Fast Object Hypotheses Generation Using 3D Position and 3D Motion

Thao Dang and Christian Hoffmann
Institut für Mess– und Regelungstechnik
University of Karlsruhe
76185 Karlsruhe, Germany
Email: {dang, hoffmann}@mrt.uka.de

***Abstract*— This contribution proposes a method to generate object hypotheses from stereo obstacle detection and image motion. Our algorithm is a general approach since it does not require any *a priori* information about the shape of the observed objects but relies on the basic assumption that the objects are rigid. The algorithm has two processing stages: First, obstacles are detected using stereo vision. Second, each obstacle is segmented into clusters of consistent motion in 3D space. The clustering process explicitly accounts for measurement uncertainties of stereo disparity and 2d motion. Our system may serve as a general feature for higher-level object detection and classification.**

I. Introduction

Object detection is one of the key abilities of modern driver assistance systems. A vast literature on this subject exists and different sensors (RADAR, LIDAR, monocular and stereo cameras, etc.) have already been employed to approach object detection. Among these sensors stereo cameras may have particularly strong potential because they combine a broad variety of different cues, such as, e.g., disparity, displacement, texture, color, and shape.

It has been shown that stereo cameras allow efficient detection of obstacles in the path of a vehicle (e.g. [1]). However, the limited depth resolution of stereopsis imposes a strong restriction for object detection. To illustrate this problem, consider the cyclist in Figure 1. Using stereo depth information alone, the bicycle can hardly be separated from the parking car standing next to it. Since the depth accuracy decreases at least quadratically with the distance, this problem becomes even more apparent if the distance between the object and the observing cameras is large.

Several approaches exist to overcome this problem: If we have *a priori* knowledge of the objects in the scene, we can combine this information with stereo object detection. E.g., if we are to design a detection algorithm specialized for vehicles, we could use image features such as shadows, symmetry, bounding boxes, etc. (e.g. [2]). Another important feature for object detection is motion. Recalling our previous example in Figure 1, the cyclist and the parking car can easily be separated if we consider the 2d motion (or displacement) between corresponding points in two consecutive frames of one camera.

Motion is a very general feature for object detection and is even an important cue in human visual perception (e.g., we can detect well camouflaged objects as soon as they start to move). For computer vision, 3D object motion is very appealing. If we detect objects by finding segments of equal 3D motion, the only constraint we rely on is rigidity of the observed objects. This assumption holds true for most "simple" objects such as cars, trucks, etc. and at least for parts of more complex objects (e.g., the head of a pedestrian). However, the measurement of general object motion comprising 3D translation and 3D rotation from stereo image streams is difficult since it relies on noisy disparity and 2d motion data. Thus, tracking procedures are required. A recursive scheme has been proposed in [3], which allows determination of rigid motion with six degrees of freedom (dof) and improves depth resolution by integrating stereopsis and 2d displacement. We envision that 3D motion will form a valuable and general clue that supports specialized algorithms for detecting certain classes of objects.

The goal of this contribution is to generate initial object hypotheses. Our presented approach combines both 3D position and 3D motion. We propose to simplify motion

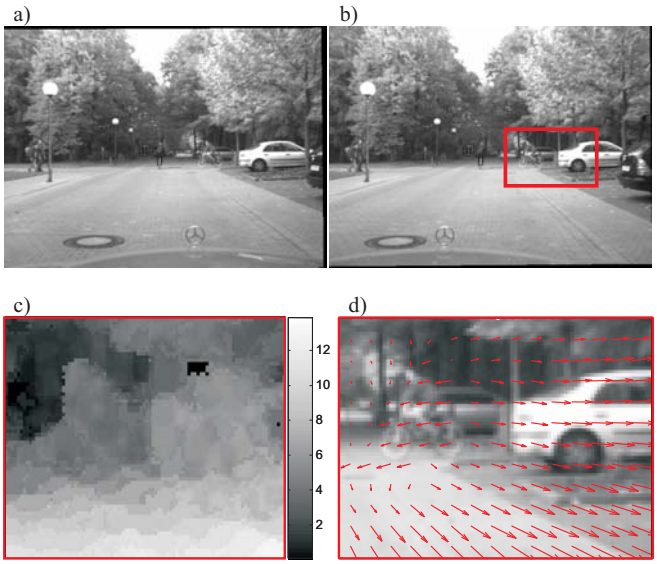


Fig. 1. Sample images and 2d correspondence data: a,b) rectified left and right stereo images, c) disparity (pixels) of marked region, d) 2d motion (pixels) of marked region. The bicycle and the parking vehicle can hardly be separated from disparity alone, but image displacement can clearly distinguish between both objects.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

