A Cognitive Approach for Object Discovery

Simone Frintrop, Germán Martín García and Armin B. Cremers
Institute of Computer Science III,
Rheinische Friedrich-Wilhelms-Universität Bonn
53117 Bonn, Germany
Email: {frintrop,martin,abc}@iai.uni-bonn.de

***Abstract*—Object discovery is the task of detecting unknown objects in images. The task is of large interest in many fields of machine vision, ranging from the automatic analysis of web images to interpreting data of a mobile robot or a driver assistant system. Here, we present a new approach for object discovery, based on findings of the human visual system. Proto-objects are detected with a segmentation module, generating perceptually coherent image regions. In parallel, a saliency system detects regions of interest in images and serves to select segments, depending on their saliency. We obtain very good results on a database of salient objects and on real-world office scenes.**

I. Introduction

One essential task in many machine vision applications is to automatically and quickly detect objects in the environment. This topic is of interest for many applications, for example automatically processing web images (thumbnailing, resizing, etc.), analyzing video data from devices such as Google Glass, or finding and manipulating objects with an autonomous robot. 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 pre-knowledge. Thus, the system addresses the question "what is an object?".[1] Object discovery is a challenging task for machine vision and belongs to the open problems in the field. The reason is the 'chicken-and-egg property' of the problem: how to search for an object before knowing how it looks like?

While difficult for machines, detecting objects is effortlessly, even unconsciously, done by humans. Thus, it is worth investigating how the human visual system achieves this task. We investigated the findings of psychology and neurobiology on object perception (cf. Sec. III) and developed a biologically inspired strategy that finds objects in a two step approach (cf. Fig. 1): first the image is segmented into perceptually coherent parts, called proto-objects; second, a saliency map is computed and proto-objects are selected depending on their saliency. The result are *object hypotheses* or *object proposals*.

Our contributions in this paper are twofold. First, we propose an improved saliency system that outperforms 7 state-of-the-art saliency models. Second, we propose a new approach for object discovery that is based on concepts from human perception and is applicable to web images as well as to real-world video data.

II. Related work

While object recognition is a well established field, object discovery still involves many challenges. Especially the discovery of objects in 2D images or videos, in which no depth information is available, is difficult. However, several people have investigated this problem and suggested promising approaches; a survey can be found in [30]. Many methods base on the fact that objects are consistent over several images while background is not, and identify regions across images that are visually similar [4], [22]. A related idea is to regard a sequence over time and detect changes, since it is likely that these changes correspond to objects [14]. In the approach of Manén et al. [21], the image is segmented into superpixels [9] and then, connected superpixels are grouped randomly by sampling partial spanning trees that have high sums of edge weights. Other approaches apply machine-learning techniques to learn which aspects of an image might correspond to an object [3]. This idea bases on the fusion of several feature channels, similar as in the field of salient object detection [5]. These approaches are often designed for web images and use often assumptions such as objects are large and central in an image (photographer bias). Instead, we propose here a generally applicable approach that works for single 2D images as well as for real-world image sequences.

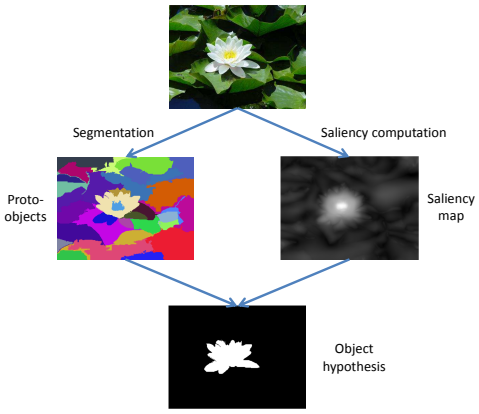

Fig. 1. Simplified overview of our object discovery approach for web images: Saliency selects the relevant proto-objects to form object hypotheses.

