Noname manuscript No.
(will be inserted by the editor)

Attentional Scene-Exploration and Object Discovery in Image and RGB-D Data

Germán Martín García · Thomas Werner · Simone Frintrop

Received: date / Accepted: date

Abstract In this paper, we summarize our project work of the last two years, where we addressed the tasks of visually exploring a scene with visual attention mechanisms based on saliency computation, and of locating unknown objects in the environment. The latter is also called object discovery and consists in finding candidate objects without previous knowledge about the objects themselves or the scene. We follow an approach motivated from human perception and combine saliency and segmentation to generate object candidates. We show results on 2D images as well as on 3D sequences obtained from an RGB-D camera.

1 The Project: Situated Vision to Perceive Object Shape and Affordances

This report describes our progress within the DFG-funded project *Situated Vision to Perceive Object Shape and Affordances* which started in February 2012 and is a cooperation with the groups of Barbara Caputo from the university of Rome, Bastian Leibe from RWTH Aachen, and Markus Vincze from TU Vienna.

The objective of this project is to develop the visual capabilities a service robot must have when operating in a home environment (correspondingly, our internal project name is "Vision@home"). More specifically, we plan to provide models and methods to detect, recognize, and categorize the 3D shape of everyday objects and their affordances in homes. The research is based on the Situated Vision paradigm, which starts from the premise that vision has the task of delivering information for the cognitive robot situated in its environment.

Institute of Computer Science III, Universität Bonn, 53117 Bonn, Germany
E-mail: {martin,wernert,frintrop}@iai.uni-bonn.de

2 Work Package Bonn: Visual Attention and Object Discovery

The part of the project that is tackled by our group is the task to visually explore a scene and to locate objects in the environment. Exploring a scene means to prioritize the huge amount of incoming visual data, which we address by visual attention mechanisms that are inspired from human vision (Section 2.1). The localization of unknown objects is also known as object discovery and will be introduced in Section 2.2. This report summarizes our work in these areas which has been published in [14,10,9].

2.1 Guiding Scene Exploration by Visual Attention

In human vision, the prioritization of sensory input is performed by mechanisms of selective attention [24]. These mechanisms consist of bottom-up parts that direct attention to regions that are salient, and top-down aspects that guide the processing to regions that are behaviourally relevant. Many computational models of visual attention have been developed during the last fifteen years, and they have been surveyed in depth in [11,5]. While top-down information is important in human perception, such information is not always available for computational systems. Here, we concentrate on bottom-up attention, which is commonly modelled by saliency maps.

During the last decade, there has been increased interest in saliency computation within the computer vision community. There are approaches that are based on the spectral analysis of images [15,29], models that base on Bayesian theory [17,35], or on decision theory [13,12], and those that use machine learning techniques

[20,3]. Because of the overwhelming number of different approaches, it is hard to keep an overview and to see the differences, and, more importantly, the similarities of the methods.

As we have outlined in [9], the most essential element of saliency methods is a center-surround contrast computation, since a high contrast in some feature dimension is an intrinsic property of a salient item: by definition, something that "stands out relative to its neighbours" (cf. Wikipedia: "Salience (neuroscience)", Jan. 2014). Basically all saliency methods compute such a value (although not always in a center-surround manner), and the most important difference between them is the way this contrast is computed.

Cognitive models compute the center-surround contrast usually by Difference-of-Gaussian filters (DoG), since these are known to model best the concentric cells of the human visual system, e.g., retinal ganglion cells [28]. Also other approaches, such as the Bayesian surprise model [17] or the decision-theoretic model of [13], use DoG and Gabor filters to compute contrasts. Some approaches compute the contrast not based on pixels but on patches [32,6] or on previously segmented regions, e.g., superpixels [25,36]. Instead of computing local contrasts, some approaches compute global contrasts by considering the whole image as surrounding region, e.g., [1] or [7]. Note, however, that while global contrasts are quicker to compute, they are not able to capture local saliencies (see [9]). The contrast computation can also be extended to the spatial domain by computing depth contrasts [22,4] or to the temporal domain, where it computes the change of the visual data over time [17].