segmentation using an assumption that may be considered natural in automotive applications: We impose the constraint that object motion is mainly parallel to the ground plane. The resulting object hypotheses may serve as a starting point for subsequent tracking as proposed in [3], but we suppose the presented method will also be beneficial for other object detection approaches.

The algorithm presented in this contribution can be divided into two main stages: In the first stage, obstacles in our path are found by stereo vision only. To accomplish this, we identify the ground plane on which our vehicle is traveling. Using the location of the ground plane, we can classify all detected 3D points into one of the four categories "ground plane", "obstacle", "negative obstacle" (e.g., a ditch in the road) and "irrelevant" (i.e. all points located well above the vehicle). Obstacle points lying close together are then grouped with a flood filling algorithm. The purpose of the second stage is to separate these groupings into clusters of consistent rigid 3D motion. Since in this contribution, we focus on generating initial object hypotheses and do not use tracking, we do not account for full 3D motion with all degrees of freedom. Instead, we simplify our problem by considering only 3D motion parallel to the ground plane. Regions with similar motion components parallel to the road surface are segmented using a divisive clustering algorithm. The proposed algorithm yields promising results on first real-world image data.

The paper is organized as follows: Section II presents the algorithm to locate the ground plane from stereo data, the classification of 3D object points and grouping of obstacle points. In Section III, we describe how visual motion can be used to refine the results of the stereo obstacle detection. A divisive clustering algorithm is used to generate object hypotheses. Throughout the paper, the proposed algorithms are demonstrated on the sample data from Fig. 1. Section IV concludes this paper with a short summary of our results.

II. Obstacle Detection

A. Locating the ground plane

We define obstacles as 3D structures that arise from the road surface. In fact, as depicted in Fig. 2, given the position of the ground plane, we can categorize all observed 3D points into four groups: (a) *irrelevant* points that are located well above the traveling vehicle, (b) *obstacles* in front of the observer, (c) points belonging to the *ground plane*, and (d) *negative obstacles* such as ditches in the road. Thus, reliable location of the road surface is a vital part of obstacle detection.

In this contribution, stereo disparity is computed via block matching. The Zero-Mean Sum of Squared Differences (ZSSD) between two blocks is used as distance measure for matching. As introduced by [4], the matching errors of neighboring windows are also included in our matching criterion. Using this method, we obtain valid disparities even in image regions with low texture. Our current stereo matching procedure requires 80 ms for each run on a Intel Pentium M processor with 2 Ghz.

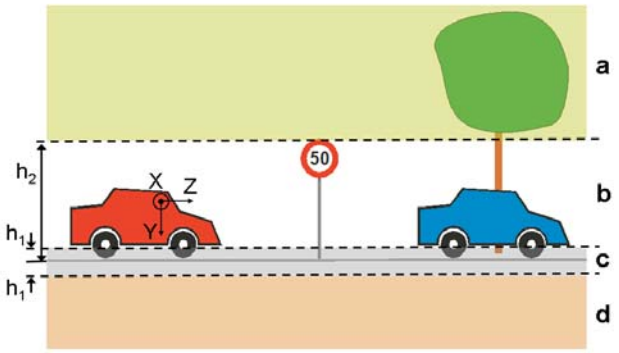


Fig. 2. Obstacle detection. Given a 3D point cloud from stereo vision and the location of the ground plane on which the vehicle is traveling, we can classify all object points into four categories: irrelevant (a), obstacles (b), ground plane (c) and negative obstacles (d).

