17^{th} *Computer Vision Winter Workshop*
Matej Kristan, Rok Mandeljc, Luka Čehovin (eds.)
Mala Nedelja, Slovenia, February 1-3, 2012

Bottom-Up Visual Attention System based on Reflectivity and Range Images

Robert Schwarz, Peter Einramhof and Markus Vincze
Automation and Control Institute (ACIN), Vienna University of Technology
Gusshausstrasse 27 - 29 / E376, A - 1040 Vienna, Austria
{rs,pe,vm}@acin.tuwien.ac.at

Abstract. *This paper presents a stimuli-driven attention system based on reflectivity and range data provided by a 3D range sensor.*

In comparison to the common approach of using 2D color images for the attention system, we incorporate 3D features derived from the range data to increase the influence of the scene structure like object boundaries and transitions. Our architecture is based on three stages, the input is decomposed into feature maps, which result in activation maps, and then those are combined into a saliency map.

We show the benefit of jump and roof edges in addition to orientation and intensity features and evaluate different combinations of the features on real data captured by a tilting laser scanner.

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 this paper we investigate how an attention mechanism can benefit from additional 3D information. The common approach of decomposing the input data in different features and combining them into a useful saliency map is extended. Additional 3D edge features derived from the range data are used to incorporate information about object boundaries and transitions, so that the structure of the scene effects the attention system. Input is provided by a tilting laser scanner, which yields range as well as reflectivity (also known as reflectance) data in a single 3D scan pass. Both data modalities are transformed into 2D images and fed into our visual attention system. In the range and in the reflectivity image, the system detects regions that are salient according to intensity and orientations as well as regions that are salient to jump and roof edges.

The rest of the paper is organized as follows: Section 2 presents an overview of different attention approaches and 3D sensors. In Section 3 our attention system based on 3D data is presented, and in Section 4 the beneficial influence of 3D edges is evaluated.

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

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.

2.1. Visual Attention

Attention is defined as "the concentration of awareness on some phenomenon to the exclusion of other stimuli" [2], visual attention in computer vision deals with modeling methods to focus on interesting regions in scenes, and how to determine what is interesting and not.

Most of the early research covers models for 2D color images, e.g. Itti et al.([10]) consider the local appearance of color, intensity and orientation. A good overview of other attention models can be found in Frintrop et al ([6]). Beside features based only on 2D information, Frintrop et al. and Wolfe ([23]) 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 Bjrkman and Eklundh [1] use hue and 3D size on stereo camera data, Bruce and Tsotsos [2] combine attention based on Gabor maps and spatial frequencies from the left and right stereo camera input, and Maki et al. [11] incorporate disparity and flow information in their work. Frintrop et al. use orientation and intensity cues on depth as well as on reflectivity data from a 3D laser scanner ([6]) and Ouerhani et al. use 3D cameras to integrate depth based on conspicuity maps([14]).

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. Wolfes 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 ([22]). The gist (semantic category) of a scene such as office scene or street guides eye movements ([21]) and can be computed from the feature channels ([19]). 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 ([15]). More advanced approaches are biasing the feature types ([13]) or tuning conspicuity maps ([7]). Other approaches inhibit target irrelevant regions ([4]) or excite target-relevant regions or use both ([12]). In order to imitate human-like behaviour, bottom-up saliency (uniqueness) and top-down saliency (target relevance) have to be fused ([16]).

2.2. Range Sensors

Today 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 from different viewpoints of the same scene are combined into a disparity map, which represents the depth information. 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.