Recently, especially with the upcoming RGB-D sensors, several groups have investigated object discovery in 3D data. Karpathy et al. find objects on the 3D meshes obtained from RGB-D data [19]. Johnson-Roberson et al. do object segmentation on full point clouds [17]. The segmentation is seeded at salient points in the image that are mapped to the full point cloud. In [26], 3D object models are built by matching scans from partial views from which they subtract

[1] When referring to objects, we follow a definition from psychology: Objects are "manipulable units with internal coherence and external boundaries" [31].

points that correspond to planar surfaces: floor, walls, etc. In [10], objects were detected in RGB-D data by observing a scene over time and incrementally updating 3D object models. Generally, such 3D approaches have the advantage that they can exploit depth information which is a very helpful feature for object discovery. In this paper, we focus instead on 2D approaches for object discovery in which no depth information is available.

III. Human Object Perception

Object perception is deeply rooted in the human visual system which enables a fast and effortless detection of objects. Even objects of completely unknown appearance are easily recognized as objects, even by young infants [28]. It is not yet completely understood how object perception works in the human brain, but many findings are well known. We will concentrate here on the findings which are important for our framework of computational object discovery.

Physiologically, object detection and recognition take place in the *ventral stream* of the human visual system. This stream is also called *what pathway* since it is strongly involved in color and form processing and is responsible for deciding *what* is visible in a scene. This is opposed to the *dorsal* or *where pathway* that processes mainly motion and depth cues and is responsible for object localization [13]. The ventral visual pathway starts its processing as early as the retina, goes on through the LGN, V1, V2, and V4, until it ends in the inferotemporal cortex (IT), responsible for object recognition.

Many cells in these visual areas have a center-surround structure: they respond excitatorily to light at the center of their receptive field[2] and inhibitorily to light at the surround or vice versa. This means, they have the strongest response if the center is bright and the surround dark (ON-OFF cells) or vice versa (OFF-ON cells). Cells are divided into three types, organized in three channels: the luminance channel, the red-green channel, and the blue-yellow channel [12]. These channels lead from the retina to higher brain areas.

Cells exist with concentric receptive fields and with elongated ones. It has been shown that the concentric fields are modeled best with a two-dimensional Difference-of-Gaussian (DoG) function [25], while the elongated fields are modeled best with Gabor filters [18]. Both types of filters are frequently used in computer vision, because the blob and edge detection that they perform is equally important there as in human vision.

Coming back to object detection, there is evidence that the individuation of objects, which addresses the question of what is an object, takes place before object recognition [23]. The decision of which parts of the visual scene belong to objects results from perceptual organization rules, especially from segmentation processes that bundle parts of the visual input. Such segmentation mechanisms are believed to exist on all levels of the visual system [27] and the bundling is based on concepts such as similarity, proximity, and other processes described already early by the Gestalt principles. A recent review about the history of the Gestalt laws as well as new findings can be found in [32].

[2]The receptive field of a cell is the collection of other cells that influences the output of the cell.

The result of these segmentation processes are so called "proto-objects" [24]. They describe the local scene structure of a spatially limited region and might correspond to objects, but they might also be object parts or collections of several objects. Rensink [24] describes them as "volatile structures of limited spatial and temporal coherence", meaning that they are regenerated constantly and not stored in visual memory. Later on, proto-objects are combined by focused attention to form coherent objects. This is an important step, since it enables to decide which segments an object consists of.

IV. Computational Object Discovery

Formally, object discovery means we are interested in an algorithm that can answer the question of whether a given pixel set corresponds to an object or not. But even if we had a method to answer this question reliably, the problem would be complex: an image of $w \times h = n$ pixels consists of 2^n possible subsets that could potentially form an object (due to partial occlusions, object parts do not necessarily have to be connected). Tsotsos has proven that the related problem of unbounded visual search, that means search for an object whose features are unknown, is NP-hard [29]. And even when restricting the problem to a rectangular bounding box, the problem is still demanding: $O(n \cdot w \cdot h)$ subwindows have to be tested for their objectness, since at each pixel, subwindows of all possible sizes have to be tested. Depending on how computationally expensive the objectness measure itself is, this can easily take several seconds or even minutes which makes the approach inapplicable for real-time applications.

