The 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI 2011)
Nov. 23-26, 2011 in Songdo ConventiA, Incheon, Korea

Range Image Analysis for Controlling an Adaptive 3D Camera

Peter Einramhof[1], Robert Schwarz[2] and Markus Vincze[3]
[1,2,3] Automation and Control Institute, Vienna University of Technology,
Gusshausstrasse 27-29, 1040 Vienna, Austria
(Tel : +43-1-58801-376663; E-mail: einramhof@acin.tuwien.ac.at)

Abstract – Human vision is *the* reference when designing perception systems for cognitive service robots, especially its ability to quickly identify task-relevant regions in a scene and to foveate on these regions. An adaptive 3D camera currently under development aims at mimicking these properties for endowing service robots with a higher level of perception and interaction capabilities with respect to everyday objects and environments. A scene is coarsely scanned and analyzed. Based on the result of analysis and the task, relevant regions within the scene are identified and data acquisition is concentrated on details of interest allowing for higher resolution 3D sampling of these details. To set the stage we first briefly describe the sensor hardware and focus then on the analysis of range images captured by the hardware. Two approaches – one based on saliency maps and the other on range image segmentation – and preliminary results are presented.

Keywords – Adaptive Camera, Foveation, Range Images.

1. Introduction

To be safe and useful, service robots must be able to perform a variety of tasks: obstacle avoidance and self-localization in dynamic, uncontrolled and cluttered environments, detecting and manipulating or transporting objects, and human-robot interaction. For each of these tasks the robot's perception system needs to be supplied with appropriate sensor data. Depending on the task, the required measurement range, frame rate, resolution and field of view of the sensor data strongly varies. Due to the lack of adaptability, no single commercial sensor currently available is sufficient, which makes using combinations of such sensors necessary. Also, none of the current sensors directly supports higher-level perception.

In [1] Thielemann et al. introduce a novel 3D camera that is currently being developed and that aims at addressing these challenges. It comprises micro-mechanical scanning elements for two-dimensional beam-steering of fast, laser-based single-point time-of-flight distance measurement, and software for fast scene analysis whose output is used in a feedback loop to control data acquisition by the hardware. The time-of-flight hardware emits a pulsed laser beam that is moved across the scene by controllable micro-mirrors to create raster-scan range images comparable to those of conventional tilting laser scanners, however, at camera-like frame rates. The use of novel quasistatic MEMS scanning mirrors [2, 3] enables the system to rapidly control the beam direction and thus adjusting the spatial and temporal resolution of the acquired data, that is, to foveate. Inspired by visual attention systems [4], features are extracted from range images and combined into saliency maps. The latter indicate the relative importance of each region in the range image. They are the basis for computing a new plan for the mirror trajectory, where the most salient parts of the range image (by binarisation of the saliency map i.e winner take all) are sampled at higher spatial density.

2. Approach

There are two kinds of features to derive saliency maps from. Firstly, dynamic features: changes (or motion) are detected and the sensor foveates on regions with the greatest change. To achieve that, a number of consecutive frames are considered; this aspect has been elaborated in [5]. And secondly, static features: each frame is treated as originating from a static scene, and based on the geometry of the scene and the task, regions of interest are determined. In classic attention systems according to Itti and Koch [6] saliency maps are computed as linear combinations of feature activation maps followed by a winner-take-all step to determine where to foveate. The features used are color, intensity and edge orientations. Instead of color the sensor under development provides range data, which encode the three-dimensional structure of the scene. The static features we use are step and roof edges, smooth and planar patches, where horizontal and vertical planar patches play a special role in the context of robotics. The direction of the vertical axis (direction of gravity) is derived from the known geometry and kinematics of the setup consisting of sensor and robot, and optionally via inclinometers.

Horizontal planar structures such as the ground or table tops play the role of support planes on which the robot moves and on which obstacles or objects are located. *Vertical planar structures* originate from walls, the faces of closets or the bodies of objects. Especially walls define the boundaries of the indoor environment and serve as features for the robot's self-localization. *Step edges* occur at object boundaries, at the transition between foreground and background. We consider step edges as part of fore-ground objects. *Roof edges* occur at the transition between parts of objects or between object and support plane (concave), and also at top rims (convex) and high-curvature surfaces.

Before extracting static features, mixed pixels at range discontinuities are detected and removed (Fig. 1). For that we adapt the bearing angle proposed by Harati et al. [7]: $max\left(|\cos BA_{i,j}^{hor}|, |\cos BA_{i,j}^{ver}|\right)$ is thresholded and only

accepted if the local standard deviation within a 3x3 neighborhood of the range values lies above the noise level. This is followed by single outlier and noise reduction (median, Gauss). Using a second order derivative on the preprocessed range image, and histograms, step and roof edges are determined. The remaining pixels of the image either belong to planar or at least smooth surface patches. Planar patches are determined via split-and-merge based region growing. Figure 2, top row, shows range images of three scenes and the mid row shows their labeling based on the extracted features.

