Fast detection of moving objects in complex scenarios

Clemens Rabe, Uwe Franke, and Stefan Gehrig
DaimlerChrysler AG
70546 Stuttgart, Germany
{clemens.rabe,uwe.franke,stefan.gehrig}@daimlerchrysler.com

***Abstract*— More than one third of all traffic accidents with injuries occur in urban areas, especially at intersections. A suitable driver assistance system for such complex situations requires the understanding of the scene, in particular a reliable detection of other moving traffic participants. This contribution shows how a robust and fast detection of relevant moving objects is obtained by a smart combination of stereo vision and motion analysis. This approach, called 6D Vision, estimates location and motion of pixels simultaneously which enables the detection of moving objects on a pixel level. Using a Kalman filter attached to each tracked pixel, the algorithm propagates the current interpretation to the next image. In addition, a Kalman filter based ego-motion compensation is described that takes advantage of the 6D information. This precise information enables us to discriminate between static and moving objects exactly and to obtain a better prediction. This speeds up tracking and a real-time implementation is achieved. Examples of critical situations in urban areas exhibit the potential of the 6D Vision concept which can also be extended to robotics applications.**

Fig. 1. Typical scene causing segmentation problems to standard stereo systems.

I. Introduction

According to current accident statistics more than one third of all accidents with injuries occur at intersections. Most of them are caused by distraction, nonattention or misinterpretation of the situation [1]. Looking at collision partners it turns out that the vast majority of accidents is "colliding with moving objects". A suitable driver assistance system for such highly complex situations requires the complete understanding of the traffic scene. Besides the perception of the infrastructure it must detect other moving traffic participants and measure their movement precisely to predict potential collisions.

Using a stereo camera system the three-dimensional structure of the scene is easily obtained. This information is commonly accumulated in an evidence-grid-like structure [2]. We refer to it as the birdview map. Since stereo does not reveal any motion information, usually this map is segmented and detected objects are tracked over time in order to obtain their motion. The major disadvantage of this standard approach is that the performance of the detection depends highly on the correctness of the segmentation. Especially moving objects in front of stationary ones – e.g. the bicycle in front of the parking vehicles shown in Figure 1 – are often merged and therefore not detected. This causes dangerous misinterpretations and requires more powerful solutions.

Argyros et al. describe a method to detect moving objects using stereo vision in [3]. Comparing the normal flow of the right camera image with the normal flow between the left and the right images of the stereo cameras they detect image regions with independent object motion as inconsistencies in the flow data. Heinrich [4] proposes a similar approach defining the so called flow-depth constraint. He compares the measured optical flow with the expectation stemming from the known ego-motion and the 3D stereo information. Independently moving objects do not fulfil the constraint and can easily be detected. However, this approach turns out to be very sensitive to small errors in the ego-motion estimation, since only two consecutive frames are considered. In addition, both approaches lack a precise measurement of the detected movements.

For a direct measurement of the objects movement Waxman and Duncan analyse the relation between the optical flow fields of each camera and defined the so called relative flow in [5]. Using this information the relative longitudinal velocity between the observer and the object is directly determined.

Direct optical flow analysis provides fast detection results, but is limited with respect to robustness and accuracy due to the immanent measurement noise. To get more reliable results, an integration of the observations over time is neccessary. The Kalman filter solves this in an elegant manner. Each measurement is used to improve the current estimate of the systems state [6]. In [7] Dang et al. combine stereo and motion information obtained for an object in a single Kalman filter and estimate the object's position and movement. This method expects a precise object segmentation.

The core algorithm of the system presented in this contribution follows the principle of fusing optical flow and stereo information given in [12]. The basic idea is to track points with depth known from stereo vision over two and more consecutive frames and to fuse the spatial and temporal information using Kalman filters. The result is an improved accuracy of the 3D-position and an estimation of the 3D-motion of the considered point at the same time. Since we get a rich 6D-state vector for each point we refer to this method as 6D-Vision. Taking into account the motion information, the above mentioned segmentation problem can be solved much more easily and robustly. In addition, using the 3D-motion information a prediction of the objects movement is possible. This allows a driver assistance system to warn and react to potential collisions in time.

The fusion implies the knowledge of the ego-motion. In our system we compute it from image points found to be stationary using a new Kalman filter based approach. This allows a fast calculation using all information already aquired by the system including inertial sensor data.

