Proceedings of the RAAD 2011
20th International Workshop on Robotics in Alpe-Adria-Danube Region
October 5-7, 2011, Brno, Czech Republic

Real-time foveation system based on dense 2.5D data

Robert Schwarz[1], Peter Einramhof[1], and Markus Vincze[1]

[1] *Automation and Control Institute*
Vienna University of Technology
Gusshausstr. 27-29, 1040 Vienna, Austria
{schwarz, einramhof, vincze}@acin.tuwien.ac.at

Abstract. This paper presents a real-time task-based foveation system using data from 2.5D sensors that have recently gained increasing popularity in robotics. Typical tasks in this field are navigation, object detection and grasping.
Compared to the "classic" approach of using only 2D laser scanners, 2.5D sensors provide a much higher amount of data resulting in increased computational complexity. To address this challenge, our system uses the concept of foveation (inspired by the human attention mechanism) to focus only on regions relevant to the task and reduce the density of the data points outside these regions, e.g. for grasping a cup, the points of the cup are more important than those of the table.
Motivated by Itti et al. single feature activation maps are calculated and combined into a saliency map (see Fig. 1). In contrast to their approach to use 2D features (orientation, intensity and color) we use 3D cues like jump edges, roof edges and planar patches as features. The foveation system adjusts the density of the data points based on the saliency map. Additionally, locally dominant features are used to segment the range image.
Keywords. 3D sensors, visual attention, robot perception

1. Introduction

Recent research indicates a near exponential growing of robotics, the expected market size in 2025 of robots in the home environment will be twice the size as in classic manufacturing industry[1]. For mobile robots the shift from organized, well-known environments to private homes is challenging. Besides the requirements to the embodiment (hardware), the perception system needs to cope with the less structured environment. While in the industrial field of application the tasks e.g. localization and obstacle avoidance can be solved with a 2D laser scanner, the tasks in a home environment are more challenging because of cluttered scenes, formless surfaces like curtains, and protruding surfaces like table tops – in short, the three-dimensional character has to be taken into account.
To deal with this challenge, robots must be able to perceive the surroundings accordingly. 3D sensors (stereo vision, structured light, TOF-cameras) provide us with the required (range) data. In comparison to 2D laser scanners, 3D sensors supply robots with high amounts of data, but the problem lies in this very aspect: the processing of the whole data can be computationally expensive, especially for high resolution sensors. An idea to overcome this problem is the concept of attention, inspired by nature where living creatures face also the problem of high amounts of sensory data and limited processing capacities. Using only the "relevant" aspects of the input reduces the requirements to the brain. The same mechanism can be applied to sensors for mobile robots.
In that context Thielemann et al. (Thielemann et al. 2010) have introduced a new 3D sensor concept of the TACO project[2]. The sensor incorporates a single-beam laser range measurement unit, micro-mirrors which enable to deflect the beam to interesting regions, and an attention mechanism. Like the human eye which foveates on objects of interest, the sensor scans the whole scene with a constant resolution, decides on these data which regions are interesting for a given task, and acquires these regions in the next step at higher resolution, that is, it foveates. The

[1] Source: Japan Robotics Association (www.jara.jp)

[2] http://www.taco-project.eu/

laser measurement unit is based on the time-of-flight principle; the expected sampling rate is one million samples per second. Each sample consists of the measured range and the mirror deflection angles, and represents one point in 3D.
The major contribution of this paper is the foveation system for such a sensor. The system is able to analyze unfoveated data from the sensor and calculate the task-dependent interest regions. The result is used to control the trajectory of the mirrors and to get higher resolutions in these regions. The attention system will work in the mirror control loop for the sensor which is under development.
The approach can also be used for existing high resolution 3D sensors as a part of the processing chain: the high resolution 3D data is sub-sampled, subjected to fast image analysis and high resolution data is only kept in regions of interest.

The rest of the paper is organized as follows: section 2 gives an overview of computational visual attention systems inspired by the human vision. In section 3 we introduce our attention system and discuss each processing step in detail. Section 4 shows how representative test data was recorded, since the actual sensor is under development. Results of the attention system are shown in section 5, followed by the conclusion and an outlook in section 6.

2. Related Work

This section is divided into two parts. The first part gives an overview of visual attention systems; the second part presents existing 3D sensors that might provide input data to our attention system.