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 lighting 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, Thilemann et al. develop 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 micromirrors instead of a rotating (macro) mirror and tilting unit ([20]). The expected output of the sensor can be

[2] Encyclopaedia Britannica. Encyclopaedia Britannica Online. Encyclopaedia Britannica Inc., 2011. Web. 13 Dec. 2011. http://www.britannica.com

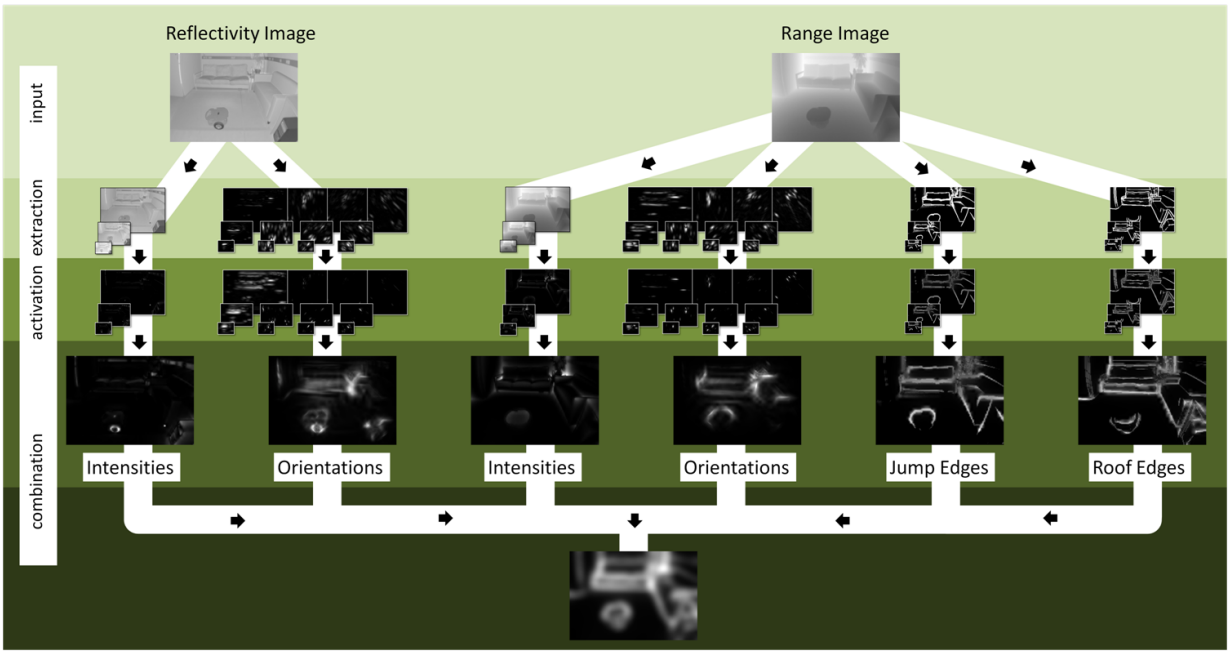


Figure 1. Overview of the bottom-up attention system, organized in main stages (horizontal) and in channels (vertical)

adjusted, higher resolutions can be provided at lower frame rate and vice versa.

With increasing resolution and frame rate of 3D sensors the amount of sensory output raises, and with it the computationally requirements to process the data. Attention mechanism can be used in a preprocessing step to overcome this problem.

3. Attention for Reflectivity and Range Images

The architecture of our bottom-up attention system is inspired by Itti et al.'s attention model ([10]), which decomposes the input image in different features and combines them into a saliency map. Wolfe [23] and Frintrop et al. [5] notice that beside Itti's selection of features (color, intensity, and orientation) there are more cues which influence attention, one of them is depth information. In our approach we combine 2D features from the reflectivity image with 3D features (particularly jump and roof edges) from the range image. The idea is to amplify saliency on object boundaries and transitions, so that the scene structure becomes more influentially. The input data is provided by a tilting laser scanner, which yields those reflectivity and range images.

Our system is shown in Figure 1 and is organized into three stages:

- **Extraction:** extract feature maps at locations over the input image
- **Activation:** form activation maps using the feature maps
- **Combination:** first combine activation maps for each feature and then combine these into a saliency map

The input is divided into similar processing chains for the reflectivity and range image. Each processing chain contains the stages mentioned before. The following channels (processing chains) are used in our system:

- Intensity channel on reflectivity image
- Orientation channel on reflectivity image
- Intensity channel on range image
- Orientation channel on range image
- Jump edge channel
- Roof edge channel

The reflectivity image is decomposed into two different channels: the intensity channel and the orientation channel. The first stage for the intensity channel is to **extract** the feature maps from the input image. This step is done on different scales on a Gaussian image pyramid (5x5 Gauss kernel and 4 scales, each subsampled by selecting every second pixel vertically and horizontally). In the intensity channel the feature maps for each scale are equally to the input reflectivity image on the corresponding scale.