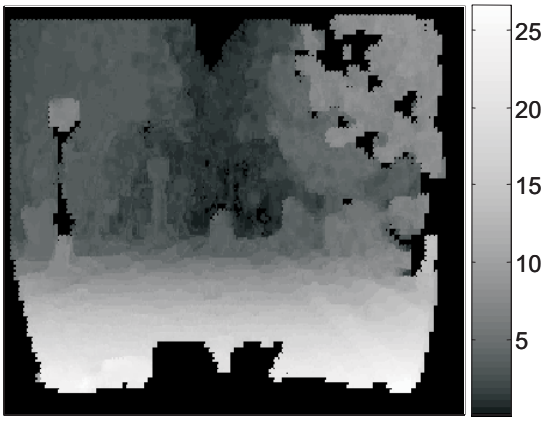


Fig. 3. Extracted disparity (in pixels) for the images in Fig. 1. Disparity is obtained by matching blocks of 9×9 pixels.

Following the work of [5], the ground plane can be detected from stereo disparities using the well known *v-disparity* concept. Each row in the v-disparity image is given by the histogram of the corresponding row in the disparity image. Labayrade et al. found that each tilted plane in 3D space (with zero roll angle) becomes a straight line in the v-disparity image (see Fig. 4). This line can easily be detected using, e.g., a Hough or Radon transform. We have decided to employ a Radon transform for line detection since it does not require binarization of the input image and it can be implemented efficiently using the central slice theorem [6]. The advantages of the v-disparity method are its simplicity and its robustness. However, there are also two drawbacks: First, the determination of the roll angle from the v-disparity is difficult. Second, if we are given sub-pixel accurate disparity estimates, the computational burden of the ground plane detection increases drastically. Several extensions to our current road surface location are possible: We could apply a Total-Least-Squares (TLS) algorithm to refine the ground plane estimation and include a roll angle. Furthermore, accounting for the dynamics of the vehicle by tracking as proposed in [7] will improve our results. However, these extensions have not yet been included in our system.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

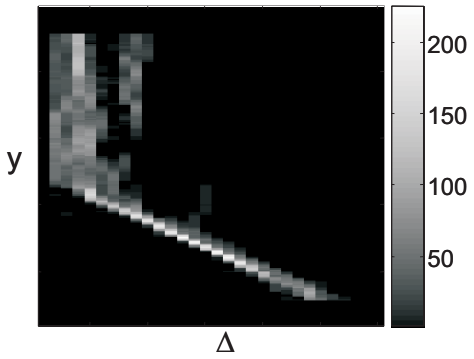
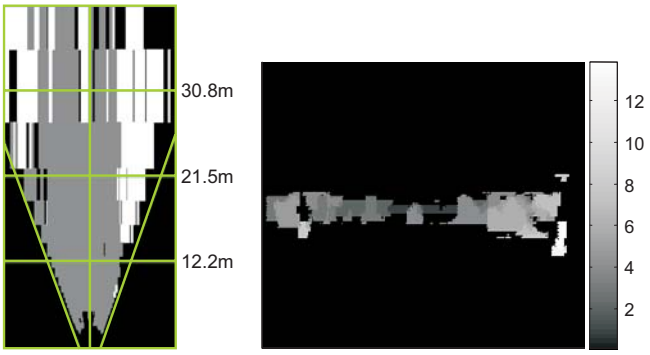


Fig. 4. V-disparity and detected ground plane.

B. Identifying and grouping obstacles

Given the position and orientation of the road surface, we are now ready to classify the image points. Using their 3D position obtained with stereo matching, we can compute their height above or below the ground plane and label them accordingly. Figure 5(a) shows the classification results as a bird's eye view.



(a) Bird's eyeview (b) Obstacles

Fig. 5. Classification of 3D points using the extracted ground plane. The left figure shows a bird's eye view in which trafficable road is plotted in gray, obstacles are denoted by white color and black marks image regions where no valid information is available. Negative obstacles were not present in this example. The image on the right shows the disparities of all points in the obstacles category.

Depicted in Fig. 5(b) are the disparities of all relevant obstacle points in our path. As a first guess for objects in the scene, we can now group connected regions in Fig. 5(b) as obstacles. Connected regions are found using a standard flood filling algorithm (e.g., taken from Intel's Integrated Performance Primitives (IPP) Library). This algorithm connects two neighboring pixels if the distance between their disparities is equal to or less than one pixel.

The result of the flood filling algorithm is shown in Fig. 6. We obtain mostly satisfying results, but as expected the cyclist and the parking car were merged due to their similar disparity.

