MOTION DETECTION BY CLASSIFICATION OF LOCAL STRUCTURES IN AIRBORNE THERMAL VIDEOS

M. Kirchhof, E. Michaelsen, K. Jäger
FGAN-FOM Research Institute for Optronics and Pattern Recognition,
76275 Ettlingen, Germany - kirchhof@fom.fgan.de

KEY WORDS: Detection of moving objects, classification, sensor pose estimation

ABSTRACT:

In this paper we present a method to efficiently detect moving objects namely vehicles in airborne thermal videos. The motion of the sensor is estimated from the optical flow using projective planar homographies as transformation model. A three level classification process is proposed: On the first level we extract interest location applying the Foerstner - operator. These are subject to the second finer level of classification. Here we distinguish four classes: 1. Vehicles cues; 2. L-junctions and other proper fixed structure 3. T-junctions and other risky fixed structure. 4. A rejection class containing all other locations. This classification is based on local features in the single images. Only structures from the L-junctions class are traced as correspondences through subsequent frames. Based on these the global optical flow is estimated that is caused by the platform movement. The flow is restricted to planar projective homographies which highly reduces the computational time. This opens the way for the third classification. The vehicle class is refined using motion as feature. Inconsistency with the estimated flow is a strong evidence for movement in the scene. This is done by computing a difference image between two sequent frames transformed by a homography to be taken from the same position. The difference images are pre-processed using vehicle properties and velocity.

1. INTRODUCTION

Detection of moving objects even if the observer moves is a very important task for every creature in the world. Think on a predator and a prey. In civilisations this is very important in traffic and security aspects. Even for sensor calibration or sensor pose estimation this is important because the movement between two sequent frames is often so small, that it can not be detected by RANSAC (Fischler & Bolles 1981) or similar outlier detection algorithms. Thermal images provide unique opportunities to detect vehicles and reveal their activity in urban regions at any season in the year and at day or night. Independent of the colour or type all vehicles appear similar with respect to their size and outer conditions (Ernst 2005). Depending on the resolution and aspect active vehicles or parts of them may appear as hot spot (e.g. the exhaust). The exterior of the body of fast vehicles takes the temperature of the surrounding air. Often they will appear as cold spots on the warmer road surface. On a sunny day high temperature differences occur due to shadow and sun. So the appearance of vehicles in such data varies strongly with the exterior condition.
Because the videos are taken from a moving platform the optical flow caused by this movement has to be estimated in order to distinguish the two types of motion in the images: Flow caused by the platform motion versus motion caused by moving objects in the scene.
Estimations of the flow caused by the sensor platform usually assume the scene to be stationary. Therefore, vehicles may cause substantial systematic error. If some of them move in the same direction and cause correspondences with residual movement below the threshold used for outlier decision they may have a considerable impact and spoil the precision of the estimation. For data from urban terrain with a lot of traffic this is not unlikely. The main purpose of this contribution is to solve both problems at the same time: Refine pose estimation by reducing the systematical error caused by moving objects and use the pose estimation namely the estimated homography to improve the detection of moving objects in the image.

2. APRIORI CLASSIFICATION OF INTEREST LOCATIONS

In order to exclude as many unreliable structures from the flow estimation as possible we propose a two level classification of image locations prior to it. The flow estimation is then followed by a third step. We use only this third step of classification to exclude moving objects from the flow estimation. In the first level homogenous and boundary locations are detected and excluded from further consideration. Only a few interest locations remain for which the second level is performed which consumes much more computation per location. Fig. 1 gives an overview of the structure of our classification hierarchy. It is described in more detail in the sections below.

