

# Object Detection and Tracking in Range Image Sequences by Separation of Image Features

Esther B. Meier, Frank Ade

*Abstract*— We propose a new approach for a driver assistance system based on range image sequences to increase the safety on motorways. This control system is capable of automatically keeping the car at an adequate distance or warning the driver when he gets too close to other cars.

The knowledge of the sensor geometry with respect to the road is used to separate image features into ground and road obstacles. We assume that the road is flat but possibly inclined. To distinguish between obstacles and road pixels we use a separating plane that is slightly steeper than the current road model. Pixels lying above this plane are assumed to belong to obstacles, and pixels below the plane are assumed to belong to the road. To update the road model, we fit a plane through all pixels marked as ground. The partition of the detected obstacles into their different objects is performed using connected component analysis. Components are then merged into objects according to proximity. For the tracking we employ Kalman filtering to provide most likely estimates of the state of each object.

We show results of some experiments with simulated traffic scenes and toy traffic scenarios. Real data will be available later in the project.

*Keywords*— range images, model-based segmentation, Kalman filtering, object tracking.

## I. INTRODUCTION

SOCIETY has to cope with ever increasing traffic densities, which lead to a growing number of potentially harmful situations. Therefore driver assistant systems and traffic surveillance systems are of increasing interest. This task is tackled by a number of research groups in university and in industry. Most approaches are based on black and white or color imagery, where obstacles are detected by their shape [1], [2], [3], by stereo [4], [5], or by motion [6]. Other types of sensors, such as radar, infrared, ultra-sound and range sensors [7], [8], [9], are also used.

For safety and surveillance applications it is important to know the geometrical description of the observed scene. Consequently, range sensors have attracted considerable attention recently. The MINORA project (in the Swiss priority program OPTIQUE II, funded by the ETH Council) aims at developing a new optical matrix range sensor working in the near infrared which is fast, cheap and can supply 3D information with high accuracy. Necessary trade-offs however, mean that the sensor provides range images which are

Esther B. Meier is with the Communication Technology Lab, Image Science, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. E-mail: ebmeier@vision.ee.ethz.ch .

Frank Ade is with the Communication Technology Lab, Image Science, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. E-mail: frank.ade@vision.ee.ethz.ch .

coarse (resulting from the need for inexpensive sensor and computing hardware), incomplete (due to insufficient or saturated reflections from targets), and inconsistent (in the case of multiple reflections). Furthermore, the developed algorithms will be integrated in the sensor on a driver and signal processing chip, where limited computational resources make it difficult to analyze and interpret the data in real-time.

In this paper we introduce a new method for the segmentation and tracking of obstacles based on coarse range image sequences. We make the assumption that the road can be modeled as a plane and then separate image features into ground and obstacles using a distance map. The obstacles are further segmented into objects which are detected and tracked over time. The modeled location of the road with respect to the sensor is updated based on the ground information. Then the resulting new parameters are used to separate the image features in the next frame (see Fig. 1).

Similar systems have been proposed, that also use a road model. [4] and [5] use the disparity between pairs of stereo images to distinguish features painted on the road and obstacles lying on the ground. The lane markings are detected and used to update the road model and the geometric parameters of the stereo cameras. In comparison to these approaches we process much smaller images and 3D information for each pixel is directly available. On this basis, all the extracted 3D ground pixels are used to update our simple road model. Given our coarse images, it is not feasible

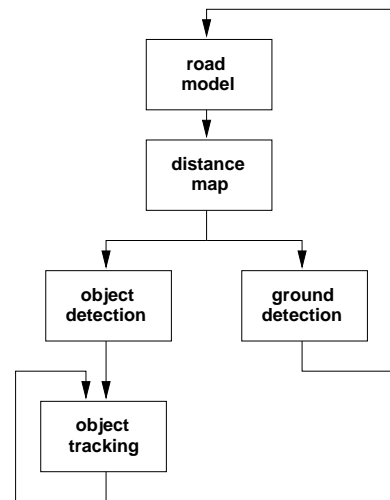


Fig. 1. Information flow of the image features separation scheme

to use a more complex road model like [10].

