Proceedings of ICPR 2002
August 2002, Quebec, Canada, to appear

Scene Classification from Dense Disparity Maps in Indoor Environments

Darius Burschka and Gregory Hager

Computational Interaction and Robotics Laboratory
Johns Hopkins University, Baltimore, USA
E-mail: {burschka|hager}@cs.jhu.edu

Abstract

We present our approach for scene classification in dense disparity maps from a binocular stereo system. The classification result is used for tracking and navigation purposes. The presented system is capable of foreground-background separation classifying room structures. The 3D model of the scene is derived directly from the disparity image. This approach is used for initial target selection and scene classification in mobile navigation. It is used on our mobile system for target tracking, but can also be used for localization as described in this paper.

We describe the basic principles of our object detection and classification using disparity information from a binocular stereo system. The theoretical derivation is supported by results from the binocular stereo sensor system on our mobile robot.

1. Motivation

Scene classification is an important task in navigation systems. It helps in sensor-based 3D model generation to discriminate between objects interesting for missions (*foreground*) and *background* objects relevant merely for localization. It is also used to trigger different behaviors of the robot depending on the environment structure. Typical classification results in this domain are: *empty space, junction, hallway, corner*. The third important application is target selection and classification in tracking. The target selection task is a challenging part of the tracking system and can be implemented as a manual or automatic process. Examples in 2D image space are described in [8, 6] in more detail. Interesting targets like single standing objects in the scene need to be separated from the supporting planes of the floor and walls that are merely relevant for path planning.

Single standing objects are categorized as *foreground*. They need to be separated from the room structure (*background*) first. In an additional step the remaining *foreground* objects are classified according to their shape, extension and movement relative to the scene. The *background* structures are used in a subsequent classification process to classify the room structure according to the criteria described above.

We structure this paper as follows. In the following section (section 2) we describe briefly the work done in the field of ground plane detection. In section 3 we describe the basic idea of the fast plane segmentation in disparity images followed by the description of the localization possibilities and our approach for scene classification. In the results section (section 4) we present a short excerpt of the results in the variety of applications of the presented system. We conclude in section 5 with a few remarks about the system and a description of future work.

2. Related Work

Ground Plane Obstacle Detection (GPOD) using stereo disparity was first reported by Sandini et al. [2], and refined by Mayhew et al. [5] and by Brady et al. [7]. These approaches use orthogonal regression techniques to estimate the parameters of the ground plane. Approaches like the one from Brady et al. [7] use line features grouped in a Hough transform to detect obstacles in the environment. Our approach uses in contrast directly the disparity information in the image. We present an approach that fits multiple planes into a dense disparity image to allow calibration, localization and object classification from a single image.

3. Approach

Our approach generalizes the ground plane detection from the approaches described in the previous section to a generic segmentation of the room structure from dense disparity images. In the following text we assume a rectified pair of stereo images from a non-verged binocular camera system.

The correct foreground-background separation and room structure classification are the two major problems addressed in the following sections. We assume an operation in indoor environments where background can be approximated with planar surfaces.

3.1. Plane Segmentation

In the first step the supporting surfaces representing the room structure need to be removed from the disparity image to reveal the position of single standing objects in the scene. The ground plane and walls connect all image elements to one continuous region. Their removal isolates single standing objects in the scene.

3.1.1 Imaging Properties of a Parallel Camera System

In indoor environments the room structure can be approximated with planar surfaces $\mathcal{P}_r$.

$$\mathcal{P}_r : a_r x + b_r y + c_r z = d_r \tag{1}$$

In a stereo system with parallel cameras the image planes are coplanar. In this case, the disparity value $D(u, v)$ of a point (u,v) in the image can be estimated from its depth z to

$$D(u, v) = \frac{B}{z}, \tag{2}$$

with B describing the distance between the cameras of the stereo system [1].

