# Saliency-Guided Object Candidates Based on Gestalt Principles

Thomas Werner[1,2], Germán Martín-García[2], Simone Frintrop[2]

[1] Fraunhofer Institut für Intelligent Analyse- und Informationssysteme
Schloss Birlinghofen, 53757 Sankt Augustin, Germany

[2] Rheinische Friedrich-Wilhelms Universität, Computer Science Dept. III
Römertsr. 164, 53117 Bonn, Germany

**Abstract.** We present a new method for generating general object candidates for cluttered RGB-D scenes. Starting from an over-segmentation of the image, we build a graph representation and define an object candidate as a subgraph that has maximal internal similarity as well as minimal external similarity. These candidates are created by successively adding segments to a seed segment in a saliency-guided way. Finally, the resulting object candidates are ranked based on Gestalt principles. We show that the proposed algorithm clearly outperforms three other recent methods for object discovery on the challenging Kitchen dataset.

## 1 Introduction

The human ability to detect arbitrary objects fast and reliably is a very important competence for everyday life. It enables us to interact with our environment and reason about it. Also computational systems would strongly profit from such capabilities, for example by being able to detect new objects without explicit instructions. This would increase the autonomy of such systems while decreasing the interaction time spent with them.

Recently, interest in this detection of arbitrary, previously unknown objects, called *object discovery*, has increased strongly. Several methods have been proposed that address this problem in the computer vision community [2, 14, 5] as well as in the field of robotics [13, 11, 16, 8]. We follow here the approach of Manén et al. [14], which is a recent approach that has been very successful. The method starts from an over-segmentation of the image and iteratively grows object hypotheses by adding segments to a random seed segment. This is done by representing the segmented image as a graph and then randomly sampling partial spanning trees that correspond to object candidates.

We modify and extend Manén's approach [14] in several ways to improve the detection quality. First, we add the processing of depth data from an RGB-D device to the processing pipeline. Second, we modify the random selection strategy of Manén to an informed search based on saliency. Saliency detection, as the bottom-up part of visual attention, is an important ability of human perception that guides the processing to regions of potential interest [15]. The saliency information affects the selection of the seed segment as well as the selection of iteratively added segments. Third, we adapt the computation of edge weights of the graph to integrate a new feature called *common border saliency*. Fourth, we adapt the termination criterion that determines when to stop the growing of a candidate by considering the internal similarity as well as the external difference of the current candidate to a new segment. This idea is borrowed from the segmentation method of [4] and fits very well here. And finally, we add an SVM-learned ranking of the resulting object candidates based on Gestalt principles. These principles are descriptive rules from psychology that aim to explain how humans segregate objects from background, especially, which visual properties of objects support our perception [17]. Gestalt principles have recently been successfully used in machine vision approaches to evaluate the shape of objects, e.g., in [16, 11, 8, 13], and we show that ranking based on these principles clearly improves the detection quality.

We have tested our method on the recently introduced Kitchen dataset for object discovery [8] that contains several challenging real-world sequences and show that our approach clearly outperforms the approach from [14] as well as several other recent methods for object discovery in terms of precision and recall.

## 2 System Overview

This section describes the proposed approach in detail (overview in Fig. 1). Given an input RGB-D image, the data is first pre-processed, including the conversion to an opponent colorspace as well as an inpainting of missing depth values. We use the colorspace of [12], but shifted and scaled to the range $[0, 1]$. Next, a saliency map and an over-segmentation are generated from the color data. From the oversegmented map, a graph is constructed which has segments as vertices and stores the similarity of neighboring segments in edge weights. Then, we introduce the saliency-guided Prim's algorithm that generates object candidates by iteratively adding segments to a set of salient seed segments. Finally, we rank the candidates by a combination of Gestalt principles which is learned with an SVM. The output of the system is a list of object candidates sorted by objectness.

### 2.1 Saliency Computation

For saliency computation, we use the recently introduced VOCUS2 saliency system [7][3], which is a re-implementation of the VOCUS system [6]. The main structure is similar to traditional attention systems such as the one from Itti [9]: the system computes intensity and color features by Difference-of-Gaussian filters (center-surround ratio: 2 : 4) on different scales (here: 2) and octaves (here: 5) before fusing them to a single saliency map (example in Fig. 2). We chose this system since it has shown to outperform many state-of-the-art methods for salient object segmentation, is real-time capable, and works on cluttered real-world scenes [7].