2.1.1 Saliency Computation

While enormous effort has been put into introducing novel concepts for computing such contrast, we have recently shown that within a well-designed framework, it is still possible to achieve state-of-the-art performance with traditional Difference-Of-Gaussian filters [10]. While the system in [10], called Simple CoDi, was a modification of the CoDi saliency system [19], we have by now a new implementation of the saliency system[1] that follows the structure of traditional saliency systems like our previous VOCUS system [8] or the the well-known iNVT system of Itti and colleagues [18]: we compute the contrast in different feature dimensions within a scale-space given by image pyramids.

In contrast to [18,8], where simple Gaussian pyramids are used, we build a more sophisticated scale-space structure with several scales on each level (octave) of the pyramid as in [21]. Since the levels of the scale-space correspond to the size of the regions that can be detected, such a structure enables a larger variety of regions it responds to and a higher precision in the saliency maps. Such a Gaussian pyramid is constructed for each of the features. We have implemented intensity, color, and orientation as features. However, for the task of object discovery, the orientation feature has shown to be less useful and we have only used intensity and color in this setting. To enable a processing of red-green and blue-yellow contrasts for the color feature channel, the system operates on an opponent color space. We have investigated the Lab color space and the opponent space from [19], with the latter giving better performance in most cases.

For each map of the scale-space, we compute a center-surround difference based on Difference-of-Gaussian filters. A DoG contrast image can be computed quickly by subtracting two layers of the image pyramid. However, this restricts the computations to smoothing factors that exist in the pyramid (usually powers of 2 with respect to the original image). We have achieved better performance by explicitly computing a "surround image" for each image of the pyramid which is then used for subtraction. That means, for each pyramid image I_s, we consider this image as center image and compute a corresponding surround image by $S_s = I_s * G$, where $'*'$ denotes the convolution and G a Gaussian. Then, we compute the DoG contrast image C_s as usual by subtracting the surround image S_s from the center image: $C_s = I_s - S_s$. Note that by choosing G appropriately, arbitrary center-surround ratios can be obtained. After computing such a DoG contrast image for each layer of the pyramid, we sum the contrast maps up, first to one conspicuity map for each feature, and finally to a saliency map.

We could show that this system, which is simple in structure, quick to implement, and fast in execution time, outperforms 7 state-of-the-art saliency systems in terms of precision and recall [10]. In contrast to many other saliency systems which are designed to detect one prominent object which is centered in the image, our approach is able to cope with natural cluttered scenes as they would be obtained from a mobile robot or a head-mounted camera. Figure 1 shows exemplary some saliency maps obtained from our saliency system and from the baseline methods.

2.1.2 Visual Attention in Video Sequences

If attention operates on sequences instead of images, new challenges occur. Regions of interest have to be fixated for a certain time, then withdrawn from this lo-

[1] Code will soon be available on our webpages

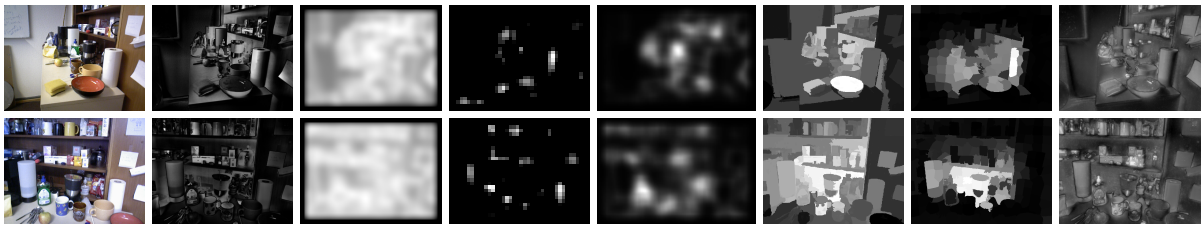

Fig. 1: Saliency maps from AC [2], AIM [7], SaliencyToolbox [31], HZ [16], HSaliency [33], Yang [34], and our saliency system (Fig. from [10])

cation, and a new fixation has to be generated. When computing the new fixation, it has to be considered which regions have been attended recently. In human vision, this problem is solved by inhibition of return (IOR) mechanisms that inhibit an attended region for a while to prevent returning to this location [26]. In computational attention systems, the inhibition of return mechanism is usually solved by inhibiting regions in the saliency map [18,8]. However, this works only in static images. In sequences, the inhibition does not inhibit the correct region any more as soon as objects or the camera moves. Why then does the inhibition of return mechanism work in human vision, although eyes, head, and objects usually move? Posner and Cohen found that the IOR mechanism works in spatial, and not in retinotopic coordinates [26].