The homographic projection onto the rectified parallel camera images preserves the planar properties of the projected pixels. We estimate the disparity $D(u_r, v_r)$ of the plane $\mathcal{P}_r$ at an image point (u_r, v_r) using the unit focal length camera (f=1) projection to

$$\begin{aligned} \forall z \neq 0 : \quad a_r \frac{x}{z} + b_r \frac{y}{z} + c_r &= \frac{d_r}{z} \\ a_r u + b_r v + c_r &= k \cdot D(u_r, v_r) \\ \text{with } u = \frac{x}{z},\ v = \frac{y}{z}, \quad k &= \frac{d_r}{B} \end{aligned} \tag{3}$$

The vector $\mathbf{n_r} = (a_r\ b_r\ c_r)^T$ is normal to the plane $\mathcal{P}_r$ and describes the orientation of the plane relative to the camera.

The equation (3) can be written in the form

$$D(u_r, v_r) = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix} \cdot \begin{pmatrix} u_r \\ v_r \\ 1 \end{pmatrix} = \mathbf{n_r^*} \cdot \begin{pmatrix} u_r \\ v_r \\ 1 \end{pmatrix} \tag{4}$$

$$\text{with} \qquad \rho_1 = \frac{a_r}{k},\ \rho_2 = \frac{b_r}{k},\ \rho_3 = \frac{c_r}{k}$$

This form uses modified parameters ρ_1, ρ_2, ρ_3 of the plane P_r relating the image data u_r, v_r to $D(u_r, v_r)$.

3.1.2 Plane Fitting in Dense Disparity Images

Using at least 3 points from the disparity image, we calculate the plane parameters for a local area by solving a set of linear equations for (ρ_1, ρ_2, ρ_3). The selected points must not be co-linear to prevent a reduction of the rank of the matrix in the linear system.

The correct plane reconstruction depends on correct selection of the points on the surface of the supporting plane. This selection may be disturbed by occlusions from single standing objects in the scene. The trivial solution to this problem is an exhaustive search for all possible plane combinations between the disparity values in the dense disparity map. The complexity of this solution is inapplicable on a mobile system using the sensor data for tracking and navigation purposes.

In our current implementation we expect the room structure to consist of horizontal (floor, ceiling) and vertical (walls) planar surfaces. We assume that the supporting plane is dominant in its area of appearance. It means that a majority of the pixels in this area belong to the given supporting plane.

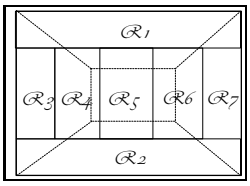

Figure 1. Regions in the disparity image used to estimate the plane equations for P_r.

We subdivide the disparity image in regions of expectation for a specific room structure R_s (fig. 1). We search for horizontal planes in the regions R_1, R_2 and for vertical planes in the remaining regions.

In case of a horizontal plane in region R_1 or R_2 the histogram for a single image line of the disparity image should be a single value (Fig. 2 left) according to equation (4). In this case the ρ_1 component of the vector n_r^* is zero making the disparity D(u,v) independent of the horizontal image coordinate u. In the reality, the cameras are never aligned so exactly. Fig. 2 right shows that in the regular case the dominant plane covers a small disparity range in the histogram, but it is still significant for this line. The same is true for vertical surfaces in regions $\{R_{s\in\{2,3,4,5,6\}}\}$ due to $\rho_2 = 0$ in equation (4).

All regions are processed sequentially. A processing of each region consists of two steps: estimation of plane parameters and removal of the matching pixels from the entire disparity image.

In our approach, histograms over the entire image row or column are calculated for 10 different lines in regions $\{R_1, R_2\}$ or columns in regions $\{R_{s\in\{2,3,4,5,6\}}\}$ (fig. 1). RANSAC [3] method is used to estimate a valid set of plane parameters $\{\rho_1, \rho_2, \rho_3\}$ in a given region.

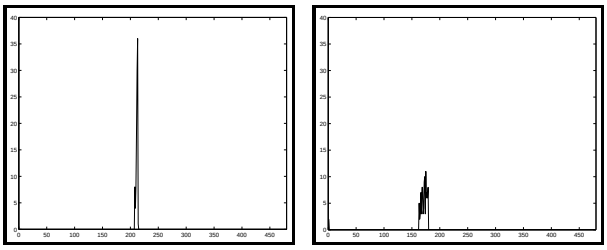

Figure 2. Disparity range covered by the floor plane in a single row of the disparity image: (left) almost perfect alignment (right) roll angle of 30°

We include a "sanity" check rejecting false supporting plane estimations resulting from possible dominant occlusions in the scene. The check is based on the orientation of the reconstructed norm vector and the distances to the surfaces. This is especially important for the on-line recalibration process mentioned at the end of this section.

