23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, December 2016.

Semantic Segmentation Priors for Object Discovery

Germán M. García*, Farzad Husain†‡, Hannes Schulz*, Simone Frintrop§, Carme Torras‡ and Sven Behnke*
*Institute of Computer Science VI, University of Bonn, Germany
{martin,schulz,behnke}@ais.uni-bonn.de
† Catchoom, Barcelona, Spain
‡Institut de Robòtica i Informàtica Industrial, CSIC-UPC. Llorens i Artigas 4-6, 08028, Barcelona, Spain
§Department of Informatics, University of Hamburg, Germany

***Abstract*—Reliable object discovery in realistic indoor scenes is a necessity for many computer vision and service robot applications. In these scenes, semantic segmentation methods have made huge advances in recent years. Such methods can provide useful prior information for object discovery by removing false positives and by delineating object boundaries. We propose a novel method that combines bottom-up object discovery and semantic priors for producing generic object candidates in RGB-D images. We use a deep learning method for semantic segmentation to classify colour and depth superpixels into meaningful categories. Separately for each category, we use saliency to estimate the location and scale of objects, and superpixels to find their precise boundaries. Finally, object candidates of all categories are combined and ranked. We evaluate our approach on the NYU Depth V2 dataset and show that we outperform other state-of-the-art object discovery methods in terms of recall.**

I. Introduction

Object discovery is the task of finding the objects that are present in a scene before knowing about their specific category. One possible application for fast object discovery methods is robotic manipulation [1]. For complex indoor scenarios, object discovery still presents a challenge to current methods [2, 3].

In recent years, semantic segmentation methods have been dramatically improved, and at this point provide reliable results in realistic indoor scenarios. In this work, we propose a method which exploits prior information provided by semantic segmentation methods in order to improve an existing object discovery method. This is, to our knowledge, the first attempt to integrate semantic segmentation into a general object discovery method. Pixel-wise semantic segmentation can be helpful in two ways: First, it provides information on where to expect objects, resulting in fewer false positives. Secondly, the pixel-wise segmentation partitions the image into semantic regions where the candidate generation process can be applied independently, resulting in more precise object candidates.

Our proposed method builds on the saliency-based generic object candidates of Martín García et al. [3] by incorporating semantic segmentation into the pipeline. An overview of our approach is depicted in Fig. 1. We use saliency as a cue to locate the presence of objects, and colour and depth bottom-up segmentations are used to precisely delimit their boundaries. We improve this method by incorporating prior information on the category of the image regions using a semantic segmentation algorithm [4]. Husain et al. [4] train a convolutional neural network to produce a pixel-wise classification of the image into four coarse categories: Floor, structures such as walls, furniture, and movable objects. The pixel-precise semantic maps are used to label the bottom-up segments for each category. Then, the process of object proposal generation is performed independently for each of these categories. Finally, object candidates of all categories are combined and ranked.

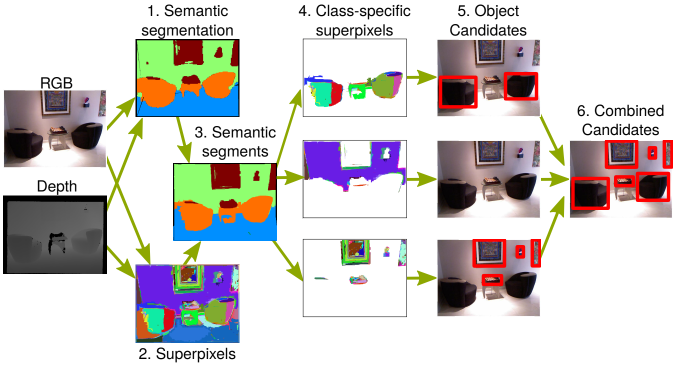

Fig. 1. Overview of our approach: 1. Semantic segmentation and 2. superpixel segmentation of the scene are computed. 3. Superpixels are assigned to semantic classes by majority vote. 4. and 5. Generic candidates are extracted separately in each set of superpixels. 6. Candidates are combined and ranked.

To summarize, this paper shows that semantic priors can successfully guide the generation of generic object candidates. Our main contribution is a simple yet effective algorithm that incorporates semantic prior information into the process of generic object candidate generation. We evaluate our approach on the NYU Depth V2 dataset [5]. Our experiments demonstrate that introducing semantic priors improves both recall and precision with respect to our baseline, and outperforms other state-of-the-art methods in object discovery in terms of recall.