[3] Code: http://www.iai.uni-bonn.de/∼frintrop/vocus2.html

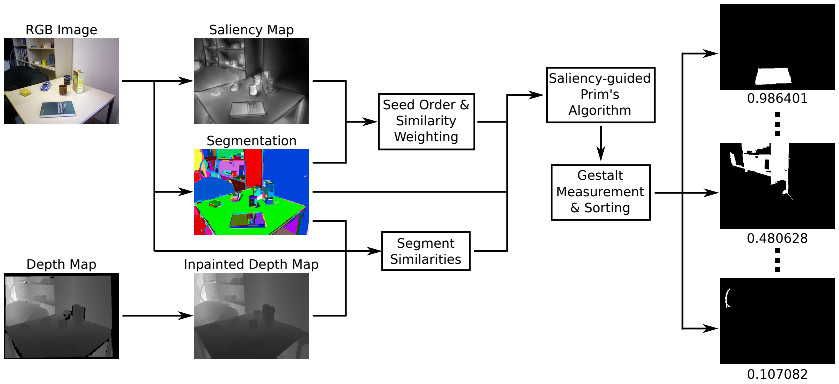


**Fig. 1.** Overview of the proposed approach: Given an RGB-D input, the algorithm produces a list of object candidates, sorted by quality, here from 0.986401 (best) to 0.107082 (worst).

In the following, the saliency map is used to determine the seeds of the candidate generation process and as a feature for determining object boundaries.

### 2.2 Graph Representation

Since our candidate generation method is based on finding minimum spanning trees in graphs, it is necessary to transform the input image into an appropriate graph representation. The graph representation is generated from an oversegmentation $S$ of the image, obtained with the algorithm of [4].

We construct the graph $G = (V, E)$ so that each segment $s_i \in S$ becomes a vertex $v_i \in V$, and for each neighboring pair of segments $s_i$ and $s_j$, an undirected edge $e_{i,j} \in E$ is introduced. To each edge $e_{i,j}$, a weight $w_{i,j}$ is assigned that represents the similarity of the corresponding segments (cf. Fig. 3).

The appearance similarity of two neighboring segments $svm(f_{i,j})$ is evaluated using an SVM that receives a feature vector $f_{i,j}$ extracted from the corresponding segments $s_i$ and $s_j$. It computes the likelihood that both segments are part of the same object. The $4 + 4 + 1 = 9$ features that are used are:

- Intersection of the normalized 16 bin histograms of the four feature channels (3 color + 1 depth)
- Absolute differences of average value per feature channel
- Common-border ratio as defined in [14]

To overcome possible problems that come with inconsistent appearances of objects, an additional feature which relies on saliency is used to measure the similarity of segments. The measure is called *common border saliency* and is based on the observation that along boundaries of perceptually different regions (e.g., object/background boundaries) a center-surround based saliency operator produces a low response (cf. Fig. 2). It is defined as the average saliency along the common border of the two segments. This information can be utilized later

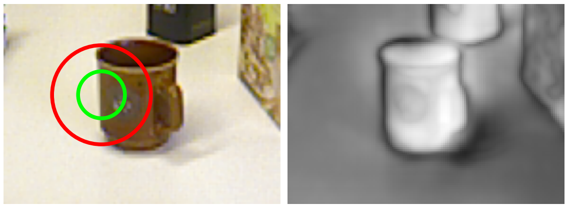


**Fig. 2.** Common-border Saliency: The Difference-of-Gaussian (DoG) operator that is used to computed center-surround contrasts has a low response when located at object boundaries (green = center, red = surround of the DoG filter).

to enforce the candidate generation process to respect such boundaries. The final edge weight/similarity value is defined as

$$w_{i,j} = svm(f_{i,j}) \cdot cbs(s_i, s_j), \tag{1}$$

which is the likelihood of belonging to the same object $svm(f_{i,j})$ weighted by their common border saliency $cbs(s_i, s_j)$.