In the second step all pixels matching the current plane assumption (ρ_1, ρ_2, ρ_3) are removed from the disparity map if their disparity $D(u, v)$ matches the prediction from equation (4). The search is performed on the entire disparity image. In this way it is not necessary to adapt the region boundaries in Fig. 1 to the surface boundaries of the supporting planes. In case a surface spawns several regions in the disparity image, the histograms in the following cycle will not generate any significant peaks leading to new plane assumptions and the second step of the processing (removal of the matching pixels) will be skipped.

The remaining pixels in the disparity image are considered *foreground* and they are processed in the final step. A clustering algorithm is run over the disparity image. It scans the image horizontally. It grows clusters $\mathcal{C}_i$ sufficing the following constraints:

- **compactness in the image space** - we require that the distance between neighboring elements p_i, p_j of an object $\mathcal{O}_i$in the image should not exceed a given threshold ϵ_c

$$|p_i - p_j| < \epsilon_c \tag{5}$$

- **compactness in the object space** - in real images, areas of an object $\mathcal{O}_i$ may not be detected correctly due to texture properties of the surface in this area. Therefore, we allow a maximum distance ϵ_d between two closest disparity values (d_k, d_m) of a single cluster $\mathcal{C}_i$.

$$\exists p_k, p_m \in \mathcal{O}_i : |d_k - d_m| < \epsilon_d \tag{6}$$

3.1.3 Plane Localization

Each plane $\mathcal{P}_r$ extracted in the above process can be described by three parameters: the distance to the plane H_r, the horizontal Θ_{rH}, and vertical Θ_{rV} orientation of the plane respective to the camera plane. The plane parameters ρ_1, ρ_2, ρ_3 are used to estimate the distance H_r and the orientations Θ_{rH} and Θ_{rV}.
From the equation (1) we can write for H_r

$$H_r = \frac{d_r}{\sqrt{a_r^2 + b_r^2 + c_r^2}}. \tag{7}$$

Using equations (3) and (4) we can re-write it to

$$\begin{aligned} H_r &= \frac{B \cdot k}{\sqrt{(k\rho_1)^2 + (k\rho_2)^2 + (k\rho_3)^2}} \\ &= \frac{B}{\sqrt{\rho_1^2 + \rho_2^2 + \rho_3^2}} \end{aligned} \tag{8}$$

The orientation angles can be estimated to

$$\begin{aligned} \Theta_{rH} &= \arctan \frac{\rho_2}{\rho_3} \\ \Theta_{rV} &= \arctan \frac{\rho_1}{\rho_3} \end{aligned} \tag{9}$$

We apply the presented algorithm to recalibrate on-line the extrinsic camera parameters $H_r, \Theta_{rH}, \Theta_{rV}$ based on the plane parameters of the ground plane (fig. 3). This helps to estimate the current orientation of the camera on a pan-tilt head during a tracking process.

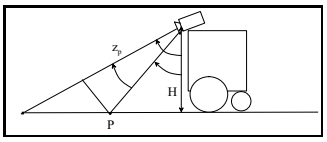

Figure 3. Calibration of the extrinsic camera parameters based on the presented plane localization.

A second important application is localization in a local area. Given two non-coplanar vertical planes, we can estimate the (x, y, φ) pose of the robot.

3.2. Classification

In our experiments the subdivision of the indoor environment in foreground and background objects brings several advantages. In typical scenarios the background consists of large plain surfaces usually representing walls, floor, ceiling, and big cabinets. These objects are detected and represented in their planar description as norm vector and distance from the origin $(\mathbf{n}^*, d^*)$. They are used for localization and path planning, but they do not represent targets interesting for tracking.

3.2.1 Foreground Objects

In our approach all disparities corresponding to *background* surfaces are removed first, before a target selection begins (Fig. 4).