II. Related Work

Object candidate generation methods have slowly changed the standard pipeline in computer vision for object detection: Instead of performing an exhaustive search at every possible location and scale for objects of a given category, a set of object candidates is generated where the object recognition algorithms are subsequently applied. For a survey on current methods see Hosang et al. [6]. The reduction in the number of queries that need to be made is significant: From potentially millions of queries in the sliding-window paradigm, we now have a set of only hundreds or thousands of potential objects.

Several successful object detection algorithms already use such a proposal generation stage in their pipeline [7–10]. In robotics, where the execution time is crucial, the use of object candidate generation methods becomes highly relevant, especially when facing complex realistic scenarios where the objects are not easy to segment.

A common approach for generating object candidates in the robotics community has been to find fixation points laying at the centre of objects, where a segmentation process then can extract the actual object boundaries. Examples are the method of Kootstra and Kragic [11], who use 2D symmetry points as fixation points and several Gestalt principles to rank the generated object candidates. Also Mishra and Aloimonos [12] use fixation points to locate the centre of objects and make use of the boundary-ownership concept in order to segment the objects. Potapova et al. [13] rely on symmetry points in the depth map to find the centre of objects. However, as we showed in [3], methods such as the one of Potapova et al. [13] have problems recalling objects in realistic cluttered scenes. Recently, Martín García et al. [3] proposed a method that uses saliency as a cue to locate the presence of objects and segmentation of the scene in the colour and depth modalities independently to find object precise boundaries. In this work, we build on this approach by adding semantic prior information in order to improve the candidate generation process. Horbert et al. [2], introduced the idea of sequence-level object proposals: by tracking frame-level object candidates over time, the authors could identify good candidates based on several frame and sequence-level features.

In the computer vision literature, we find several methods for generating object candidates/proposals. In a recent evaluation, Hosang et al. [6] show that one of the most successful methods in terms of recall is the Selective Search approach of Uijlings et al. [14]. The method is based on a colour segmentation of the image at different scales, and merges segments based on their similarity. The saliency-inspired convolutional neural network by Erhan et al. [10] generates class-independent bounding box proposals. The network is modeled such that it outputs a fixed number of bounding boxes and gives a confidence value for the object contained inside by directly regressing a fixed number of proposals, whereas we use the network output as a simple prior to improve the bottom-up object discovery. Other popular methods are the Objectness measure of Alexe et al. [15] and the Randomized Prim proposals of Manén et al. [16]. The problem with many of these methods is that often they require a high number of object candidates to achieve high recall values. In a robotics context, thousand potential objects are not practical if a robot has to interact with them.

Our approach for object discovery uses semantic segmentation as prior information to improve the candidate generation process. We use the semantic segmentation method of Husain et al. [4], which is a feature learning approach similar to Eigen and Fergus [17] and Long et al. [18]. Other approaches for semantic segmentation introduce hand-crafted features in their model such as gradient, colour, local binary pattern, depth gradient, spin, surface normals by Wu et al. [19] and pixel value comparison and oriented gradients by Hermans et al. [20]. Our method combines the final segmentation result and does not rely on any particular feature, hence it is compatible with any approach that clearly separates the object classes.

III. OBJECT CANDIDATE GENERATION

Our method for generating object candidates has two main components: A bottom up object discovery method and a semantic segmentation method. For bottom-up processing, we use the approach of Martín García et al. [3]. Here, saliency is used to locate the objects, while colour and depth segmentations are used to define their precise boundaries. For semantic segmentation, we chose the deep convolutional neural network approach of Husain et al. [4], which produces state-of-the-art segmentation results. The semantic segmentation induces a partitioning of the superpixel set into semantic categories, to which the object discovery method can be applied separately.

