AUTOMATIC VELOCITY ESTIMATION OF TARGETS IN DYNAMIC STEREO

Norbert Scherer, Michael Kirchhof, Rolf Schäfer

FGAN-FOM Research Institute for Optronics and Pattern Recognition, Gutleuthausstr. 1, 76275 Ettlingen, Germany

KEY WORDS: camera calibration, homography, parameter estimation, position estimation, velocity estimation, structure from motion

ABSTRACT:

In this paper we focus on a stereo system consisting of one stationary and one moving platform. The stationary platform detects motions and gives cues to the moving platform. When an object is in the field of view of both sensors the object's position, velocity, and acceleration are estimated from stereo. To prove the results a second stationary sensor platform is available to compute a reference from traditional stereo. We describe investigations about the accuracy and reliability concerning the automatic velocity estimation of targets in image sequences of multi-ocular systems with emphasis on moving sensor platforms. In order to make the assessment more comprehensible, we discuss a stereo-system consisting of one stationary and one moving platform. Therefore in this context dynamic stereo means that the stereo basis and relative orientation are varying within the image sequence.
We present image sequences originating from different sensors at different positions including one moving sensor that were recorded during the same time span. Since all-day capability is required for warning systems predominantly infrared (IR) sensors were employed. We took image sequences of scenarios with one or more vehicles. The vehicles moved with typical velocities in natural surrounding. Their distances to the sensors was in the close range. The trajectories of all sensors and vehicles were registered with GPS-recorders to obtain ground-truth data.

1 INTRODUCTION

Modern automatic surveillance and warning systems need to ensure the safety of high value targets. To achieve results that guarantee adequate protection against different attacks high quality classifications of threats are essential. Incorporating the movement patterns of active objects appearing in the field of view of the surveillance system into the analysis of potential threats improves the reliability of the classification results. This is true even more, if the information about the movement of the objects is of high quality. The required high quality information could be obtained by a reconstruction of the trajectory of the objects in time and space, which can be done with multi-ocular stereo vision. Furthermore, a more precise definition of the threat is obtained, if the accuracy of the reconstructed trajectory allows the derivation of the three-dimensional velocity and acceleration of each object. Especially for objects moving directly towards the sensor, the three-dimensional position, velocity or even acceleration of the object are of highest interest, since they result in more robust features than shape, size, texture or intensity for the analysis of the threat. Since also moving objects need protection, e.g., convoys of vehicles, the necessity to discuss a moving sensor platform arises. In order to make the assessment more comprehensible, we discuss a stereo-system consisting of one stationary and one moving platform.

The analysis starts with detection and tracking of objects in each individual image sequence. Additionally the trajectory of the moving sensor is estimated by Structure-From-Motion methods (Hartley and Zisserman, 2000), (Kirchhof and Stilla, 2006). The accuracy of the trajectory and relative orientation is improved by sliding bundle adjustment over up to 100 subsequent frames. The resulting trajectory is then matched with the corresponding GPS-track of the vehicle carrying the moving sensor. This ensures that all cameras can now be described in the same coordinate system. Afterwards the correspondence problem is solved and the three-dimensional trajectories and velocities of the observed objects are reconstructed (Scherer and Gabler, 2004). The accuracy of the reconstructed trajectories in space and time is assessed by comparison to the recorded GPS-data. Furthermore to analyze possible performance degradations arising from the movement of sensors, we compare the results of the moving stereo system with the results of a stereo system consisting of two stationary sensors. The solution to the correspondence problem and the trajectory and velocity estimation of the observed object is identical for the stationary and the dynamic case.

1.1 Related Work

Typical approaches for the estimation of the trajectory and velocity of moving objects assume that the objects move on an observed plane. In this case a monocular image sequence is sufficient to determine the trajectory and velocity of a moving object. One typical example may be (Reinartz et al., 2006). Reinartz et al. compute a georeferenced image sequence which allows to compute the position and velocity directly from image displacements. Nistér presented quite different work about dynamic stereo (Nistér et al., 2004) introducing a system of a stereo camera mounted on a vehicle. The advantage of this system is that the stereo basis and relative orientation remain constant over the sequence although the sensors are moving. The approach presented here is based on (Scherer and Gabler, 2004) where the range and velocity of objects at long ranges were computed from triangulation. Caused by the long range application the contribution focuses on discretization effects in the range and velocity estimation. For such applications the relative rotations between the sensors is very important while the relative positioning error can be very large without effecting the results. Therefore positioning the sensors with GPS was sufficient in that work.