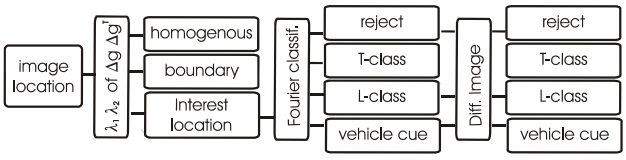

Figure 1. Classification hierarchy for image locations

2.1 First Level: Interest Locations, Boundary Locations and Homogenous Locations

It is not possible to localize correspondence between different frames if the image is homogenous in that location. If an edge or line structure is present at a location in the 2-d image array there may still be an aperture problem. Secure correspondence can only be obtained at locations where a corner, crossing or spot is present. It is proposed to use the averaged tensor product of the gradient g of the grey-values (Förstner, 1994)

$$\nabla g \nabla g^T = \begin{pmatrix} g_x^2 & g_x g_y \\ g_y g_x & g_y^2 \end{pmatrix} \quad (1)$$

where x and y indicate the directions in the image. The discrete version of this matrix is obtained by convolution of the original image with three masks successively (two for the directional derivatives with inherent smoothing and one Gaussian for averaging the squared gradient). For better precision we recommend to use an hourglass like filter in the last averaging step, oriented on the gradient direction (Köthe, 2003). We do not further treat the interest-operator in this contribution.

Note that the remaining pixels have both eigenvalues significantly non-zero and are called *interest pixels* from this on. For image materials like the one presented in Section 3 around one ‰ of the pixels are classified as interest pixels. For further reduction of computational complexity without losing precision of the result we perform non maximum suppression using the eight pixel neighbourhood.

2.2 Second Level: Spots, Corners, T-Junctions and other Structure

As already proposed by Förstner (1994) the pixels around the interest pixels class are grouped in clusters according to proximity and local paraboloids are fitted to these clusters to determine a unique location for each such cluster with sub-pixel accuracy. The result is the set of *interest locations*.

In order to make sure that the interest locations are distributed over the whole image and do not cluster too much in densely structured areas, we set an equally spaced grid (e.g. ten by ten) over the image. Inside of each sub-image only a limited number of interest locations (e.g. maximal fifteen) is accepted. If there are more interest locations available only those with the best value for the structure tensor will be accepted.

Felsberg & Sommer (2001) theoretically derive the recommendation of using polar coordinates once interest locations have been detected by local energy maximizations like the structure tensor operator presented above. As practical consequence there is a research line particularly concerning structure based correspondence evaluation for stereo based on this decomposition of local structure around interest locations (Krüger & Felsberg 2004).

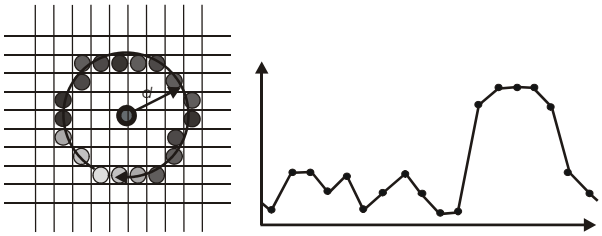

Figure 2. The circular grey-level function: Left pixel grid with interest location and circle around it; right grey–level function along this path

Following this we define a cyclic one-dimensional function g_d along a circular path of radius d around each interest location containing the grey-values as function on the interval $[0,2\pi)$. Fig. 2 shows the principle. Note that while the grey-values are associated to discrete pixel positions the centre of the circle is located at the interest location which is determined with sub-pixel accuracy. In order to avoid strong dependence of the results on the parameter d we repeat the circular sampling at all radii $1 \le d \le d_{max}$ and obtain several functions over the interval $[0,2\pi)$. Figs. 3 display such sampling in the upper left position under 'Original' whereby the functions are appended one after the other so that the total domain is $[0,d_{max}2\pi)$. This function is normalized (byte to one) and average-free so that it takes positive and negative values.