A. BOTTOM-UP SALIENCY OBJECT CANDIDATES

As a baseline, we rely on the method of Martín García et al. [3] for generating purely bottom-up generic object proposals. The approach uses saliency as a cue to estimate the location and extent of the objects, and performs segmentation in order to find their precise boundaries.

a) Saliency computation: We use the VOCUS2 [21][1] method for computing the saliency map $\mathrm{sal}(u, v)$. In VOCUS2, saliency is computed as centre-surround contrast on different feature channels and at different scales of observation. There is one feature channel for intensity and two for colour. The method uses an opponent colour space which is based on the opponent theory of human perception [22]. Contrast is computed as difference of Gaussians, which is implemented with the help of a twin-pyramid structure (one Gaussian pyramid for the centre and one for the surround). This allows for arbitrary centre-surround ratios and makes the whole system very fast (about 60 ms for a 640×480 image on a standard desktop computer).

b) Extracting salient regions: After computing the saliency map, the next step is to find the set of local maxima $\{l_1, \ldots, l_n\}$. Then, we perform region growing seeded on the maxima to extract a set of salient regions, $R = \{r_1, \ldots, r_{2n}\}$, from the saliency map. The region growing algorithm recursively explores the neighbourhood of the local maxima: For every explored pixel $p = (u_p, v_p)$, if $\mathrm{sal}(u_l, v_l) \geq \mathrm{sal}(u_p, v_p) \geq t\,\mathrm{sal}(u_l, v_l)$ holds, with $0 < t < 1$, the pixel is added to the region. This process is repeated for three values of t (we use 0.3, 0.4 and 0.5) and results in a set of salient regions R.

c) Defining boundaries with bottom-up segmentation: The actual boundaries of the objects are defined by bottom-up segments $S = \{s_1, \ldots, s_m\}$ that partition the image. Given a segmentation S, and a set of salient regions R, the algorithm now iterates for every salient region $r_i \in R$ and finds the

[1] The VOCUS2 code is available online at http://www.iai.uni-bonn.de/~frintrop/vocus2.html

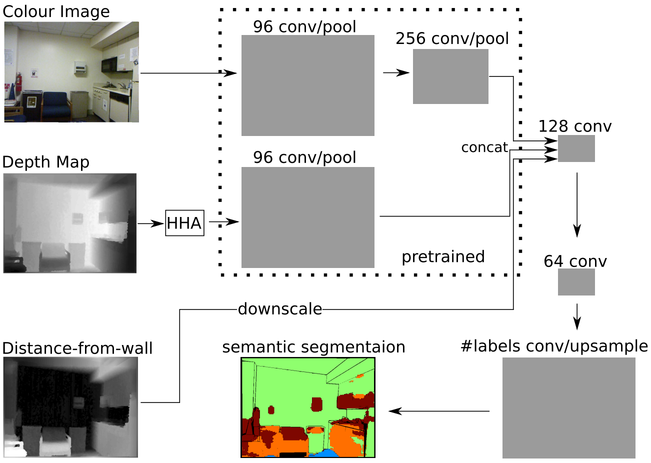

Fig. 2. Illustrating the top-down semantic segmentation as proposed by Husain et al. [4]. Inputs are the colour image and the depth map. The depth image is transformed with the HHA encoding [26]. Afterwards, feature maps are extracted using a deep network. The network parameters are learned using a pixelwise cross entropy loss.

Algorithm 1 Generate Object Proposals

```
Input: Image $I$, Depth map $D$
Input: $Y$                    ▶ A set of semantic labels, e.g. $\{furniture, prop\}$
Output: A sequence of object proposals $(C'_1, \dots, C'_{2n})$
$M_{\mathrm{TD}}$ := SemanticSegmentation($I$,$D$)
$R$ := ExtractSalientRegions($I$)
$S_1 = \{s^1_1, \dots, s^1_{m_1}\}$ := OverSegment($I$)
$S_2 = \{s^2_1, \dots, s^2_{m_2}\}$ := OverSegment($D$)
for $s^1_j \in S_1$, $s^2_k \in S_2$ do
    $y^1_j$ := MaxLabel($s^1_j$,$M_{\mathrm{TD}}$)
    $y^2_k$ := MaxLabel($s^2_k$,$M_{\mathrm{TD}}$)
for $y \in \{Y\}$ do
    for $r_i \in R$ do
        for $s^1_j \in S_1$ do
            if $y^1_j = y$ and $|r_i \cap s^1_j| > \gamma \cdot |s^1_j|$ then
                $C_i := C_i \cup \{s^1_j\}$
        for $s^2_k \in S_2$ do
            if $y^2_k = y$ and $|r_i \cap s^2_k| > \gamma \cdot |s^2_k|$ then
                $C_{n+i} := C_{n+i} \cup \{s^2_k\}$
$(C'_1, \dots, C'_{2n})$ := Ranking($\{C_1, \dots, C_{2n}\}$)
```

segments $s_i \in S$ that overlap at least a fraction γ of the area of s_i (we set it to $\gamma = 0.3$ in our experiments).