### 2.3 Candidate Generation Process

Motivated by [14], the candidate generation process is formulated as the problem of finding partial maximal spanning trees using the Prim's algorithm. The differences to [14] are: (i) instead of random vertices the seeds are chosen according to their saliency, which enforces the inspection of the most promising regions first; (ii) segments are chosen in a greedy manner based on the similarity measure of Eq. 1; and finally, we introduce a deterministic stopping criterion compared to the randomized one from [14].

Given the graph $G$ of an image, we extract several partial maximal spanning trees that serve as object candidates. The main idea is to select a starting segment (seed) $s_{h_0}$ as the initial candidate $h_0 = \{s_{h_0}\}$ and to iteratively add the most similar neighboring vertices until a termination criterion is met. After $t$ iterations, this results in candidate $h_t$ consisting of segments $\{s_{h_0}, ..., s_{h_t}\}$. The termination predicate, inspired by the one in [4], takes into account the *internal dissimilarity* between the segments of the candidate, the *external dissimilarity* of the candidate to a given segment, and the size of the candidate.

Given a candidate $h_t$ at iteration $t$ and a vertex (segment) $v_i$, the external dissimilarity is recursively defined as

$$Ext(h_t, v_i) = 1 - \frac{1}{|E_{v_i \leftrightarrow h_t}|} \sum_{e_{j,k} \in E_{v_i \leftrightarrow h_t}} w_{j,k}, \tag{2}$$

where $E_{v_i \leftrightarrow h_t}$ is the subset of edges connecting vertex $v_i$ with any vertex in $h_t$. In other words, the external dissimilarity is the average dissimilarity of the vertex $v_i$ to the current candidate $h_t$. The final termination predicate is defined as

$$Ext(h_t, v_i) > \min\left(k, Int(h_t) + \frac{k}{t+1}\right), \tag{3}$$

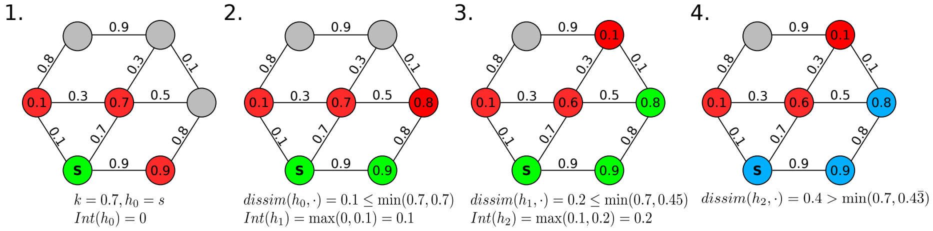

1.
2.
3.
4.
S
$k = 0.7, h_0 = s$
$Int(h_0) = 0$
$dissim(h_0, \cdot) = 0.1 \leq \min(0.7, 0.7)$
$Int(h_1) = \max(0, 0.1) = 0.1$
$dissim(h_1, \cdot) = 0.2 \leq \min(0.7, 0.45)$
$Int(h_2) = \max(0.1, 0.2) = 0.2$
$dissim(h_2, \cdot) = 0.4 > \min(0.7, 0.4\bar{3})$

**Fig. 3.** Small example of the candidate generation process on an artificial graph: Green = current candidate, red = pool of next segments, blue = final candidate. The vertex weights are their external similarity to the current candidate and the edge weights reflect the similarity of the connected vertices.

which compares the external dissimilarity to the next vertex with current maximal external dissimilarity. As in [4], parameter $k$ regulates how much extra internal variability is accepted within the candidates; the less segments the candidate has, the more influence $k$ has. If the predicate holds, the vertex $v_i$ is rejected and the growing process stops. Otherwise, the vertex is accepted and added to the current hypothesis. In such case, the internal dissimilarity $Int(h_{t+1})$ of the candidate at time $t + 1$ is updated as

$$Int(h_{t+1}) = \max\left(Int(h_t), Ext(h_t, v_i)\right), \tag{4}$$

which is the maximal external dissimilarity obtained so far.

We generate object candidates for several values of $k$: we use $k = 0.6$ to $k = 0.9$ in steps of 0.05. A small example of the candidate generation process on an artificial graph is shown in Fig. 3. There it can be seen that 3 vertices are accepted due to their high similarity whereas the next candidate vertex is not similar enough and therefore rejected.