III. Motion Segmentation and Clustering

The objective of this section is to refine the obstacle detections results. We incorporate image motion to check the consistency of each connected obstacle region from Sec. II-B and segment each obstacle into separate object hypotheses if necessary. A divisive clustering method is used that explicitly accounts for the measurement uncertainties of disparity and 2D motion.

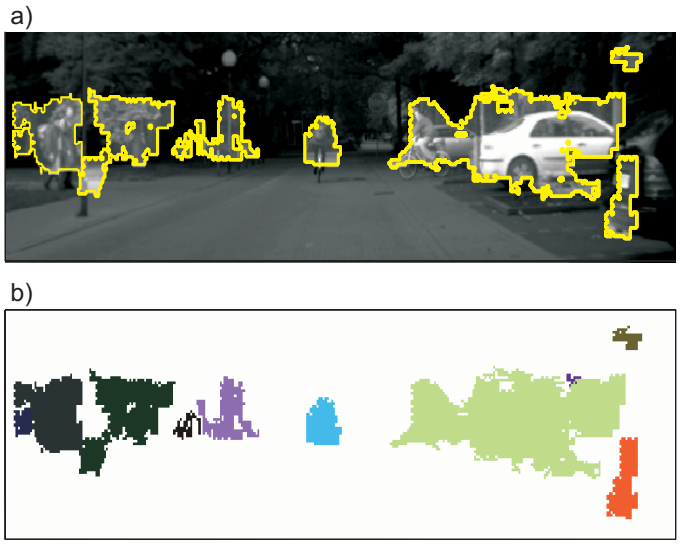


Fig. 6. Obstacle detection results (a: object contours, b: color coded objects). Note that the cyclist and the parking vehicle could not be separated.

A. Image displacement and 3D motion

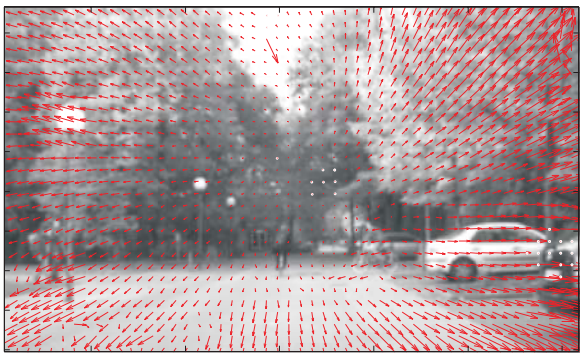
In this contribution, we determine image displacement with a hierarchical Lukas-Kanade-algorithm as proposed in [8]. To reduce computation time, 2D motion is not determined for all image pixels. Instead, we confine 2D motion estimation and clustering to a subset of equally-spaced image points with distances of 3 pixels in both horizontal and vertical direction. In a post-processing stage, each remaining point is assigned to a cluster using nearest-neighbor interpolation. Figure 7 depicts the result obtained for our sample images from Figure 1.

Let us consider an observed object point with 3D coordinates $(X, Y, Z)^T$ and denote its image position by $(x, y)^T$. For simplicity, we assume that the cameras are fully calibrated and all entities within the images are given in normalized coordinates, i.e. we suppose that the images were acquired by ideal cameras with focal lengths $f = 1$ and image centers located at coordinates $(0, 0)^T$. Please note that the measured disparity Δ and image displacement $(u, v)^T$ are also specified in normalized camera coordinates. With this camera setup, the distance Z of the object point and its disparity are related by

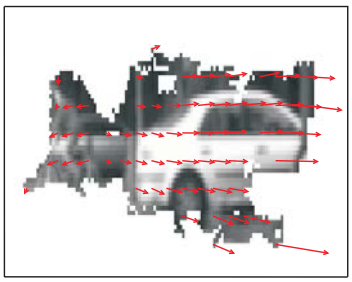
$$Z = \frac{b}{\Delta}\,, \tag{1}$$

where b indicates the base length of the stereo rig.