To deal with the complexity of the object discovery problem, we follow the strategy that nature developed and find objects in a two step approach: first the image is segmented into perceptually coherent parts (proto-objects [24]); second, a saliency map is computed and segments are selected depending on their saliency. Thus, the saliency system is responsible for prioritizing the data processing by providing reasonable regions of interest.

For generating proto-objects, we use the segmentation approach of Felzenzwalb and Huttenlocher [9] (cf. Fig. 1, left). This is a graph-based segmentation method that is based on two important Gestalt principles: the similarity and proximity of pixels. The method creates, as the authors state, "perceptually important regions". The second step addresses the question of which segments belong together to form objects. According to Rensink [24], we let attention select the relevant proto-objects. This is done by computing a saliency map that highlights regions of potential interest: the brighter a pixel in the saliency map, the more salient this region is and the larger the probability to contain perceptually relevant data. While in human vision, bottom-up as well as top-down cues play an important role for attention, top-down knowledge is not always available, and in absence of a task, bottom-up saliency is often the best that can be used. Therefore, we use here a pure bottom-up saliency map to select proto-objects, but if top-down information is available, a top-down map can equally well be used.

To compute the saliency map, we use the CoDi saliency system [20] since it is real-time capable, computes precise saliency maps, and works for web images as well as real-

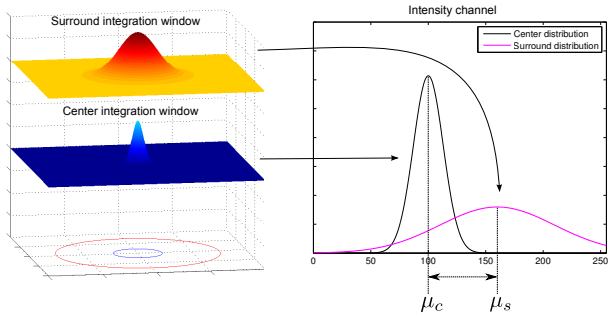

Fig. 2. Visualization of the center-surround computations in the original CoDi saliency system [20] and in our adapted version. Center and surround regions in the image (red and blue ellipses) are weighted with Gaussian windows (left) and feature distributions are determined (right: intensity distribution). The contrast was originally computed with the W_2 metric; here we use the Manhattan distance on the mean values of the distributions (right). The figure shows how this approach corresponds to a Difference of Gaussian approach since each mean value corresponds to the weighted mean of values in the corresponding ellipse.

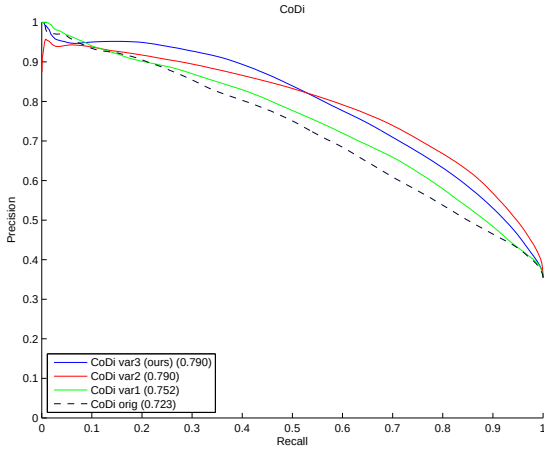

Fig. 3. Comparison of the original version of the CoDi-Saliency system [20] with 3 improvements that we suggested (see text for details). AUC values in parentheses. Evaluation done as described in Sec. V-A.

world images[3]. The CoDi system has shown to outperform many other saliency methods in [20] and source code is openly available[4]. The idea of the CoDi-Saliency is to compute center-surround contrast by comparing normal distributions that represent the feature statistics in the corresponding image regions. Distributions are compared with the W_2-distance (Wasserstein metric based on the Euclidean norm). This concept is visualized in Fig. 2, left. This center-surround measure is embedded into a scale-space structure to enable the detection of objects of different sizes. The computations are performed for intensity and color features, where the latter operates on an opponent-color space with one red-green and one blue-yellow axis. These dimensions correspond to the opponent color channels of the human visual system (cf. Sec. III).

