Monocular Video serves RADAR-based Emergency Braking

Andreas Wedel and Uwe Franke
DaimlerChrysler AG
70546 Stuttgart
Email: {andreas.wedel, uwe.franke}@daimlerchrysler.com

Abstract— **Much work was carried out recently for emergency braking based on radar signals. The key step for emergency braking is the reliable detection of obstacles. Moving objects are verified as such by tracking the radar signal. However, discarding so-called phantom objects remains a challenge for stationary objects. This leads to the question of sensor fusion for more reliable verification of obstacles. In this paper we propose a novel method using a monocular camera, such as the night view camera in the Mercedes S class.**

Our two goals in this paper are the verification of obstacles and the detection of obstacle boundaries. This allows to analyze the situation for carrying out emergency braking. The verification of obstacles is done by analyzing the scaling of obstacles as they get closer to the camera. The perspective image motion of the ground plane serves as a counter hypothesis to detect phantom objects. Obstacle boundaries are found by graph cut segmentation on these two motion fields.

I. Introduction

Time has come that collision avoidance and emergency braking are leading research and technology areas in driving assistance. Obstacle detection is a prerequisite to warn the driver or actively control a vehicle in hazardous situations such as depicted in Figure 1. Intelligent cars today are equipped with a radar system for adaptive cruise control keeping the distance to the moving vehicle ahead. While such systems work well in practice, the risk of false alarms for emergency braking at stationary obstacles in the driving path cannot be denied. The cause of such false alarms can be manifold and includes parked cars on the hard shoulder, railway roads and low bridges just to name a few. Typically heuristics have to be used or hazardous situations need to be checked by other sensors than radar. The more sophisticated way is to use a second, preferable independent sensor to decrease the uncertainty of the obstacle hypothesis and thus to enable autonomous emergency braking.

In this paper we investigate the use of a computer vision for serving radar emergency braking. Computer vision is an uprising field in driver assistance and an increasing number of upper class vehicles are equipped with a camera system. The new Lexus uses a stereo camera pair to directly verify radar distances whereas the Mercedes-Benz S-class is equipped with a monocular Night Vision camera. Monocular cameras are used also for lane detection in a variety of other vehicles. Therefore, in this paper we address the challenge of combining radar distance measures with a monocular camera.

Fig. 1. **Stationary obstacle** in 12 m distance. The image was taken at ego-vehicle speed of 33 km/h therefore remaining 1,3 seconds until collision. The obstacle boundaries and size are successfully determined by monocular camera without imposing shape knowledge in twilight conditions.

II. Obstacle Detection in Monocular Vision

The advantage of a camera in comparison to radar is the high spatial resolution of the camera chip. However, distance measurements are not directly possible using a monocular camera. Therefore more advanced techniques have to be used. As motivated by Bertozzi [1] and Sun [10], obstacle detection in monocular vision can be split into methods employing a-priori knowledge and others based on relative image motion. Typically, a-priori information consists of the appearance of observed objects and has to be learned from many examples. In controlled and isolated scenarios with limited obstacle classes such methods work well. Things get more complicated when we consider a more realistic scenario with arbitrary obstacles, as inevitable for emergency braking. A model free approach demands the use of relative image motion and multiple images, also known as structure from motion. From a mathematical view, this can be understood as integration of information over time. The results are expected to depend on time steps (time between successive frames) and the used algorithm accuracy. In this paper the focus lies on driver assistance systems that can be achieved with current computer hardware. Therefore the computational complexity is of special consideration, too.

A basic realtime structure from motion approach is Kalman filter based depth estimation as described in [4]. See Fig. 2 for an example. Interest points are tracked and multiple Kalman filters for each interest point are run in parallel to

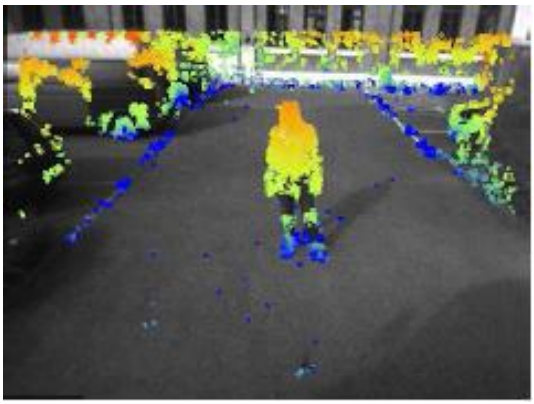

Fig. 2. **Estimation result for a person** after 10 frames based on Kalman filtered feature tracks. The color warmth (gray value brightness) denotes obstacle height. Note that due to camera setup the focus of expansion is located above the image.

