropriate filtering.

The second algorithm for detection needed no stabilization. Based on a dense optical flow field, the homography between every two frames was calculated and outliers were processed as possible moving objects. The detection process was followed by a tracking algorithm based on a Kalman filter. Heuristic rules have been successfully implemented to reduce false alarms.

The performance of both algorithms was tested with in-flight image sequences. Both algorithms were reliably able to detect and track moving objects. The detection process without stabilization depended on a clear outline of the objects and on a significant intensity difference between the object and the adjacent background. The procedure with preceding stabilization was more robust against changes of parameters and false alarms.

Even if the used object-model was rather basic, the achieved results were encouraging. The automated choice of the parameters as well as the optimization of the computing time is intended for further work.

ACKNOWLEDGEMENTS

The authors wish to thank the microdrones GmbH.

7. REFERENCES

[1] Medioni G., Cohen I., Brémond F., Hongeng S. und Nevatia R., “Event Detection and Analysis from Video Streams,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 873 – 889, 2001.

[2] Araki S., Matsuoka T., Yokoya N., Takemura H., “Real-Time Tracking of Multiple Moving Object Contours in a Moving Camera Image Sequence,” *IEICE Transactions on Information and Systems*, vol. E83-D (7), pp. 1583 – 1591, 2000.

[3] Sugaya Y. und Kanatani K., “Extracting Moving Objects from a Moving Camera Video Sequence,” *Memoirs of the Faculty of Engineering, Okayama University*, vol. 39, pp. 56 – 62, 2005.

[4] Han M., Kanade T., “Reconstruction of a Scene with Multiple Linearly Moving Objects,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 542–549.

[5] Ollero A., Ferruz J., Caballero F., Hurtado S., Merino L., “Motion compensation and object detection for autonomous helicopter visual navigation in the COMETS system,” in *Proceedings of the IEEE Conference on Robotics and Automation*, 2004, pp. 19–24.

[6] Fischler M. A. and Bolles R. C., “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

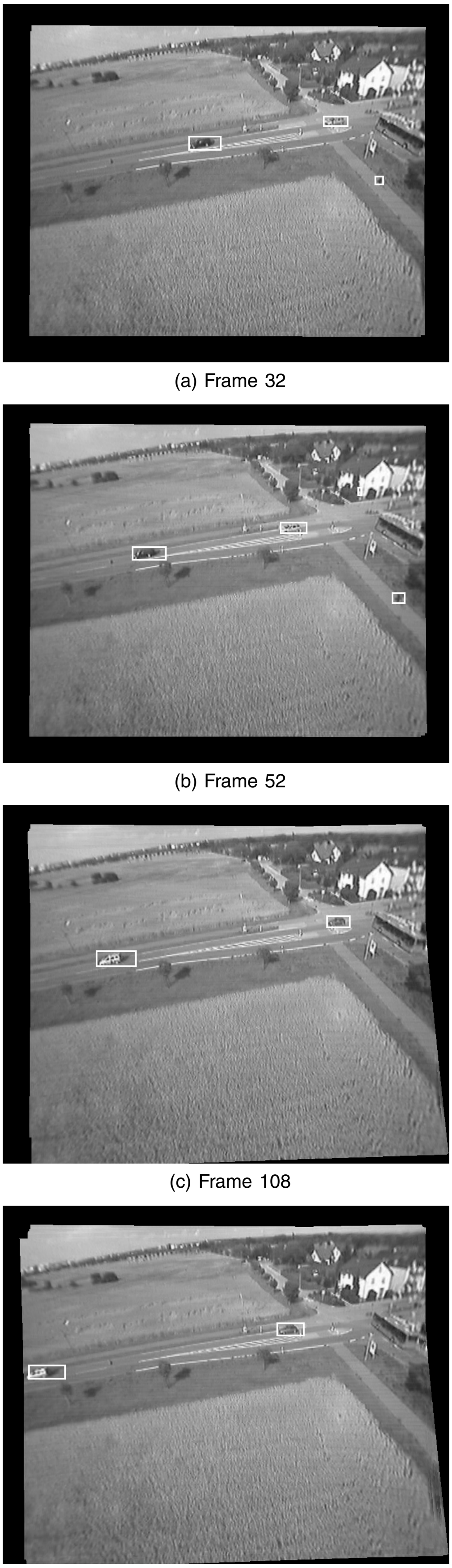
[7] Lucas B. D., Kanade, T., “An Iterative Image Registration Technique with an Application to Stereo Vision,” in *Proceedings of Imaging Understanding Workshop*, 1981, pp. 121–130.