In our real-time application we track about 2000 image points. So far, the best results are obtained using a version of the well-known Kanade-Lucas-Tomasi (KLT) tracker [8] that was optimized with respect to speed. The depth estimation is based on a hierarchical correlation based scheme [9]. However, any comparable optical flow estimation and any other stereo algorithm can be used.

The paper is organized as follows: Section II gives the system overview followed by a description of the Kalman filter model used for the fusion of optical flow and stereo information in Section III. To improve the estimation results an image based ego-motion compensation is introduced and described in Section IV.

II. System Overview

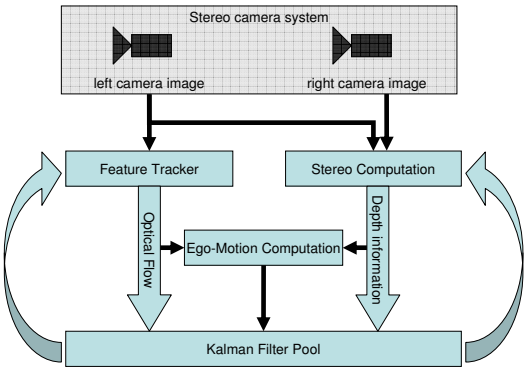

Fig. 2. The 6D-Vision system.

The block diagram in Figure 2 shows the main components of the proposed system. Each cycle a new stereo image pair is obtained and the left image is first analysed by the tracking component. It identifies small distinctive image regions called features and tracks them over time. In the current application we use a version of the Kanade-Lucas-Tomasi tracker [8] which provides sub-pixel accuracy and tracks the features robustly for a long sequence of images. It was optimized with respect to speed allowing the complete system to analyse up to 2000 features in real-time (cycletime of $40 - 80$ms).

The displacement of the same feature in the left and the right image is called disparity. It is determined for all features in the stereo module. Here a hierarchical correlation based algorithm is used [9]. After this step the current 3D-position of each analysed feature is known.

In combination with the set of 3D-positions of the last frame these features are used to compute the observers ego-motion. One option to accomplish this is to match the clouds of static world points using optimal rotation and translation estimation [10]. The movement needed to match these point clouds corresponds to the observed camera motion. Instead of using an image-based ego-motion calculation it can be reconstructed using inertial sensor data only. However, todays cars have only sensors for the speed and the yaw rate. Other rotational components such as pitch or roll are not measured and thus not compensated for, which results in a less accurate estimation of the 3D motion. Our novel approach presented in this paper utilizes both image features and inertial sensors combined in a Kalman filter. This results in an extremely fast computation of the ego-motion.

The measurements of the tracking and the stereo module together with the calculated ego-motion are given to the Kalman filter system. For each feature one Kalman filter estimates the 6D state vector consisting of the 3D-position and the 3D-motion vector. A detailed description of the underlying models is given in [12] and is recapitulated shortly in the following section. In addition, the covariance matrix for each state vector is available representing the uncertainty of the estimation. This information is important for further processing steps to build up probabilistic models of the perceived world.

For the next image pair analysis, the already aquired 6D information is used to predict the image position of the features in the tracker. This yields to a better tracking performance with respect to speed and robustness. In addition, the predicted depth information is used to improve the stereo calculation.

Features get lost over time as they move out of the image or get covered by other image portions. To replace these, the feature detector searches each image for regions that are good to track. In our case a gradient based tracker is used and therefore the eigenvalues of the gradient matrix are evaluated according to [11]. As we want to concentrate mainly on moving objects and determine their motion quickly and accurately it is preferable to have as much information as possible about these objects. Therefore the feature detector increases the density of features in image areas known to have object motion.

III. Fusion of Optical Flow and Stereo

In the following we use a right handed coordinate system with the origin on the road. The lateral x-axis points to the left, the height axis y points upwards and the z-axis represents the distance of a point straight ahead. This coordinate system is fixed to the car, so that all estimated positions are given in the coordinate system of the moving observer. The camera is at $(x, y, z)^T = (0, height, 0)^T$ looking in the positive z-direction.