The next stage is to get **activation** maps from these feature maps. This is done with the use of center-surround mechanisms which compute the intensity differences between image regions and their surroundings. The center c is given by a pixel in the feature map. The surrounding s is calculated as the average of the surrounding pixels for two different sizes of surrounds. The value of the surrounding can be determined by averaging two convolutions of the feature map with Gaussian kernels with different σ (σ represents the size of the surrounding). The center-surround difference $d = |c - s|$ is a measure for the intensity contrast in the specified region. Three intensity feature maps from the scales $[2\ 3\ 4]$ yield to three activation maps in the intensity channel.

The **combination** of these three maps results in one activation map for this channel. This step corresponds to the first phase of the combination stage.

The second channel of the reflectivity image is the orientation. To extract the feature maps, Gabor filters are used to detect bar-like feature of orientations $\{0^\circ\ 45^\circ\ 90^\circ\ 135^\circ\}$ in different scales. The activation and the combination is done similar to the previous intensity channel. The four orientation maps from from each scale $[2\ 3\ 4]$ result into one activation map for this orientation channel.

Similar to the reflectivity image the range image is divided into channels, the first two channels are equal to the channels of the reflectivity image.

Besides intensity and orientation, two edge channels are added:

Jump edges occur on object boundaries and appear in scenes where an object is occluded by another object or itself. Discontinuities in the range image represent jump edges and can be determined by the gradient magnitude (using the Sobel approximation to the derivative). **Roof edges** do not represent discontinuities in the range value, but discontinuities in the direction of the surface normal vectors and appear for example, where two differently oriented surface patches intersect. The normals are calculated according to [17], and the maximum of the dot products (of the surface normals) of the horizontal and vertical neighbouring pixels can be used to determine the location of roof edges. Figure 2 shows both edge feature maps.

The activation and the first phase of the combination is done similar to the other channels. The three jump edge feature maps (scales $[2\ 3\ 4]$) yield to one activation map for the jump edge channel. The same goes for the roof edges.

(a) (b)

Figure 2. Edge Feature Extraction (from Figure 6(a)): (a) jump edges, (b) roof edges

Finally, the six activation maps from each channel are combined to the saliency map. One common approach is to sum up the activation maps after normalizing each activation map (c.f. [6]). In addition to this additive approach we propose a multiplicative approach: the sum of the edge activation maps is pixelwise multiplied with the sum of the rest of the activation maps to inhibit regions without any edge correspondence and exhibit regions corresponding to object transitions and boundaries.

In the section 4 the different saliency maps are evaluated.

4. Experiments

Common evaluations for attention systems are to compare the output saliencies of different approaches on the basis of example images, another evaluation method is to use ground truth of eye fixations. Since there are only data sets for saliency evaluations for 2D color images available, we use our recorded test data with the generated ground truth for objects to test the performance for the area of mobile robotics.

4.1. Data Acquisition

The test data was recorded with a tilting laser scanner, we used a SICK LMS-100 which is mounted on a tilt unit (SCHUNK PW70) (see Figure 3(a)). The resolution of the test data is 360×500 with a field of view of $90^\circ \times 62.5^\circ$ (horizontal $\times$ vertical).

The test data contains two different sequences:

- a home environment with a robot crossing the ground plane in front of the sensor (90 frames)
- the sensor is approaching a table with a cup on it (80 frames)

For both scenes groundtruth data was generated by marking the robot in the first sequence and the cup in

the second sequence (see Figure 3(b) and 3(c)). This represents objects relevant for mobile robotics, especially the tasks of navigation (and obstacle avoidance) and grasping objects.