### 2.4 Candidate Ranking Using Gestalt Measurements

After all object candidates are generated, a scoring mechanism is applied to each candidate $h$ to evaluate its visual objectness. Several measures that correspond to different Gestalt principles are derived and evaluated on the candidates.

**Color/Depth Contrast:** since objects usually have a strong contrast to their surround, a contrast measure is derived. For each feature channel, each segment within the candidate is compared with the neighbors outside the candidate, based on the intersection of normalized 16 bin histograms and the difference of color averages of the corresponding segments.

**Good Continuation:** objects usually have smooth contours. Thus, the mean curvature of the contour of the candidate is a good indicator for its objectness. As in [13], we define good continuation as the average angular change of contour.

**Symmetry:** symmetry is a non-accidental property and is known to have influence on human object perception [10]. Based on [11], we measure the overlap $O_1$ and $O_2$ of the candidate with itself after mirroring along both of its principle axes (eigenvectors of the scatter matrix). The two measures describing the symmetry are the *maximal symmetry* ($max(O_1, O_2)$) and the *weighted average symmetry* ($\frac{1}{\lambda_1+\lambda_2}(\lambda_1 O_1 + \lambda_2 O_2)$) weighted by the corresponding eigenvalues.

**Convexity:** convexity is also an important part in human object perception [10]. We propose three measures that capture the convexity of a candidate based on its convex hull. The first one, motivated by [11], is the *average convex distance*, which is the average distance of the candidate boundary to the closest point on the convex hull. Since it depends on the size of the candidate, it is normalized by the number of points that contributed and the largest observed distance. The second and third measure are the *perimeter ratio* and the *area ratio* of the convex hull and the candidate.

**Compactness:** the compactness measure consists of three values. The *average centroid distance* is computed as the average distance of the boundary points to the center of mass of the candidate. It is normalized by the number of boundary points and the maximal observed distance. The second measure is the *circularity*, that measures the similarity to a perfect circle of the same size. The measure is based on the ratio of perimeter P and area A of the candidate and is defined as $\frac{4\pi A(h)}{P(h)^2}$. The last measure is the *eccentricity* and is defined as the ratio of extensions $\lambda_2, \lambda_1$ along both principle axes by $\sqrt{1 - \frac{\lambda_2}{\lambda_1}}$.

**Combination of measures:** After all measures are computed, their concatenation is fed to an SVM that evaluates the objectness of the corresponding candidate and assigns a real value to it. Based on the objectness, the list of object candidates is sorted so that those that are likely to correspond to a real object appear first. Finally, non-maxima suppression is applied to remove duplicate candidates with lower objectness. Whether a candidate is a duplicate, is determined using the *Intersection-over-Union (IoU)* measure and a threshold of 0.5 [3]. Output of our system is a list of object candidates, sorted by their decreasing objectness value.

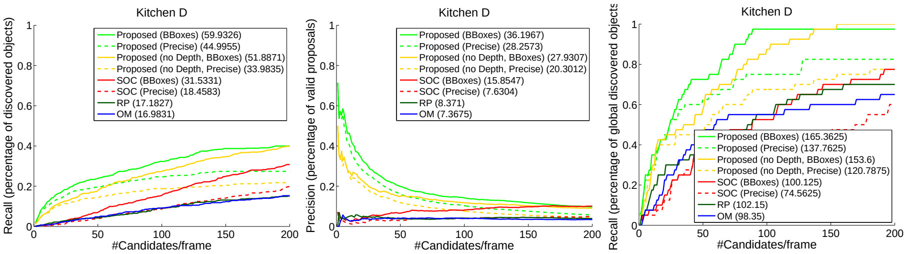
## 3 Training, Evaluation and Results

In this section, we evaluate the performance of our algorithm and compare it to other state-of-the-art methods. We use the Kitchen dataset [8], which consists of five video sequences showing different cluttered, real-world indoor scenes. The sequences contain 600 frames and 80 objects on average. Ground truth labels are available for every $30th$ frame. Furthermore, the labels are consistent over the sequences, making it possible to evaluate the candidates on a sequence level.