Other approaches for range images are described by [7] and [9]. Both attempt to localize obstacles. [7] first determines the radial slope, and [9] begins by calculating the surface normals. Such features are extracted on a local basis, and segmentation is derived via clustering, similar to our approach.

In [8] we describe a procedure that operates on registered range and reflectance images using data fusion to combine the physical and spatial properties. A hierarchical region growing scheme is proposed where first the reflectance image is segmented and then the result is refined by using the range image. In our approach, however, it is difficult to distinguish between potential obstacles and the ground. This problem is addressed in this paper, where we investigate the task of road extraction in range images.

## II. DESCRIPTION OF THE RANGE SENSOR

Our work is based on novel types of active ranging techniques [11] with which we can measure distances of up to 150m. The accuracy depends on the distance, and is  $\pm 0.2\text{m}$  at a distance of 100m. Typically, the range sensor has a field of view of  $2.38^\circ \times 10.0^\circ$  and an image size of  $16 \times 64$  pixels (see Fig. 2). In our application, it will be installed behind the windshield or next to a front light. The basic principle of the sensor is to determine the distance to the scene for each pixel by measuring the relative phase of the reflected modulated laser beam.

Unfortunately, the images of the sensor have some practical imperfections such as coarse, incomplete and inconsistent data. It might be possible that small objects, like a road post for example, are not visible in every frame. Due to insufficient or saturated reflections the data might be incomplete, with some image pixels having an undefined value. Inconsistencies arise through multiple reflections; for example if the emitted light is reflected by a mirror.

The sensor is also capable of producing registered reflectance images, where each image pixel represents the amount of light reflected by the target. This information is not, however, used in this approach.

## III. TEST DATA

Until the prototype is fully operational, we have written a simulation package to generate artificial data,

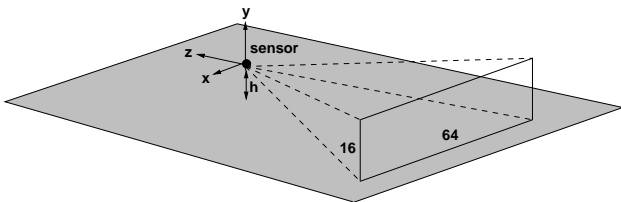


Fig. 2. The geometry of the range sensor

that can be used to develop and test the segmentation and target tracking algorithms. We simulate a virtual world in which a car containing the range sensor is driven through scenery that includes roads, hills, trees and houses, as well as other vehicles. The range sensor is placed in the reference car and calculates distance and reflectance images at given time intervals. Distances are calculated by tracing rays for each image pixel and computing the intersection with the nearest object. The reflectance properties of the materials are simulated by using the Phong illumination model [12]. To create more realistic data camera nodding and noise can be added.

In Fig. 7 simulated range images are shown. The distance data are represented by different grey levels, where black represents undefined pixels. As the car in front of the house has a dark color and is far away, some pixels are undefined due to insufficient reflectance. Additionally, some pixels near the horizon are undefined.

The Institute for Computer Science and Applied Mathematics at the University of Bern, Switzerland, provides us with range image sequences of toy traffic scenarios. They use an ABW range scanner (ABW GmbH, Germany) which is based on structured light. To obtain range image sequences of low resolution, they subsample and crop the original images (see Fig. 10).

## IV. SEPARATION OF IMAGE FEATURES

Our approach is to separate image features into ground and road obstacles by a separating plane. We make the assumption that the road can be locally modeled as a plane and determine a distance map of a slightly steeper plane than the current road model (rotated by  $0.5^\circ$ ). The center of rotation is placed vertically under the range sensor (see Fig. 3). At the beginning the road model is assumed to be horizontal and the sensor has a height  $h$ . Using an inclined plane has the advantage that noisy ground pixels and small road elevations are less likely erroneously identified as obstacles. It is better to use a rotated instead of a shifted plane as our expected noise depends on the distance. Consequently, close and potentially dangerous obstacles are completely detected whereas objects far away might be lost. In a distance of 100m obstacles which are smaller than 87cm are not detected. Such a simple road model is sufficient as the range and the field of