A. System Model

Let $\vec{p}_k = (X, Y, Z)^T$ be the 3D position of an observed world point and $\vec{v}_k = (\dot{X}, \dot{Y}, \dot{Z})^T$ its associated velocity vector at the time step k. Assuming a constant motion during the time interval Δt the 3D position at the time step $k+1$ is given by

$$\vec{p}_{k+1} = R\vec{p}_k + \vec{T} + \Delta t R\vec{v}_k \quad (1)$$

Here the rotation matrix R and the translation vector $\vec{T}$ give the motion of the scene, that is the inverse camera motion. The camera motion components are either measured using the inertial sensor data or computed by the image-based ego-motion module.

The new velocity vector of the observed point is described by

$$\vec{v}_{k+1} = R\vec{v}_k \quad (2)$$

Combining the location $\vec{p}_k$ and the velocity $\vec{v}_k$ in the 6D state vector $\vec{x}_k = (X, Y, Z, \dot{X}, \dot{Y}, \dot{Z})^T$ the time discrete linear system model is given by

$$\vec{x}_k = A_k\vec{x}_{k-1} + B_k + \vec{\omega} \quad (3)$$

with the state transition matrix

$$A_k = \begin{bmatrix} R_k & \Delta t R_k \\ 0 & R_k \end{bmatrix} \quad (4)$$

and the control matrix

$$B_k = \begin{bmatrix} \vec{T}_k \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (5)$$

The noise term $\vec{\omega}$ is assumed to be Gaussian white noise with covariance matrix Q.

B. Measurement Model

The measurement consists of two pieces of information: the image coordinates u and v of a tracked feature and the disparity d delivered by stereo vision working on rectified images. Assuming a pin-hole camera the non-linear measurement equation for a point given in the camera coordinate system is

$$\vec{z} = \begin{bmatrix} u \\ v \\ d \end{bmatrix} = \frac{1}{Z}\begin{bmatrix} Xf_u \\ Yf_v \\ bf_u \end{bmatrix} + \vec{\nu} \quad (6)$$

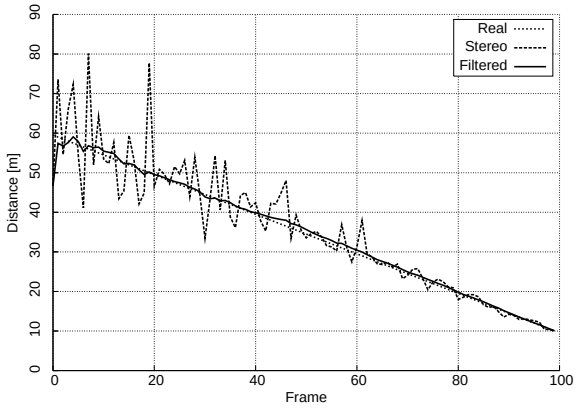

Fig. 3. Estimation results of the presented Kalman filter. The considered world point is at the initial position $(10.0\,\mathrm{m}, 1.0\,\mathrm{m}, 60.0\,\mathrm{m})^T$. The observer moves at a constant speed of $v_z = 10\,\frac{\mathrm{m}}{\mathrm{s}}$ in positive z-direction (20 fps).

with the focal lengths f_u and f_v in pixel and the baseline b of the stereo camera system. The noise term $\vec{\nu}$ is assumed to be Gaussian white noise with covariance matrix R.

To improve the Kalman filters rate of convergence a multi filter system is used. It consists of multiple differently initialized and parameterized Kalman filters running in parallel. By analysing the innovation of each filter the best matching estimation is chosen. A detailed description of this approach is given in [12].

C. Simulation Results

The benefit of filtering the three-dimensional measurement is illustrated by Figure 3. It shows the estimated relative distance of a simulated static world point measured from an observer moving at a speed of $10\,\frac{\mathrm{m}}{\mathrm{s}}$. The initial position of the point is $(10.0\,\mathrm{m}, 1.0\,\mathrm{m}, 60.0\,\mathrm{m})^T$. White gaussian noise was added to the image position and the disparity with a variance of $1.0\,\mathrm{px}^2$. The dashed curve shows the unfiltered 3D position calculation which suffers from the additive noise. The continuous curve represents the excellent result of the filter.