Since we consider each 3D point individually, we cannot account for 3D rotation. Thus, our rigid motion model is fully specified by the velocity $T = (U, V, W)^T$. The Longuet-Higgins equations (cf. [9]) describe the relation between 3D velocity and 2D displacement. For the case of purely



(a)



(b)

Fig. 7. Extracted 2D motion for the whole image (top) and detected obstacle "cyclist and parking car" (bottom).

translational 3D motion, we obtain

$$\begin{aligned} x &= X/Z\,, \\ \Rightarrow\ \dot{x} &= \frac{\dot{X}}{Z} - \frac{X}{Z^2}\dot{Z} = \frac{U - xW}{Z}\,, \qquad (2) \\ y &= Y/Z\,, \\ \Rightarrow\ \dot{y} &= \frac{\dot{Y}}{Z} - \frac{Y}{Z^2}\dot{Z} = \frac{V - yW}{Z}\,. \qquad (3) \end{aligned}$$

Combining Eqs. (1–3), we get

$$u = \frac{\Delta}{b}(U - xW)\,, \qquad (4)$$

$$v = \frac{\Delta}{b}(V - yW)\,. \qquad (5)$$

Eqs. (4) and (5) constitute two constraints for the three unknown velocity components $(U, V, W)^T$. Consequently, we need to impose further simplifications before we can evaluate the velocity of the observed point. Several possibilities exist:

1) *Known ego motion:*
This simplification (presented in e.g. [10], [11], though with a different derivation as given here) was adapted for automotive applications in [11]. Assuming that the ego velocity $(-U_0, -V_0, -W_0)^T$ of the vehicle is known, the relative velocity of a stationary object is given by $(U_0, V_0, W_0)^T$. Using Eqs. (4) and (5), we obtain

$$\frac{u}{\Delta} = \frac{U_0}{b} - \frac{W_0}{b}\,x\,, \qquad (6)$$

$$\frac{v}{\Delta} = \frac{V_0}{b} - \frac{W_0}{b}\,y\,. \qquad (7)$$

Eqs. (6) and (7) form two linear relations between $(x, \frac{u}{\Delta})$ and $(y, \frac{v}{\Delta})$, respectively. A stochastic test on these linear constraints can be employed to decide whether or not an image point $(x, y)^T$ belongs to a stationary object. In [11], this test was extended to account for 6d ego motion (translation and rotation) and good performance in practice was reported. The advantages of this method are its simplicity and robustness. On the other hand, it requires knowledge of the vehicle's ego motion (which probably will not constitute a problem in automotive applications). Furthermore, it only reaches a binary decision whether or not the object we are looking at is moving independently from the observer. Thus, it cannot distinguish between two different non-stationary objects.

2) *Motion parallel to the retina $W = 0$:*
Another possible approach is to consider 3D motion components parallel to the image plane only, i.e. $W = 0$. Using this assumption, we can determine the remaining velocity components from our disparity and displacement measurements:

$$U = \frac{b\,u}{\Delta}\,, \qquad (8)$$

$$V = \frac{b\,v}{\Delta}\,. \qquad (9)$$

Subsequently, we could generate object hypotheses by clustering points with similar parallel motion $(U, V)^T$. The same result was obtained by [12], where dynamic objects are also assumed to move mainly parallel to the image plane and 2D projection of the motion component W in Eqs. (4–5) was imposed being approximately constant, i.e. $xW = yW = k$. Talukder et al. have also implemented a test based on the constraints (8–9) to detect independently moving objects in the scene [12].

3) *Motion parallel to the ground plane:*
We impose another assumption that may be considered natural in automotive applications. Namely, we consider object motion being parallel to the ground plane. To illustrate this simplification, let us first analyze (without loss of generality) the case where the road surface is parallel to the X-Z-plane of the camera system. With both roll and pitch angle equaling zero, motion parallel to the ground plane implies $V_{||} = V = 0$. Thus, from the Longuet-Higgins Eqs. (4–5) follows that

$$U_{||} = U = -\frac{b}{\Delta}\,\frac{vx - uy}{y}\,, \qquad (10)$$

$$W_{||} = W = -\frac{b}{\Delta}\,\frac{v}{y}\,. \qquad (11)$$

In the general case with non-zero pitch angle θ and roll angle ρ, $(U_{||}, V_{||}, W_{||})^T$ denote the velocity of the observed point with respect to a coordinate system that is aligned with the ground plane. For parallel motion,