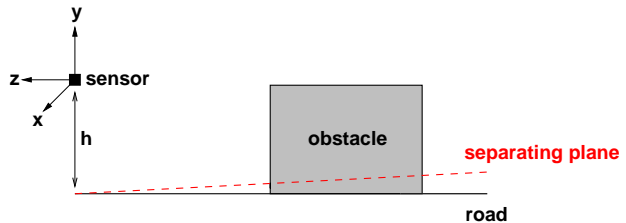


Fig. 3. Separation of image feature

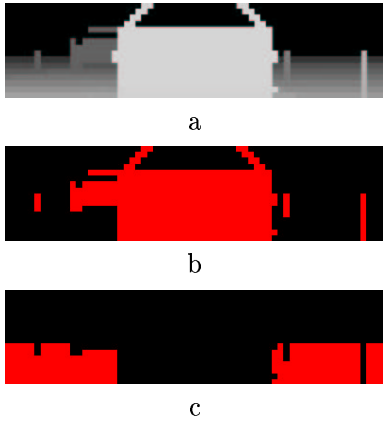


Fig. 4. **a**: Distance image; **b**: Obstacle detection; **c**: Road/ground detection

view of the sensor are limited.

We determine a distance map by calculating the distances the sensor would measure if only the separating plane was seen. By comparing the actual distances with those from the distance map, we can separate image features into ground and obstacles. Smaller distances belong to an obstacle and larger distances belong to the ground (see Fig. 4).

#### A. Ground Detection and Updating the Road Model

So far, we have assumed that the road model is known. However, the car suspension and the road inclination may alter its geometry, with small differences potentially producing large errors in the separation of the image features. Therefore, we update the parameters by fitting a plane through all pixels marked as ground.

In the local coordinate system of the sensor, a plane is determined by the equation

$$n_x \cdot p_x + n_y \cdot p_y + n_z \cdot p_z + h = 0 \quad (1)$$

where  $\vec{n}$  is the normal vector,  $\vec{p}$  is a point on the plane and  $h$  is the distance from the sensor to the plane. Only three parameters are independent since  $|\vec{n}| = 1$ . We divide all coefficients by  $n_y$  (it will never be zero) and substitute the fractions  $\frac{n_x}{n_y}$ ,  $\frac{n_z}{n_y}$  and  $\frac{h}{n_y}$  with the variables  $a$ ,  $b$  and  $c$ :

$$\frac{n_x}{n_y} \cdot p_x + \frac{n_z}{n_y} \cdot p_z + \frac{h}{n_y} + p_y = 0 \quad (2)$$

$$a \cdot p_x + b \cdot p_z + p_y + c = 0. \quad (3)$$

To find the parameters of the plane we insert the known coordinates of the extracted ground pixels and solve the equations for the three unknowns  $a$ ,  $b$  and  $c$ . As we generally have more than three ground pixels, the system is overdetermined and not generally solvable, but deviations or residuals of the individual equations can be minimized. A well-known approach

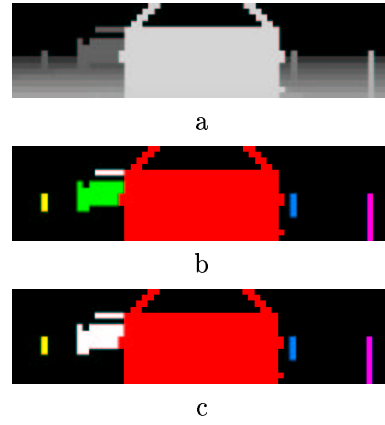


Fig. 5. **a**: Distance image; **b**: Result after looking for connected components; **c**: Result after combining close regions

is the least squares method which minimizes the sum of the squares of the residuals.

As we are dealing with incomplete data, we might not always have enough pixels to get a robust result. Even if we have enough pixels to perform the fit ( $> 3$ ) it can still be insufficient through incorrectly classified pixels. Therefore, we calculate a measure that provides us with additional information. This measure  $q$  is the mean distance of the extracted ground pixels to the fitted plane

$$q = \sqrt{\frac{\sum_{i=1}^n d_i^2}{n}}, \quad (4)$$

where  $n$  is the number of defined ground pixels and  $d_i$  is the distance of each pixel to the plane. If the measure is higher than a specified threshold, the plane fit is insufficient and the update of the geometric parameters is not reasonable. In our approach the threshold is set to 0.5m.