The cognitive process of selectively concentrating on one aspect of the environment while ignoring others is called attention. In nature evolution has favored the concepts of selective attention to overcome the problem of high amounts of sensory input and limited processing capabilities. One example of attention is the so-called cocktail-party problem: in a room full of different voices it is possible for humans to focus on one certain person and to follow the conversation. The similar concept exists also for the visual sense and is an intensely studied topic within psychology and cognitive science – but it also made entrance into computer vision and robotics.
Inspired by the human attention system, Itti et al. (Itti et al., 1998) proposed an attention system model with the ability to simulate the pre-attention of humans. The system attempts to predict for a given scene which areas of the image will draw attention. The input image is decomposed into a set of feature maps which represent the local appearance of color, intensity and orientation discontinuities. The model combines these feature maps into a single saliency map that highlights regions of interest.
In a winner-take-all (WTA) manner the most salient region draws the focus of attention towards it, by subsequent inhibition the path of the focus is determined.
Besides the three features Itti et al. use, Frintrop et al. (Frintrop et al., 2010) and Wolfe (Wolfe, 1994) give clues that these three are only a selection of useful cues. The third dimension can be used as a depth cue to guide and modulate the deployment of attention. Depth data from various sensor modalities has been used as input for attention systems: for example (Maki et al. 2000), (Bruce et al. 2005) and (Björkman et al. 2006) use stereo cameras. Frintrop et al. use depth as well as reflectivity data from a 3D laser scanner (Frintrop et al., 2010) and Ouerhani et al. use 3D cameras (Ouerhani and Hügli, 2000).
The classic approach of combining single feature maps to a saliency map does not take the task into account, only simulates the pre-attentive, stimulus-driven attention system. The counterpart to the bottom-up approach is to incorporate the task in a top-down manner. Wolfe´s model considers information of the goal by selecting features with high differences between the target and the rest of the scene; only features that are useful for the task are considered.
Top-down information can be incorporated in various ways: searching for salient regions can be restricted to certain regions, e.g., the street when searching for persons, but ignore the sky (Torralba, 2003a). The gist (semantic category) of a scene such as “office scene” or “street” guides eye movements (Torralba, 2003b) and can be computed from the feature channels (Siagian and Itti, 2009). If prior knowledge about a target is to be used to perform visual search, the target similarity of the most salient regions in bottom-up saliency maps can be investigated (Rao et al., 2002). More advanced approaches are biasing the feature types (Navalpakkam and Itti, 2006) or tuning conspicuity maps (Hamker, 2005). Other approaches inhibit target irrelevant regions (Choi et al., 2004) or excite target-relevant regions (Hamker, 2005) or use both (Navalpakkam and Itti, 2005). In order to imitate human-like behaviour, bottom-up saliency (uniqueness) and top-down saliency (target relevance) have to be fused (Rasolzadeh et al., 2010).

Nowadays there are four different sensor systems available which provide range data: inspired by the human vision, *stereo vision systems* are working on the same principle as the human depth perception; images from two cameras with a different viewpoint of the same scene are combined into a disparity map, which represents the depth information. The calculation can be done on embedded-systems, on the CPU or on the GPU, depending on the actual stereo system. In areas where no correspondence can be determined between the two images, no valid range data can be calculated and holes occur. Due to its

passive sensing the system depends on the lighting conditions in the environment.
In contrast to passive stereo vision systems, structured *light based systems* replace the second camera by a projector that projects a known light pattern. The depth information is calculated from the distortion of this pattern due to the 3D structure of the scene. One novel representative of this principle is the Kinect[3], which provides camera frame rates.
Since camera and projector in structured light systems and both cameras in stereo systems view the scene from slightly different viewpoints, regions without valid data can result due to occlusion.
The *time-of-flight cameras* emit modulated light and measure the time it takes for the reflected light to return to the sensor. These systems are vulnerable to background lightning and interference with other sensors of the same kind.
In comparison to these three concepts, the *tilting laser scanner* measures only one data point at a time; a conventional 2D laser scanner is mounted on a tilting unit to introduce the third degree of freedom. Due to the sequential measurement and the relatively slow speed, the video frame rate is low compared to the three other 3D sensor concepts.
Similar to the tilting laser scanner, the TACO project develops a 3D sensor system which uses also one laser beam, but in contrast to the tilting laser scanner the deflection of the beam is achieved by micro-mirrors instead of a rotating (macro) mirror and tilting unit (Surmann et al., 2001). The expected output of the sensor will be one million samples per second, which is equivalent to a resolution of 250x160 at a frame rate of 25 Hz. While this is the default resolution of the sensor, higher resolutions can be provided at lower frame rate and vice versa.