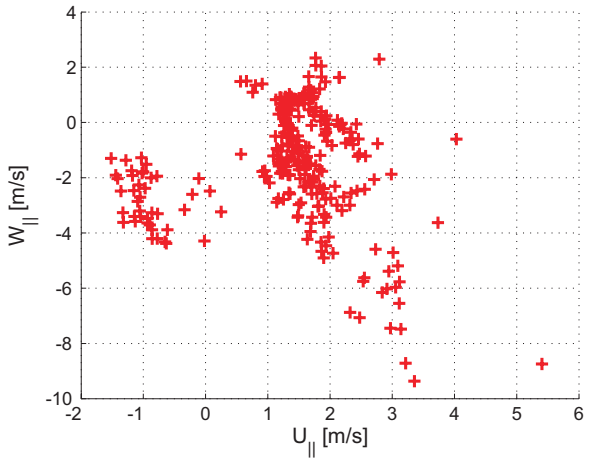


Fig. 8. 3D motion parallel to the ground plane for the detected obstacle ("cyclist and parking car") from Fig. 1.

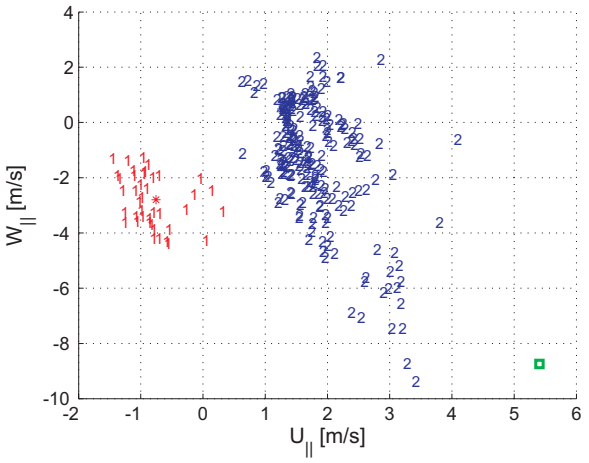


Fig. 9. Clustering result for Fig. 8 using Mahalanobis distances. Our algorithm found two clusters and removed one velocity estimate from the original set.

we have $V_{||} = 0$. Then, the velocities $(U, V, W)^T$ and $(U_{||}, V_{||}, W_{||})^T$ are related by:

$$U = U_{||} \cos\rho + W_{||} \sin\rho \sin\theta \quad (12)$$

$$V = U_{||} \sin\rho - W_{||} \cos\rho \sin\theta \quad (13)$$

$$W = W_{||} \cos\theta \quad (14)$$

and using the Longuet-Higgins equations, we obtain

$$U_{||} = -\frac{b}{\Delta} \frac{u\, c_2 + v\, c_3}{c_1}\,, \quad (15)$$

$$W_{||} = -\frac{b}{\Delta} \frac{u \sin\rho - v \cos\rho}{c_1}\,, \quad (16)$$

where
$$c_1 = x \sin\rho \cos\theta - y \cos\rho \cos\theta - \sin\theta\,,$$
$$c_2 = \cos\rho \sin\theta + y \cos\theta\,,$$
$$c_3 = \sin\rho \sin\theta - x \cos\theta\,.$$

The specific application will have to determine whether roll and pitch are negligible and it is safe to use Eqs. (10–11) instead of the general formulation (15–16). Fig. 8 depicts the estimated horizontal motion components for the detected obstacle "cyclist and parking vehicle". Visually, we can distinguish two dominant velocities and thus two different objects. Clustering all points with consistent motion will be discussed in the next section.

B. Segmentation and Clustering

To segment different objects from Fig. 8, we group points with similar horizontal velocity $\mathbf{T}_{||} = \left[U_{||}, W_{||}\right]^T$. However, the accuracy of the velocity vector will not be equal for all obstacle points. In fact, the uncertainty of $\mathbf{T}_{||}$ depends on the 3D position of the object: For closer objects, the precision of the horizontal velocity will be significantly higher than for objects that are far away. A clustering algorithm must account for this property. We therefore use Mahalanobis distances to evaluate the similarity of two velocity vectors.