As proper rotation invariant features we chose the Fourier power coefficients of each function. The first ten coefficients i.e. $d_{max}10$ features are used in a nearest neighbour classifier. To obtain a consistent metric we normalised the features to be in the interval $[0,1]$. The feature vectors are displayed in the Figs. 3 upper right under 'Discrete Fourier transform'. Also the grey-values around the interest locations are shown in a window of $(2d_{max})^2$ pixel size (d_{max} was chosen to be 7 here). The lower right part indicates the corresponding location in the image using a black arrow.

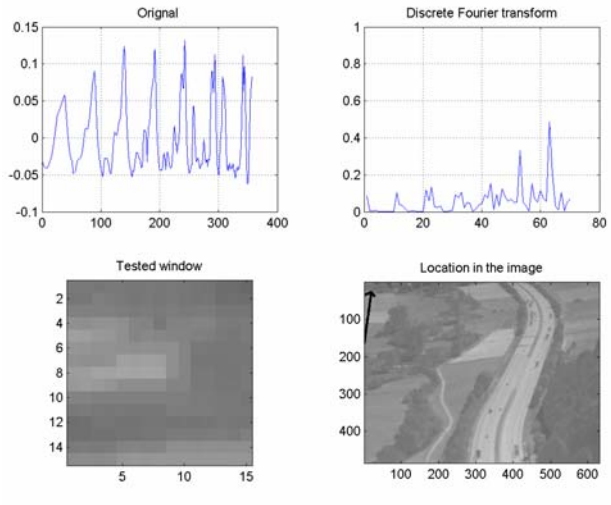

Figure 3. Example for an interest location of the L-class For each radius more than one of the lower frequencies appears strongly. For the larger radii this gets stronger

The following three classes were distinguished:

Class 1 – L-class: The main purpose of the classification is to distinguish among the interest locations those that may contribute reliably to the homography estimation from those that do not. We call this class the *L-class* because it contains things like the corners of terrain regions of different temperature. In urban terrain this includes also building vertices. So also Y-junctions belong to this class.

Class 2 – T-class: One source of systematic error for the homography estimation are image structures that result from partial occlusion. We call this class the *T-class* because typically the occluding part crosses an occluded boundary like a T. Such structure will often not be detected as outlier by the sub-sequent RANSAC analysis described in Sect. 3. because the systematic error is below the thresshold. We include also other unreliable structure like those locations that do not provide robust correspondence – e.g. for lack of curvature.

Class 3 – vehicle cues: Vehicles often move and thus violate the assumptions made for the homography estimations. If their movement is small they may not be detected as outliers by the RANSAC search just like the T-class objects and cause systematic deviation of the estimation. On this stage we can only detect vehicles that appear sufficiently small, so that they

can be discriminated as being spot-shaped by our features within the radius d_{max}.. Larger vehicles in the foreground will often end up in the L-class, because they are to big to be recognized by local features. But these are usually fast enough to be recognized later as not consistent with the homography.

Fig 4. shows the result after the first two classification steps. You can see many falsely detected vehicles because all spot shaped objects were put in this class.

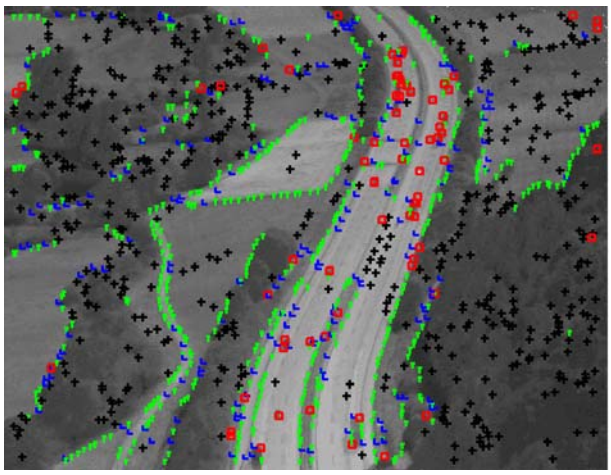

Figure 4. A priory classification of interest location: +=reject, T=T-class, L=L-class, □=vehicle cue class