Our application is in close range where the relative positioning error induces very large disparities. Therefore the registration of the stationary sensor positions was improved by adjustments over many GPS-measurements supported by distance measurements. Additionally the registration of the orientation was done by the comparison of the sensor's view with the view of a virtual camera computed from laser measurements taken from a helicopter.

1.2 Notation

We describe experiments in which we employed different vehicles ($\mathcal{V}_1$, $\mathcal{V}_2$ and $\mathcal{V}_3$). The vehicles $\mathcal{V}_1$ and $\mathcal{V}_2$ where monitored with three different cameras ($\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_V$). Cameras $\mathcal{C}_1$ and $\mathcal{C}_2$ were stationary, whereas camera $\mathcal{C}_V$ was mounted on top of vehicle $\mathcal{V}_3$, which followed $\mathcal{V}_1$ and $\mathcal{V}_2$. $\mathcal{V}_1$ was driving in front of $\mathcal{V}_2$. The positions of the vehicles were tracked with Global Positioning Systems (GPS). The GPS-receivers are described by the letter $\mathcal{G}$. The connection to the vehicle, which's position is measured, is established by the use of an index that corresponds to the vehicle, i. e. the GPS-system in the vehicle $\mathcal{V}_1$ is depicted by $\mathcal{G}_1$. To obtain further ground-truth information we used the information of a GPS-system $\mathcal{G}_H$ mounted on a helicopter $\mathcal{V}_H$, which performed laser measurements of the area in which our experiments took place.

2 DATA ACQUISITION

The data described in this paper had been recorded during a measurement campaign realized by the authors and other members of FGAN-FOM. The campaign took place at the end of September 2006.

Three infrared cameras ($\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_V$) were used as imaging sensors. The cameras $\mathcal{C}_1$ and $\mathcal{C}_2$ were from AIM INFRAROT-MODULE. Camera $\mathcal{C}_V$ was from FLIR Systems. The technical details of the cameras are summarized in table 1.

Ground truth data were recorded using the Global Positioning System (GPS). The vehicles $\mathcal{V}_1$ and $\mathcal{V}_2$ employed portable GPS-systems GPSMAP 76CSx ($\mathcal{G}_1$ and $\mathcal{G}_2$). $\mathcal{V}_3$, which carried the Camera $\mathcal{C}_V$, was equipped with a GPS-mouse system ($\mathcal{G}_3$). The portable GPS-systems were also used to determine the positions of the stationary cameras $\mathcal{C}_1$ and $\mathcal{C}_2$ and some additional outstanding points in the terrain.

Furthermore the terrain was scanned with a Riegl LMS Q650 laser scanner, which was mounted on a Helicopter ($\mathcal{V}_H$) of type Bell UH-1D. The scans produced several overlapping stripes containing height profiles of the terrain. $\mathcal{V}_H$ was equipped with an Inertial Measurement Unit (IMU) and a GPS-antenna ($\mathcal{G}_H$). The signals from both sensors were processed by a special computer in such a way that position, orientation and acceleration of the helicopter are known during the data acquisition phase. Further details of the helicopter equipment can be found in (Hebel et al., 2006).

3 TRIANGULATION AND VELOCITY ESTIMATION

We are now going to describe our general approach. First we show the general procedure of obtaining a three-dimensional track for the case of two stationary cameras. Second the necessary modifications to expand the approach to the case of moving sensors are depicted.

Generally the approach is divided into two steps. In the first step the image sequences of each sensor are processed separately. The results of this step are then used as input to the second step. The second step combines the results of the analysis of the two image sequences and constitutes the desired three-dimensional track of the object of interest.

Camera	FOV	Spectrum	No. of Pixels
$\mathcal{C}_1$	$17.5^\circ \times 13.3^\circ$	4.4 - 5.2 μm	640×480
$\mathcal{C}_2$	$18.7^\circ \times 14.1^\circ$	2.0 - 5.3 μm	384×288
$\mathcal{C}_V$	$20.0^\circ \times 15.0^\circ$	3.0 - 5.0 μm	320×256

Table 1: Technical Data of the used IR-cameras.