3. Approach

In this section we give an overview of our approach of using range data for a foveation system and describe the particular steps of the processing chain in detail.

3.1. Overview

As mentioned in section 2, Itti et al.'s original architecture of bottom-up visual attention (Itti et al., 1998) is based on the idea that an input image is decomposed into a set of feature maps (colors, intensities and orientations) which are then combined into a so-called saliency map.
Instead of using 2D images, our work focuses on range images, so we adopt the principal architecture of Itti and instead of using 2D features we replace them with 3D features derived from the range image.
Fig. 1 shows the structure of our foveation system, the input is a range image which is equivalent to a structured point cloud that maintains the initial neighborhood information. Depending on the sensor model the data is filtered to reduce the noise and to deal with outliers.
After filtering individual feature activation maps are calculated. Each feature activation map provides the location of the respective feature within the scene. The resolution of the map is the same as the one of the input range image; the values are normalized to [0…1].
The feature maps are then combined into a saliency map that has again the same resolution as the feature maps. The combination of the feature maps is influenced by the task; only task-relevant feature maps which model the task-condition, are considered and used for the calculation of the saliency map.
Based on the saliency map the foveation system focuses only on regions of interest and outputs the foveated point cloud.
Additionally, the comparison of the feature activation maps is used to create a "label map" in which each pixel represents the locally dominant feature.
Aside the 3D features another aspect is different to Itti's architecture; while his system is a task-independent model of the visual attention, we incorporate the task and the context of the scene, which influences the selection and the combination of the features.

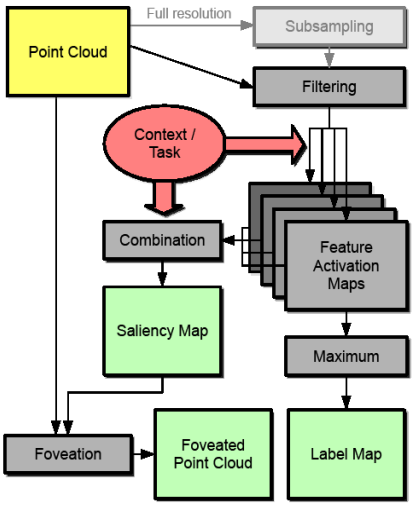

Fig. 1. Overview of the attention system concept; input data (yellow), processing chain (grey), output (green), context (red)

3.2. Noise reduction

The original range image is available in the form of three 2D arrays corresponding to the range and the two mirror deflection angles of the measurements.
Before the feature extraction can be done, filtering of the sensor data is necessary. Using the advantage of the structured data in the range image (neighborhood of each pixel is known) filter mechanism similar to that used for 2D images can be applied. A Gaussian filter with a 3×3 kernel is used to reduce the Gaussian

[3] www.xbox.com/kinect

noise, additionally the adaptive Wiener filter (Jain, 1989) is applied, which is based on statistics estimated from a local 3×3 neighborhood of each pixel.

3.3. Feature extraction

After filtering the input range image is decomposed into different feature activation maps, which represent the appearance of a feature; e.g. a feature activation map of the feature "color red" shows only response to red areas of the image. Since we are dealing with range images instead of RGB images, the classic features (color, orientation and intensity) are not feasible.

For our attention system we consider features relevant to a given task. For grasping (a cup on a table) the interesting parts are objects on a support plane; this means transitions from vertical to horizontal surfaces as well as jumps in the depth values indicate object boundaries. Additionally the geometry of a scene can be used for further cues, e.g. grasping objects defines a specific height range which is covered by the gripper, or for obstacle avoidance only the nearest objects are interesting. Features designated by these aspects are jump and roof edges (object boundaries), vertical and horizontal surfaces (object properties), and distance and height ranges (geometry of the scene).