D. Real world results

First we concentrate on the crossing situation already shown in Figure 1. The result of the velocity estimation is given in Figure 4. The cyclist drives in front of parked vehicles while the observer moves towards him at a nearly constant speed of $4\frac{\mathrm{m}}{\mathrm{s}}$. The arrows show the predicted position of the corresponding world point in 0.5s projected into the image. The colors encode the estimated lateral speed; the warmer the colour the higher the velocity. In order to prove the results, the right image in Figure 4 shows the same situation 0.5s later. As can be seen, the prediction shown in the left image was very accurate.

Figure 5 shows the estimation results for a typical oncoming traffic situation in which the observer moves at a constant speed of $14\frac{\mathrm{m}}{\mathrm{s}}$. Here the color encodes the absolute velocity of the tracked points. The prediction matches the real position shown in the right image.

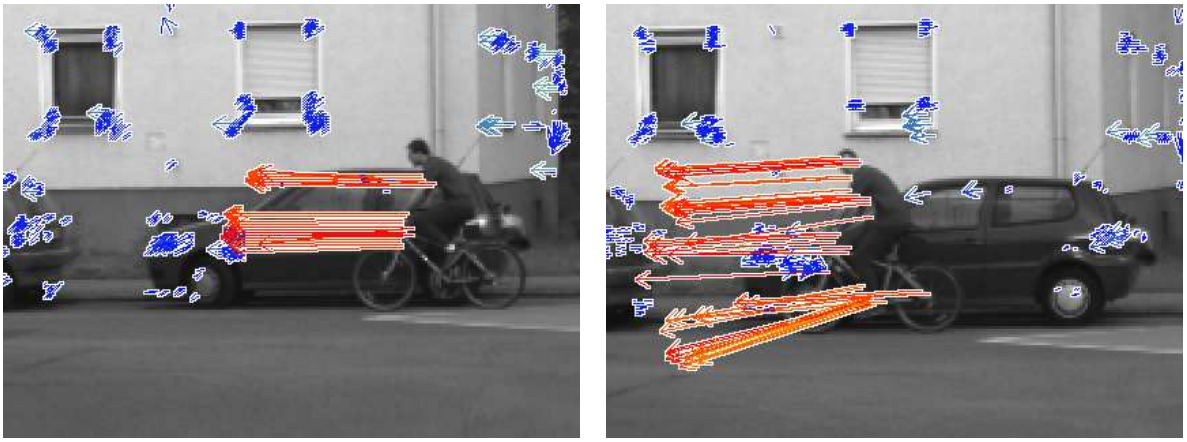

Fig. 4. Velocity estimation results for cyclist moving in front of parking cars. The arrows show the predicted position of the world point in 0.5s. The right image was taken 0.5s later allowing a comparison of the estimation from the left image. Blue encodes stationary points.

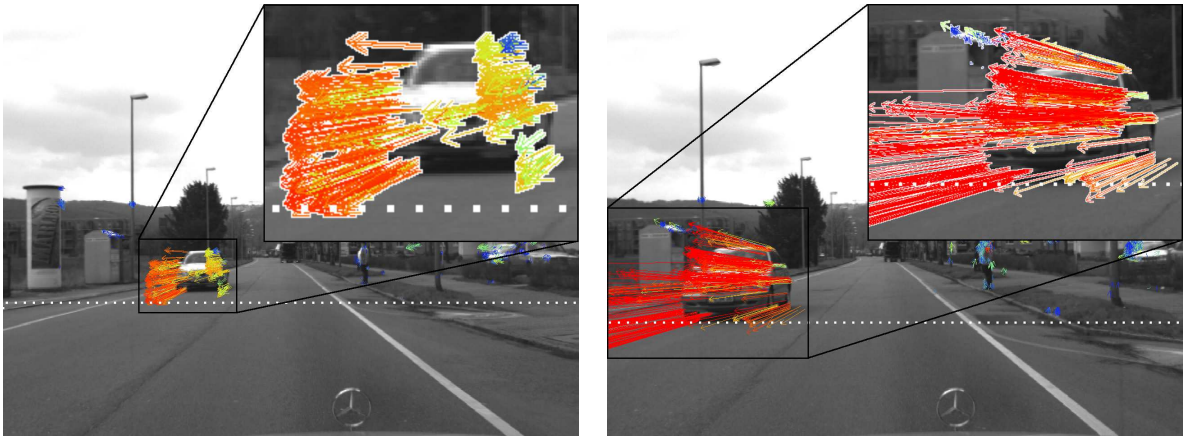

Fig. 5. Velocity estimation results for an oncoming car. The observer moves at a constant speed of $14\frac{\mathrm{m}}{\mathrm{s}}$.

