Stereo-based Real-time Scene Segmentation for a Home Robot

Peter Einramhof[1], Markus Vincze[1]
[1]Vienna University of Technology, Automation and Control Institute, Gusshausstrasse 27-29, 1040 Wien, Austria
einramhof@acin.tuwien.ac.at

***Abstract*—This paper proposes a real-time scene segmentation method based on stereovision and intended for the use on a home service robot. In the first step of our approach the input disparity image is replaced by a lower resolution image. Its pixel disparity values are the result of building histograms over small neighbourhoods in the original image and selecting the maxima. This significantly reduces noise as well as wrong matches, and allows for meaningful local processing. Using constraints derived from the geometry and kinematics of the system camera - robot, ground plane pixels in the lower part of the image are selected. A least squares plane fit is applied to these points to determine the parameters of the ground plane and the rest of the pixels is labelled either as ground point or object point. Removal of the ground leaves single standing clusters of disparities corresponding to objects, which serves as both input for obstacle avoidance as well as object classification.**

I. Introduction

In classic mobile robotics the main focus lies on safe navigation, that is, moving along a pre-planned path while avoiding obstacles, and on self localisation. In service robotics these capabilities alone are too little since there is no immediate benefit for the user– they are merely a basic prerequisite. Also, the way mobile robots perceive their environment significantly differs from how humans do, thus human-robot-interaction suitable for the ”average user” without technical training is challenging: where the user sees a location in front of his couch, the robot might just see a pose (x, y, θ).

The predominant sensor type is the 2D laser range finder scanning parallel to the ground plane. In well-structured environments such as laboratories and offices, where vertical surfaces (unobstructed walls, closets, file drawers) are dominant, and together with 2D maps of the environment, obstacle avoidance and self localisation work quite well with these sensors. Domestic environments are typically more cluttered and less structured, so that - except for simple automatic vacuum cleaners and lawn mowers - there are no service robots in the mass market.

The project robots@home[1] aims at developing safe and robust navigation methods that set the case for using robots in homes everywhere, and at developing a vision-based perception system for learning and mapping of the rooms and classifying the main items of furniture (chair, table, couch, cupboard and door). As main sensor modality stereovision has been chosen as it provides both 2D data (e.g., for colour segmentation, shape-based approaches) as well as 3D data (distances of objects, geometry of the scene).

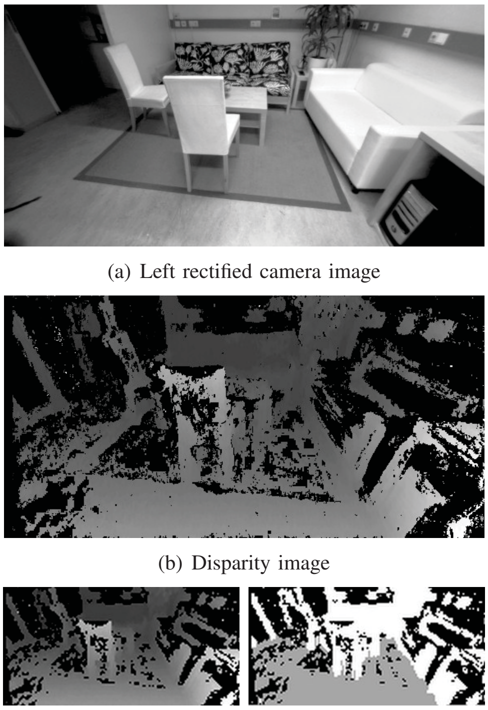
This paper proposes a method for segmenting indoor scenes into ground and objects based on stereo data in real-time. Since stereovision allows to detect the three-dimensional structure of the environment, better obstacle detection (especially protruding surfaces such as table tops) is possible. Furthermore, isolated 3D point clouds stemming from single-standing disparity clusters can serve as input for object classification.

The rest of the paper is structured as follows: in section II we discuss related work. Section III provides an overview of our approach and section IV describes the test setup and shows results.