We made several changes on the CoDi system to improve performance. The effect of each of the changes is visualized in Fig. 3 and Fig. 4. First, we adapted the size of the integration window for the center and the surround distribution from $\sigma_c = 1$ versus $\sigma_s = 10$ to $\sigma_c = 1$ versus $\sigma_s = 5$. The latter fits better to human perception [7] and it achieved better performance also in our experiments. We call CoDi with this improvement **variant 1**. Second, we changed the Difference of Gaussian pyramid to a Gaussian pyramid (**variant 2**, includes improvements of variant 1). This makes sense because the DoG operation computes contrasts, which is anyway done by the center-surround operation that is applied to each layer later on. So, it is reasonable to restrict the contrast computation to one place and operate directly on the Gaussian pyramid. This change had the largest visible effect from our improvements since it produces much preciser saliency maps (cf. Fig. 4).

The third change (**variant 3**, includes improvements of variants 1 and 2) affects the computation of the center-surround difference itself. The original CoDi system computes the W_2-distance of normal distributions. However, we found that using instead the much simpler Manhattan distance achieves basically the same results with less computational effort and results in cleaner saliency maps. Interestingly, the Manhattan distance which compares only the mean values of the normal distributions and ignores the variance, corresponds to a Difference of Gaussian approach which is the traditional way to simulate human ganglion and simple cells which are responsible for contrast detection in the human visual system [7]. The reason is the following: the normal distributions computed in CoDi are maximum-likelihood estimates of the center or the surround region, weighted by a Gaussian integration window. Thus, the mean of the normal distribution of a center region centered at pixel position (x,y), is defined as

$$\hat{\mu}_c(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} w(x-i, y-j) F(x-i, y-j), \quad (1)$$

for a $k \times k$ Gaussian window centered at (x,y) with variance ${\sigma_c}^2$ and resulting weights w; F contains the values of the corresponding feature channel, e.g., intensity or 2D color values. The mean of the surround region $\hat{\mu}_s$ is obtained in the same way with a σ_s that is larger than σ_c (as mentioned above, we used $\sigma_c = 1$ versus $\sigma_s = 5$). $\hat{\mu}_c$ and $\hat{\mu}_s$ are either single values (intensity), or two-dimensional vectors (color). Thus, by simply subtracting $\hat{\mu}_c$ from $\hat{\mu}_s$ or vice versa, we obtain the traditional Difference-of-Gaussian method. Since this can be done exactly in the same framework as the distribution-based version, it enables a direct comparison of the methods. This idea is visualized in Fig. 2.

While the AUC value did not change when switching from W_2 to Manhattan distance (cf. Fig. 3), the system is faster and obtained cleaner saliency maps (there are less bright borders around objects, cf. Fig. 4, right). The latter aspect resulted in considerably better performance when combining the saliency maps with segmentation. We call this variant 3 of CoDi **"simple CoDi"**, since it is simpler and faster to compute while producing cleaner saliency maps than the original CoDi system.

Selecting proto-objects based on saliency is then done by combining all segments in which at least $k\%$ of the pixels are

[3]Many other recent approaches for saliency computation are only suitable for web images, since they make several assumptions on images, such as objects are large and central in an image and do not intersect with the image borders

[4]http://www.iai.uni-bonn.de/∼kleind/

Fig. 4. From left to right: Original image, saliency maps of the original version of the CoDi-Saliency system [20], of CoDi variant 1, of CoDi variant 2, and of CoDi variant 3 ("simple CoDi") (see text for details). We used "simple CoDi" in this work.