### 3.1 Parameter Evaluation

We use the first of the sequences in [8] as training data and for parameter estimation. The rest are used as test sequences.

**Saliency System:** Following the method of [1], we evaluated saliency maps for several sets of parameters using the training sequence's ground truth. The optimal parameters are introduced in Sec. 2. A detailed description of the saliency evaluation and the results can be found in [18].



| | Kitchen A | | Kitchen B | | Kitchen C | | Kitchen D | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Recall | Prec. | Recall | Prec. | Recall | Prec. | Recall | Prec. | Recall |
| Ours (B) | 14.3 | **81.4** | **30.3** | **89.9** | 12.2 | **90.8** | **36.2** | **59.9** | **23.2** | **80.5** |
| Ours (P) | 8.3 | 43.7 | 24.4 | 75.2 | 8.4 | 66.8 | 28.3 | 45.0 | 17.4 | 57.7 |
| Ours (no Depth, B) | 9.8 | 60.9 | 25.3 | 85.5 | 9.6 | 81.6 | 27.9 | 51.9 | 18.2 | 70.0 |
| Ours (no Depth, P) | 5.8 | 33.6 | 19.6 | 64.4 | 6.8 | 55.6 | 20.3 | 34.0 | 13.1 | 46.9 |
| SOC (B) | **19.3** | 74.3 | 21.5 | 68.0 | 11.3 | 75.2 | 15.9 | 31.5 | 17.0 | 62.3 |
| SOC (P) | 12.8 | 50.9 | 14.9 | 53.4 | 7.9 | 54.3 | 7.6 | 18.5 | 10.8 | 44.3 |
| RP | 12.4 | 51.5 | 14.0 | 47.9 | **14.1** | 66.2 | 8.4 | 17.2 | 12.2 | 45.7 |
| OM | 4.7 | 27.0 | 6.2 | 24.5 | 3.3 | 27.7 | 7.4 | 17.0 | 5.4 | 24.0 |

**Fig. 4.** Evaluation on the Kitchen dataset sequences with three reference methods RP [14], OM [2] and SOC [8]. Top: Frame-based recall (left), precision (middle) and scene-based recall (right) on the Kitchen D sequence. Bottom: Overview of area under curve (AUC) values for precision and recall on all kitchen sequences. Highest values shown in bold. B: bounding boxes, P: pixel-precise.

**SVMs:** Within the proposed method two SVMs are used. The first one is trained to estimate the visual similarity of two segments (Sec. 2.2), a problem which is treated here as a classification problem. Training is performed as follows: (i) Given the training sequence, an over-segmentation as described in Sec. 2.2 is generated, (ii) positive and negative segment pairs are extracted, (iii) the corresponding feature vector from Sec. 2.2 is extracted and (iv) the feature vectors and positive/negative class labels are fed to the SVM. A positive segment pair consists of two neighboring segments belonging to the same object, and a negative pair is a set of two neighboring segments that either belong to different object or one belongs to an object and the other to the background. A segment is part of an object if its area is covered by the object by at least 50%. To find the best parameters, the training of the SVM is done using a grid search in the parameter space and 10-fold cross-validation. The best parameter set is the one that has the lowest average error over all 10 rounds.

The second SVM is used to evaluate the objectness of object candidates (Sec. 2.4), which is treated as a regression problem. Training is performed as follows: (i) As before, an over-segmentation of the training data is produced, (ii) all ground truth objects are extracted by forming candidates of all covered segments, (iii) for each candidate the feature vector introduced in Sec. 2.4 is extracted and the IoU with the ground truth is measured, and (iv) for each candidate the feature vector is the input to the SVM, which regresses on the IoU. Like before, the best parameter set is obtained using a grid search and 10-fold cross validation. A detailed description of the training process and the results can be found in [18].