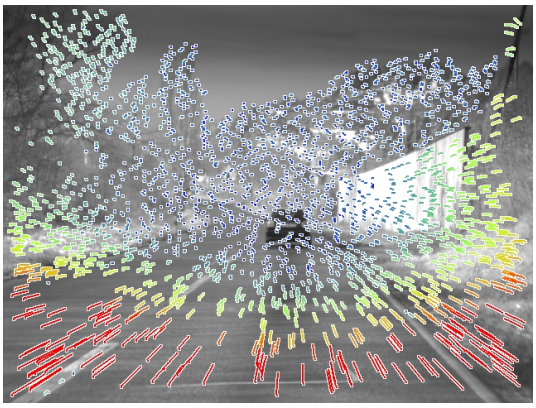

Fig. 3. **Flow Field of a stationary scene.** Note that close to the focus of expansion and for distant objects the induced flow is very small.

estimate depth and obstacle height. A goodness-of-fit test fuses the state of the different filters in an optimum manner. The test not only leads to distance estimation but also allows to distinguish between static and moving obstacles. However, the algorithm requires feature points to be extracted and tracked throughout the video sequence. Realtime optical flow algorithms such as [3], [7], [9] and on these algorithms based obstacle detection such as the Kalman based feature tracker [4] integrate displacement between successive image pairs. This leads to the problem of drifts. Especially close to the focus of expansion (compare with Fig. 3) the inaccuracy of the tracker compared to the flow vector length is high and therefore measurements are error-prone. In the application of obstacle detection for emergency braking this is the very image region we are interested in, thus leading to the situation that other algorithms, suited to the special situation, have to be used. Our recent investigation shows how scale change [12] and dense transformation classification [11] can be used to verify obstacle hypotheses. By the introduction of scale and thus possible tracking of regions throughout multiple images of a video sequence drift problems are elininated. As a result, distance measurements near the focus of expansion become possible. Mathematically, the integration of information is replaced by a single step. In addition, pixelwise segmentation allows to accurately estimate obstacle boundaries and lateral position.

III. Algorithm Outline

Assume we are given a distance hypothesis for a stationary obstacle from radar. To start with a simple example, consider the two trucks in Fig. 4 as a radar hypothesis. Our goal is to verify or discard this hypothesis by means of computer vision. The feature we are going to use is image scale. As the vehicle approaches the obstacles, the image of the obstacles taken by a forward-looking camera will grow in size. This principle is well known for humans as it is simply based on the perspective transformation of our eye or - in computer vision context - of the camera lens. The principle of distance estimation by relative scale in camera sequences is well illustrated by the theorem on intersecting lines (compare Figure 6). The quantity that relates scale to distance is the covered distance of the ego-vehicle. If the car travels half the distance to any obstacle, the size of the imaged obstacle will double. On the other hand, if the scale and traveled distance are known, obstacle distance can be computed. Surely we don't want to wait for the scale factor do double to estimate distances. But as scale can be efficiently computed in images, small scale changes already allow for distance estimates. In this paper we measure scale changes by automatic tracking of template regions.

A problem arises if the scale of such a template region does not originate from obstacles and therefore leads to false results. Truly, such problem can arise if no obstacle is contained in the image and for instance figures painted on the street are tracked. The small patch in the middle of Fig. 5 is such an outlier and also gets larger while approaching. This leads to the question of how one can distinguish such template regions on the street from others on obstacles. Both, a distant obstacle and the street, are a plane (planar surface) in first approximation. Under perspective transformations planes undergo homographic transformations (eight degrees of freedom). Homographies contain the normal of mapped planes and therefore the homography of the street is different to that of obstacles. Because the visible surface of an obstacle is approximately parallel to the camera plane, its transformation can be modeled by translation and scale (similarity transformation). The key to distingush template regions on the street from others on obstacles lies in checking if the transformation is described by a similarity transformation or by the homography generated by the street. In Sect. IV and Sect. V we describe the mathematics needed to calculate distances and track template patches. Section VI describes the hypothesis verification algorithm in detail.

While verification of radar hypotheses decreases the uncertainty of measurements to enable emergency braking it makes sense to detect obstacle boundaries, too. We can identify boundaries by checking for the transformation of image regions but we encounter a problem in that regions may

Fig. 4. **Example for a radar hypothesis for a stationary obstacle**.