An update of the current road model is not necessary for each frame. The following three parameters are used to decide if an update should be made:

1. the number of ground points
2. the quality of the plane fit
3. the change of the normal vector.

Noise and unevenness of the road should not cause an update of the geometric parameters. Therefore, a small change in the normal vector does not trigger an update. The particular threshold is set at an angle of  $0.3^\circ$ .

#### B. Object Detection

All distances which are smaller than the distances of the separating plane characterize obstacles which we want to group into different objects. This is done by

1. searching for connected components and
2. combining close object regions.

These two steps reduce the number of objects and the subsequent reduction in our data set facilitates easier tracking. For example in Fig. 5 the body and the roof of the oncoming car are merged to one object.

### B.1 Finding Regions through Connected Components

We use spatial coherence to identify regions from the depth information. First, we are looking for connected components in the 8-neighborhood of the image dimensions. In the depth dimension a neighboring pixel is connected if the difference in depth is less than a threshold  $\epsilon$ :

$$\bar{X}_{N+1} = \frac{N}{N+1} \bar{X}_N + \frac{1}{N+1} f(x, y) \quad (5)$$

$$\epsilon > |\bar{X}_N - f(x, y)|. \quad (6)$$

$\bar{X}$  is the mean distance of a region, which is updated whenever a new pixel is added. The distance value of the current pixel is  $f(x, y)$  and  $N$  is the number of pixels which already belong to the region.

Basically, the growing starts at a seed pixel and new neighboring pixels are included if they fulfill the criterion in Eq. 6.

### B.2 Combining Regions

As we have to deal with incomplete data and occlusion, it is possible that an object will be split into different disjoint regions. If two regions belong to the same object, they must be close to each other in 3D space. For each pair of regions we want to determine a probability measure  $s$ , that gives the likelihood that the regions belong to the same object. To calculate  $s$  we first look at  $d_0$ , the difference between the mean distances:

$$d_0 = \left| \frac{\sum_{x,y \in R_1} f(x, y)}{|R_1|} - \frac{\sum_{x,y \in R_2} f(x, y)}{|R_2|} \right| \quad (7)$$

where  $R_i$  is a region,  $|R_i|$  is the number of pixels in the region and  $f(x, y)$  the value of the current distance pixel.

Then the minimal distance  $d_1$  between the bounding boxes around the regions is determined. We measure the distances  $d_x$  in x-direction, respectively  $d_y$  in y-direction between the bounding boxes using min-max-comparisons of the corners (see Fig. 6). If the bounding boxes overlap in a direction the distance is set to zero.  $d_1$  is then  $\sqrt{d_x^2 + d_y^2}$ . Both  $d_0$  and  $d_1$  are calculated in meters.

To describe the *distance similarity* and the *proximity* of two objects a measure is determined which is decreasing with increasing values:

$$p_i = \begin{cases} 1 & : d_i < 1 \\ \frac{1}{d_i} & : \text{otherwise.} \end{cases} \quad (8)$$

The final probability measure  $s$  to combine close regions is set to

$$s = \phi \cdot p_0 + (1 - \phi) \cdot p_1 \quad (9)$$

where  $\phi$  controls the weight given to the distance similarity and the proximity. In our application we choose  $\phi = 0.2$ .

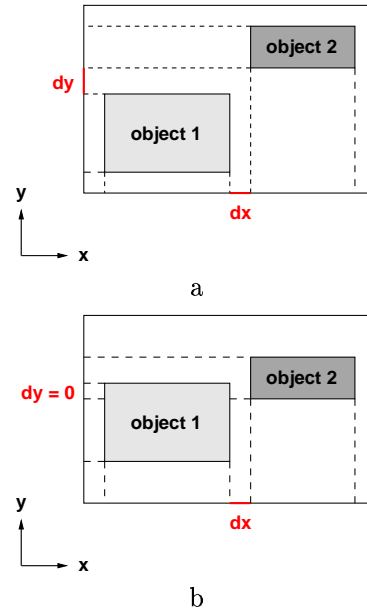


Fig. 6. a: Calculate the distances  $d_x$  and  $d_y$  between two objects; b: The distance  $d_y$  is zero as the bounding boxes overlap in y-direction

A high probability means that the considered regions presumably belong together. Two regions are merged if  $s$  is larger than a threshold  $\theta$  (0.85 in our application).