IV. Kalman filter based Ego-motion calculation

In order to obtain the best results in the given fusion process, the observer's ego-motion must be known. The inertial sensors installed in today's cars measure the current speed and the yaw rate. However, this information is not sufficient for a full description of the observer's motion as it lacks important components such as the pitch and the roll rate. Not compensating these influences results in a wrong 3D motion estimation.

This is illustrated in Figure 6. Here the ego-motion was computed using inertial sensors only. As the car undergoes a heavy pitch movement, the world seems to move downwards.

Using the estimated 6D information static world points are easily identified. Assuming they remain static the observers ego-motion is determined by comparing the predicted world position with the measured one. We use a Kalman filter to accumulate all these measurements and estimate a state vector containing all ego-motion parameters. In addition, the inertial sensor data is integrated as an additional source of information.

This calculation of the ego-motion fits well into the proposed system as it uses the already aquired information including the knowledge of the 6D filtering.

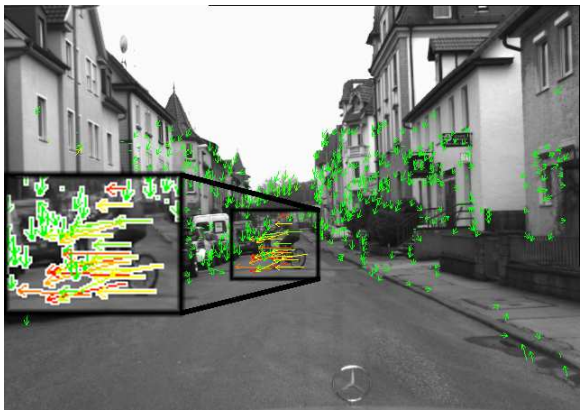

Fig. 6. Velocity estimation results without image based ego-motion compensation.

A. System model

We use a bicycle model for the car and assume that the pitch and the roll angle change only slowly. The rotational parameters of the ego-motion are the pitch angle α, the

yaw angle ψ and the roll angle γ. Their derivative with respect to time are the pitch rate $\dot{\alpha}$, the yaw rate $\dot{\psi}$ and the roll rate $\dot{\gamma}$. The translational parameters are the components of the observers velocity vector $\vec{v} = [v_x\ v_y\ v_z]^\top$. Together with the acceleration a in z-direction and a scale factor β, neccessary to compensate systematic errors of the internal speed sensor, the state vector of the system is $\vec{x} = \left[\dot{\alpha}\ \dot{\psi}\ \dot{\gamma}\ v_x\ v_y\ v_z\ a\ \alpha\ \gamma\ \beta\right]^\top$.

To transform the state vector $\vec{x}_{k-1}$ of the previous time step into the current one, the discrete system model is given by

$$\vec{x}_k = A_k \vec{x}_{k-1} + \vec{\omega} \tag{7}$$

with the state transition matrix defined as

$$A_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \Delta t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \Delta t & 0 & 0 & 0 & 0 & 0 & 0 & \tilde{g} & 0 & 0 \\ 0 & 0 & \Delta t & 0 & 0 & 0 & 0 & 0 & \tilde{h} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

and the noise term $\vec{\omega}$ which is assumed to be Gaussian white noise with covariance matrix Q.

The terms $\tilde{g}$ and $\tilde{h}$ are

$$\tilde{g} = 1 - g\Delta t, \qquad 0 < g < 1 \tag{9}$$

$$\tilde{h} = 1 - h\Delta t, \qquad 0 < h < 1 \tag{10}$$

where g and h are damping factors to reduce effects of divergence of the integrated pitch and roll angle.

B. Measurement model

The measurement vector contains the components of the measured optical flow Δu and Δv, and the change of disparity from the stereo module Δd of each point found to be static. In addition, the translational velocity components s_x, s_y and s_z calculated using the inertial sensor data speed s_{sensor} and yaw rate $\dot{\psi}_{\text{sensor}}$ are used. Therefore, the stacked measurement vector - using only one image point for simplicity - is $\vec{z} = [\Delta u\ \Delta v\ \Delta d\ s_x\ s_y\ s_z]^\top$.