Fig. 5. **Distance Verification by scale checking**. The Distance from scale for the highlighted regions corresponds to the given obstacle distance. Note the outlier in the middle of the image on the ground where figures on the street show approximately the same scale as the obstacles in the given distance.

not necessarily fully lie on obstacles or the street no matter how small we make the region. This conceptually difficulty is overcome by pixel-wise checking for the better-suited transformation. Checking each pixel for itself would lead to cluttered results, demanding to take into consideration the neighborhood as well. By transforming the decision problem in an energy minimization problem, a global optimal solution for binary segmentation of the image into obstacle and non-obstacle regions is found by graph cut algorithms [2]. Results are shown in Section VII.

IV. Distance from Scale near Focus of Expansion

To verify obstacle hypotheses from radar with computer vision, first of all understanding relative image motion is necessary. We'll therefore briefly review distance estimation for obstacles from monocular video. The formulas have been introduced by Longuet-Higgins in [6] and have been the basis for many monocular vision algorithms. We show that the relative motion of obstacles is well described by image scale.

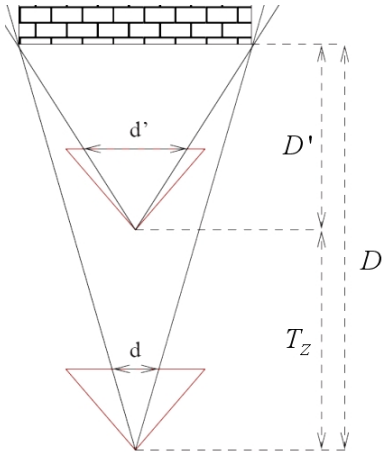

Fig. 6. **The underlying principle for depth from scale** is illustrated by the theorem on intersection: $D = T_Z \frac{d'}{d}$.

Details on accuracy and implementation can be found in [12].

We assume a static world (for moving obstacles distance cannot directly be estimated by relative motion). A world point $(X, Y, Z)^\top$ in camera coordinates is projected into the image point $\vec{x}$ via a perspective transformation with camera focal length f,

$$\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} X \\ Y \end{pmatrix}. \tag{1}$$

The video sequence is represented as 2-D gray value fields $I_t(x, y)$ at time t and image points $\vec{x} = (x, y)^\top$. As the camera translates by $(T_X, T_Y, T_Z)^\top$ in camera coordinates from frame I_t to frame I_{t+1}, the world point at time $t + 1$ will be projected to

$$\begin{aligned} \vec{x}_{t+1} &= \frac{f}{Z + T_Z} \begin{pmatrix} X + T_X \\ Y + T_Y \end{pmatrix} \\ &= \underbrace{\frac{Z}{Z + T_Z}}_{s} \underbrace{\frac{f}{Z} \begin{pmatrix} X \\ Y \end{pmatrix}}_{\vec{x}} + \frac{f}{Z + T_Z} \begin{pmatrix} T_X \\ T_Y \end{pmatrix}. \end{aligned} \tag{2}$$

Thus, the distance Z of the point can be deduced from the scaling s of $\vec{x}$ with respect to the focus of expansion (the additive component in 2). The principle is illustrated in Figure 6. With the additional assumption that all image points on an obstacle have the same depth (also known as *weak perspective assumption*), scale can be calculated by tracking a number of points in an image region over multiple frames. This allows for very accurate distance estimates at interactive frame rates within 50m as the examples in Fig. 7 verify.

V. Transformation of Planar Ground

While the transformation of an obstacle for standard camera installation can be approximated by the weak perspective model, the transformation of the planar ground is more complicated. With the assumption of planarity however, the transformation of the ground under perspective transformation is a homography. A first approximation of the homography

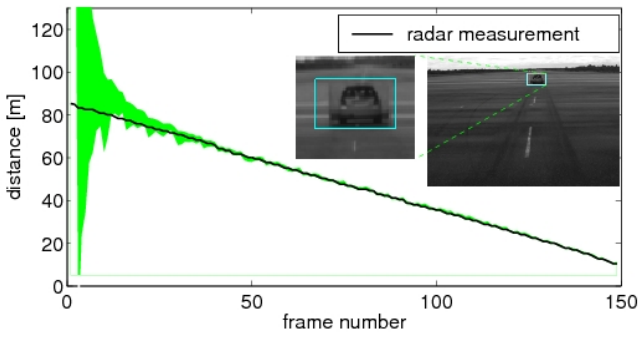

Fig. 7. The plot shows **accurate distance estimation by image scale** in comparison to radar measurements. The estimated obstacle distance plus it's standard deviation is represented by the green area. Note that the obstacle is near the focus of expansion. Nevertheless, distance estimation by image scale proves to give accurate results provided sufficient vehicle translation.