[8] Tomasi C., Kanade T., “Detection and Tracking of Point Features,” Carnegie Mellon University, Pittsburgh, Tech. Rep. CMU-CS-91-132, 1991.

[9] Hartley R., Zisserman A., *Multiple View Geometry, Second Edition*. Cambridge: Cambridge University Press, 2003.

[10] Horn B., *Robot Vision*. Cambridge: The MIT Press, 1986.

[11] Welch G., Bishop G., “An Introduction to the Kalman Filter,” Department of Computer Science, University of North Carolina at Chapel Hill, Tech. Rep. 95-041, 2004.



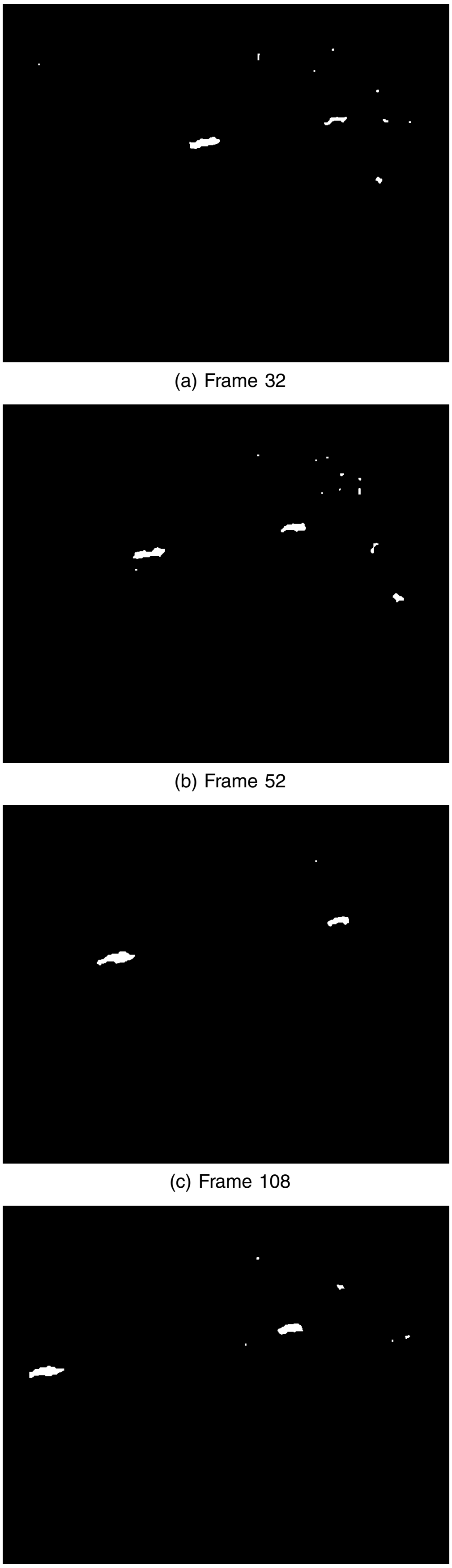
(a) Frame 32

(b) Frame 52

(c) Frame 108

(d) Frame 128

Fig. 12: Tracking results of the processing of an in-flight sequence with preceding stabilization