(a) (b) (c)

Figure 3. (a) tilting laser scanner (b) and (c) groundtruth

4.2. Evaluation

For the evaluation of our attention system we considered two different perspectives:

- different combinations for our attention system
- our system in comparison to approaches mentioned in section 2

Figure 6 shows saliency maps from different combinations of our attention system: (a) shows the input range image, (e) the input range image. In the first row only combinations without edge channels are plotted: (b) combination of intensity and orientation channels of the reflectivity image, (c) intensity and orientation of the range image, (d) intensity and orientation from both images. In (i) the edge channel is presented, the additive combination of the edge channel for (b)-(d) is shown in (f)-(h), the multiplicative combination in (j)-(l).

The advantage of the edge channel is that the saliency is amplified in regions where object boundaries occur, with the result that the saliency map corresponds more to the actual structure of the scene.

A numerical evaluation is done with the ground truth data from the two test sequences, true-positive rates and false-positive rates are calculated based on a threshold for the binarization of the saliency map. $(binarymap \wedge groundtruth)$ results in true positive, $(binarymap\neg \wedge groundtruth)$ in false positive, and so on. The resulting average (over all frames of the sequence) ROC curves are shown in Figure 4 and 5.

In the first test sequence the reflectivity channels increase the performance, which can be explained that the object (robot) marked as ground truth has a high contrast in the reflectivity image. In the second sequence the object (cup) has much lower contrast to the surrounding, so that in this case the influence of the edge channels come into effect. Table 4.2 contains the area under ROC curve (AUC), higher values indicates better performance. In the first sequence the additive combination of the edge channel decreases the AUC more than the multiplicative combination. In both sequences the values of the additive combination are lower than of the multiplicative combination.

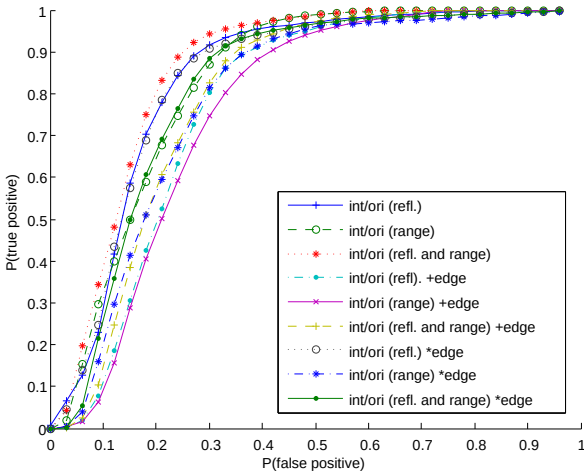


Figure 4. ROC-curve from test data sequence #1

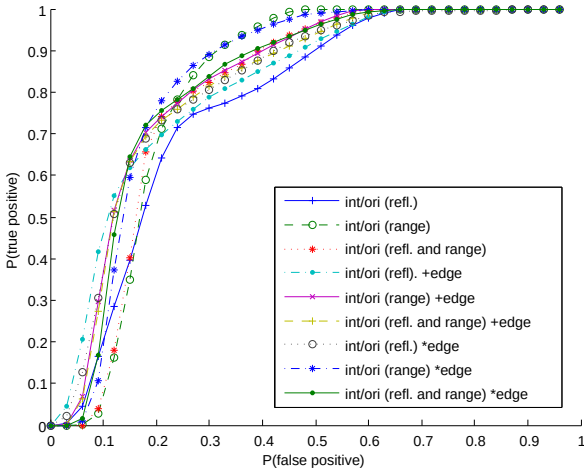


Figure 5. ROC-curve from test data sequence #2

The ROC curves and the AUC confirm a higher score for the multiplicative combination compared to the additive. The attention system benefits from the edge channels in sequences with low contrast in the reflectivity image.