can be derived from the camera installation and camera motion. However, this approximation has to be refined. This is related to computing ego-motion [5]. Based on brightness constancy, one can apply an incremental warping technique as originally proposed for translational motion in [7]. We will briefly describe the warp update for homographic motion fields in this section. Note that the computation of scale and translation is a special case of the general homography keeping all other parameters fixed.

A point $\vec{x}$ in the current frame corresponds to the point

$$\vec{x'} = H(\vec{h}, \vec{x}) = \begin{pmatrix} \frac{h_{1,1}\cdot x + h_{1,2}\cdot y + h_{1,3}}{h_{3,1}\cdot x + h_{3,2}\cdot y + 1} \\ \\ \frac{h_{2,1}\cdot x + h_{2,2}\cdot y + h_{2,3}}{h_{3,1}\cdot x + h_{3,2}\cdot y + 1} \end{pmatrix}$$

in the previous frame, where $\vec{h} \in \mathbb{R}^8$ are the eight homographic motion parameters. Applying this homography leads to a warped image I^w_{t-1} such that the planar region coincides with the region at time t. Given an estimate $\vec{h}^0$ for h, Taylor-expansion leads to an estimate of the warped frame for the parameters $\vec{h}^0 + \Delta\vec{h}$:

$$\begin{aligned} I^w_{t-1}(\vec{h}^0 + \Delta\vec{h}, \vec{x}) \;\approx\; & I_{t-1}(H(\vec{h}^0, \vec{x})) \\ & + \nabla I_{t-1}(H(\vec{h}^0, \vec{x}))\, \frac{dH(\cdot,\cdot)}{d\vec{h}}\bigg|_{\vec{x}, \vec{h}=h_0} \Delta\vec{h} \end{aligned}$$

Minimizing the sum of squared differences for the region R

$$\sum_{\vec{x}\in R} \left(I_t(\vec{x}) - I^w_{t-1}\left(\vec{h}^0 + \Delta\vec{h}, \vec{x}\right)\right)^2 ,$$

can be done by setting the derivative w.r.t. $\Delta\vec{h}$ to zero and one can solve for the update (with simplified notation):

$$\begin{aligned} \Delta\vec{h} = & \sum_{\vec{x}\in R} \left(\frac{dH}{d\vec{h}}(\vec{x})^\top \nabla I_{t-1}(\vec{x})^\top \nabla I_{t-1}(\vec{x}) \frac{dH}{d\vec{h}}(\vec{x}) \right)^{-1} \\ & \cdot \sum_{\vec{x}\in R} \left(I_t(\vec{x} - I^w_{t-1}(\vec{h}^0, \vec{x})\right) \nabla I_{t-1}(\vec{x}) \frac{dH}{d\vec{h}}(\vec{x}) \end{aligned}$$

Such warping scheme is also known as the Gauss-Newton method and allows the estimation of homographies without knowledge of point-correspondences. From this homography estimate the road normal and camera translation up to scale can be deducted [8]. Therefore, ramps in front of cars are detected and results can be cross-checked with vehicle ego-motion.

VI. Verification of RADAR Measurements

Given an obstacle from radar by distance measurements, different verification techniques are possible using the outlined vision algorithms in this paper. To compare distance by image scale with the hypothesis distance given by radar is straightforward. In a second step, the transformation of the obstacle hypothesis region is checked against the transformation of the planar ground to get a more reliable statement. Determination of obstacle boundaries will be discussed in the next section.

A. Distance Verification

Given a distance to an obstacle, the associated image region for the obstacle can be computed with known camera parameters. However, due to poor spatial resolution of commonly used radar sensors and only approximately determinable camera parameters, a somewhat bigger region of the image has to be investigated. Since blowing up the region implies that the obstacle covers only a part of the interest region, it is subdivided into several subregions, each of preset size depending on obstacle distance. Fig. 5 shows the verification result for the example from the introduction (Figure 4). These regions are randomly inserted, hence overlapping is not prohibited. The regions are tracked independently (by warping onto the reference template as described in Sect. V) throughout the sequence to extract respective scale and translation. If any obstacle (of significant size) is present in the hypothesized distance, one would expect

- the scale of at least one image region to be related to obstacle distance and own velocity,
- no lateral translation of the obstacle and hence no unfeasible translation of the image region,
- the obstacle distance to decrease continuously, imposing continuous increase of the scale.