In the following we describe the individual features that we have implemented:

Jump edges (Fig. 2c) appear in scenes where an object is occluded by another object or itself; discontinuities in the depth values occur at the object boundaries. For the calculation of the jump edges the neighborhood of the data points is useful, the difference of the left and the right neighbor can be used to detect jump edges in row direction, the same can be done for columns with the upper and lower neighbor. To combine both values the gradient magnitude is calculated (Eq. 1). In the following we used the notation of *i* and *j* as row- and column-index, *r* represents the range value.

$$jump_{i,j} = \left\| \begin{pmatrix} r_{i,j+1} - r_{i,j-1} \\ r_{i+1,j} - r_{i-1,j} \end{pmatrix} \right\| \quad (1)$$

Roof edges (Fig. 2d) do not represent discontinuities in the range value, but discontinuities in the direction of the normal vectors and appear for example two differently oriented surface patches intersect.

The first step is to calculate the surface normals which can be done in different ways (PCA, least-square fit on points in a defined distance (Rusu et al. 2008)). For our implementation we decided on an approach only using the direct neighboring data points in the data array. In comparison to the approach of using a neighborhood of a defined spatial extent, the benefit is that no threshold (e.g. 5cm radius) is needed to decide if a point is considered for the surface normal estimation or not. The downside is a greater sensitivity to noise; however, the latter was reduced in the filtering step.

For the calculation of the normals we use the cross product of the vectors $\overrightarrow{p_{i,j-1}p_{i,j+1}}$ and $\overrightarrow{p_{i-1,j}p_{i+1,j}}$, where *p* is the 3D data point *(x,y,z)* computed from the associated range data and mirror deflection angles. To reduce the noise we repeated this with the diagonal points and calculate the mean of both values (Eq. 2-3).

$$\vec{v}_{i,j} = mean(\overrightarrow{p_{i,j-1}p_{i,j+1}} \times \overrightarrow{p_{i-1,j}p_{i+1,j}}, \overrightarrow{p_{i-1,j-1}p_{i+1,j+1}} \times \overrightarrow{p_{i-1,j+1}p_{i+1,j-1}}) \quad (2)$$

$$\overrightarrow{normal}_{i,j} = \frac{\vec{v}_{i,j}}{\|\vec{v}_{i,j}\|} \quad (3)$$

The dot product of the normal vector and his right neighbor indicates discontinuities of the surface orientation in rows; the same applies to columns with the neighbor below. The gradient magnitude corresponds to the actual strength of roof edges (Eq. 4).

$$roof_{i,j} = \left\| \begin{pmatrix} \overrightarrow{normal}_{i,j} \cdot \overrightarrow{normal}_{i+1,j} \\ \overrightarrow{normal}_{i,j} \cdot \overrightarrow{normal}_{i,j+1} \end{pmatrix} \right\| \quad (4)$$

The classification in *horizontal and vertical surfaces* (Fig. 2d, 2f) makes only sense if the pose of the sensor is known. The pose can be estimated from previous frames (e.g. RANSAC) or from additional sensors (IMU). With the given pose the categorization into two main directions can be done by calculating the angular deviation (Eq. 5) of the normal vectors of the data points from the vertical direction $\overrightarrow{d_v}$.

$$\alpha_{i,j} = acos(\overrightarrow{normal}_{i,j}, \overrightarrow{d_v}) \quad (5)$$

Vertical surfaces have values of $\alpha_{i,j}$ close to 0°, horizontal ones close to 90°.

The *height* range (Fig. 2b) is also used as a feature for our attention system. For some tasks it is practical to inhibit irrelevant regions. For the task of searching cups on a table points far away from an expected height can be ignored, only points in the height range of the table top are considered. The notion of "height" only makes sense if the sensor pose is known.

The range (Fig. 2a) itself as a feature can be used to raise saliency on near object. For grasping, objects near the robot are more interesting than objects far away, for self localization the opposite applies.