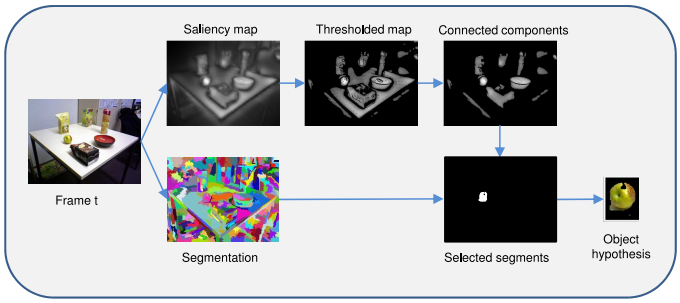

Fig. 5. Object discovery in real-world images.

above a saliency threshold t (we used $k = 25$ and $t = 112$). From these selected segments, all connected components form an object hypothesis (see Fig. 1).

While the described approach works very well on many web images, even without using assumptions about the location of objects (e.g. center-bias), real-world applications are more challenging in many aspects. When interpreting data from an autonomous mobile robot or a mobile device like Google Glass, images are, on the one hand, usually of lower quality due to illumination changes, motion blur, and cheaper cameras, but on the other hand much more complex in content because they contain more objects and clutter. To deal with several objects, we have to determine which proto-objects belong to which object hypothesis. Therefore, we have extended our approach for object discovery as follows. Here again, the saliency map is computed with our "simple CoDi" system. Then, adaptive thresholding[5] (OpenCV method) thresholds the saliency map with help of a local Gaussian kernel, and connected components are found in the resulting map and ranked by average saliency. Finally, the overlap of each proto-object with these salient components is determined and all proto-objects that are covered by at least $k\%$ of a salient component are chosen to belong to the current object candidate. Thus, each salient component results in an object hypothesis and the precise boundaries are obtained by the segmentation process. Fig. 5 visualizes the process.

V. Experiments and Results

Our experiments are divided into three parts: first, we evaluate the improvements on the CoDi-saliency system. Second, we show the performance of the proposed object discovery approach on a database of salient objects. Finally, we show that the approach is also applicable to challenging real-world settings with many objects and clutter.

[5]In most recent work, we obtained even better results with region growing instead of adaptive thresholding. Please check our newest publications at http://www.iai.uni-bonn.de/∼frintrop

A. Saliency evaluation

We have compared our new adaption of the CoDi-saliency system with 6 other saliency systems: HSaliency [34], Yang 2013 [35], AC 2010 [2], HZ [15], AIM [6], and the SaliencyToolbox (ST) [33] which is a reimplementation of the Itti-system [16]. They have been chosen due to their popularity and frequency of citations [6], [15] or due to their recency and very good results on similar tasks [2], [34], [35], and due to the availability of source code.

We have evaluated the results on images from the *coffee machine sequence* which was also used in [11]. The sequence has 600 frames and shows a complex office scene. Each frame contains between 20 and 50 objects. Object ground truth was annotated on every 30-th frame. We chose this setting for evaluation instead of the commonly used benchmark datasets with web images, because we want to test the ability of the systems to deal with challenging real-world scenes that contain many objects. The images were evaluated according to the procedure proposed in [1]: thresholding the saliency maps with an increasing $k \in [0, 255]$ results in binarized maps. Then, each of these maps is matched against the ground truth to obtain precision and recall.

The results of the comparison are displayed in Fig. 6, some of the saliency maps are displayed in Fig. 7. It can be seen that the "simple CoDi" saliency system clearly outperforms all other systems in terms of precision and recall. Furthermore, the system is with 0.098 sec. on an 320×240 image (Intel Core i3-2330M, 4 x 2,2 GHz, 32bit, 4GB RAM) close to real-time on non-optimized code. Parallelization could further improve the speed of the system.

B. Object discovery on web images

In this section, we evaluate our object discovery approach on the MSRA-1000 database of salient objects [1]. The images contain objects that were marked as salient by 2 out of 3 users. Fig. 8 shows several examples from our approach for object discovery, and Fig. 9 shows how the new approach outperforms the CoDi-saliency method without segmentation. It can be seen that the curve drops considerably later when the recall values grow.