3.1 Stationary Case

The first step of creating a three-dimensional track applies an Infrared Search and Track (IRST) algorithm to each of the image sequences. This algorithm starts with pre-processing the images to correct for sensor specific inhomogeneities. Afterwards the image sequences are integrated to increase the signal-to-noise-ratio. In the resulting sequences point like objects are tracked, so that two-dimensional tracks of these objects are created in each image sequence.

Figures 1, 2 and 3 show examples of the 'point-like objects' as seen from the cameras $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_V$. The images have been taken at the same time. The point-like objects found by the application of the IRST-algorithm are marked with rectangles. Please notice that not all of the marked points in one image must have a corresponding mark in any of the other two images. On the other hand the blue rectangle in each image marks a point that has correspondences in the other images. That point belongs to the back of vehicle $\mathcal{V}_1$. An example of a two-dimensional track resulting from one object is given in figure 4.

The second step uses the two-dimensional tracks that have been created by the IRST-algorithm and reconstructs the three-dimensional trajectories of objects by combining corresponding two-dimensional tracks from the image sequences. For this reconstruction the knowledge of the camera's position and orientation are important. Further details and a theoretical discussion of the accuracy and reliability of this approach can be found in (Scherer and Gabler, 2004).

3.2 Dynamic Case

In the dynamic case one camera is moving during the observation. Therefore the second step of the analysis procedure is modified in such a way that the possible changes of the positional information (position and orientation) of the camera are considered. These information are obtained with Structure from Motion methods as described later in 4.3.

Since these methods only return relative positional informations, they have to be transformed into our reference frame by an Euclidean transformation. This is done by fitting the whole track of the moving camera $\mathcal{C}_V$ to the whole track of the sensor carrying vehicle $\mathcal{V}_3$ obtained from the GPS-system $\mathcal{G}_3$.

Due to the variance of the positional information obtained by our GPS-receivers a final translation of the whole track of the moving camera is needed. This translation is obtained by comparing the position of the moving camera at one point in time with the position where it is seen in the corresponding image of one of the stationary cameras.

4 CALIBRATION AND SYNCHRONIZATION

The measurement of the velocity of vehicles requires knowledge about the time at which the vehicle is at a certain point. Since we want to estimate the velocity from different cameras and compare the results with ground-truth-data the data-streams of all sensors need to be 'synchronized' not only in time but also in space.

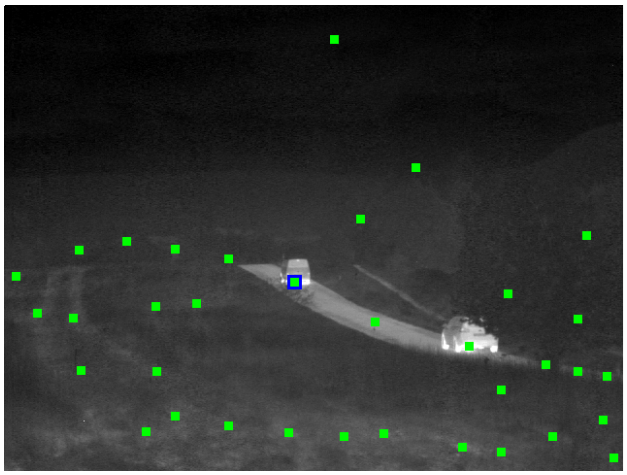

Figure 1: Image from camera $\mathcal{C}_1$. In the center of the picture vehicle $\mathcal{V}_1$ can be seen. Vehicle $\mathcal{V}_2$ follows $\mathcal{V}_1$. The rectangles mark 'objects' for which two-dimensional tracks had been created by the IRST-algorithm.

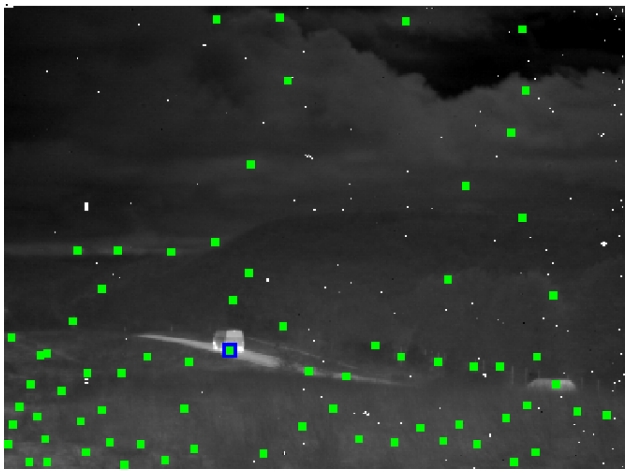

Figure 2: Image from Camera $\mathcal{C}_2$ taken at the same time as the image shown in figure 1. Rectangles mark 'objects' that had been found by the application of the IRST-algorithm.