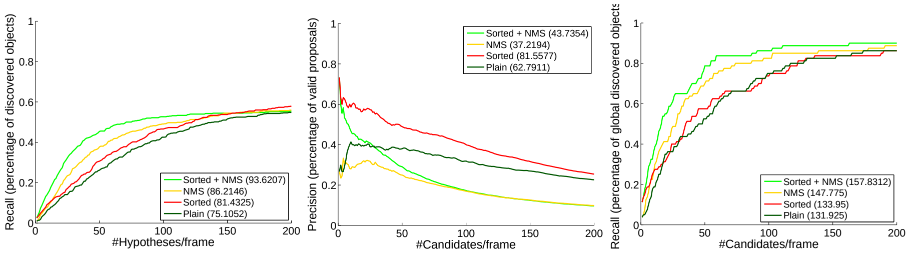


**Fig. 5.** Evaluation of the effect of different ranking methods. Plain: ranking by average saliency; sorted: ranking by Gestalt features; NMS: ranking after non-maxima-suppression (removes dublicates). Each individual step as well as their combination increases the quality of the candidates.

### 3.2 Comparison to other Methods

We compare our approach to the methods *RP* [14], *OM* [2] and *SOC* [8]. We measure precision and recall in terms of the number of valid candidates (IoU $\geq$ 0.5). Additionally, we measure global, scene-based recall as the number of objects that were found throughout the whole sequence. We use the four sequences that were not used for training of the Kitchen dataset [8] as test data.

Although the dataset is annotated with pixel-precise ground truth, all evaluations are also performed using bounding boxes. This way a fair comparison of the methods is guaranteed, since *RP* and *OM* only produce bounding boxes. Our method is evaluated with and without depth information, since the reference methods are also developed to work only with color data.

Fig. 4 shows the evaluation results exemplarily for the Kitchen D sequence. It contains recall, precision and global recall (from left to right) along with the corresponding area under curve values for each method. The proposed method outperforms all reference methods in terms of recall as well as precision. The recall plot shows that around 40% of the objects in a frame are detected and that almost all objects in the sequence are detected at least once. The precision plot shows that when taking few candidates, e.g. 20, many of them match an object. The results on the other sequences are consistently good and can be seen in the table in Fig. 4: our method has on average the highest precision and recall, and outperforms all the other methods in each sequence in terms of recall.

In Fig. 5 we compare different ranking strategies: the default ranking strategy according to the average saliency ('Plain' in the figure), sorting according to the Gestalt measures ('Sorted'), and the non-maxima suppression ('NMS'). The results can be explained as follows: sorting the candidates by their objectness will cause good candidates to appear first which explains the high precision for few candidates and the early increase of the recall. On the other hand, the non-maxima suppression removes duplicate candidates (also good ones) which explains the overall high recall as well as the drop in precision. For the recall the removal of duplicates has only positive effects, since duplicates will at most only re-discover objects or will not discover any object at all.

In Fig. 6 the ten best candidates per method are shown. It can be seen that two of the reference methods (RP and OM) generally produce very large candidates that capture either multiple objects or very large structures. The proposed method on the other hand adequately produces candidates for the individual objects.