C. Object discovery on real-world scenes

Finally, we have applied the object discovery approach to real-world images obtained from the office sequence mentioned before. Some example images are shown in Fig. 11: on

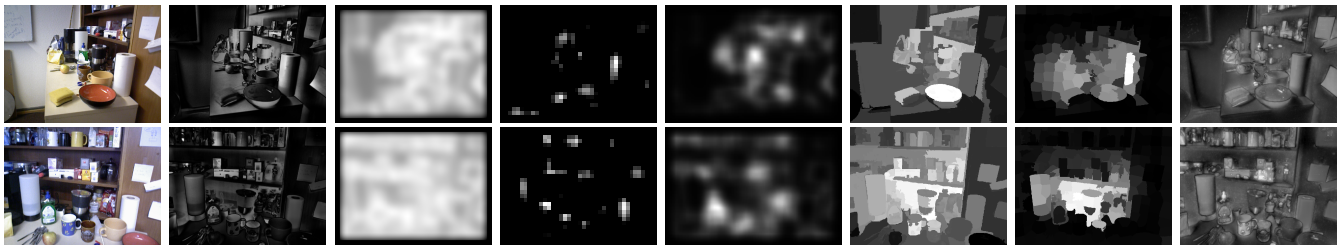

Fig. 7. Saliency maps from AC [2], AIM [6], SaliencyToolbox [33], HZ [15], HSaliency [34], Yang [35], and our "simple CoDi" saliency system

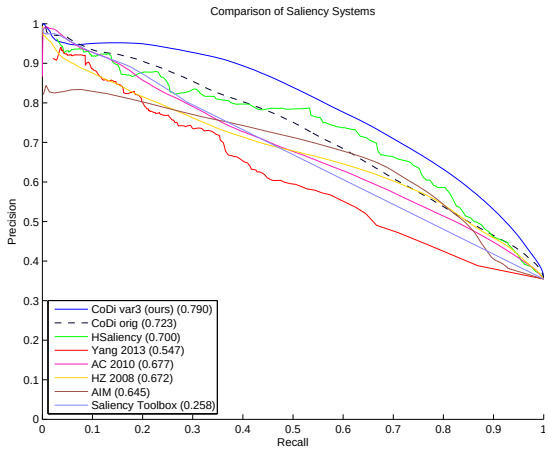

Fig. 6. Comparing our "simple CoDi" saliency system to 7 state-of-the-art methods: CoDi orig [20], HSaliency [34], Yang [35], AC [2], HZ [15], AIM [6], and the SaliencyToolbox [33]. AUC values in parentheses.

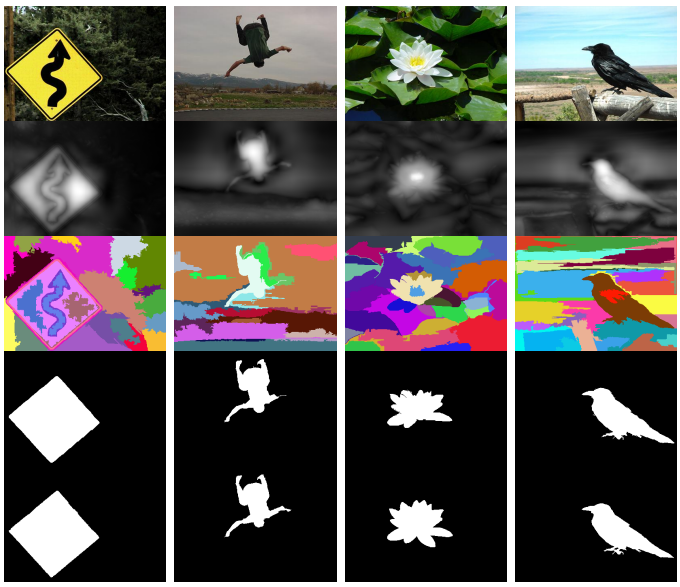

Fig. 8. Several examples of our object discovery. From top to bottom: original images, saliency maps, segmentations, object hypotheses, ground truth.

the left, a simple table-top scene to illustrate the idea (not used for the quantitative analysis), in the middle and on the right two examples for the office database (used for quantitative analysis).

