CLASSIFICATION OF LOCAL STRUCTURES IN AIRBORNE THERMAL VIDEOS FOR VEHICLE DETECTION

E. Michaelsen[1], M. Kirchhof[1], K. Jäger[1], U. Stilla[2]
[1] FGAN-FOM Research Institute for Optronics and Pattern Recognition,
76275 Ettlingen, Germany - mich@fom.fgan.de
[2] Photogrammetry and Remote Sensing, Technische Universitaet Muenchen,
80290 Muenchen, Germany - stilla@bv.tum.de

KEY WORDS: Detection of vehicles, interest points, classification

ABSTRACT:

In this paper airborne thermal videos are used to detect vehicles. The movement of the camera is estimated from the optical flow using projective planar homographies as transformation model. A three level classification process is proposed: On the first level the eigenvalues of the squared averaged gradient are used to extract interest locations that can be put into correspondence with other such locations in subsequent frames of the video. 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 namely Fourier coefficients. 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. 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.

1. INTRODUCTION

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. 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 be 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.

A strong evidence for vehicles is movement. But if the video is 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.

2. CLASSIFICATION OF INTEREST LOCATIONS

In order to exclude as many vehicles and other unreliable structure from the flow estimation we propose a two level classification of image locations prior to it. 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.

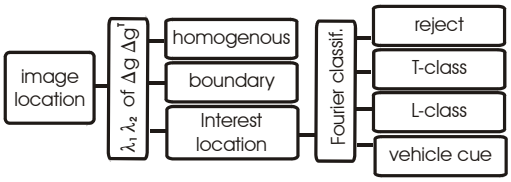


Figure 1. Classification hierarchy for image locations

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). A pixel will be classified as homogenous if the sum of the eigenvalues of this matrix $\lambda_1+\lambda_2$ is smaller than a threshold. A proper

threshold t_h is chosen so that the vast majority of the pixels belong to this class which is excluded from further consideration. t_h may also be calculated from statistic models for image noise and content (Förstner, 1994).
Among the remaining locations boundaries along a straight edge exhibit one positive eigenvalue $\lambda_1>0$ and the other one disappearing $\lambda_2=0$. Förstner (1994) recommends using

$$\frac{\lambda_1\lambda_2}{(\lambda_1+\lambda_2)^2} \le t_b \qquad (2)$$

where the decision threshold t_b is related to the maximum curvature that is permitted for a boundary. Boundary pixels can be further processed by thinning, contour chaining and approximation using straight line segments. We do not further treat them in this contribution.
The remaining pixels with both eigen-values being significantly non-zero are called *interest pixels*. For image material like the one presented in Section 3 around one ‰ of the pixels are classified as interest pixels. For reducing complexity without losing precision of the result we perform non maximum suppression on 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).

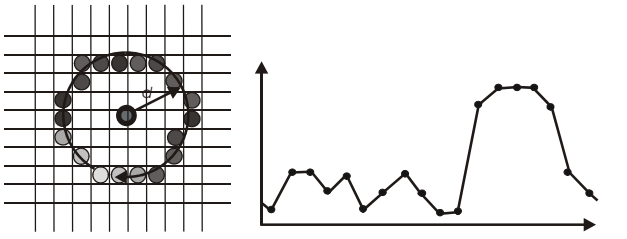


Figure 2. The circular grey-level function: Left pixel grid with interest location and circle around it; right grey–level function along this path

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. 3-5 display such sampling in the upper right 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. 3-5 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.

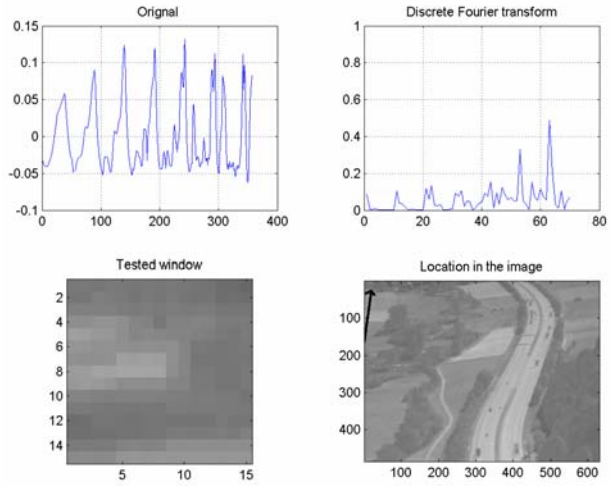


Figure 3. Example for an interest location of the L-class For each radius more than one of the lower frequencies appears strongly. For the larger radii this gets stronger