II. Related Work

The literature proposes several approaches for finding planes in stereo data and in real-time. The RANSAC method [1] searches for parameters of the plane with the largest number of supporting 3D points (calculated from disparity images). Konolige et al. use this approach to detect the ground in outdoor scenes [2]. In [3] Yu et al. apply RANSAC to fit a plane directly in the disparity domain. Although RANSAC is robust against outliers, it will produce undesired results if the ground is not the dominant feature, especially in cluttered indoor scenes. In [4] Labayrade et al. propose the concept of ”v-disparity” that has gained popularity in the driving safety assistance system community. For each disparity image row, a histogram over the disparities is built. Given that the camera baseline is parallel to the ground plane, each row's histogram shows a distinct peak. The peaks of all rows lie on a straight line that can be detected using e.g. the Hough transform. If the longitudinal profile of the ground is not flat, the inital straight line breaks up into several connected line segments. Problematic are roll of the camera as well as clutter since the distinct histogram peaks disappear. A number of modifications and extensions of the initial concept have been proposed [5]–[8]. In [9] Burschka et al. present a method that is similar to both the v-disparity approach and ours. It selects points for ground plane estimation via the histogram peaks over ten selected rows in the lower part of the image and they also apply constraints derived from the geometry of the setup to identify outliers.

[1] http://robots-at-home.acin.tuwien.ac.at/



(a) Left rectified camera image

(b) Disparity image

(c) Filtered image

(d) Labelled image

Fig. 1. The two top images show the left rectified camera image and the disparity image. The left lower image is the output of the first stage of our approach. The right lower image is the result of the labelling process (ground gray, objects white)

III. Approach

A. De-noising and data reduction

The disparity values of pixels within small neighbourhoods that correspond to planar patches such as the ground plane should lie within a very narrow interval. However, measurement noise, missing or wrong matches make local considerations difficult.

To mitigate this problem, we build histograms over the disparity values in neighbourhoods of $n*n$ pixels (in our case $n=4$). In parallel to the normal histogram we use a second one, whose bins serve as accumulators for the disparity values that voted for a certain bin including subpixel resolution. Using a sliding window in which the sum of the bins' vote counts is calculated, we scan through the histogram in order to find the maximum. The sliding window is two bins wide, the scanning starts at the minimum disparity value. To be accepted, a maximum must have got at least n votes. Then, the average disparity of the pixels that contributed to the maximum is calculated. The result is an images whose pixels are assigned the average disparity of the respective peaks (Fig. 1c). It has (far) less noise and wrong matches than the original image. The width and height of this image are $\frac{1}{n}$ of the original image, which means a considerable reduction of data that has to be evaluated.

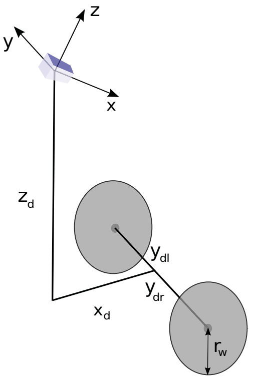


Fig. 2. Geometry of the setup and camera coordinate system. x points in the viewing direction of the camera, y is parallel to the robot's coordinate system y-axis that lies on the line connecting the drive wheel centers

B. Selection of ground plane candidates

The stereo camera is mounted on top of the robot and tilted downwards, so that the lower portion of the camera image covers the immediate space in front of the robot. Since one of the tasks of the robot is obstacle avoidance, there is a good chance that the ground is at least partially visible. However, we cannot blindly rely on that. Based on the geometry and the kinematics of our setup (Fig. 2), the pose of the ground plane with respect to the camera coordinate system can be calculated. The plane is represented by its normal vector

$$\vec{n} = \begin{pmatrix} cos(\rho)sin(\theta+\gamma) \\ -sin(\rho) \\ cos(\rho)cos(\theta+\gamma) \end{pmatrix} \tag{1}$$

and one point on that plane

$$\vec{P} = \begin{pmatrix} -(z_d sin(\gamma) - x_d cos(\gamma) + r_w sin(\theta+\gamma)) \\ y_d \\ -(z_d cos(\rho) + x_d sin(\rho) + r_w cos(\theta+\gamma)) \end{pmatrix} \tag{2}$$

with

$$y_d = y_{dl} \ for \ \rho \geq 0, \ y_d = -y_{dr} \ for \ \rho < 0 \tag{3}$$

γ is the angle by which the camera is tilted towards the ground, θ and ρ are the (dynamic) pitch and roll angle of the mobile robot. r_w is the radius of the robot's drive wheels, x_d and z_d are the displacements of the camera center with respect to the odometry center and y_d is the displacement with respect to the left ($\rho \geq 0$) or right ($\rho < 0$) drive wheel center.