We compute the recall, i.e., the percentage of objects which are found by our approach, and the precision, i.e., the percentage of valid object hypotheses that really represent an object[6]. For this, we consider a match as valid if the Pascal measure is satisfied (intersection-over-union > 0.5) [8]. We compare our method with two other approaches: the "objectness" measure of Alexe et al. [3], and the object discovery method of Manén and colleagues [21]. Since our approach assigns a saliency value to each detected proposal and the two other methods have a ranking for their proposals, we have a fair way of comparing the best N object candidates of all three approaches. This is often of advantage for real-time systems that have to prioritize processing capacities. Therefore, we sort the detected objects by their quality and evaluate the performance of the systems depending of the number of object hypotheses per image that are considered. Since the objectness measure returns bounding boxes instead of precise regions, we represent the ground truth also by boxes for their approach and evaluate our measure once with pixel-precise regions (green curve) and once with boxes (red curve) to enable a fair comparison.

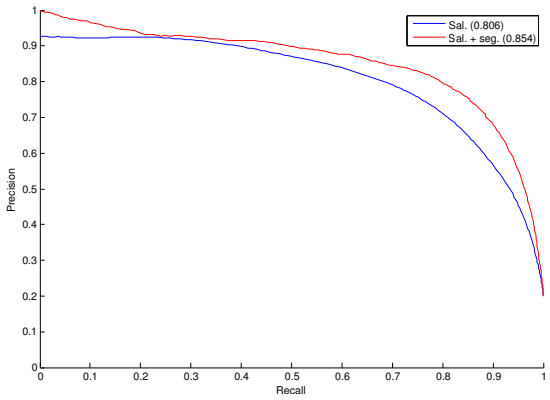

Fig. 9. Object discovery on the MSRA database. Blue curve (Sal.): CoDi-Saliency [20]; red curve (Sal. + Seg.): new combination of saliency and segmentation. AUC values in parentheses.

The results of the quantitative evaluation are shown in Fig. 10. It shows that our method outperforms the objectness measure clearly. Although it is also visible that the approach still misses many objects (there is no current method that can detect all objects in such challenging scenes), it can also be seen that the detected object hypotheses have a good quality and are good candidates as input for object recognition modules or for manipulation by a mobile robot. In the future, we plan to track proposals over time to improve the quality of the approach.

[6]Note that recall and precision measure different qualities here than in Sec. V-A

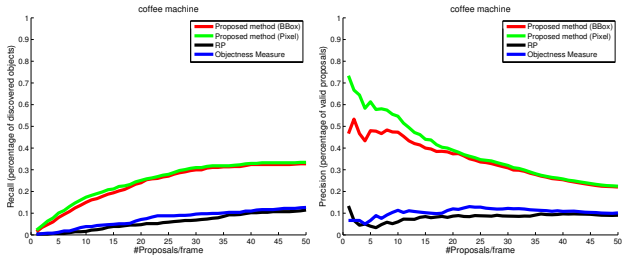

Fig. 10. Comparison of our object discovery method (once with pixel-precise regions and ground truth (green curve), once with bounding boxes (red curve) with the objectness measure from [3] (blue curve). Left: the percentage of discovered objects per frame (recall), right: the percentage of valid proposals (precision). Performance is plotted depending on the number of object proposals that were considered (best N proposals per frame).

Fig. 11. Top: some examples of our object discovery method on real-world office scenes. Each colored contour shows one detected object hypothesis. Bottom: separately displayed object hypotheses of the above images.

VI. Conclusion

We have presented a cognitive approach for object discovery that is based on several findings from human object perception. Perceptually coherent regions are detected with a segmentation method and saliency serves to select and combine segments to form object hypotheses. We have shown that the approach is able to detect objects in web images, which is useful for applications such as thumbnailing or automatic resizing, as well as to operate on real-world data as a mobile robot or a head-mounted camera would obtain. In future work, we will add Gestalt principles such as symmetry or convexity to evaluate whether the obtained object hypotheses are valid.