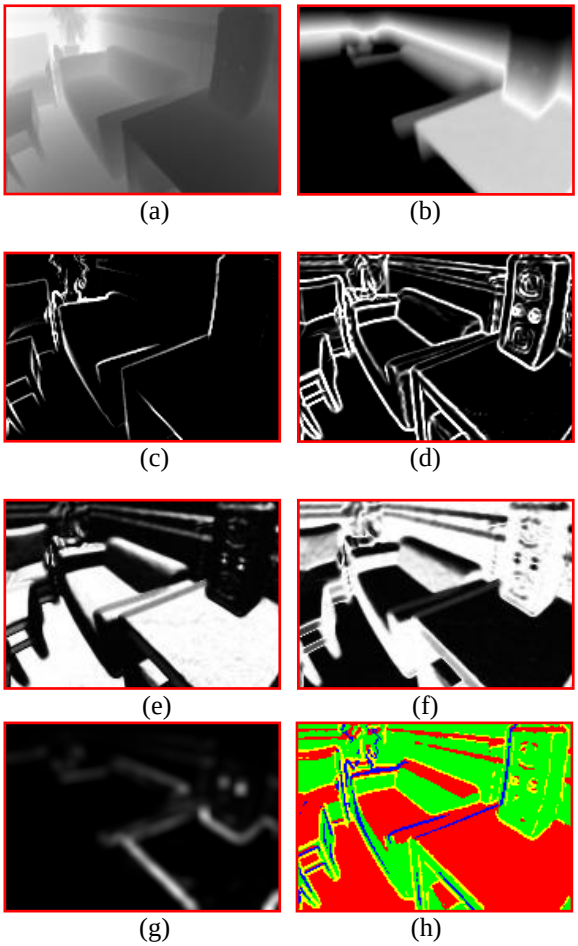

Fig. 2. Examples for the feature activation maps; (a) range image (b) height range feature (c) jump edges (d) roof edges (e) horizontal (f) vertical surface feature (g) saliency map (h) label map

3.3. Saliency calculation and label map

The classic approach for attention used by Itti uses only weighted summarization of the feature maps. Motivated by Wolfe´s model, which applies task-dependent selection of the feature maps and combines only task relevant features, our system uses a similar approach: The task influences the parameters of the height and range feature maps (e.g. expected height range) and only feature maps relevant to the task are combined into the saliency map (Fig. 2g). The selected feature maps are multiplied so that only regions are salient, where all feature maps show activation; each cell in the array of the saliency map is the result of the multiplication of the corresponding normalized values from the feature activation maps, and hence the values also lies in the range of [0…1].

Additionally, of the selected feature maps the corresponding pixel positions are compared. The locally dominant feature (at a pixel position) is the one with the highest activation. In case two or more feature activations have the same value, the following prioritization scheme is applied: jump edges before roof edges, then vertical surfaces and finally horizontal surfaces. The result is a label map (Fig. 2h), which separates the scene in segments with the same dominant feature.

4. Data acquisition

The expected output of the TACO sensor will look similar to the data of a tilting laser scanner; the principle of the measurement is the same, only the sampling rates are different. To emulate the TACO sensor we used a SICK LMS-100 which is mounted on a tilt unit (SCHUNK PW70). The sensor itself is mounted on top of a robot in a height of approximately 1.2m to allow capturing table scenes as well as providing data for navigation (see Fig. 3).

One scan takes about 20s, which is 500 times slower than the expected 25Hz of the TACO sensor. The resolution of the test data is 360×500 with a field of view of 90°×62.5° (horizontal×vertical).

Different scenarios were recorded (table top scenes, kitchen scene, navigation trough offices, grasping scenario), the data contains about 2000 frames.

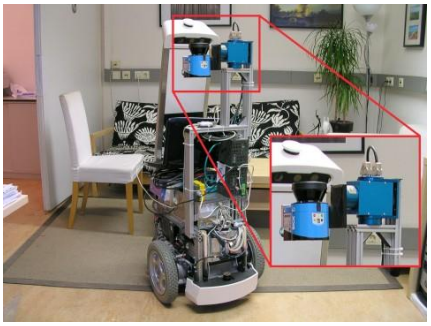

Fig. 3 Tilting laser scanner mounted on top of a mobile robot. This setup was used for data acquisition

5. Results

As input for our attention system the test data is sub-sampled to a resolution of 180×125; the higher resolution is used to simulate the foveation ability of the sensor and allows zooming in by the factor 2 horizontally and by the factor 4 vertically.

Fig. 4 shows detailed results for three different scenarios: objects on a table (left column), opening a door (middle column) and obstacle detection (right column). The final saliency maps are marking regions which are interesting for the task *grasping* (cup on a table and door handle); for the task *obstacle detection* the saliency map represents nearby objects on the floor, which are used for obstacle avoidance (right).