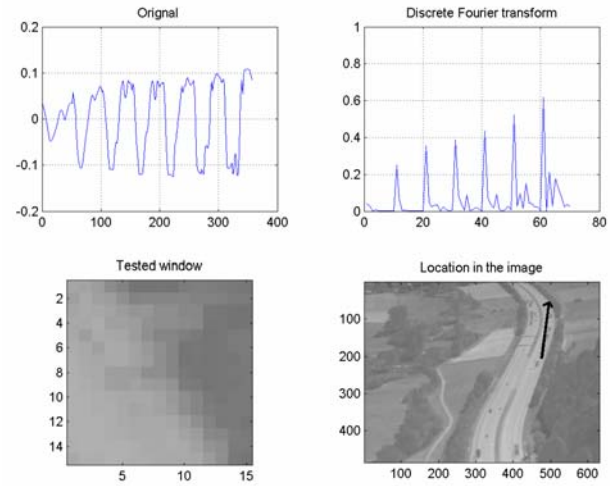


Figure 4. Example for an interest location of the T-class: For each radius the lowest frequency dominates (180° structure). For the larger radii this gets stronger

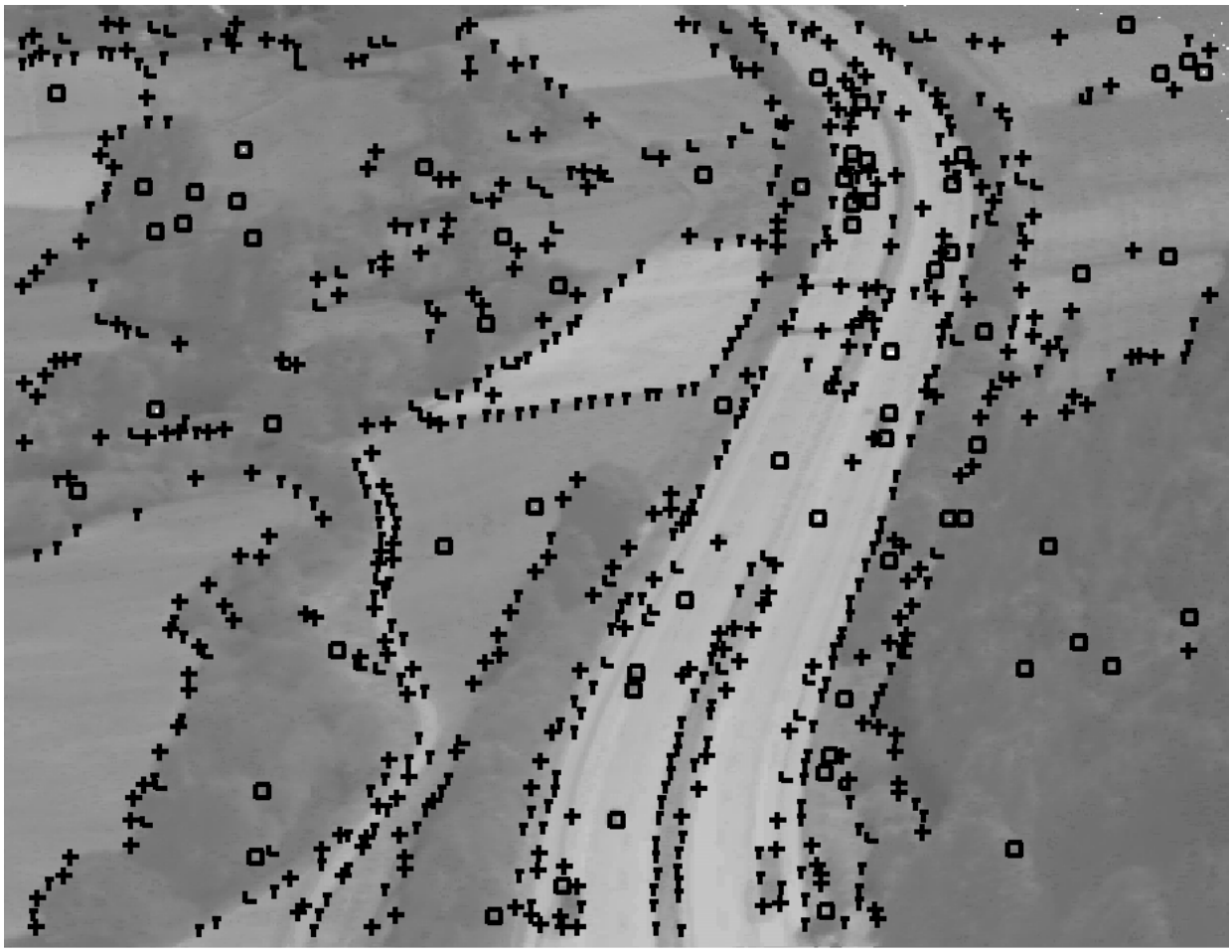


Figure 6. A priory classification of interest location: +=reject, T=T-class, L=L-class, □=vehicle cue class

Class 1 – L-class: The main purpose of the classification is to distingiush 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 temparture. 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 out-lier by the sub-sequent RANSAC analysis described in Sect. 2.3 because the systematic error is below the thresshold. We include also other unreliable structure like those locations that do not provide save 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.