3. MOTION DETECTION

The main feature for the recognition of vehicles is their movement in the scene. However movements in the aerial video result from two sources, the vehicle movement and the movement of the platform. We are interested in oblique views from sufficiently high moving planes. The appropriate model for the optical flow caused by platform movements is therefore a planar projective homography (Hartley & Zisserman 2000) which is estimated from the video itself.
Any correspondence trace of interest points not consistent with this mapping is resulting from a moving vehicle with high significance or objects far outside the main scene plain. The homography estimation must be robust and precise. If the precision is too bad many interest points may violate the homography mapping and will falsely be detected as vehicle.

3.1 Robust Homography estimation and difference image

We therefore basically use the RANSAC method (Fischler & Bolles 1981). But there are some modifications and improvements. At first we perform a standard adaptive RANSAC with guided matching (Hartley & Zisserman 2000). This means that the algorithm computes the number of samples on its own. Based on the lowest outlier rate this is done by assuring that with a probability of 99% one of the samples consists of inliers only. Then a least square solution is computed and the outliers are proved on this hypothesis. To increase the precision by redundancy we compute three homographies for each time step *t*, i.e. frame pairs *(t,t+1)*, *(t+1,t+2)* and *(t,t+2)*. Taking only correspondences that were tracked through all three images we perform bundle adjustment for homographies (Kraus 1994, Hartley & Zisserman 2000) using minimal parameterization. To this end we compute the quadric decomposition of the homography as proposed by Faugeras (1993). To select the right one of the two solutions we build pairs of solutions with consistent normal vector of the scene plain and compute the back-projection-error as last criterion. The constraint that all correspondences are located on the same 3D-plane reduces the problem from 24 to 14 parameters describing the homographies uniquely. To further improve the homography estimation we only use those view correspondence-triple that have –after initialising over the first triple – the first interest point classified as L-class after the last step of classification.
Figs. 5 and 6 show exemplarily that prior classification of the interest locations and restriction of the input to the estimation process only to the most reliable class – the L-class - leads to high improvement of the flow estimation. The first image was transformed with the computed homography to match the same sensor position as the second but to a sequent time period. Then the absolute differences between the transformed first image and the second image of a pair is computed and coded in grey in Fig. 6.