4.1 Spatial Registration

The positions of the stationary cameras $\mathcal{C}_1$ and $\mathcal{C}_2$ were established by combing all available position measurements (GPS, position information derived from the laser scans of the helicopter and some additional distance measurements), stating a minimization problem for the position and distance differences between the measurements and solving it with the Levenberg-Marquardt algorithm.

As a result of this procedure we obtained the camera positions within a precision less than half a meter, which is much better than the variance of one of our single GPS-measurements.

Now that the position of the stationary cameras had been fixed we obtained the orientation of the cameras with the virtual overlay technique. By this we use the data of the height profiles from the laser scanner to produce 'images' of a camera with a virtual reality tool. These pictures are then compared with the real camera image. The comparison is done in an overlay of the real and the virtual image. The parameters of the virtual camera are then manually modified until a reasonable conformance between both images is reached. An example of an overlay image is seen in figure 5.

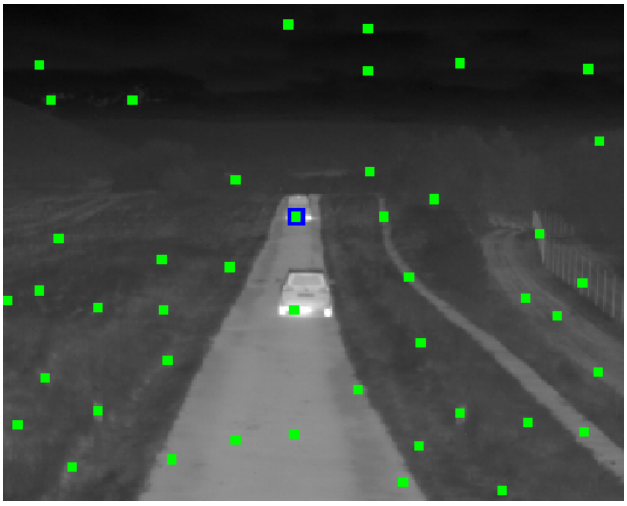

Figure 3: Picture from camera $\mathcal{C}_V$ taken at the same time as the picture 1. 'objects' resulting from the processing of this video stream by the IRST-algorithm are again marked as rectangles.

Figure 4: Same image as in figure 1. Here a two-dimensional track, as it resulted from the application of the IRST-algorithm to the whole image-sequence of camera $\mathcal{C}_V$, is overlayed. The track belongs to the 'object' that had been marked with a blue rectangle in the center of figure 1.

4.2 Temporal Registration

For the temporal registration, we need to synchronize our cameras to a global time, e. g. GPS-time. Fortunately in order to achieve the synchronization we only need to determine one constant time-shift for each camera, since each image sequence is equipped with timestamps of a local clock. We identified this constant at that parts of the image-sequences that show starting vehicles, since this incident could be identified with high precision in the GPS-time-stream.

4.3 Structure from Motion

Since the GPS-data contains no information about the orientation (rotation) of the sensor an image-based reconstruction approach is required. The reconstruction is initially computed independently from the available GPS-data and is improved by sliding bundle adjustment which takes the GPS-data into account. We assume that the internal camera parameters are known for example by the use of self-calibration techniques like (Zhang, 1999).