The measurement model is given by

$$\vec{z} = \begin{bmatrix} \Delta u \\ \Delta v \\ \Delta d \\ s_x \\ s_y \\ s_z \end{bmatrix} = \begin{bmatrix} u_k - u_{k-1} \\ v_k - v_{k-1} \\ d_k - d_{k-1} \\ \left(c_0 - c_1 s_z^2\right) \dot{\psi}_{\text{sensor}} \\ s_{\text{sensor}} \sin\theta \\ s_{\text{sensor}} \cos\theta \end{bmatrix} \tag{11}$$

where c_0 and c_1 are constants describing the influence of the side slip angle and θ is the yaw installation angle.

In addition, the equation

$$0 = v_z - \beta s_z \tag{12}$$

Fig. 7. Velocity estimation results with image based ego-motion compensation.

describes the relation of the measured velocity and the estimated velocity using the scale factor β.

Static features used for the ego-motion computation are identified by their associated 6D vector. The estimated motion of these features has to be small and a subset is selected evenly distributed in the image. Features with low estimation covariance are preferred. The algorithm runs in less than one millisecond per frame.

C. Real World Result

The benefit of the presented image-based ego-motion compensation is demonstrated in Figure 7. Comparing to Figure 6, which showed the same scene using only inertial sensors for the ego-motion calculation, the world seems to be more stable. In fact, all previously moving image points remain static. In addition, there are more vectors on the cyclist. To suppress measurement outliers a standard 3σ-test is performed in the Kalman filter. As the pitch movement in Figure 7 is not estimated at all, the features are misinterpreted as outliers and are therefore rejected.

Looking at the same situation a few frames earlier, we see in the left image of Figure 8 the cyclist appearing behind a wall. The rear wheel of the bicycle is covered by a solid fence. At this time the cyclist is at a distance of about 32 m and covers a visible region in the image of about 30×40 pixels. Nevertheless, first reliable estimation results are available. Looking at the right image, a birdview display on the same scene is given. Here only motion vectors whose associated world point lies above 1 m are displayed to provide a better view of the situation. It can be seen, that this rich information helps detecting the moving cyclist in a following detection step and provides a first prediction of its movement at the same time.

V. Summary

The proposed fusion of stereo and optical flow simultaneously improves the depth accuracy and allows estimating position and motion of each considered point. Segmentation based on this 6D-information is much more reliable and a fast recognition of moving objects becomes possible.

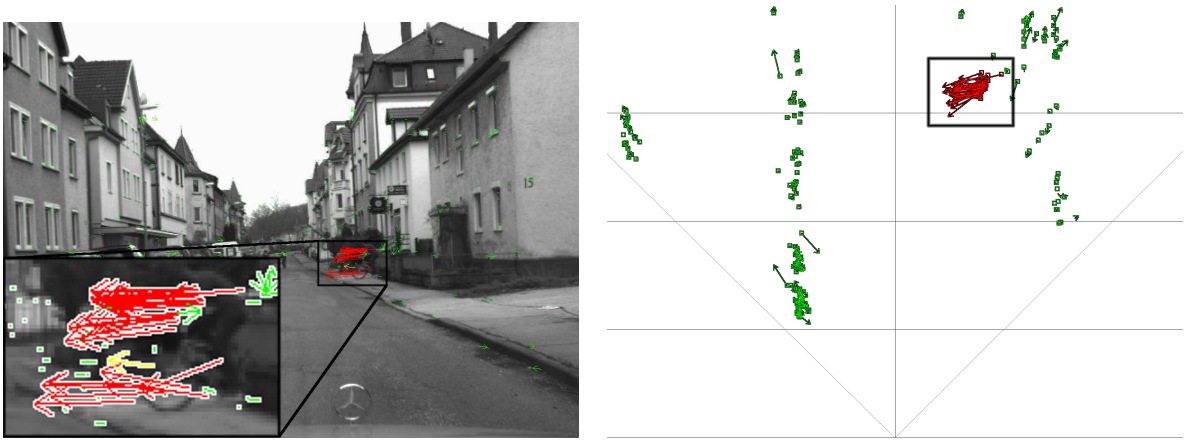

Fig. 8. Estimation results for partially occluded cyclist (rear wheel occluded). On the left image the velocity estimation is shown. On the right image the same scene is displayed in a birdview perspective with the camera center at the middle of the bottom line. The total width is 20 m and the displayed depth is 40 m.