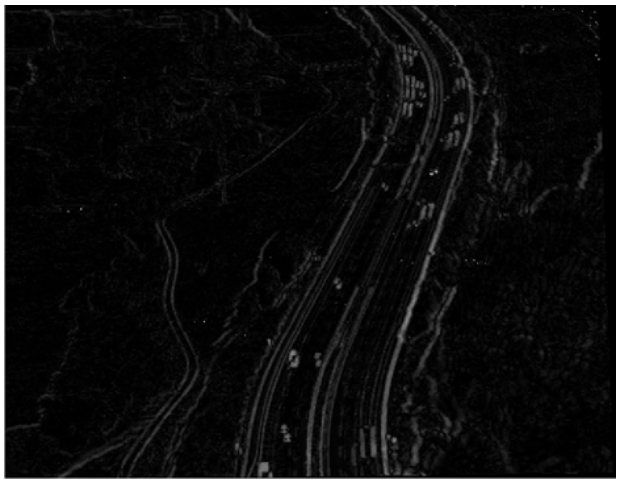

Figure 5. Homography differences without prior classification

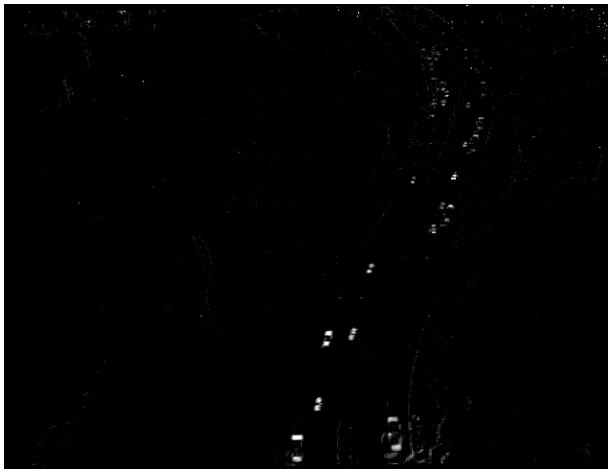

Figure 6. Homography differences with prior classification

3.2 Segmentation and physical information

In the difference image we see different types of objects. There are some systematical errors caused by the fact that the scene is no texture on a plan but truly three-dimensional. The larger the scene depth is the more systematical errors appear in the scene. Mostly we see grey value differences caused by the motion of objects like vehicles.
We have to decide between the two object classes: systematical error from motion and from three-dimensional structure. For

this decision we use some physical information about the objects we like to detect. We can not compute the exact velocity of a vehicle from one difference image but we can compute the velocity relative to the width of the vehicle. Setting a default velocity we can perform segmentation of the difference image. The computed segments can be evaluated using properties like roundness or eccentricity. Objects with a distinct three dimensional structure – like buildings – cause segments with high eccentricity which can easily be distinguished from the others. Fig. 7 and 8. show the different results after the segmentation and after using physical information. The different segments are coded in multiple colours.

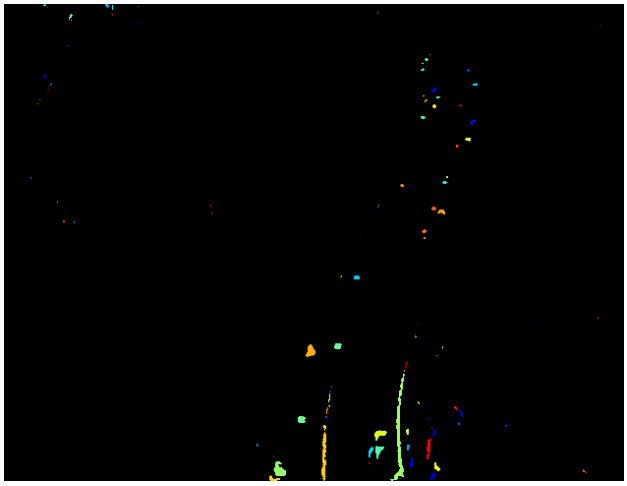

Figure 7 default segmentation result

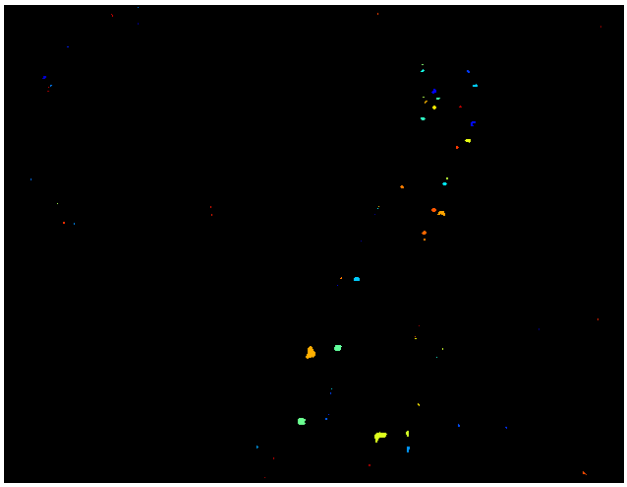

Figure 8 segmentation result with physical information

3.3 Motion detection in the context of interst location

The last step in our approach is the motion detection. In contrast to other methods we analyse only the motion of our interst locations. This highly reduces the computational time. We analyse the region around our interst location in the difference image and can set different thresholds for the size of the nearest segment for our decision based on the a priori information.
The results are shown in Fig.12 and 13.