Through each pixel (u, v) of the rectified left image (or equivalently the disparity image) we can send out a visual line

$$\vec{x}(u,v) = \begin{pmatrix} \mu(u,v) - f \\ \mu(u,v)\frac{c_x - u}{f_x} \\ \mu(u,v)\frac{c_y - v}{f_y} \end{pmatrix} \tag{4}$$

and intersect it with the model of the ground plane. This yields

$$\mu(u,v) = \frac{(x_d sin(\theta) - z_d cos(\theta) - r_w + f sin(\theta+\gamma))cos(\rho) - y_d sin(\rho)}{(sin(\theta+\gamma) + \frac{c_y - v}{f_y} cos(\theta+\gamma))cos(\rho) - \frac{c_x - u}{f_x} sin(\rho)} \tag{5}$$

Finally, we calculate the disparity of the pixel (u, v):

$$d_{Subpixel}(u,v) = \frac{Bf_x 2^{Subpixel}}{\mu(u,v) - f} \tag{6}$$

B is the baseline of the stereo camera and f the focal length of its lenses. c_x and c_y are the pixel coordinates of the principal point, f_x and f_y is the focal length divided by the pixel pitch in horizontal and vertical direction, respectively.

By setting θ and ρ in (5) to the maximally allowed negative and positive pitch and roll angle of the robot while moving, and using (6), we can calculate a disparity interval for each pixel corresponding to the ground plane. In our approach, we only (pre-)calculate such intervals for the bottom-most pixel row once during system startup.

The bottom-most pixels of the disparity image described in the previous section are Hough-transformed, but only those that lie within said interval. The rest of the pixels that correspond to clutter are not regarded, which prevents the Hough-transform from producing meaningless results if the ground plane is not dominant within the image. If there are too few ground pixels, we fall back to the (static) offline calibration.

Based on the result of the Hough transform, the bottom-most pixels belonging to the ground plane are labelled. Then, from each such pixel we scan upwards in the respective pixel column, further labelling pixels whose disparity values monotonously decrease at a certain rate. This local processing makes only sense due to the initial de-noising step.

C. *Ground plane parameters and ground labelling*

A plane in 3D corresponds to a plane in the disparity domain. Thus, the labelled pixels are part of a plane $a\tilde{u} + b\tilde{v} + c = d_{Subpixel}$, where $\tilde{u}$ and $\tilde{v}$ are the pixel coordinates of the reduced resolution disparity image and $d_{Subpixel}$ is the disparity value (in subpixels) at $(\tilde{u}, \tilde{v})$. The parameters a, b and c are determined via a least squares plane fit. We select three non-collinear points on that plane, and using

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = x \begin{pmatrix} 1 \\ \frac{c_x - \tilde{u}}{f_x} \\ \frac{c_y - \tilde{v}}{f_y} \end{pmatrix} \tag{7}$$

with

$$x = \frac{Bf_x 2^{Subpixel}}{d_{Subpixel}} \tag{8}$$

we calculate the corresponding 3D points. Based on these three points we calculate the normal vector as well as the normal distance of the camera center from the plane. The plane in disparity domain is used to label the rest of the ground within the low resolution image. a, b, c and $d_{Subpixel}$ are adjusted for the full resolution disparity image, and there the pixels belonging to the ground are removed. We do this quite conservatively as not to cut away too much from the objects. The scattered ground points that remain are removed by using the labelled image as mask (one labelled pixel for $n{*}n$ pixels).