In particular, objects with known direction and speed can directly be detected on the image level without further non-linear processing or classification steps that fail if objects occur, which were not provided in the training process.

Since the fusion is based on Kalman filters, the information contained in a number of frames is integrated. This leads to much more robust estimations than differential approaches like pure evaluation of the optical flow. The multi-filter approach adopted from our depth-from-motion work [13] speeds up the rate of convergence of the estimation, which is important for fast reactions. For example, practical tests confirm that a crossing cyclist at an intersection is detected within 4-5 frames. In addition, the novel image based ego-motion compensation improves the quality of the estimation significantly.

Even partially occluded objects are detected fast and reliably as seen in Figure 8. This demonstrates the power of the presented system not only to detect but also to measure and predict the objects movement. At the moment, no other sensor is able to provide these results at such an early stage.

The described system is implemented in our demonstrator vehicle (UTA, Mercedes Benz S-Class vehicle) on a 3.2 GHz Pentium 4. The cycle time of the complete system for analysing about 2000 image points is $40 - 80$ms. That includes the image aquisition as well as the visualisation.

The next step focuses on the segmentation of the available 6D information in order to provide reliable object hypotheses. The detected objects will be analysed subsequently with respect to their risk of collision. In a first implementation, we adopted a method used for segmenting stereo data. All points with a high risk of collision are put into a birdview map and objects are identified using a connected component analysis. In a test scenario, we used this information successfully to perform a fully autonomous emergency braking in our demonstrator vehicle. However, as this method uses only a subset of the available information, we are investigating more powerful segmentation algorithms.

REFERENCES

[1] G. V. Statistisches Bundesamt, *Fachserie 8, Reihe 7, Verkehrsunfaelle - Dez. 2005*, S. B. Wiesbaden, Ed., 2006.

[2] M. C. Martin and H. Moravec, "Robot evidence grids," Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-96-06, March 1996.

[3] A. A. Argyros, M. I. Lourakis, P. E. Trahanias, and S. C. Orphanoudakis, "Qualitative detection of 3d motion discontinuities," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'96)*, vol. 3, November 1996, pp. 1630–1637.

[4] S. Heinrich, "Real time fusion of motion and stereo using flow/depth constraint for fast obstacle detection," in *Proceedings of the 24th DAGM Symposium*, ser. Lecture Notes in Computer Science, vol. 2449, Zurich, Switzerland, September 2002, pp. 75–82.

[5] A. M. Waxman and J. H. Duncan, "Binocular image flows: steps toward stereo-motion fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 715–729, 1986.

[6] G. Welch and G. Bishop, "An introduction to the kalman filter," University of North Carolina at Chapel Hill, Department of Computer Science, Tech. Rep. TR 95-041, 1995.

[7] T. Dang, C. Hoffmann, and C. Stiller, "Fusing optical flow and stereo disparity for object tracking," in *Proceedings of the IEEE V. International Conference on Intelligent Transportation Systems*, 2002, pp. 112–117.

[8] C. Tomasi and T. Kanade, "Detection and tracking of point features," School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-CS-91-132, April 1991.

[9] U. Franke and A. Joos, "Real-time stereo vision for urban traffic scene understanding," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, October 2000, pp. 273–278.

[10] H. Badino, U. Franke, C. Rabe, and S. Gehrig, "Stereo vision based detection of moving objects under strong camera motion," in *Proceedings of the First International Conference on Computer Vision Theory and Applications*, vol. 2, February 2006, pp. 253–260.

[11] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'94)*, Seattle, June 1994.

[12] U. Franke, C. Rabe, H. Badino, and S. Gehrig, "6d-vision: Fusion of stereo and motion for robust environment perception," in *Proceedings of the 27th DAGM Symposium*, 2005, pp. 216–223.

[13] U. Franke and C. Rabe, "Kalman filter based depth from motion with fast convergence," in *Proceedings of the 2005 IEEE Intelligent Vehicles Symposium*, 2005, pp. 181–186.