3.4 Limitations

The first kind of limitations we have to take care of is shown in Fig. 12. If we have to deal with many occlusions the detected motion will often be smaller than the true and can not be detected over the whole sequence. The second limitation we have to take care of comes from the relative scene depth. Figure 9 shows the maximal relative scene depth dependent on the velocity of the air vehicle under two different depression angles and two different maximal errors in the image. If the scene depth is too strong the system can be changed to work in the motion detection step with half resolution. This is in general only the case if the flight altitude is too low and the vehicles we want to detect become very big. This means that the system also works under worse conditions.

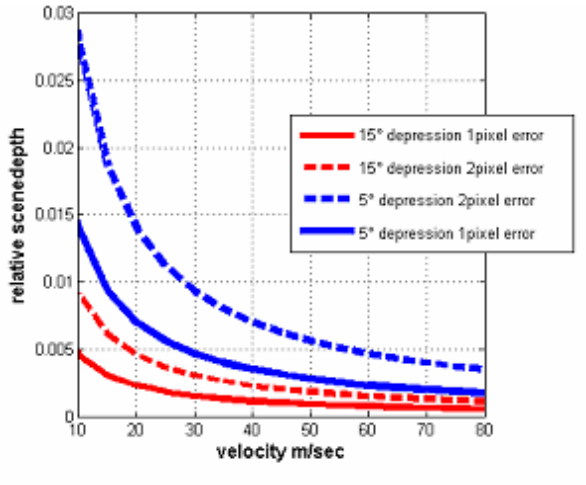

Figure 9 maximal scene depths dependent on velocity

One different way to deal with the errors caused by scene depth was presented in section 3.2. We have to remember that the main scene depth comes from extended structure and therefore lead to extended objects in the difference image. This is why in Fig. 12 false alarms are only isolated single.

4. EXPERIMENTS AND CONCLUSION

4.1 The Data

Both example videos had been obtained with an AIM camera with focal plane array sensitive in the mid thermal domain at 3-5µm. It was forward looking mounted to a helicopter platform. They were taken during day-time, so that vehicles and vegetation appear darker (i.e. colder) than the background. The road surface is quite warm. One frame is displayed in Fig. 10. This needs special care in order not to disturb the flow estimation. Some vehicles on the right lanes exhibit hot exhausts. Vehicles are of considerable different size.
One frame of the second sequence - called small road scene - is shown in Fig. 11. In this scene we had to deal with very low flight altitude which increases the scene depth and many occlusions on the road. Even our reference classification - made the human eye failed in the most frames of this sequence.
Fig. 12 shows the classification result of one frame from the autobahn scene after using the movement feature to detect the vehicles. Note that in contrast to the prior classification presented in Fig. 4 almost no false alarm appears off the road. On the other hand almost all vehicles are correctly marked now.

Figure 10. One frame of the autobahn scene (grey values have been adapted for good visibility here)

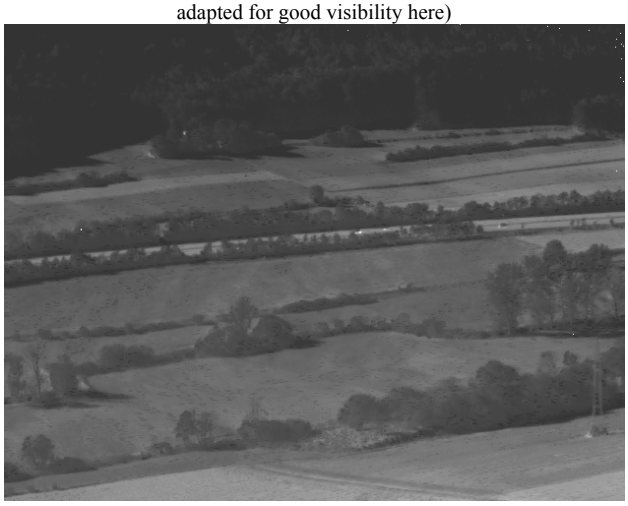

Figure 11. One frame of the small road scene (grey values have been adapted for good visibility here)

4.2 Results

Table 1 and 2 show the classification results of our approach on two randomly picked frames of each scene as absolute values.