## V. TRACKING

After detecting the different objects we track them in consecutive frames. An optimal estimator is a computational algorithm that processes measurements to deduce a minimum error estimate of a system by utilizing: knowledge of system and measurement dynamics, assumed statistics of system noises and measurement errors and initial condition information. The optimal estimator for a quadratic error function is the Kalman filter [13], [14].

The model of a linear dynamic system is defined by

$$x_{k+1} = \Phi_k x_k + w_k, \quad w_k \sim N(0, Q_k) \quad (10)$$

where the transition matrix  $\Phi_k$  models the evolution of the state vector  $x_k$  at time  $k$  and the measurement model

$$z_k = H_k x_k + v_k, \quad v_k \sim N(0, R_k) \quad (11)$$

determines the measurements  $z_k$  as a function of the state  $x_k$ .  $H_k$  is called the measurement sensitivity matrix. The system noise  $w_k$  and measurement noise  $v_k$  are zero-mean Gaussian sequences with given covariance-matrices  $Q_k$ , respectively  $R_k$ .

We employ Kalman filtering to provide most likely estimates of the state of each object. In our system the state vector contains the distance  $d_t$ , the relative velocity  $\dot{d}_t$ , the horizontal angle  $\psi_t$  and its change  $\dot{\psi}_t$ ,

the vertical angle  $\theta_t$  and the corresponding change  $\dot{\theta}_t$ , where all measurements are taken at time  $t$ :

$$x_t^T = [d_t \dot{d}_t \psi_t \dot{\psi}_t \theta_t \dot{\theta}_t]. \quad (12)$$

The state predictions are mapped with the measurements (position, horizontal and vertical angle of an obstacle) from the segmentation step. Note that the Kalman filter provides us with the relative velocity and the change of angles.

The complexity of the matching process can be reduced by restricting the search spaces. If several objects are detected within the search interval, the most probable object is matched.

Objects can disappear in some frames due to the coarseness of the data. This problem is solved by updating their parameters with the estimated values from the Kalman filter over several frames. The object can thus be recognized when it reappears within a certain time or otherwise it will be removed from the management.

New objects are added to the management if not all detected objects of the segmentation could be matched.

## VI. RESULTS ON RANGE IMAGES

We have tested our system on several range data and present the results of five sequences.

In Fig. 7, 8 and 9 we see different simulated traffic scenes where the detected objects are represented by their bounding boxes. These 100-frame sequences have a frame rate of 25 images per second, an image size of  $16 \times 64$  pixels and a field of view of  $2.38^\circ \times 10.0^\circ$ . In the first two sequences the sensor is mounted at a height of 0.6m, whereas in the third sequence the height is 1.3m and the sensor is inclined about  $1.5^\circ$ .

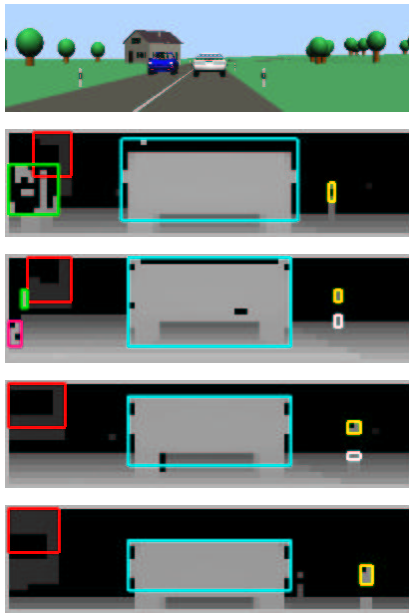


Fig. 7. Simulated range images with sensor oscillation where the top image shows the view from the driver's seat