For the evaluation of the foveation mechanism we generated ground truth for seven sequences. In sequence #001 - #003 an obstacle on the floor is marked (cf. Fig. 4 right column), in the rest of the sequences objects on different tables are used as reference. Tab. 1 shows the mean and the standard

deviation of saliency values on the object vs. outside the object. The saliency is up to 4 times higher on the objects than on the rest of the scene.
The time consumption of the attention system is about 60ms, which correlates to a frame rate of 16Hz.

		saliency on object		saliency outside object		time consumption [s]	
Seq.	Frames	*mean*	*std*	*mean*	*std*	*mean*	*std*
#001	90	0,105	0,012	0,042	0,000	0,064	0,003
#002	67	0,104	0,014	0,032	0,002	0,065	0,004
#003	60	0,101	0,011	0,034	0,001	0,063	0,002
#004	105	0,218	0,015	0,067	0,003	0,064	0,003
#005	80	0,264	0,006	0,068	0,004	0,064	0,003
#006	70	0,299	0,005	0,086	0,002	0,065	0,004
#007	60	0,133	0,005	0,093	0,001	0,066	0,004

Tab. 1 Mean and standard deviation of saliency on ground truth object and outside the object; Time consumption on Intel Quad Core @ 2.4GHz

6. Conclusion and outlook

In this paper we presented a foveation system based on dense 2.5D range data. Our main focus hereby was on demonstrating how to combine features derived only from 3D data for computing meaningful saliency maps for tasks from service robotics such as detecting objects on a table or obstacles. The resulting saliency maps can be used to control the TACO sensor to foveate on regions of interest and furthermore to provide higher resolution.
Future work includes improving real-time performance through C++ implementation. Also additional features will be implemented and the combination of the feature maps will get more complex as presented here.

7. Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 248623.

8. References

Björkman M. and Eklundh J. 2006. Vision in the real world: Finding, attending and recognizing objects. *International Journal of Imaging Systems and Technology*, 16:189–208

Bruce N. and Tsotsos J. 2005. An attentional framework for stereo vision. *Computer and Robot Vision*, 0:88–95

Choi S., Ban S. and Lee M. 2004. Biologically motivated visual attention system using bottom-up saliency map and topdown inhibition. *Neural Information Processing-Letters and Review*

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

Hamker F. 2005. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *Computer Vision and Image Understanding, 100(1-2):64–106, Special Issue on Attention and Performance in Computer Vision*.

Itti L., Koch C. and Niebur E. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 11, pp. 1254-1259

Jain A.K. 1989. Fundamentals of digital image processing. Prentice-*Hall, Inc. Upper Saddle River*

Maki A., Nordlund P. and Eklundh J. 2000. Attentional scene segmentation: Integrating depth and motion from phase. *Computer Vision and Image Understanding*, 78:351–373

Navalpakkam V. and Itti L. 2005. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231

Navalpakkam V. and Itti L. 2006. An integrated model of top-down and bottom-up attention for optimizing detection speed. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2 of CVPR '06, pp 2049–2056

Ouerhani N. and Hügli H. 2000. Computing visual attention from scene depth. *International Conference on Pattern Recognition*, Vol. 1, pages 375–378.

Rao R., Zelinsky G., Hayhoe M. and Ballard D. 2002. Eye movements in iconic visual search. *Vision Research*, 42(11):1447–1463

Rasolzadeh B., Björkman M., Huebner K. and Kragic D. 2010. An active vision system for detecting, fixating and manipulating objects in the real world. *Int. J. Rob. Res*., 29:133–154

Rusu R.B, Marton Z.C., Blodow, N., Dolha, M. and Beetz, M. 2008. Towards 3D Point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56:11:927-941

Siagian C. and Itti L. 2009. Biologically inspired mobile robot vision localization. *Trans. Rob*., 25:861–873