	Moving objects	Non moving objects
Detected motion	38 …|… 39	1 ...|... 5
No motion detected	0 ...|... 3	1045 ...|... 995

Table 1: Detected motion in the autobahn scene

	Moving objects	Non moving objects
Detected motion	2 ...|... 16	0 ...|... 5?
No motion detected	0 ...|... 1	995 ...|... 887

Table 2: Detected motion in the small road scene

Here the non moving objects are all inters points from scene fixed structure. You can see the relative results averaged over ten randomly picked frames of each sequence in Table 3. Note that for the small road scene even the human classification failed so that some results are shown with a "?" to remind on the inaccuracy.

	Moving objects	Non moving objects
Detected motion	91.30% | 88 ±4%	0.30% | 0.8 ±0.7%
No motion detected	8.70% | 12 ±4%	99.70% | 99.2 ±0.7%

Table 3: Left: autobahn scene , right: small road scene

Figure 12. Posterior classification for autobahn scene

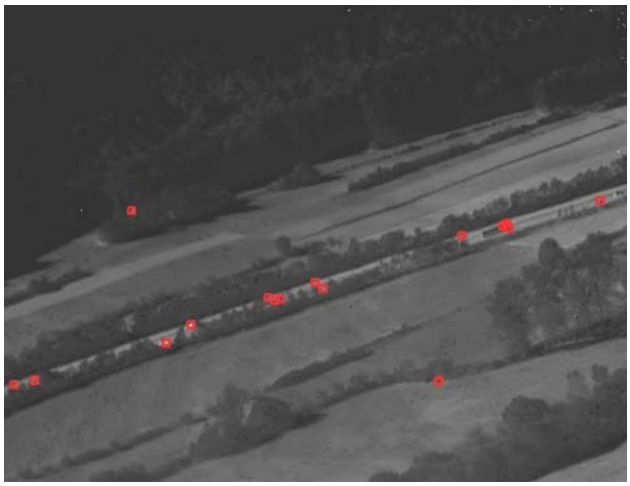

Figure 13. Posterior classification for small road scene

4.3 Discussion and Future Research Lines

Learning samples –about ten for the first three classes- have been taken from the first image of the 400 frames of the first video. For the experiment we used then only the last 200 frames of this video to ensure that we made no self-classification. For the second experiment we used the same trainings data. The prior classification may suffer from inadequate samples if the parameters of the scene (daytime, season, wheather) or of the camera change. On the other hand the posterior classification opens the way for automatic adaptation of the learning example set. We may use constantly appearing moving vehicles as new learning examples for the vehicle class and remove those old ones that could not be affirmed. Also we may use the residual error after estimating the flow as criterion to assess old and new learning samples for the L- and T-class. For experiments following this line of research we need a larger data set.

Deviations in the optical flow from the proper homography may not only result from movement but also from violation of the planarity assumption. Actually, the terrain around the Autobahn shown in Fig. 10 is not flat at all. Yet, in this example the flight altitude is so high compared to differences in the terrain (scene depth) and the time interval for the correspondences between the frames so short, that such effects hardly matter. The 3d structure can be understood as texture on the main plane.

We could also show that for more difficult terrain like in Fig. 11 the systematical error from scene depth could be suppressed by the segmentation described in section 3.2. Particularly, for low flying platforms over urban terrain with high buildings, however, they will. Albeit violating the homography transform such non-planar structure flow must fulfil another weaker constraint – the epipolar condition which is best captured in the fundamental matrix. This can also be estimated from correspondences (Hartley & Zisserman 2000). In such situations an additional classification of interest locations is recommended – consistent with the epipolar constraint versus violating it. The latter must really be moving, while nothing can be asserted about the former. Those may be moving in the direction of the epipolar line. Further studies in this direction are intended. They require appropriate example videos.