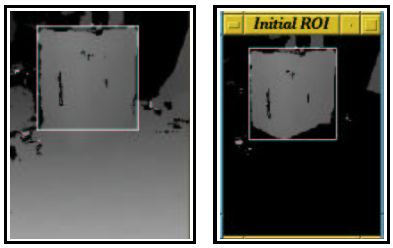

Figure 4. Distinction between foreground and background at the example of the floor.

Remaining regions found in the clustering step (section 3.1.2) are considered *foreground* and saved in a local obstacle map. They are represented by their: position in the image p_i, size of the region s_i, disparity range describing the extension of the object $[d_{min}, d_{max}]$, shape represented by the ratio of height to width, and the compactness of the region described by the percentage of the pixels that belong to the disparity range $[d_{min}, d_{max}]$.

In a typical application single standing objects close to the robot are inspected for movements in consecutive frames. The shape and depth range derived from the classificator helps to identify objects of interest for the tracking algorithm (usually another mobile robot in our lab) and to distinguish it from, e.g., humans walking in the room. The foreground-background separation reduces the number of the monitored objects significantly.

3.2.2 Background Objects

The planes P_r removed from the disparity image are used for a subsequent classification of the room structure. We store the actual extension in pixel coordinates for each plane P_r removed from the disparity image together with the region information. The region information describes in which regions the plane was found. The currently detected structures are:

- *hallway* - both vertical boundary regions R_3, R_6 (Fig. 1) report a significant vertical plane, the absolute value of the scalar product of the normalized vectors $|n^*_{ri} \cdot n^*_{rj}| > \cos angle_{max}$ with $angle_{max}$ describing maximum deviation in the orientation from the coplanar arrangement, and $H_{ri} + H_{rj} < d_{max}$ with d_{max} describing the maximum width of a passage to still be considered a hallway;

- *junction* - both vertical boundary regions R_3, R_6 report the same plane P_r with $H_r < d_{empty}$. In this configuration the robot considers an approaching wall as a junction. The behaviors in both cases are similar and did not require additional subdivision. It is possible to add additional hallway search in the middle regions (see previous point) to distinguish between wall and junction;

- *corner* - the vertical boundary regions R_3, R_6 report two planes with $|n^*_{ri} \cdot n^*_{rj}| < \cos angle_{corner}$;

- *empty space* - both vertical boundary regions R_3, R_6 report the same plane P_r with $H_r > d_{empty}$.

4. Results

In our experiments we used a Nomad Scout as a mobile robot with a PentiumIII@850MHz notebook running Linux-OS. The system was equipped with SRI's MEGA-D Megapixel Stereo Head with 8mm lenses. The cameras were mounted in a distance of 8.8cm from each other.

The typical tilt angle of the camera system during the experiments was $\Theta = 53°$. The system was mounted $H = 48.26cm$ above the ground. This configuration robustly detected obstacles in front of the robot while still allowing viewing up to 4m in front of the robot.

In this configuration the system was running with a frequency of 7.2 Hz for the entire obstacle detection and classification cycle.

4.1. Quality of the Background Segmentation

Background segmentation is fundamental to the entire approach. An example of the calibration based on the ground floor is shown in Fig. 5. Each image triple shows the real image in the top left corner, the computed disparity image in the top right corner and the detected obstacles in the bottom part. It shows the resolution of the system, which is capable of distinguishing between the ground plane and objects as low as 1cm above the ground at a distance up to 2m. The newspaper disappears as an obstacle as soon as it lays flat on the ground.

The background detection was tested on different types of floor. We modified the tilt angle of the camera in a range $45° < \Theta < 70°$ in $5°$ steps. The number of pixels that could not be removed correctly, lied at $0.6 \pm 0.01\%$ of the total number of pixels in the image. All these remaining pixels were wrong depth estimations of the stereo algorithm. In case of the white plane in the bottom image, no disparity values were obtained and a warning was generated, because the size of the resulting information gap without valid disparity values was larger than the specified maximum size.