D. *Obstacle point computation*

Using the results of the previous section, the non-ground pixels of the reduced resolution disparity image are transformed into 3D points within the robot coordinate system. The x-axis (pointing into the robot's motion direction) is divided into cells of 5cm, the maximum x-coordinate considered is 5m. For each cell exists a linked list. For each pixel column the respective 3D points are added to one of these linked lists if their x-coordinate falls into the respective cell. The cells are scanned through (starting at $x = 0$) and the first cell (for each pixel column) with at least two points in a height dangerous for the robot is selected. Within this cell the point with the smallest x-coordinate is selected as 2D representative for the respective pixel column (the z-coordinate is skipped). All points together form a ”virtual laser scan” (Fig. 4a). As stated in the introduction, the 2D representation is compatible to current mobile robotics' obstacle avoidance algorithms.

IV. Test Setup and Results

A. *Test Setup*

The custom-built stereo camera consists of two monochrome USB UI-1226LE[2] camera modules. The baseline is 12cm and the focal length of the S-mount lenses is 2.5mm. The modules' native resolution is 720x480 pixels, after rectification a resolution of 586x295 pixels remains. A third module (Bayer pattern) is located half way between the monochrome cameras but was not used in our experiments.

The stereo engine was developed by the Safety and Security Department[3] of the Austrian Institute of Technology. It is based on the Census Transform [10]. In our experiments we used a maximum disparity of 80 with 16 subpixels per disparity. To reduce false matches in regions with little or highly repetitive texture, the stereo engine allows to set texture and confidence thresholds (8 bits each). We used the default values of 30 for confidence and 10 for texture.

Our differential-drive robot (Fig. 3) was manufactured by the Swiss company BlueBotics[4]. A superstructure made of aluminum profiles as well as an additional on-board PC were mounted onto the robot. The stereo camera is mounted in a height of 132cm above the ground and tilted downwards 32 degrees.

B. *Results*

Our approach was tested on a notebook with Core Duo T2250 (1.73GHz) of which only one core (single-threaded) was used. Computation of the reduced resolution disparity image takes 4ms on average. The Hough transform, scanning for the initial ground point candidates, the least squares plane fit, determination of the 3D plane parameters and the final

[2] http://www.ids-imaging.de

[3] http://www.smart-systems.at/

[4] http://www.bluebotics.ch/

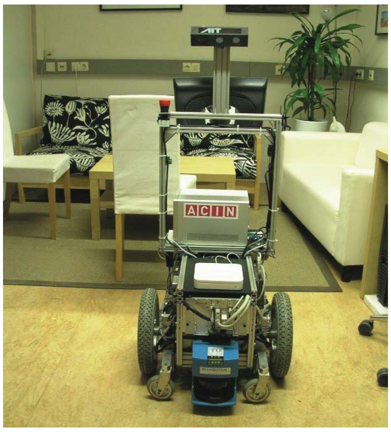
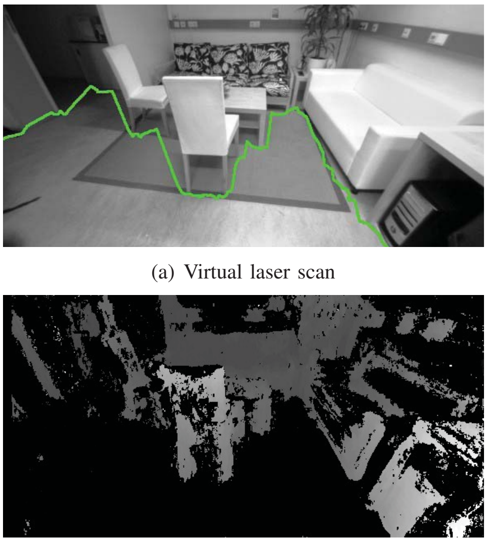