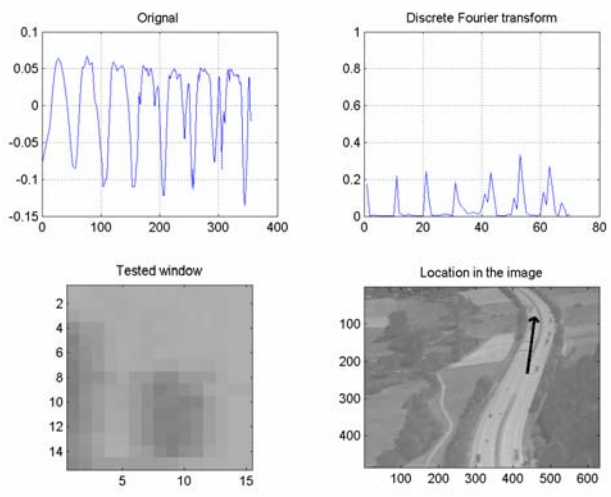


Figure 5. Example for an interest location of the vehicle cue class: Spot structure leads to reduction of features for larger radii, for the very large radii disturbance from surrounding occurs

Figure 10. Posterior classification of vehicles using movement differences from homography flow estimation

3. PLATFORM AND VEHICLE MOTIONS

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.

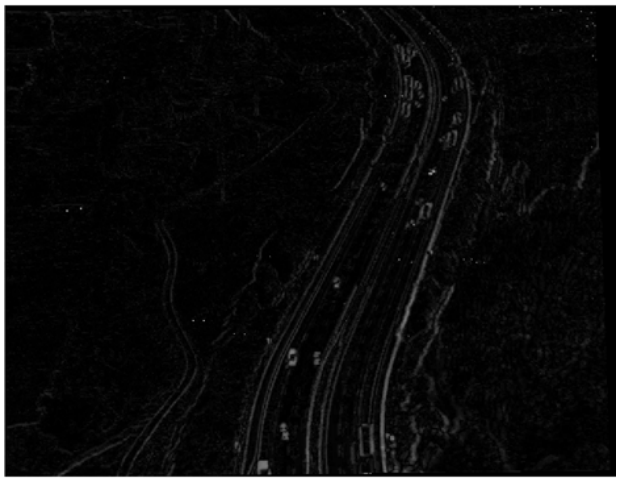
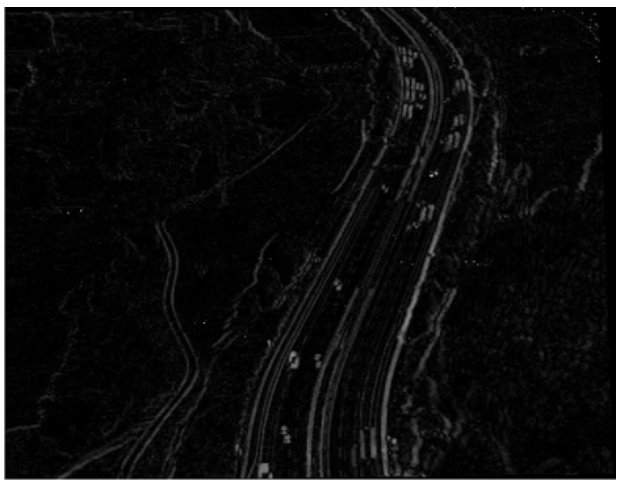


Figure 7. Homography differences with prior classification

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.



b)

Figure 8. Homography differences without prior classification

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. 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 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.
Figs. 7 and 8 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 improvement of the flow estimation. The absolute differences between the transformed first image and the second image of a pair are coded in grey.

4. EXPERIMENTS AND CONCLUSION

4.1 The Autobahn Example Video

The example video has 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. The video was 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. 9. Notice some 'dead pixels' in the upper right corner. 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.

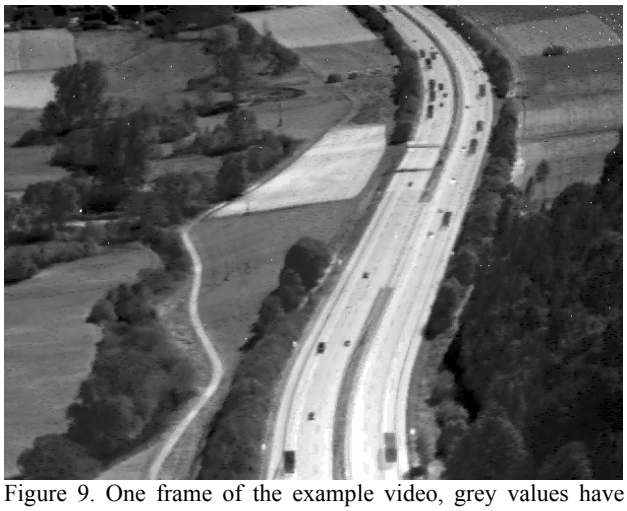


Figure 9. One frame of the example video, grey values have been adapted for good visibility here

Fig. 10 shows the classification after using the movement feature to detect the vehicles. Note that in contrast to the prior classification presented in Fig. 6 almost no false alarm appears off the road. On the other hand almost all vehicles are correctly marked now.

4.2 Discussion and Future Research Lines

Learning samples have been taken not from the same image that was used for the experiment, but admitandly from the same video sequence. So 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 corpus.
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 Figure 9 is not flat at all. Yet, in this example the flight altitude is so high compared to differences in the terrain 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. 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. 9 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.

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.

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.