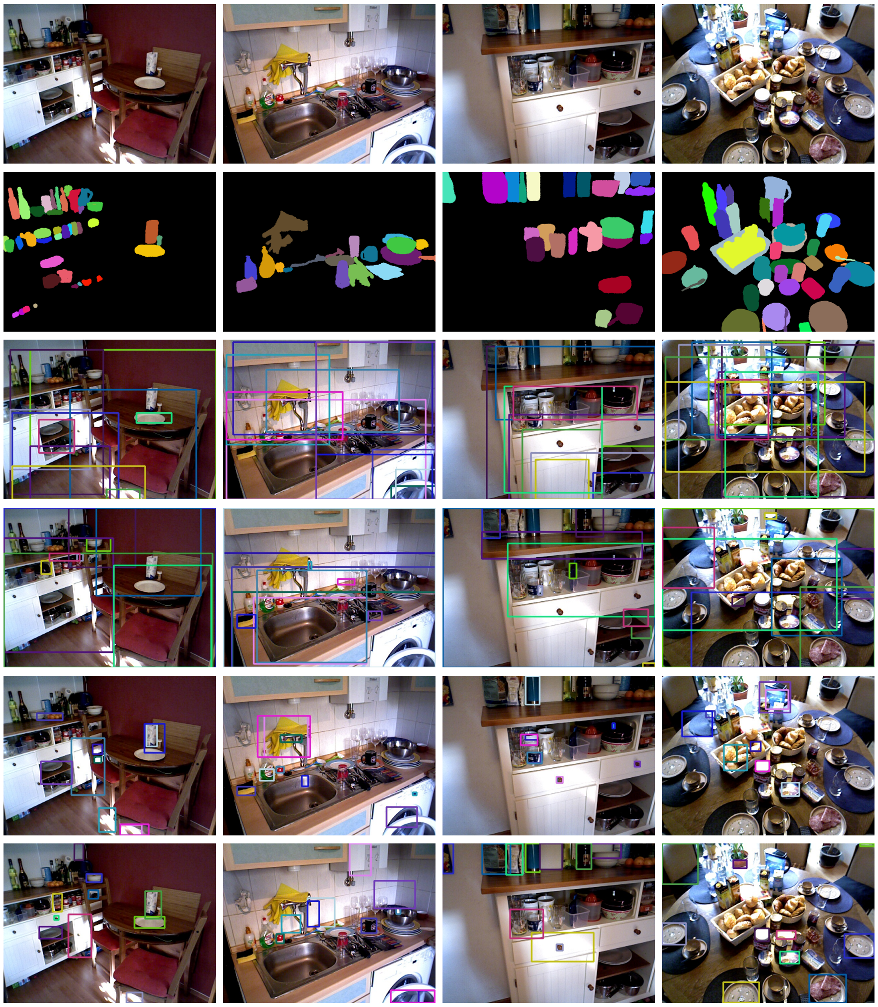


**Fig. 6.** Top 10 object hypotheses for some images. From top to bottom: Input image, ground truth, OM [2], RP [14], SOC [8], Ours (sorted + NMS).

## 4 Conclusion

We have presented a method for finding arbitrary, unknown objects in RGB-D data that utilizes principles of human object perception – visual attention and Gestalt psychology. This enables a fast and reliable generation of object candidates even for complex scenes containing many objects. We have shown that the presented method outperforms several state of the art methods and is able to detect more than 50% of the objects in a frame and more than 90% of the objects visible in the scene.

## References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: Proc. CVPR (2009)
2. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. Trans. on PAMI 34(11) (2012)
3. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. of Computer Vision 88(2) (2010)
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-Based Image Segmentation. Int J. of Computer Vision 59(2) (2004)
5. Frintrop, S., Martín García, G., Cremers, A.B.: A cognitive approach for object discovery. Proc. of ICPR (2014)
6. Frintrop, S.: VOCUS: A Visual Attention System for Object Detection and Goal-directed Search, Lecture Notes in Artificial Intelligence (LNAI) (PhD thesis), vol. 3899. Springer (2006)
7. Frintrop, S., Werner, T., Martín García, G.: Traditional saliency reloaded: A good old model in new shape. In: Proc. of CVPR (2015)
8. Horbert, E., Martín García, G., Frintrop, S., Leibe, B.: Sequence Level Object Candidates Based on Saliency for Generic Object Recognition on Mobile Systems. Proc. of ICRA (2015)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Trans. on PAMI 20(11) (1998)
10. Kanizsa, G., Gerbino, W.: Convexity and symmetry in figure-ground organization. Vision and artifact (1976)
11. Karpathy, A., Miller, S.: Object discovery in 3D scenes via shape analysis. Proc. of ICRA (2013)
12. Klein, D.A., Frintrop, S.: Salient Pattern Detection using $W_2$ on Multivariate Normal Distributions. In: Proc. of (DAGM-OAGM) (2012)
13. Kootstra, G., Kragic, D.: Fast and bottom-up object detection, segmentation, and evaluation using Gestalt principles. Proc. of ICRA (2011)
14. Manén, S., Guillaumin, M.: Prime Object Proposals with Randomized Prim's Algorithm. Proc. of ICCV (2013)
15. Pashler, H.: The Psychology of Attention. MIT Press, Cambridge, MA (1997)
16. Richtsfeld, A., Zillich, M., Vincze, M.: Implementation of Gestalt principles for object segmentation. Proc. of ICPR (2012)
17. Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M., Singh, M., von der Heydt, R.: A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychological bulletin 138(6) (2012)
18. Werner, T.: Saliency-driven object dicovery based on gestalt principles (2015)