Based on these findings, we introduced in [14] a spatial attention mechanism that operates on 3D data from an RGB-D camera and that performs the attentional saccade-fixate cycle in spatial coordinates. A 3D environment map is created based on the KinectFusion algorithm [23] and the information about when an object was attended last is stored directly in the voxels. Based on an inhibition weight that stores the time since the last fixation of this voxel and an inhibition flag that determines whether a voxel should be currently inhibited, we raycast the inhibition data from the spatial map to the current viewpoint. The resulting 2D inhibition map can then be used for inhibiting the values in the saliency map (see [14] for details). This 3D attention mechanism will be used in our 3D object discovery method, described in Sec. 2.2.2

2.2 Object Discovery

Object discovery is the task to find all the objects in a scene without knowing how they might look like or what category they belong to. In contrast to object recognition or classification, the types of objects are not known in advance, there is no training phase, and the system starts without any prior knowledge. Following the phrasing of [3], such a system addresses the question "what is an object?". This topic is of interest for many applications, ranging from automatically cropping the most interesting thumbnail from your holiday pictures up to collecting a database of objects with an autonomous service robot that explores a new environment.

The notation for the task to detect and localize unknown objects in a scene varies strongly among communities. While the robotics community calls the problem *object discovery* or *general object detection*, in computer vision the problem is commonly known as *object proposal generation*. Literature in cognitive science and psychology usually speaks about *object detection* or *object perception*. We call the problem "object discovery" since we think that the term best describes the fact that objects are not known in advance. Also the resulting object candidates have different names in the communities. To disambiguate the notation, we list the terminology in Table 1.

Here, we briefly summarize our work on object discovery that has been published in [10] for 2D images and in [14] for 3D data from an RGB-D camera. For more details please refer to the corresponding publications.

2.2.1 Object Discovery in 2D Images

Our method bases on the idea of proto-objects that originates from psychological research where it was introduced by [27]. As mentioned there, proto-objects are object candidates, which correspond to visual structures that result from early segmentation processes. According to Rensink, attention then "acts as a hand to grasp proto-objects to form coherent objects". Following this idea, we find objects in a two step approach: first the image is segmented into perceptually coherent parts; second, a saliency map is computed and segments are selected depending on their saliency. The concept is visualized in Figure 2.

This simple approach already works very well on many web images, as the ones from the MSRA Dataset

Community:	Computer Vision	Robotics	Cognitive Psychology
Task:	Object proposal generation	Object discovery/ General object detection	Object detection Object perception
Results of segmentation:	Segments/ Superpixels	Segments	Proto-objects
Final results:	Object proposals Object candidates Object hypotheses	Object candidates Object hypotheses	Proto-objects Object candidates Object hypotheses

Table 1: Disambiguation of terminology in different communities (Table from [9]).

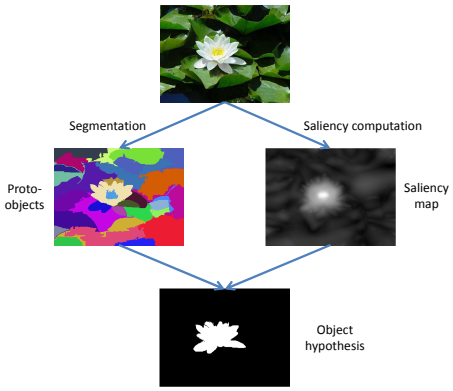

Fig. 2: Simplified overview of the object discovery approach for web images: saliency (right) selects the segments (left) that compose an object hypothesis (bottom). Fig. from [10].

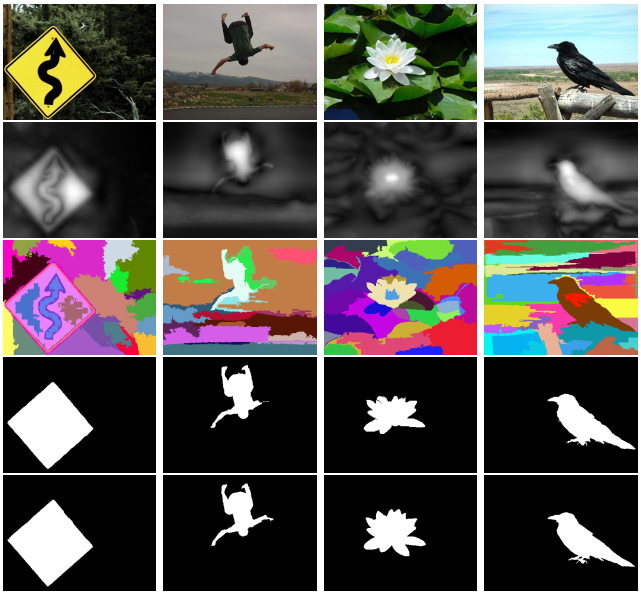

Fig. 3: Several examples of our object discovery method on web images (MSRA dataset). From top to bottom: original images, saliency maps, segmentations, object hypotheses, ground truth. Fig. from [10]