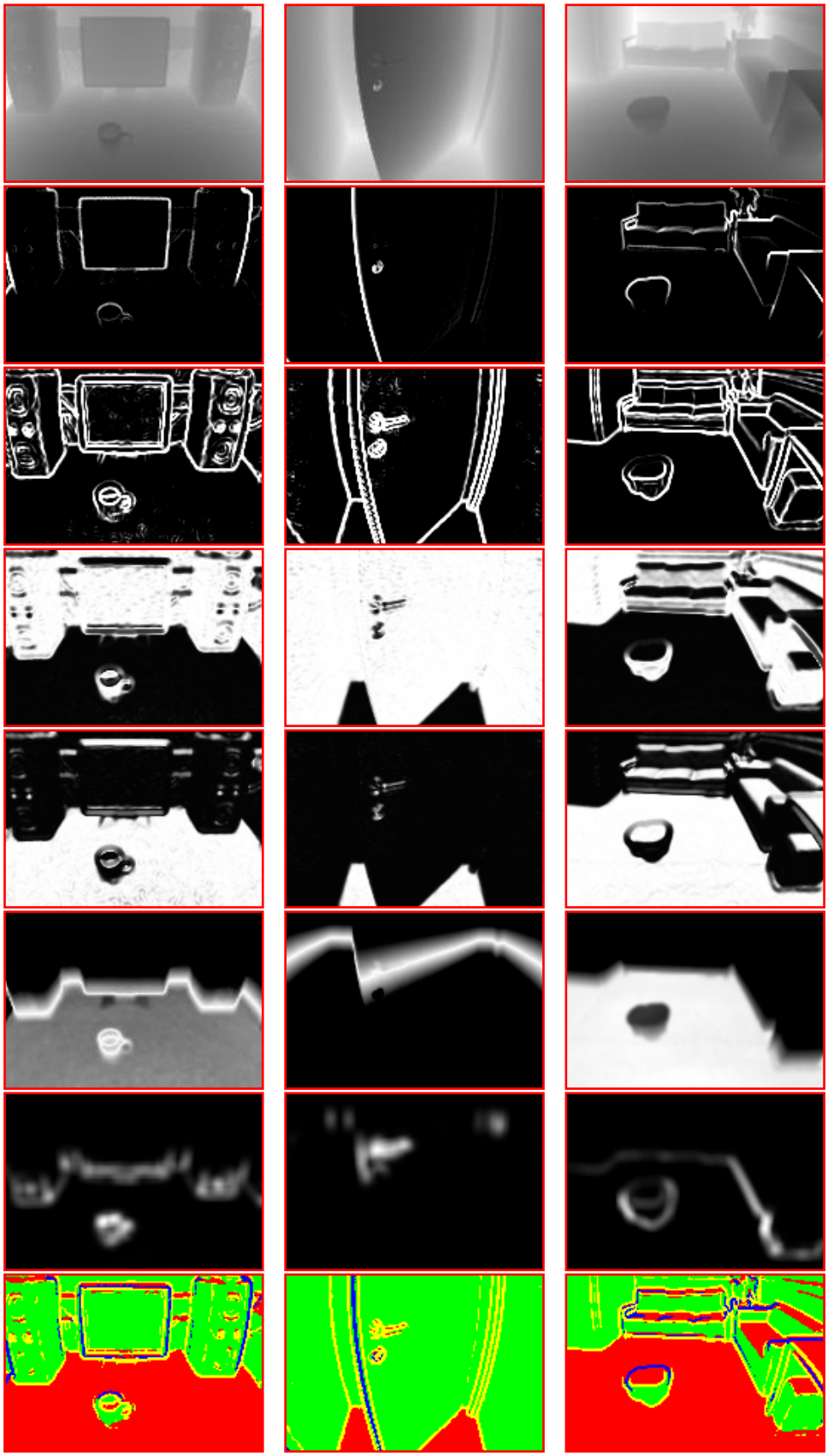

Fig. 4 Results of the attention system from three different scenes: objects on a table (first column), door handle (second column), obstacles on the ground floor (third column). The rows are showing following feature activation maps (top to bottom): range image, jump edge, roof edge, vertical surfaces, horizontal surface, height range (different for each scene), saliency map and label map.

Surmann H., Lingemann K., Nüchter A. and Hertzberg J. 2001. A 3d laser range finder for autonomous mobile robots. *International Symposium om Robotics (ISR)*, Vol. 1, pp 153–158

Thielemann J., Sandner T., Schwarzer S., Cupcic U.,Schumann-Olsen H., and Kirkhus T. 2010. TACO: A three-dimensional camera with object detection and foveation. *SAB Workshops*

Torralba A. 2003a. Modeling global scene factors in attention. *JOSA - A*, 20:1407–1418

Torralba A. 2003b. Contextual priming for object detection. *IJCV*, 53:169–191

Wolfe J. 1994. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238