Acknowledgment

The authors would like to thank DFG for financing this research and Thomas Werner for his help with the experiments.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *Proc. CVPR*, 2009.

[2] R. Achanta and S. Süsstrunk. Saliency Detection using Maximum Symmetric Surround. In *Proc. of ICIP*, 2010.

[3] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the Objectness of Image Windows. *Trans. on PAMI*, 34(11):2189–2202, 2012.

[4] H. Arora, N. Loeff, D. Forsyth, and N. Ahuja. Unsupervised segmentation of objects using efficient learning. In *Proc. of CVPR*, 2007.

[5] Ali Borji, Hamed R. Tavakoli, Dicky N. Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency modeling. In *ICCV*, 2013.

[6] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *J. of Vision*, 9(3):1–24, 2009.

[7] C. Enroth-Cugell and J.G. Robson. The Contrast Sensitivity of Retinal Ganglion Cells of the Cat. *J. Physiol.*, 1966.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2), 2004.

[10] G. Martín García, S. Frintrop, and A. B. Cremers. Attention-based detection of unknown objects in a situated vision framework. *German Journal of Artificial Intelligenz, Springer*, 2013.

[11] Germán Martín García and Simone Frintrop. A computational framework for attentional 3D object detection. In *Proc. of the Annual Conf. of the Cognitive Science Society*, 2013.

[12] K. R. Gegenfurtner. Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, 4:563–572, 2003.

[13] M.A. Goodale and A.D. Milner. Separate visual pathways for perception and action. *Trends Neuroscience*, 15(1), 1992.

[14] E. Herbst, P. Henry, X. Ren, and D. Fox. Toward object discovery and modeling via 3-d scene comparison. In *ICRA*, 2011.

[15] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. In *Advances in Neural Information Processing Systems*, 2008.

[16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[17] M. Johnson-Roberson, J. Bohg, M. Björkman, and D. Kragic. Attention-based active 3D point cloud segmentation. In *IROS*, 2010.

[18] J.P. Jones and L.A. Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.*, 58(6):1233–1258, 1987.

[19] A. Karpathy, S. Miller, and L. Fei-Fei. Object discovery in 3D scenes via shape analysis. In *ICRA*, 2013.

[20] D. A. Klein and S. Frintrop. Salient Pattern Detection using W_2 on Multivariate Normal Distributions. In *Proc. of (DAGM-OAGM)*, 2012.

[21] S. Manén, M. Guillaumin, and L. Van Gool. Prime Object Proposals with Randomized Prim's Algorithm. In *ICCV*, 2013.

[22] N. Payet and S. Todorovic. From a set of shapes to object discovery. In *Proc. of ECCV*, 2010.

[23] Z. W. Pylyshyn. Visual indexes, preconceptual objects, and situated vision. *Cognition*, 80(1-2):127–158, June 2001.

[24] R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17–42, 2000.

[25] R. W. Rodieck. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research*, 1965.

[26] M. Ruhnke, B. Steder, G. Grisetti, and W. Burgard. Unsupervised learning of 3D object models from partial views. In *ICRA*, 2009.

[27] B. J. Scholl. Objects and attention: the state of the art. *Cognition*, 80:1–46, 2001.

[28] E. S. Spelke. Principles of object perception. *Cog. Science*, 14, 1990.

[29] J. K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423–445, 1990.

[30] T. Tuytelaars, C. Lampert, M. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *Int. j. of computer vision*, 88(2), 2010.

[31] C. von Hofsten and E.S. Spelke. Object perception and object-directed reaching in infancy. *Journal of Experimental Psychology*, 144(2), 1985.

[32] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt. A Century of Gestalt Psychology in Visual Perception: I. perceptual Grouping and Figure-Ground Organization. *Psychological Bulletin*, 2012.

[33] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 2006.

[34] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical Saliency Detection. In *Proc. of CVPR*, 2013.

[35] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency Detection via Graph-based Manifold Ranking. In *Proc. of CVPR*, 2013.