In the first step points of interest are detected in the first frame (Förstner and Gülch, 1987). These points are then tracked

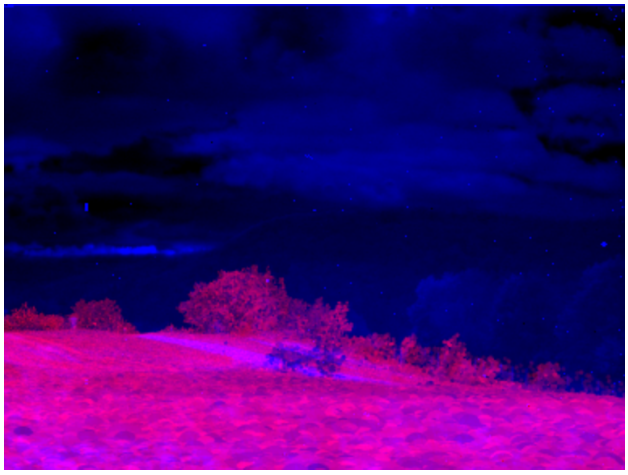

Figure 5: Example of an overlay picture used to obtain the orientation of camera $\mathcal{C}_2$. Reddish scene parts belong to the image as seen from a virtual camera with data based on the laser measurements. Bluish parts belong to the IR-image from the camera $\mathcal{C}_2$.

through the sequence based on template matching (Lucas and Kanade, 1981). Outliers of the tracking process are detected with RANSAC (random sample consensus) (Fischler and Bolles, 1981) for homographies based on our previous work (Kirchhof and Stilla, 2006). Every tenth frame the tracked point set is improved by applying the point of interest detection to the current frame.

The relative orientation can now be computed from the essential matrix using the five point algorithm of Nist'er (Nistér, 2003). This relative orientation enforces the triangulation of the corresponding 3d-point set. Subsequent frames can now be stitched to the reconstruction by linear regression followed by nonlinear minimization of the reprojection error using Levenberg-Marquardt (McGlone et al., 2004). The 3d-point set can now frequently be updated in a robust way by retriangulation again using RANSAC.

As mentioned above we refine the reconstruction with bundle adjustment over the latest one hundred frames using Levenberg-Marquardt taking the GPS-data into account. Although we used a tracking and matching strategy the computed tracks may be corrupted by slightly moving objects or drifts of the tracker. The Huber robust cost function (Huber, 1981) reduces the influence of such errors while it is still convex. Therefore no additional local minima are induced by it.

5 EXPERIMENTS

For the comparison of the stationary case with the dynamic case we tracked vehicle $\mathcal{V}_1$ with all three cameras. The figures 1 to 3 show images of the sequences. Within these images the blue rectangles mark the objects in the images that we used to reconstruct the three-dimensional trajectory of $\mathcal{V}_1$. For the camera $\mathcal{C}_1$ the two-dimensional track of the object marked with a blue rectangle is shown in figure 4.

5.1 Stationary Case

The result of the evaluation of the stationary cameras $\mathcal{C}_1$ and $\mathcal{C}_2$ is visible in figure 6 as a blue line. Because of the field of view of the camera $\mathcal{C}_2$ only the last part of the track is visible. In that range the position of the track coincides very well with the real track of the vehicle.

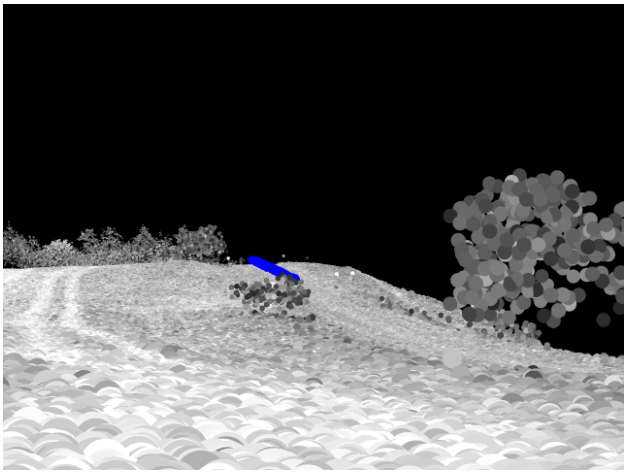

Figure 6: Reconstruction of the three-dimensional trajectory of vehicle $\mathcal{V}_1$ (blue line) based on the images from the stationary cameras $\mathcal{C}_1$ and $\mathcal{C}_2$.