Fig. 3. Mobile robot with the stereo camera mounted on top



(a) Virtual laser scan

(b) Ground plane removed

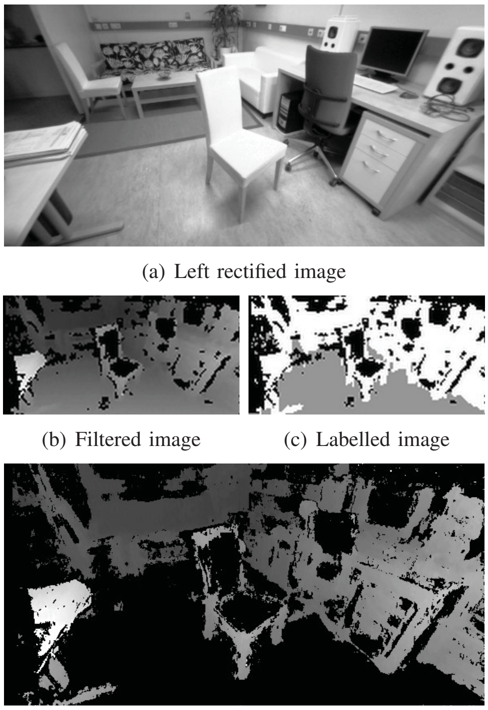
Fig. 4. Processing results intended as input for obstacle avoidance and object classification

labelling of the reduced resolution image take 1ms. Calculation of the 3D points from the reduced disparity image, their derotation and the obstacle point computation take 2ms. This means that after around 7ms the input for obstacle avoidance can be provided. Removing the ground plane from the full resultion disparity image takes an additional 4ms.

Fig. 4 shows the results for the scene depicted in Fig. 1. In the upper image the virtual laser scan is overlayed onto the left rectified camera image. Fig. 4b and Fig. 5d show that result of removing the ground plane in the full resolution disparity images.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Sixth Framework Programme (FP6/2003-2006) under grant agreement no FP6-2006-IST-6-045350 (robots@home).



(a) Left rectified image

(b) Filtered image

(c) Labelled image

(d) Ground plane removed

Fig. 5. Processing result intended as input for object classification

REFERENCES

[1] Fischler, M. and Bolles, R.: Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. Commun. ACM., 24:381-395, 1981.

[2] Konolige, K., Agrawal, M., Bolles, R. C., Cowan, C., Fischler, M. and Gerkey, B.: Outdoor mapping and navigation using stereo vision. Proceedings of the International Symposium on Experimental Robotics, Brazil, 2006.

[3] Yu, Q., Araujo, H., Wang, H.: Stereo-Vision Based Real time Obstacle Detection for Urban Environments. The 11th International Conference on Advanced Robotics, 1671-1676, Coimbra, Portugal, June 30 - July 3, 2003

[4] Labayrade, R., Aubert, D., Tarel, J.-P.: Real Time Obstacle Detection on Non Flat Road Geometry through "V-Disparity" Representation. 2002 IEEE Intelligent Vehicle Symposium, 646–651, Versailles, France, 18-20 June 2002.

[5] Lemonde, V., Devy, M.: Obstacle detection with stereovision. 2004 IEEE Mechatronics and Robotics, 919–924, Aachen, Germany, 13-15 September 2004

[6] Hu, Z., Uchimura, K.: U-V-Disparity: an efficient algorithm for stereovision based scene analysis. 2005 IEEE Intelligent Vehicles Symposium, 48–54, Las Vegas, USA, 6-8 June 2005

[7] Zhao, J., Katupitiya, J. and Ward, J.: Global Correlation Based Ground Plane Estimation Using V-Disparity Image. 2007 IEEE International Conference on Robotics and Automation, 529–534, Roma, Italy, 10-14 April 2007

[8] Soquet, N., Aubert, D., Hautiere, N.: Road Segmentation Supervised by an Extended V-Disparity Algorithm for Autonomous Navigation. 2007 IEEE Intelligent Vehicles Symposium, 160–165, Istanbul, Turkey, 13-15 June 2007

[9] Burschka, D., Lee, S. and Hager, G.: Stereo-Based Obstacle Avoidance in Indoor Environments with Active Sensor Re-Calibration. 2002 IEEE International Conference on Robotics and Automation, 2066–2072, Washington DC, USA, 11-15 May 2002.

[10] Zinner, C., Humenberger, M., Ambrosch, K. and Kubinger, W.: An Optimized Software-Based Implementation of a Census-Based Stereo Matching Algorithm. Lecture Notes in Computer Science, 216–227, Vol. 5358, 2008.