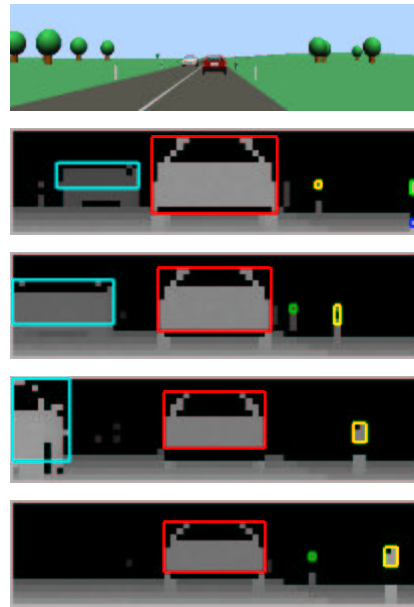


Fig. 8. Simulated range images with Gaussian noise

In Fig. 7 the fluctuations given by the sensor nodding can be observed. The results of a noisy data sequence is shown in Fig. 8. The expected noise of the sensor is only 0.2m for a range of 100m. Due to coarse data the second road post in Fig. 8, third image from the top, is not visible in the fourth image. By updating the position of detected objects over some time steps it can be recognized later (see fifth image). The proposed approach handles these problems very robustly.

In sequence 9 the sensor is mounted behind the windshield and is inclined. Although the traffic scene is seen

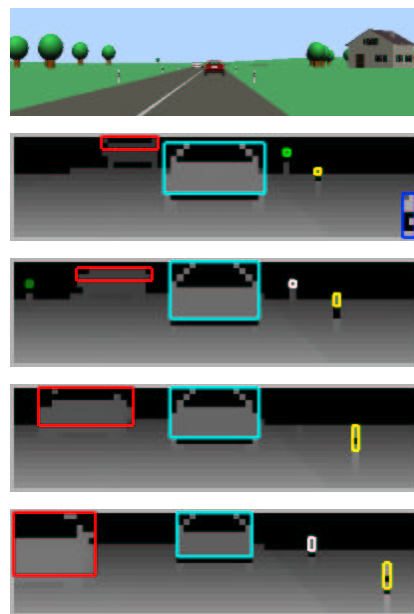


Fig. 9. Simulated range images where the sensor is mounted behind the windshield

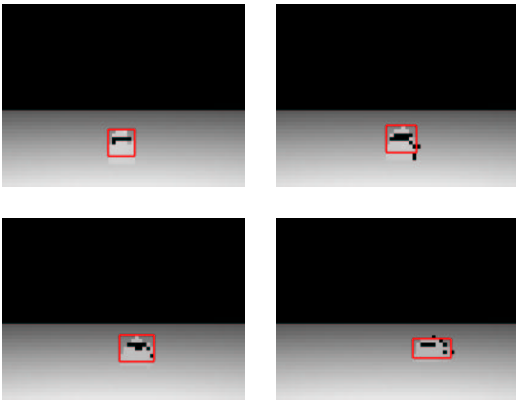


Fig. 10. Toy traffic scenario recorded with structured light sensor

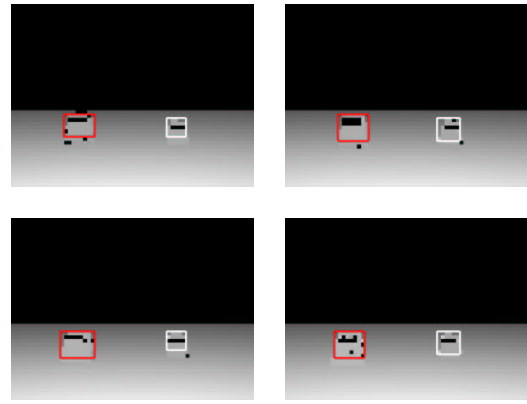


Fig. 11. Toy traffic scenario recorded with structured light sensor

from a different angle the segmentation and tracking methods work properly.

The results on range data that were recorded with the structured light sensor are shown in Fig. 10 and 11. These sequences consist of 10 frames and have an image size of  $64 \times 64$  pixels and a field of view of  $12.8^\circ \times 18.5^\circ$ . The images of these toy traffic scenarios were taken from a height of 2.1m and the sensor was inclined about  $12.3^\circ$ . The undefined pixels within the cars do not affect the segmentation and tracking process. In Fig. 10 we see a vehicle turning to the right and two vehicles can be seen in Fig. 11.