5.2 Dynamic Case

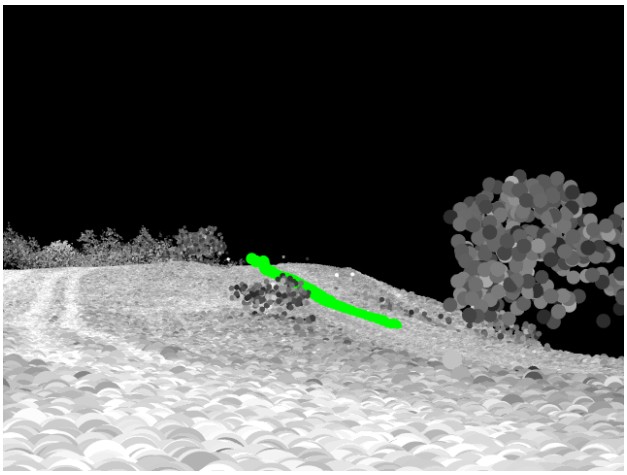

Figure 7: Three-dimensional trajectory of $\mathcal{V}_1$ (green line) reconstructed from the cameras $\mathcal{C}_1$ and $\mathcal{C}_V$.

For the case including one moving camera $\mathcal{C}_V$, which was the main purpose of our investigation, the resulting three dimensional track of vehicle $\mathcal{V}_1$ is shown in figure 7 as a green line. Again the track is in good accordance with the track of vehicle $\mathcal{V}_1$, as could be seen by comparison with the two-dimensional track shown in figure 4.

5.3 Comparison

A more quantitative comparison of the reconstruction results with the real trajectory of $\mathcal{V}_1$ as the one presented in the previous sections 5.1 and 5.2 is shown in figure 8 as a top view. The center of the depicted coordinate system coincides with the position of the camera $\mathcal{C}_1$. The vehicle $\mathcal{V}_1$ moves from the upper left corner to the lower right corner. It is obvious that in the stationary (blue dots) and the dynamic case (green dots) the reconstructed positions match well with the GPS-measurements (black dots). At the end the dynamic track shows problems arising from instabilities in computation of the position and orientation of the moving camera $\mathcal{C}_V$. These problems result in reconstructed positions showing a backward moving vehicle in contradiction to the real trajectory of $\mathcal{V}_1$. Possibly the problem arises from the same source as the fact that the pitch of the camera $\mathcal{C}_V$ needed a correction of about 1° before the data could be processed. Up to that

point the results are quite good, as seen in the transition to the track reconstructed from the stationary cameras.

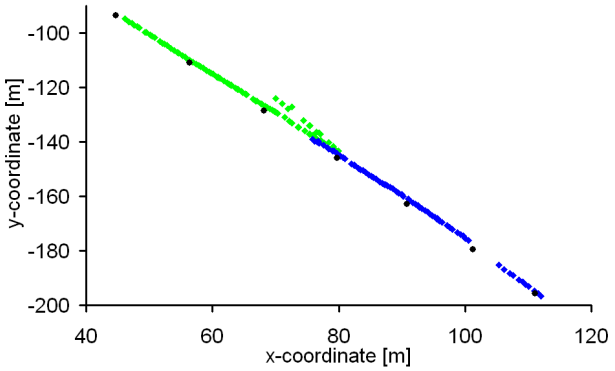

Figure 8: Comparison between the positions obtained in the stationary (marked by green dots) and the dynamic (blue dots) case with GPS-measurements. Black dots represent the positions of vehicle $\mathcal{V}_1$ as measured with $\mathcal{G}_1$.

Based on the good three-dimensional reconstruction of the position it is now possible to derive the velocity of the vehicle $\mathcal{V}_1$. The velocity is calculated as a running linear regression over 15 time points. Figure 9 shows the results as green dots for the dynamic case and blue dots for the stationary case. The mean velocity value from the GPS-data is shown as a black line. For both cases the velocities vary around the mean value obtained from the GPS-data. Obviously the velocity values from the dynamic case are distorted by the reconstruction problems of the trajectory mentioned in the previous paragraph, which start at 26.5 seconds in figure 9. On the other hand it is seen, that without these problems the stationary data are a good continuation of dynamic ones. Furthermore the figure shows that the variation of the velocity obtained from the dynamic case (up to 26.5 seconds) is equal to the variation of the velocity obtained from the stationary case.