Assuming that the measurement uncertainty of the estimated disparity is given by σ_Δ and the uncertainty of the displacement components is specified by σ_u and σ_v, respectively, we can approximate the covariance matrix $\mathbf{K}$ of our horizontal velocity vector by

$$\mathbf{K} = Cov\{\mathbf{T}_{||}\} = \mathbf{H} \begin{bmatrix} \sigma_\Delta^2 & 0 & 0 \\ 0 & \sigma_u^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{bmatrix} \mathbf{H}^T\,. \quad (17)$$

Here, the matrix $\mathbf{H}$ denotes the Jacobian of $\mathbf{T}_{||}$ as defined in Eqs. (15–16) with respect to the measurements Δ, u and v (all in normalized coordinates):

$$\mathbf{H} = \frac{\partial \mathbf{T}_{||}}{\partial \left[\Delta, u, v\right]^T} = \frac{b}{\Delta\, c_1} \begin{bmatrix} \frac{u\, c_2 + v\, c_3}{\Delta} & -c_2 & -c_3 \\ \frac{u \sin\rho - v \cos\rho}{\Delta} & -\sin\rho & \cos\rho \end{bmatrix}\,. \quad (18)$$

In the object segmentation process, we will identify each cluster by its center μ. Using the covariance matrices $\mathbf{K}_i$ associated with the elements $\mathbf{T}_{||,i}$ of a cluster C, the center can be computed as a weighted sum:

$$\mu = \left[\sum_{i \in C} \mathbf{K}_i^{-1}\right]^{-1} \left[\sum_{i \in C} \mathbf{K}_i^{-1}\, \mathbf{T}_{||,i}\right]\,. \quad (19)$$

The distance of an element to a cluster is then given by the Mahalanobis distance

$$d\left(\mathbf{T}_{||}; C\right) = \left(\mathbf{T}_{||} - \mu\right)^T \mathbf{K}^{-1} \left(\mathbf{T}_{||} - \mu\right)\,. \quad (20)$$

If we assume that the measurement errors in Δ, u and v originate from a Gaussian white noise process, the Mahalanobis distance is χ^2-distributed with two degrees of freedom. We can then decide with an error probability of 5% that a velocity vector $\mathbf{T}_{||}$ is inconsistent with a cluster C if the distance $d\left(\mathbf{T}_{||}; C\right)$ is larger than $d_{max} = 5.9915$.

To separate observed objects, we apply a divisive clustering technique that is adapted from the well-known ISODATA

method of [13]. The algorithm is described as follows:

1) Initialization: Set number of clusters $N = 1$ and initialize the first cluster C_1. Its center μ_1 is the centroid (19) of all given features $\mathbf{T}_{||,i}$.
2) Determine all unclassified features that cannot be associated with C_1, i.e. all features for which the distance (20) to the cluster center is above the threshold d_{max}.
3) While the number of unclassified features is above a given threshold (in our examples, we allow four unclassified feature points), do:
   a) Initialize a new cluster C_{N+1} and use an arbitrary unclassified feature point as its center μ_{N+1}. Set $N = N + 1$.
   b) For each feature $\mathbf{T}_{||,i}$, find the nearest cluster center using the Mahalanobis distance (20). If the minimum distance is below the threshold d_{max}, associate the feature to its closest cluster. Otherwise, $\mathbf{T}_{||,i}$ is stored in the list of currently unclassified features.
   c) Recompute the centers for each cluster.
   d) Repeat steps b) and c) until convergence is achieved, i.e. the association of the elements to the clusters does not change any more, or a maximum number of iterations is exceeded (in our example, the algorithm converged after two iterations).

Fig. 9 depicts the results of the clustering algorithm for our previous example. Two clusters were identified and one feature was rejected since it could not be associated with either of the two dominant velocities. The proposed segmentation method was applied to all detected obstacles and performs well as visualized in Fig. 10. We find that the cyclist and the parking car are now well separated into two distinct objects.

IV. Conclusion

We have presented a method to generate object hypotheses from three frames of a stereo image sequence: In a first step, obstacles that arise from the road surface are detected using disparity information of a stereo camera. This yields an initial guess for objects in the scene but in many cases will produce insufficient results due to the limited depth resolution of stereopsis.

Thus, in a second step, each detected obstacle is segmented into clusters of similar motion. We argue that motion in space parallel to the road surface is well suited for object segmentation. A divisive clustering technique is used to separate objects. Divisive clustering is computationally efficient since we expect only a small number of different objects within each obstacle region. In addition, our clustering algorithm explicitly accounts for measurement uncertainties in stereo and 2D motion data.

Our algorithms give good results on first real data experiments. We believe that the object hypotheses generation from stereo vision and image motion can be used to initialize object tracking. It could also serve as a general feature to support high-level object detection and classification. Future work should extend obstacle detection to non-flat road geometry.