As the results show, the separation of image features works well. The different objects can successfully be detected and tracked if the sensor and road parameters are updated continually.

## VII. CONCLUSION

This paper has proposed a robust system for a driver assistance system based on low-resolution range image sequences. We separate image features into ground and obstacles by determining a distance map of a slightly inclined plane. A good fit is made possible by the accurate measurements of the sensor. Although hardware limitations restrict us to the development of only a simple road model (e.g. limited field of view and range), we have shown how this is sufficient for our purpose. The system can cope with oscillations of the sensor and road inclinations. As the images are small, they are analyzed in real time. Abrupt changes in road inclination may cause a poor fit of the plane. In such cases, part of the road is erroneously identified as a distant obstacle, and the system requires a few frames before the plane is able to stabilize. Our application, however, is motorway usage where such extreme conditions do not occur.

## ACKNOWLEDGMENT

This research was partially supported by MINORA, a project of the Swiss priority program OPTIQUE II, funded by the ETH Council. Further we thank Prof.

Horst Bunke, Dr. Xiaoyi Jiang and Karin Sobottka of the Institute of Computer Science and Applied Mathematics, University of Bern, Switzerland, for providing us with range image sequences of toy traffic scenes to test our approach.

## REFERENCES

- [1] David Beymer, Philip McLauchlan, Benn Coifman, and Jitendra Malik, *A Real-time Computer Vision System for Measuring Traffic Parameters*, IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp. 495-501, 1997.
- [2] Dieter Koller, Joseph Weber, and Jitendra Malik, *Towards Realtime Visual Based Tracking in Cluttered Traffic Scenes*, Proc. 2nd IEEE Symposium on Intelligent Vehicles, pp. 201-206, 1994.
- [3] Thomas Zielke, Michael Brauckmann, and Werner von Seelen, *Intensity and Edge-Based Symmetry Detection with an Application to Car-Following*, CVGIP: Image Understanding, vol. 58(2), pp. 177-190, September 1993.
- [4] Massimo Bertozzi, Alberto Broggi, Gianni Conte, and Alessandra Fasciol, *Obstacle and Lane Detection on ARGO*, IEEE Conference on Intelligent Transportation Systems, pp. 1010-1015, 1997.
- [5] Q.-T. Luong, J. Weber, D. Koller, and J. Malik, *An Integrated Stereo-Based Approach to Automatic Vehicle Guidance*, 5th International Conference on Computer Vision, pp. 52-57, 1995.
- [6] Andrea Giachetti, Marco Cappello, and Vincent Torre, *Dynamic Segmentation of Traffic Scenes*, Proc. 3th IEEE Symposium on Intelligent Vehicles, pp. 258-263, 1995.
- [7] K. Sobottka, and H. Bunke, *Obstacle Detection in Range Image Sequences Using Radial Slope*, 3rd IFAC Symposium on Intelligent Autonomous Vehicles, pp. 535-540, 1998.
- [8] Esther B. Meier, and Frank Ade, *Tracking Cars in Range Image Sequences*, IEEE Conference on Intelligent Transportation Systems, pp. 105-110, 1997.
- [9] Martial Hebert, and Takeo Kanade, *Outdoor Scene Analysis Using Range Data*, Proc. IEEE Int. Conf. Robotics and Automation, pp. 1426-1432, 1986.
- [10] Ernst D. Dickmanns, and Birger D. Mysliwetz, *Recursive 3-D Road and Relative Ego-State Recognition*, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 14(2), pp. 199-213, 1992.
- [11] T. Spirig, M. Marley, and P. Seitz, *The Multi-Tap Lock-In CCD with Offset Subtraction*, IEEE Transactions on Electron Devices, vol. 44(10), pp. 1643-1647, October 1997.
- [12] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes, *Computer Graphics, Principles and Practice*, Addison-Wesley Publishing Company, Inc., 1993.
- [13] Arthur Gelb, *Applied Optimal Estimation*, MIT Press, 1996.
- [14] Mohinder S. Grewal, and Angus P. Andrews, *Kalman Filtering Theory and Practice*, Prentice-Hall, Inc., 1993.