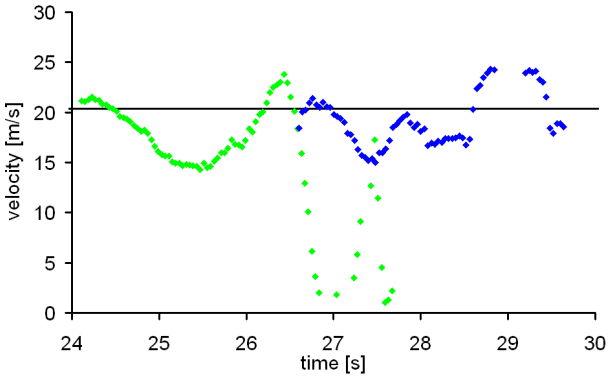

Figure 9: Comparison of the velocities obtained from the stationary (marked by green dots) and the dynamic (blue dots) case. The horizontal black line represents the mean velocity obtained from the GPS-data.

6 CONCLUSION

We described the results obtained from systems for automatic velocity estimation from stereo data. One of the systems had a fixed stereo basis, the other had a varying stereo basis since the second camera of the stereo pair was mounted on a moving vehicle. It has been shown that the described approach is applicable in principle, provided that a high quality registration of all necessary data is available.

The structure from motion methods need to be supported by additional metrical information, e. g., GPS in our case. But the obtainable results seem to depend strongly on the quality of this additional information. In our case the velocity calculation was quite good in the beginning, but failed when the position estimates obtained by the structure from motion approach break down. Further detailed investigations will be necessary to find the cause of this failure.

7 OUTLOOK

As we pointed out above the structure from motion approach is the bottleneck of the presented work. The 3d registration can be improved by considering not only the relative orientation of the monocular moving sensor but also the relative orientation between the moving and the stationary sensor. This is in general a wide base stereo approach. Therefore the used descriptors for points of interest have to be replaced by rotational and scale (and in the optimal case affine) invariant descriptors like SIFT (Lowe, 2004) - or MSR - features. Additional improvement can be obtained by detecting the moving objects and exclude them from further processing.

8 ACKNOWLEDGMENTS

The work described above has been supported by BMVg (Bundesverteidigungsministerium, German Ministry of Defense) and BWB (Bundesamt für Wehrtechnik und Beschaffung, German Office for Armament and Procurement). The assistance of the WTD 81 (Wehrtechnische Dienststelle für Informationstechnologie und Elektronik) for preparation and realization of the measurement campaign at Greding in September 2006 is gratefully acknowledged.

REFERENCES

Fischler, M. A. and Bolles, R. C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the Association for Computing Machinery 24(6), pp. 381–395.

Förstner, W. and Gülch, E., 1987. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features. In: ISPRS Intercommission Workshop, Interlaken.

Hartley, R. and Zisserman, A., 2000. Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, UK.

Hebel, M., Bers, K.-H. and Jäger, K., 2006. Imaging sensor fusion and enhanced vision for helicopter landing operations. In: J. J. G. Jacques G Verly (ed.), Signal Processing, Sensor Fusion and Target Recognition XV, Vol. 6226, SPIE.

Huber, P. J., 1981. Robust Statistics. John Wiley Publishers.

Kirchhof, M. and Stilla, U., 2006. Detection of moving objects in airborne thermal videos. ISPRS Journal of Photogrammetry and Remote Sensing 61, pp. 187 – 196.

Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. Int. J. of Computer Vision 60(2), pp. 91–110.

Lucas, B. T. and Kanade, T., 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. In: Proc. of Image Understanding Workshop, pp. 212–130.

McGlone, J. C., Mikhail, E. M. and Bethel, J. (eds), 2004. Manual of Photogrammetry. 5th edn, American Society of Photogrammetry and Remote Sensing.

Nistér, D., 2003. An efficient solution to the five-point relative pose problem. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03) 02, pp. 195.

Nistér, D., Naroditsky, O. and Bergen, J., 2004. Visual odometry. Vol. 1, pp. I–652–I–659.

Reinartz, P., Lachaise, M., Schmeer, E., Krauß, T. and Runge, H., 2006. Traffic monitoring with serial images from airborne cameras. ISPRS Journal of Photogrammetry and Remote Sensing 61, pp. 149 – 158.

Scherer, N. and Gabler, R., 2004. Range and velocity estimation of objects at long ranges using multiocular video sequences. International Archives of Photogrammetry and Remote Sensing 35(Part B5), pp. 741–746.

Zhang, Z., 1999. Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In: International Conference on Computer Vision (ICCV'99), Corfu, Greece, pp. 666–673.