For each of the equally distributed image regions, outliers (those not fulfilling above rules) are rejected and replaced by a new randomly inserted region. Inliers, fulfilling the three rules, are counted and the proportion of outliers to inliers as well as the number of inliers is used to confirm the radar hypotheses. Obstacles of vanishing size (e.g. a corner reflector) are filtered out avoiding emergency brake execution in such cases. Structures on the road, e.g. railroads, however are still inliers in the sense of fulfilling the above rules. The remaining part of this section describes an algorithm to reject such hypotheses by transformation comparison.

B. Transformation Comparison

While obstacles in a given distance show distinct scale in image space, structures on roads are transformed by a homography. However, structures on roads can be seen as low

Fig. 8. **Distance Verification and obstacle boundary detection** in monocular video. The *black line* corresponds to the distance given on the left. The *white line* shows the probability difference between obstacle (above the dashed line) and non-obstacle (below the dashed line). The images show verification and lateral boundary detection of obstacles.

obstacles and therefore cause the scale estimation to actually fit well. However, distinguishing the ground transformation from obstacle transformation, helps finding such false positives. Even more, we will deduce a probability for obstacle and non-obstacle based on comparison of the transformation. For each single tracked region the transformation caused by scale is checked against the transformation caused by road homography. Let $I(\vec{x})$ be the current gray values and $I_O(\vec{x})$, respectively $I_G(\vec{x})$ be the warped most recent images onto the current image for the obstacle and the ground plane. The probability for obstacle p_O and ground plane p_G can be expressed by

$$p_O = \exp\left(-SSD\left(I(\vec{x}) - I_O(\vec{x})\right)\right) \tag{3}$$

and

$$p_G = \exp\left(-SSD\left(I(\vec{x}) - I_G(\vec{x})\right)\right) \tag{4}$$

with SSD encoding the sum of squared gray value differences. p is maximal if the intensity differences are minimal and vice versa. Therefore, hypothesis testing boils down to checking for the transformation with higher probability.

VII. Accurate Obstacle Boundaries

Some obstacles may protrude into the vehicle pathway but not hinder in such a way that emergency braking is necessary (cars parked on the road shoulder). To analyze if the vehicle can pass besides the obstacle, object boundaries need to be detected. While hypothesis verification with vision serves radar in a way that it enforces or weakens, even discards, obstacle detection, the actual soft spot of radar, high uncertainty in lateral direction, is not yet compensated. But just here lies the actual strength of computer vision, while it's weakness is the distance measurement. Hence it is most demanding to develop concepts and algorithms that enable to carry forward spatial resolution from computer vision to serve radar. This section describes two concepts, the difference being a local and a global solution to the stated task, namely providing higher spatial resolution for obstacle detection. As demonstrated in Sect. VI transformation comparison can be used to distinguish obstacles from the ground plane. If done for each pixel, this would segment the image into an obstacle and another non-obstacle region. However, a single pixel cannot be used to calculate an approximation for translation and scale and hence, more advanced algorithms have to be used. In the following we will demonstrate how image regions and pixel-wise checking can be used to detect obstacle boundaries.

Fig. 9. **Multiple regions for obstacle verification and boundary detection**.

A. Region Transformation Comparison

For each subregion within the region of interest, obstacle and non-obstacle decision becomes possible by transformation checking. If subregions are not randomly inserted but arranged in a well-structured manner, obstacle boundaries can be deduced. In order to fulfill this requirement, we take a region of interest in the image to which obstacles in given distance with 1.1m height are mapped (compare Fig 9). This region is than subdivided into several vertical slices which are tracked individually. Tracking results are than compared and updated to reduce outliers. This results in a histogram of probability differences ($p_O - p_G$) for obstacle and non-obstacle and vertical obstacle boundaries are located at the zero-crossings. Results show rough lateral boundary detection as can be seen in Figure 8. Experimental results show good results for obstacles within 50m with a camera focal length of 840 pixels. The algorithm runs at frame rates of 15Hz on a Pentium IV 3.2GHz including tracking of the regions and verification.