Figure 7 shows saliency maps from other works for saliency calculation (Bruce and Tsotsos [3], Harel et al. [8], Hou and Zhang [9], Itti et al. [10], Seo and Milanfar [18],Zhang et al. [24]). Since those approaches are usually used on color images and our test data does not include color information, we di-

channels and combination	AUC #1	AUC #2
int/ori (refl.)	0.79964	0.73543
int/ori (range)	0.78842	0.76896
int/ori (refl. and range)	0.81989	0.75829
int/ori (refl). +edge	0.73571	0.77809
int/ori (range) +edge	0.72161	0.78723
int/ori (refl. and range) +edge	0.75410	0.77853
int/ori (refl.) *edge	0.79425	0.77951
int/ori (range) *edge	0.74798	0.79091
int/ori (refl. and range) *edge	0.77539	0.78189

Table 1. Area under ROC curve (AUC) for different channel combinations on both test sequences.

approaches	AUC #1	AUC #2	runtime [s]
our approach	0.77539	0.78189	0.45
Bruce refl. [3]	0.82228	0.76673	8.22
Bruce range [3]	0.68382	0.81224	8.22
Harel refl. [8]	0.85148	0.76474	18.54
Harel range [8]	0.76592	0.77800	18.54
Hou refl. [9]	0.52519	0.53869	0.02
Hou range [9]	0.32546	0.34732	0.02
Itti refl. [10]	0.66255	0.76524	0.32
Itti range [10]	0.82373	0.68932	0.32
Seo refl. [18]	0.83189	0.75686	2.42
Seo range [18]	0.75458	0.78468	2.42
Zhang refl. [24]	0.80433	0.66407	2.35
Zhang range [24]	0.61619	0.72008	2.35

Table 2. Area under ROC curve (AUC) on both test sequences with different approaches (each with reflectivity and range image as input)

vert the input data into range and reflectivity images. The corresponding AUC and runtimes (MATLAB on Intel Core(TM) Quad @ 2.4GHz) are presented in table 4.2, e.g. Harel et al. and Bruce and Tsotsos perform similar to our approach but need at least 20 times longer, Hou and Zhang's approach is 40 percent faster but lacks on the performance side.

The advantage of our system is a trade-off between runtime and performance.

5. Conclusion

In this paper we have presented a new approach of combining 2D and 3D features for an attention system.

Input data are provided by a tilting laser scanner, which yields range and reflectivity images. The first stage of our attention approach is the extraction of feature maps from the input data. Additional to orientation and intensity features, we incorporate jump and roof edge features to increase the influence of the scene structure on the saliency calculation. In the second stage the feature maps result into single activation maps for each feature, which are then combined in the third stage into the final saliency map.

We have shown that our system benefits from the additional edge features. In scenes with low contrast in the reflectivity image the performance increases significantly. In comparison to other attention approaches our system is a good trade-off between performance and runtime.

Further work will include the incorporation of task-dependent top-down information and a real-time implementation. The overall goal will be a flexible vision system that recognizes salient objects first, guided by attentional mechanisms in real-time.

Acknowledgements

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.

References

[1] M. Bjrkman and J.-O. Eklundh. Vision in the real world: Finding, attending and recognizing objects. *International Journal of Imaging Systems and Technology*, 16(5):189–208, 2006. 2

[2] N. Bruce and J. Tsotsos. An attentional framework for stereo vision. In *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*, pages 88 – 95, may 2005. 2

[3] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 2009. 5, 6, 8

[4] S. Choi, S. Ban, and M. Lee. Biologically motivated visual attention system using bottom-up saliency map and top-down inhibition. *Neural Information Processing-Letters and Review*, 2(1), 2004. 2

[5] S. Frintrop, A. Nuchter, H. Surmann, and J. Hertzberg. Saliency-based object recognition in 3d data. In *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2167 – 2172 vol.3, sept.-2 oct. 2004. 3

[6] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7:6:1–6:39, January 2010. 2, 4

[7] F. Hamker. The emergence of attention by population-based inference and its role in distributed