Another source of deviation from the homography flow is distortion resulting from the camera (Michaelsen et al. 2004). Particularly, cameras that scan the image using rotating mirrors and only a few sensors exhibit strong non-projective distortions. They violate the planarity assumption inherent in the pin-hole camera model. More recent and future thermal cameras feature focal plane array sensors and thus overcome the problem. The example video of this contribution was obtained by such modern device. The remaining non-projective lens distortions are a minor problem. A linear radial symmetric model for this is usually sufficient. In the estimation procedure outlined in Sect. 3 of this contribution such distortion estimation with one parameter is included.
In Sect. 2.1 we excluded boundary locations from further processing for this contribution. However, it is possible to use straight lines instead of point locations for homography estimation as well. Following this rationale the boundary locations have to be connected and prolonged into sufficiently long and straight line segments. This is going to be one of our future research topics. Freeform boundaries can also be used in this context following the approach of Akav et al. (2004). It is obvious from Fig. 10 that this opens the way to utilize much more of the information contained in such data.

References

Akav, A., Zalmanson G.H., Doythser, Y., 2004. Linear Fature Based Aerial Triangulation. In: *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXV, Part B, XXth ISPRS Congress, Commission 3, pp. 7-12.

Dreschler, L., Nagel, H.-H., 1982. Volumetric model and trajectory of a moving car derived from monocular TV frame sequence of a street scene. *CGIP*, Vol. 20, pp. 199-228.

Ernst, I., Hetschler, M., Lehmann, S., Lippok, A., Ruhé, M., 2005, Use Of GIS Methodology For Online Urban Traffic Monitoring. In: *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVI, Part 8/W27, 3rd International Symposium Remote Sensing and Data Fusion Over Urban Area URBAN2005

Faugeras, O., 1993. *Three-Dimensional Computer Vision*. MIT Press, Cambridge, Mass.

Felsberg, M., Sommer, G., 2001 *The monogenic signal*. IEEE Trans. On Signal Processing, Vol 49 (12), pp. 3136-3144.

Fischler, M. A., Bolles, R. C., 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. Assoc. Comp. Mach.*, Vol. 24, pp. 381-395.

Foerstner, W. 1994. A Framework for Low Level Feature Extraction. In: Eklundh J.-O. (ed.) *Computer Vision – ECCV 94*. vol. II, pp. 383-394.

Haag, M., Nagel, H.-H., 1999. Combination of edge element and optical flow estimates for 3D-model based vehicle tracking in traffic image sequences. *IJCV*, Vol. 35:3, pp. 295-319.

Hartley, R., Zisserman, A., 2000. *Multiple View Geometry*. Cambridge Univ. Press, Cambridge, UK.

Köthe U., 2003. Edge and Junction Detection with Improved Structure Tensor. In: Michaelis B., Krell G. (eds.) *Pattern Recognition*, DAGM 2003, LNCS 2781, Springer, Berlin, pp. 25-32.

Krüger, N., Felsberg, M., 2004, *An explicit and compact coding of geometric and structural image information applied to stereo processing.* Accepted for Pattern Recognition Letters.

Kraus, K., 1994, *Photogrammetrie Band 1/2,* Dümmler Verlag Bonn, Germany.

Michaelsen, E., Kirchhof, M., Stilla, U., 2004. Sensor pose inference from airborne videos by decomposing homography estimates. In: *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXV, Part B, XXth ISPRS Congress, Commission 3, pp. 1-6.

Michaelsen, E., E., Kirchhof, M., Jaeger, K., Stilla, U., 2005. classification of local structures in airborne thermal videos for vehicle detection. In: *Int. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XXXVI, Part 8/W27, 3rd International Symposium Remote Sensing and Data Fusion Over Urban Area URBAN2005