of Salient Objects [20], which is the most frequently used data collection for testing saliency systems. Some example results are shown in Figure 3, and quantitative results can be found in [10]. When interpreting data from a moving camera, the task is much more challenging since scenes are more cluttered and objects are not centered in the image and often intersect with the image borders. Here, it is essential to first extract salient regions from the saliency map, which can then be combined with the superpixels obtained in the segmentation step (details in [10, 9]). Some example images on real-world images from an office scene are shown in Fig. 4.

2.2.2 Object Discovery in 3D Sequences

In [14], we have presented an approach to find objects in RGB-D data from a Kinect-like camera and built 3D object models incrementally while observing a scene over time. As in human perception, where color and depth information are processed mainly independently in separate pathways [30], we have a color and a depth processing stream. The color processing stream detects object candidates according to the procedure presented in Sec. 2.2.1. The depth stream builds a 3D map of the environment using the KinectFusion algorithm [23]. Finally, the object candidates are projected into the 3D scene to form 3D object models. An overview of the method is shown in Figure 5.

To prioritize the processing, attention is guided to process the data according to their saliency. The scene is analyzed starting from the most salient item, which is regarded for several frames, before this region is inhibited and processing continues with the next item. To keep consistency of already fixated regions over frames, we root our attention process directly in the 3D data, using the 3D attention model that was described in Sec. 2.1. Storing inhibition data in the 3D voxels instead of pixels in the saliency map prevents the system from loosing track of an inhibited region when moving

Fig. 4: Top: some examples of our object discovery method on real-world office scenes. Each colored contour shows one object hypothesis. Bottom: separately displayed object hypotheses of the above images. Fig. from [10].

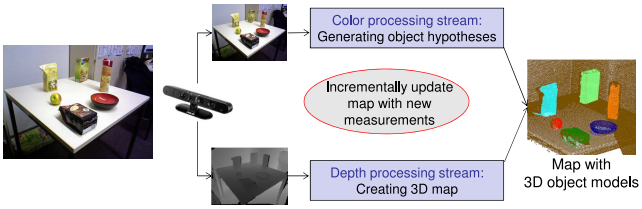

Fig. 5: 3D object discovery: RGB-D data is analyzed in two streams: a color stream processes the RGB image and generates object hypotheses and a depth stream processes the depth data of the sensor and produces a 3D map. Object hypotheses are then projected into the 3D map and data is incrementally improved over time when new measurements arrive. Fig. from [9].

the camera and results in a saccade-fixate cycle that is oriented towards novelty.

In [14], we have shown that our system is able to find many objects, even in cluttered real-world scenes, and that the detection precision is mostly very high (more than 90% for 17 out of 25 objects). An example can be seen in Figure 6. In this quite complex scene, 19 object candidates have been generated after 438 frames (13 sec.). More objects could be found by observing the scene longer.

3 Conclusion

We have presented our work on the field of object discovery, and shown how an attention system can be used for exploring a scene and finding unknown objects in it. In the future, we plan to incorporate the Gestalt principles into our framework of object discovery to select and rank candidates according to their objectness.

Fig. 6: Discovered objects in one of our sequences. Left: original scene. Right: 3D map with 19 discovered objects by the end of the sequence. The rectangles show the automatically obtained 2D object candidates from the color processing stream. Fig. from [9].