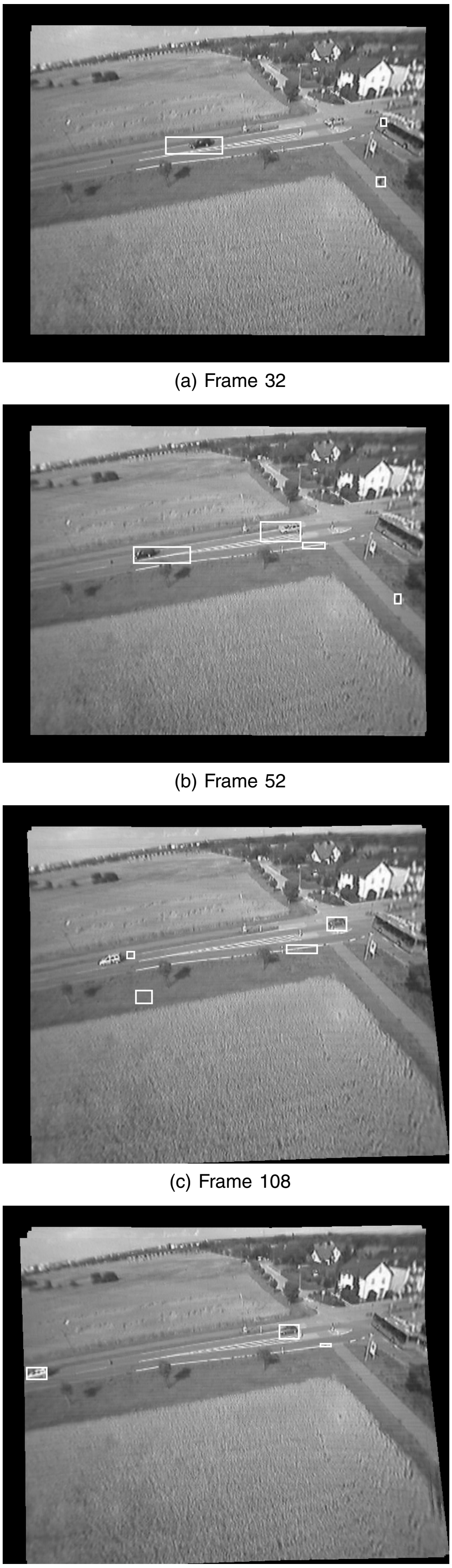
(a) Frame 32

(b) Frame 52

(c) Frame 108

(d) Frame 128

Fig. 13: Extraction of moving areas (in white) with the algorithm with preceding stabilization



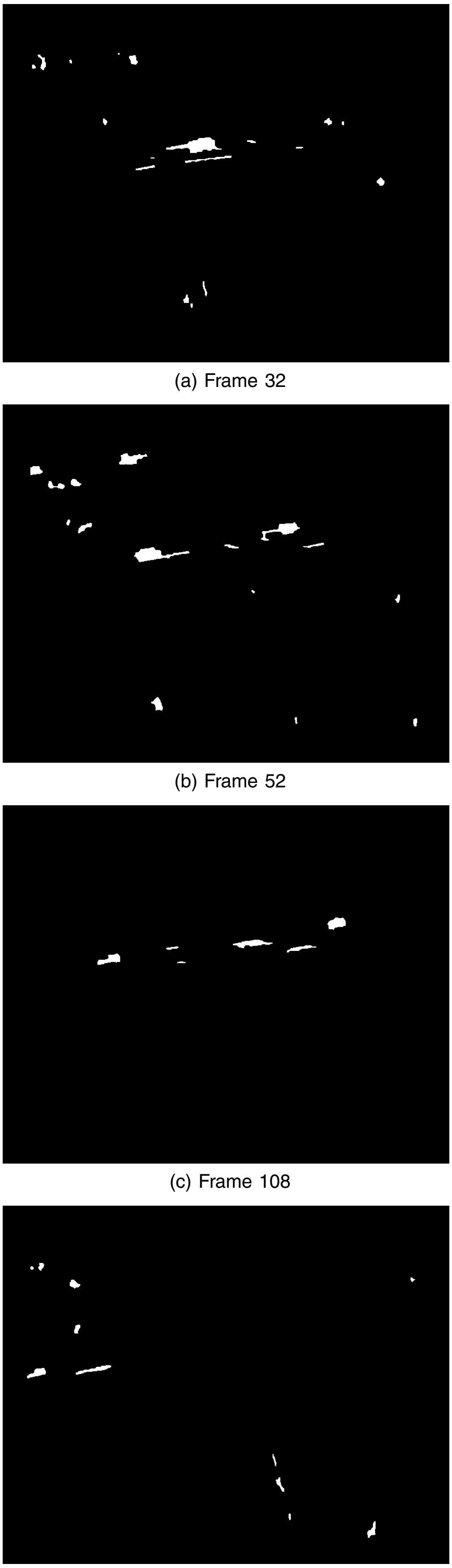
(a) Frame 32

(b) Frame 52

(c) Frame 108

(d) Frame 128

Fig. 14: Tracking results of the processing of an in-flight sequence without stabilization



(a) Frame 32

(b) Frame 52

(c) Frame 108

(d) Frame 128

Fig. 15: Extraction of moving areas (in white) with the algorithm without stabilization