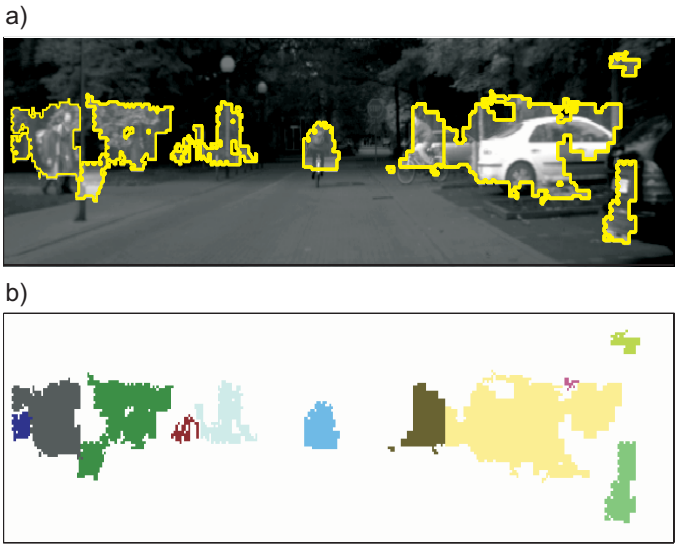


Fig. 10. Detected objects using the proposed clustering technique (a: object contours, b: color coded objects). The cyclist and the parking vehicle are now detected as two individual objects.

References

[1] U. Franke, and A. Joos, Real-time stereo vision for urban traffic scene understanding, *IEEE Intelligent Vehicles Symposium*, pp. 273–278, 2000.
[2] C. Hoffmann, T. Dang, and C. Stiller, Vehicle detection fusing 2D visual features, *IEEE Intelligent Vehicles Symposium*, pp. 280–285, 2004.
[3] T. Dang, C. Hoffmann, and C. Stiller, Fusing Optical Flow and Stereo Disparity for Object Tracking, *IEEE V. International Conference on Intelligent Transportation Systems*, pp. 112-117, 2002.
[4] H. Hirschmüller, P.R. Innocent, and J.M. Garibaldi, Real-Time Correlation-Based Stereo Vision with Reduced Border Errors, *IJCV*, Vol. 47(1/2/3), pp. 229–246, 2002.
[5] R. Labayrade, D. Aubert, and J. Tarel, Real Time Obstacle Detection in Stereovision on Non Flat Road Geometry Through "v-Disparity" Representation, *IEEE Intelligent Vehicle Symposium*, pp. 646–651, 2002.
[6] J. Beyerer and F. Puente Leon, Die Radontransformation in der Digitalen Bildverarbeitung, *Automatisierungstechnik*, vol. 50(10), pp. 472–480, 2002.
[7] M. Cech, W. Niem, S. Abraham und C. Stiller, Dynamic ego-pose estimation for driver assistance in urban environments, *IEEE Intelligent Vehicles Symp.*, pp. 43–48, 2004.
[8] J.-Y. Bouguet. "Pyramidal Implementation of the Lucas Kanade Feature Tracker.", Tech. Rep. included in the OpenCV libary, 2002 (http://www.intel.com/research/mrl/research/opencv/).
[9] H. Longuet-Higgins, and K. Prazdny, The Interpretation of a Moving Retinal Image, *Proc. Royal Society London*, vol. 208, pp. 385–397, 1980.
[10] A.A. Argyros, M.I.A. Lourakis, P.E. Trahanias, and S.C. Ophanoudakis, Qualitative detection of 3D motion discontinuities, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 1630-1637, 1996.
[11] U. Franke, and S. Heinrich, Fast Obstacle Detection for Urban Traffic Situations, *IEEE Transactions on Intelligent Transportation Systems*, vol. 3(3), pp. 173–181, 2002.
[12] A. Talukder, S. Goldberg, and L. Matthies, and A. Ansar, Real-time detection of moving objects in a dynamic scene from moving robotic vehicles, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 2, pp. 1308–1313, 2003.
[13] G. H. Ball, and D. J. Hall, ISODATA, a novel method for data analysis and classification. Tech. Rep., Stanford University, Stanford, CA, 1965.