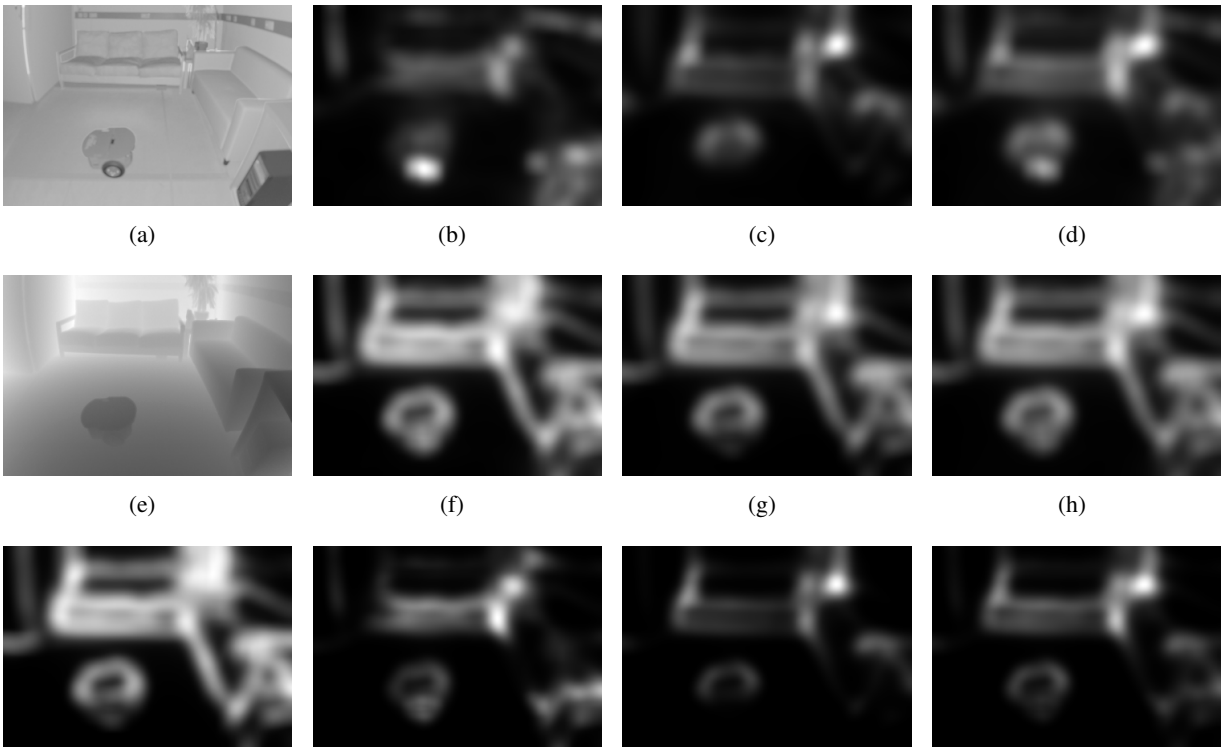


Figure 6. Saliency maps from our approach with different combinations for test sequence nr. 1. (a) reflectivity image, (e) range image, (i) combined edge channels, (b-d) intensity and orientation channels for (b) reflectivity (c) range and (d) both. (f-g) additive and (j-l) multiplicative combination of each (b-d) with edges.

processing and cognitive control of vision. *Computer Vision and Image Understanding*, 100(1-2):64 – 106, 2005. 2

[8] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552, 2007. 5, 6, 8

[9] X. Hou and L. Zhang. Dynamic visual attention: searching for coding length increments. In *NIPS*, pages 681–688, 2008. 5, 6, 8

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254 –1259, nov 1998. 2, 3, 5, 6, 8

[11] A. Maki, P. Nordlund, and J.-O. Eklundh. Attentional scene segmentation: Integrating depth and motion. *Computer Vision and Image Understanding*, 78(3):351 – 373, 2000. 2

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

[13] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimizing detection speed. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2049 – 2056, 2006. 2

[14] N. Ouerhani and H. Hugli. Computing visual attention from scene depth. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 375 –378 vol.1, 2000. 2