References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
2. Achanta, R., Süsstrunk, S.: Saliency Detection using Maximum Symmetric Surround. In: Proc. of the Int. Conference on Image Processing (ICIP) (2010)
3. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
4. Björkman, M., Eklundh, J.O.: Vision in the real world: Finding, attending and recognizing objects. Int'l Journal of Imaging Systems and Technology **16**(2), 189–208 (2007)
5. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (2010)
6. Borji, A., Itti, L.: Exploiting local and global patch rarities for saliency detection. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
7. Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision **9**(3) (2009)
8. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-directed Search, *Lecture Notes in Artificial Intelligence (LNAI)*, vol. 3899. Springer, Berlin/Heidelberg (2006)
9. Frintrop, S.: Cognitive approaches for mobile vision systems. Habilitation thesis at the University of Bonn, Germany (2014)
10. Frintrop, S., García, G.M., Cremers, A.B.: A cognitive approach for object discovery. In: Proc. of the International Conference on Pattern Recognition (ICPR). Stockholm, Sweden (2014)
11. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Trans. on Applied Perception **7**(1) (2010)
12. Gao, D., Han, S., Vasconcelos, N.: Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **31**(6) (2009)

13. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: Proc. of the IEEE International Conference on Computer Vision (ICCV) (2007)
14. García, G.M., Frintrop, S.: A computational framework for attentional 3D object detection. In: Proc. of the Annual Conference of the Cognitive Science Society. Berlin, Germany (2013)
15. Hou, X., Harel, J., Koch, C.: Image signature: Highlighting sparse salient regions. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) (2012)
16. Hou, X., Zhang, L.: Dynamic visual attention: Searching for coding length increments. In: Advances in Neural Information Processing Systems (2008)
17. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision Research **49**(10) (2009)
18. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **20**(11) (1998)
19. Klein, D.A., Frintrop, S.: Salient Pattern Detection using W_2 on Multivariate Normal Distributions. In: Proc. of the Joint Pattern Recognition Symposium of the German Association for Pattern Recognition (DAGM) and the Austrian Association for Pattern Recognition (OAGM) (DAGM-OAGM). Graz, Austria (2012)
20. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence (2009)
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int'l J. of Computer Vision (IJCV) **60**(2), 91–110 (2004)
22. Maki, A., Nordlund, P., Eklundh, J.O.: Attentional scene segmentation: Integrating depth and motion. Computer Vision and Image Understanding (CVIU) **78**(3), 351–373 (2000)
23. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohli, P., Shotton, J., Hodges, S., Fitzgibbon, A.: KinectFusion: Real-time dense surface mapping and tracking. In: Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (2011)
24. Pashler, H.: The Psychology of Attention. MIT Press, Cambridge, MA (1997)
25. Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (2012)
26. Posner, M., Cohen, Y.: Components of visual orienting. In: H. Bouma, D. Bouwhuis (eds.) Attention and Performance X, pp. 531–556. London: Erlbaum (1984)
27. Rensink, R.A.: The dynamic representation of scenes. Visual Cognition **7**, 17–42 (2000)
28. Rodieck, R.W.: Quantitative analysis of cat retinal ganglion cell response to visual stimuli. Vision Research (1965)
29. Schauerte, B., Stiefelhagen, R.: Quaternion-based spectral saliency detection for eye fixation prediction. In: Proc. of the European Conference on Computer Vision (ECCV) (2012)
30. Ungerleider, L., Mishkin, M.: Two cortical visual systems. In: D. Ingle, M. Goodale, R. Mansfield (eds.) Analysis of visual behavior, pp. 549–586. MIT Press (1982)
31. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks (2006)
32. X. Sun H. Yao, R.J.: What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
33. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical Saliency Detection. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
34. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency Detection via Graph-based Manifold Ranking. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
35. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. J. of Vision **8**(32)
36. Zhu, L., Klein, D.A., Frintrop, S., Cao, Z., Cremers, A.B.: Multi-scale region-based saliency detection using W_2 distance on n-dimensional normal distributions. In: Proc. of the Int. Conference on Image Processing (ICIP) (2013)

Germán Martín García received a Dipl-Ing. degree in Informatics from Universidad Autónoma de Madrid in 2008. In 2012 received a MSc. Degree in Computer Science from the University of Bonn, where he is currently enrolled as a PhD. student. His interest is in the fields of visual attention and object discovery.

Thomas Werner received a BSc. degree in Computer Science from the University of Bonn in 2012 and is currently enrolled there as a MSc. student. His interests include computer vision, visual attention and machine learning.

Simone Frintrop is a senior researcher at the Computer Science department at the University of Bonn and is currently heading the Cognitive Computer Vision group. She received her master and her doctoral degree from the University of Bonn in 2001 and 2005, respectively. Her research interests include computational visual attention, cognitive computer vision, and robot vision.