We compute saliency maps (Fig. 2, bottom row) in the fashion of Itti and Koch [6] from step and roof edges. Instead of using only linear combinations, also products of feature activation maps or range and height maps serve to incorporate constraints or biases. Task-dependent configurations select the features and their combination. The most salient region decides where to foveate. By inhibition of return the less salient regions are successively visited, too. If too much or nothing of the scene is salient, the attention system underperforms and no foveation is done.

Instead of using saliency maps, a more "targeted" approach based on "classic" scene segmentation is investigated, too. In the real world, a robot interacts with objects. Such objects are located on horizontal or vertical support planes (e.g. cup on table or handle on door). Thus, possible support planes are searched in the labeled range image and removed, and object candidates are determined in the remaining data. This yields a "binarized" saliency map, where all object candidates are (equally) salient. The sequence of foveating on the object candidates can be either random or derived from the task, e.g. closest object first in the case of grasping or obstacle avoidance.

3. Preliminary Results and Outlook

As sensor hardware and software are being developed in parallel, test data was generated using a conventional tilting laser scanner (SICK LMS100 mounted onto a rotary axis). 14 sequences with a total of over 2,000 scans were taken in a home-like office environment at a resolution of 360x250 pixels, covering a field of view of 90°(H)x60°(V). The software (C++) was tested on (one core of) an Intel Core i5-430M notebook. The average runtime is about 46ms per frame. Preliminary results for labeled range images and saliency maps based on the extracted features are shown in Fig. 2.

One of the main aspects currently under investigation is how to formulate and incorporate (high-level) task knowledge, e.g. in the form of suitable constraints. Furthermore, improvement of the extracted features' quality and combination as well as speedups will be addressed.

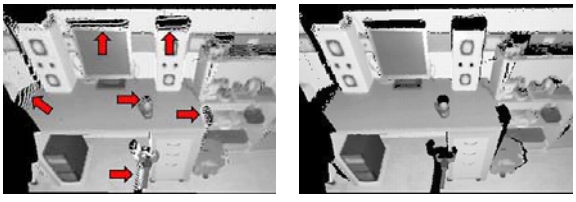

Fig. 1. Removal of mixed pixels

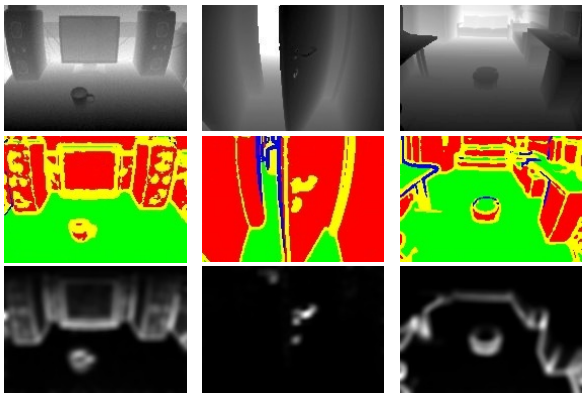

Fig. 2. *Top row*: Range images: cup on table, door handle, obstacle on floor. *Mid row*: Labeled range images. Green and red: horizontal and vertical planar patches, blue: step edges, yellow: roof edges. *Bottom row*: Saliency maps

Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°248623.

References

[1] J.T. Thielemann, T. Sandner, S. Schwarzer, U. Cupcic, H. Schumann-Olsen and T. Kirkhus, "TACO: A Three-dimensional Camera with Object Detection and Foveation", Smarter sensors, easier processing – SAB 2010 workshops, Paris, France, August 24, 2010.

[2] D. Jung, D. Kallweit, T. Sandner, H. Conrad, H. Schenk and H. Lakner, "Fabrication of 3D comb drive microscanners by mechanically induced permanent displacement", Proceedings of SPIE, volume 7208 of MOEMS and Miniaturized Systems VIII, page 72080A, San Jose, CA, USA, January 27, 2009.

[3] T. Sandner, T. Grasshoff, M. Wildenhain and H. Schenk, "Synchronized micro scanner array for large aperature receiver optics of LIDAR systems", Proceedings of SPIE, volume 7594 of MOEMS and Miniaturized Systems IX, pages 75940C112, San Francisco, California, USA, January 25, 2010.

[4] S. Frintrop, E. Rome and H. Christensen, "Computational visual attention systems and their cognitive foundations: A survey", ACM Trans. Appl. Percept., 7:6:1–6:39, January 2010.

[5] G.M. Breivik, J.T. Thielemann, A. Berge, Ø. Skotheim and T. Kirkhus, "A Motion based Real-time Foveation Control Loop for Rapid and Relevant 3D Laser Scanning", The Seventh IEEE Workshop on Embedded Computer Vision, ECVW 2011, Colorado Springs, CO, USA, June 20, 2011.

[6] L. Itti and C. Koch, "Computational modelling of visual attention", Nat. Rev. Neurosci, 2(3):194–203, March 2001.

[7] A. Harati, S. Gächter and R. Siegwart, "Fast Range Image Segmentation for Indoor 3D-SLAM", 6th IFAC Symposium on Intelligent Autonomous Vehicles, IAV 2007, Toulouse, France, September 3–5, 2007.