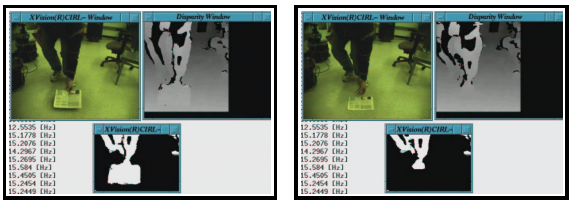

Figure 5. The newspaper is visible in the top, but it disappeared in the bottom image.

4.2 Quality of the Foreground Segmentation

The algorithm was applied in a variety of scenes and generated reliable results in all situations, where the scene contained enough texture for the stereo reconstruction algorithm. A few examples are shown in the images (Fig. 6). The typical resolution of the system is 1cm above the ground in the area covered by the sensor. All obstacles and gaps are reliably detected and avoided in the path planning algorithm that utilizes the output of the presented processing.

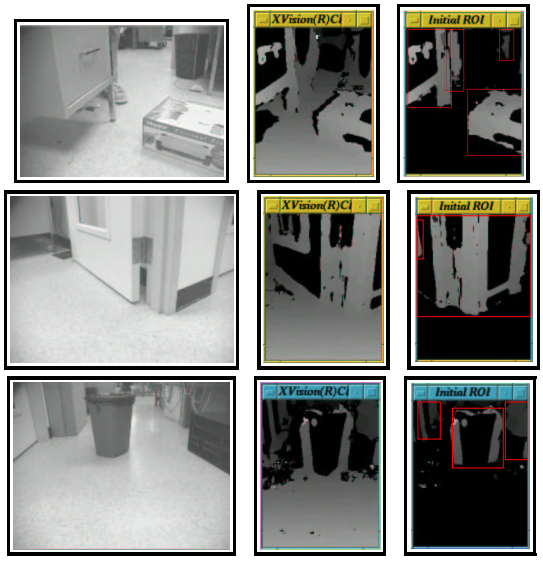

Figure 6. Examples of obstacle detection in different scenes.

5. Conclusions and Future Work

We have presented a system that is capable of automatic selection of interesting targets in the local area of a robot. It allows on-line re-calibration of extrinsic sensor parameters and additionally it allows to localize the system relative to *background* objects in the scene.

The system is used on our mobile robot for automatic target selection and following based on the perception of a binocular stereo sensor. The shortness of this paper allowed only a coarse review of the applied algorithms that are used for dynamical composition of tracking primitives dependent on the current environment structure. The classification of the scene helps to decide, which objects are interesting and should be monitored as well as which *behavior* is appropriate depending on the current scene structure. The robot can switch from wall following in hallway environments to localization based on corner structures, etc.

We want to replace the region based search function for the background planes described in section 3 by a new method derived from computer rendering algorithms in the next version of our system.

ACKNOWLEDGMENTS

This work was supported by the DARPA MARS program.

References

[1] O. Faugeras. *Three-Dimensional Computer Vision.* Massachusetts Institute of Technology, The MIT Press, Cambridge, Massachusetts London, England, 1993.

[2] F.Ferrari, E. Grosso, G. Sandini, and M. Magrassi. A stereo vision system for real-time obstacle avoidance in unknown environment. In Proc. of IEEE International Workshop on Intelligent Robots and Systems IROS'90, pages 703–708, 1990.

[3] M.A. Fischler and B.C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Comm. ACM*, 24(6):381–395, 1981.

[4] A. Hauck and N. O. Stffler. A Hierarchical World Model with Sensor- and Task-Specific Features. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS'96)*, pages 1614–1621, 1996.

[5] J. Mayhew, Y. Zheng, and S. Cornell. The adaptive control of a four-degrees-of-freedom stereo camera head. In *Natural and Artificial Low-level Seeing Systems, The Royal Society, London*, pages 63–74, 1992.

[6] C. Schmid, R. Mohr, and C. Bauckhage. Comparing and evaluating interest points. In *In Proc. of International Conference on Computer Vision, Bombay, January*, 1998.

[7] S. Se and M. Brady. Vision-based detection of kerbs and steps. In *Eighth British Machine Vision Conference BMVC '97*, pages 410–419, 1997.

[8] S. Simhon and G. Dudek. Selecting targets for local reference frames. In Proc. IEEE Int. Conf. on Robotics and Automation, pages 2840–2845, 1998.