B. Segmentation via Graph Cut

Transformation parameters (hence scale) can be more precisely estimated if obstacle boundaries are more accurate. In order to get accurate obstacle boundaries, pixel-wise segmentation instead of local decision making by region

Fig. 10. **Graph Cut Segmentation.** The obstacles in 18 m distance are accurately separated from the background by means of segmentation carried out on relative image motion.

Fig. 11. **Segmentation of moving vehicle.** As the car turns into the field of view of the radar sensor, it is verified and segmented by computer vision. Recall that segmentation is purely based on monocular vision analyzing relative image motion. This examples demonstrates the benefit of sensor collaboration in complex situations.

comparison is needed. However, since all transformation parameters cannot be estimated by a single pixel the two steps, transformation estimation and segmentation, have to be decoupled.

- In a first step transformation parameters are estimated from the whole obstacle region respectively the background region.
- In a second step a decision is made for every pixel, which transformation suits best taking into consideration neighboring pixels as well.

The second step is executed for all pixels at the same time leading to a global optimal solution for the segmentation using the graph cut algorithm. Both steps are repeated until convergence. Initial motion models are derived from ego motion and radar measurements of the obstacle distance.

Graph cut algorithms solve energy minimization problems. We therefore have to formulate an energy term for the segmentation problem. This energy consists of a data term and a smoothness term. The data term includes the gray value difference for respective transformation. The smoothness term encodes neighborhood relations and favors cuts along edges in the image. For further detail we refer to [11].

The implemented algorithm runs at 5 frames per second on quarter VGA images allowing near real-time performance. Promising results using graph cut are given in Figure 10. While originally the method was proposed for static objects, radar measurements provide obstacle distance and relative speed, such that moving obstacles can be segmented by similar means (compare Figure 11). The segmentation of moving vehicles is the focus of ongoing research.

VIII. Conclusions

We presented concepts to compensate weaknesses of commonly used radar and monocular computer vision by their mutual strengths to serve each other. Radar measurements show precise distance estimates, however, seeing phantom objects is still problematic. We investigated fusion of common radar and monocular vision and presented a method to verify obstacle hypotheses given by distance measurements from radar using image motion alone without prior knowledge on appearance. Combining distance estimation from radar with verification and obstacle boundary detection from monocular vision shows to outline any obstacle precisely and therefore enables emergency braking and obstacle avoidance strategies. Moreover, the differentiation between obstacle and background proves strong enough for accurate obstacle segmentation to visually enhance driver assistance systems.

References

[1] M. Bertozzi, A. Broggi, M. Cellario, A. Fascioli, P. Lombardi, , and M. Porta. Artificial vision in road vehicles. In *Proceedings of the IEEE*, volume 90, pages 1258–1271, 2002.

[2] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *International Conference on Computer Vision*, volume 1, pages 105–112 vol.1, 2001.

[3] A. Bruhn, J. Weickert, and C. Schnörr. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 63(3):211–231, 2005.

[4] U. Franke and C. Rabe. Kalman filter based depth from motion with fast convergence. In *Proc. IEEE Intell. Vehicles Symp.*, pages 180–185, Las Vegas, 2005.

[5] Q. Ke and T. Kanade. Transforming camera geometry to a virtual downward-looking camera: Robust ego-motion estimation and ground-layer detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[6] H. Longuet-Higgins and K. Prazdny. The Interpretation of a Moving Retinal Image. In *Proceedings of the Royal Society of London*, volume 208 of *Series B, Biological Sciences*, pages 385–397, July 1980.

[7] B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[8] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3-D Vision*. Springer, 2004.

[9] F. Stein. Efficient computation of optical flow using the census transform. In *DAGM04*, pages 79–86, 2004.

[10] Z. Sun, G. Bebis, and R. Miller. On-road vehicle detection using optical sensors: A review. In *IEEE International Conference on Intelligent Transportation Systems*, volume 6, pages 125 – 137, 2004.

[11] A. Wedel, D. Cremers, T. Brox, and T. Schoenemann. Warpcut - fast obstacle segmentation in monocular video. *To appear*, 2007.

[12] A. Wedel, U. Franke, J. Klappstein, T. Brox, and D. Cremers. Realtime depth estimation and obstacle detection from monocular video. In *Pattern Recognition (Proc. DAGM)*, volume 4174 of *LNCS*, pages 475–484, Berlin, Germany, September 2006. Springer.