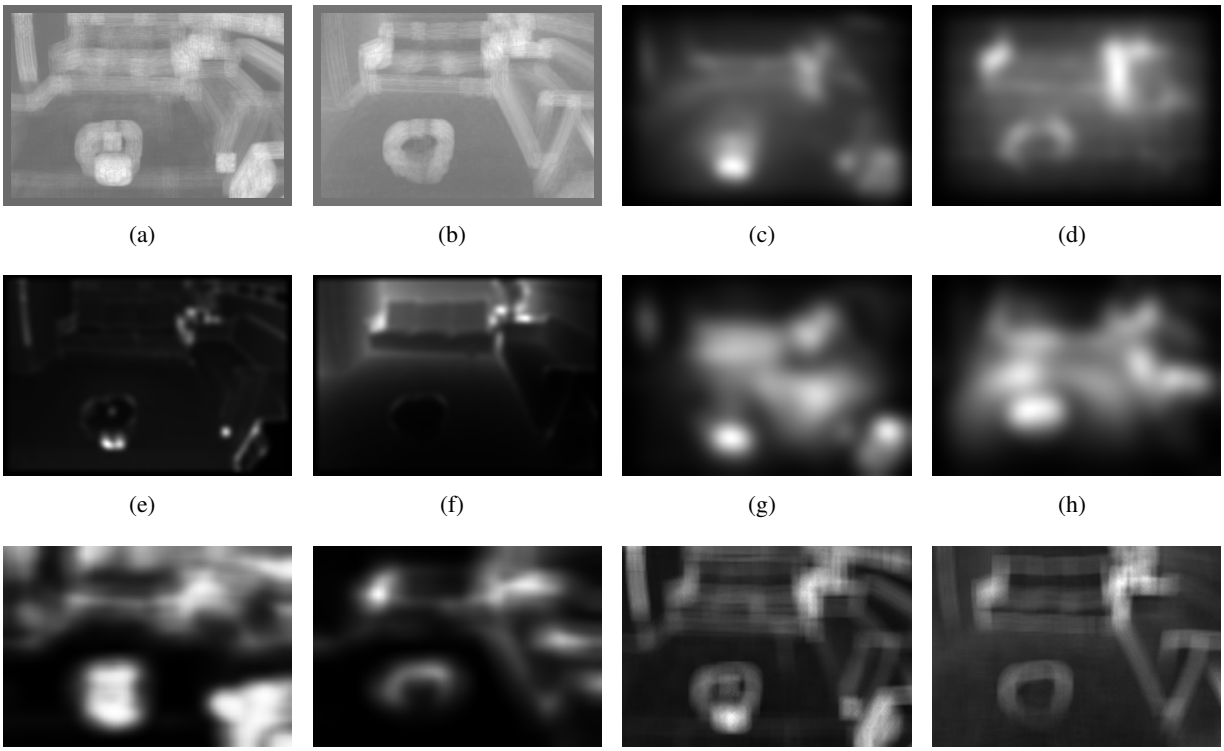
[15] R. P. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard. Eye movements in iconic visual search. *Vision Research*, 42(11):1447 – 1463, 2002. 2

[16] B. Rasolzadeh, M. Bjrkman, K. Huebner, and D. Kragic. An active vision system for detecting, fixating and manipulating objects in the real world. *The International Journal of Robotics Research*, 29(2-3):133–154, 2010. 2

[17] R. Schwarz, E. Einramhof, and M. Vincze. Real-time foveation system based on dense 2.5d data. *Proceedings of the 20th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD2011)*, (20):8, oct 2011. 4

[18] H. J. Seo and P. Milanfar. Nonparametric bottom-up saliency detection by self-resemblance. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 45 –52, june 2009. 5, 6, 8

[19] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *Robotics, IEEE Transactions on*, 25(4):861 –873, aug. 2009. 2



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 7. Saliency maps by the alternative approaches (each with reflectivity and range as input image): (a) and (b) Bruce and Tsotsos [3], (c) and (d) Harel et al. [8], (e) and (f) Hou and Zhang [9], (g) and (h) Itti et al. [10], (i) and (j) Seo and Milanfar [18], (k) and (l) Zhang et al. [24]

[20] S. S. C. U. S.-O. H. Thielemann J., Sandner T. and K. T. Taco: A three-dimensional camera with object detection and foveation. *SAB 2010*, 2010. 2

[21] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53:169–191, 2003. 10.1023/A:1023052124951. 2

[22] A. Torralba. Modeling global scene factors in attention. *J. Opt. Soc. Am. A*, 20(7):1407–1418, Jul 2003. 2

[23] J. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin and Review*, 1:202–238, 1994. 2, 3

[24] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008. 5, 6, 8