As shown in [3], the most successful approach was to generate object candidates independently with a colour and a depth segmentation of the scene. For the colour segmentation, we used the graph-based approach of Felzenszwalb and Huttenlocher [23]. The method constructs a graph based on the pixel neighbourhoods and iteratively merges groups of pixels into regions, keeping a trade-off between the internal variability of the regions and the difference between neighbouring components. To obtain surface patches based on the depth map, we use an approach similar to the first stage of Richtsfeld et al. [24]: We compute surface normals and greedily cluster them into surface patches.

d) Ranking of the proposals: The colour and depth candidates are put together and ranked according to their objectness. Several features were evaluated on the object candidates and fed into an SVM that was trained to distinguish good from bad candidates:

1-7: Hu's image moments [25]. These are well-known descriptors of shape which are rotation and scale invariant.
8: A 3D convexity measure used in [3, 13], computed as the average distance from the 3D points of the candidate to their closest face of the convex hull.
9: The normalized area of the object candidate mask.
10: The average saliency of the proposal.
11: The normalized perimeter of the object candidate.
12: The normalized average depth of the proposal.

B. *SEMANTIC SEGMENTATION*

Recently, convolutional neural networks (CNN) have emerged as state-of-the-art for semantic segmentation [17]. Their strong performance stems from using generic, multi-layered image features, which have been trained on large-scale annotated datasets [27]. We use a semantic segmentation method from Husain et al. [4], which produces state-of-the-art results on RGB-D indoor scenes.

The method uses a deep network and borrows the weights for the first two layers from another network trained on the ImageNet dataset. The architecture is shown in Fig. 2. The network takes colour and depth images and a distance-from-wall feature as input. The depth image is encoded using the geocentric HHA[2] encoding [26], which has been shown to yield superior results for object detection as compared to using the raw depth values [26]. The distance-from-wall feature is the minimum point-plane distance from the planes detected at the outermost region of a scene, saturated beyond a distance threshold. The first two layers for the colour channel and the first layer for the depth channel are borrowed from the OverFeat network [28], which is pre-trained on the ImageNet dataset. The output from the pre-trained layers and the distance-from-wall feature is concatenated and fed to a randomly initialized three-layer convolutional network. The pre-trained layers are kept fixed while the newly added layers are updated during training. During learning, we minimize the weighted, multiclass cross entropy loss [17]

$$L = \sum_{i \in \text{pixels}} \sum_{b \in \text{classes}} \alpha_b y_{i,b} \ln(\hat{y}_{i,b}), \tag{1}$$

where $\alpha_y = \text{median-freq} \,/\, \text{freq}(y)$, $\hat{y}$ is the predicted class distribution, and y is the ground truth class distribution. The factor α weighs each class y according to the frequency $(\text{freq}(y))$ with which it appears, and median-freq is the median of all the frequencies.

[2] height above the ground, horizontal disparity, and angle with respect to gravity

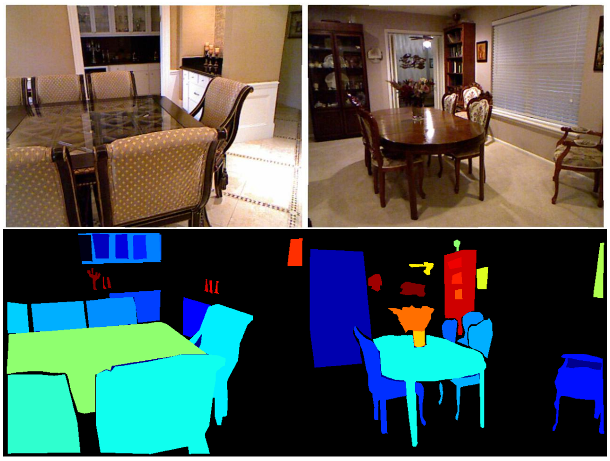

Fig. 3. Example of our ground truth. The different colours correspond to different object instances.

C. INTEGRATING SEMANTIC PRIORS INTO THE OBJECT CANDIDATE GENERATION

The output from the semantic segmentation network is a pixel-precise map containing probabilities for the semantic categories. Algorithm 1 shows how this information is combined with the bottom-up segmentations S from Section III-A.c. For each segment s_i in the bottom-up segmentation $S = \{s_1, \ldots, s_n\}$, we find the label with the highest occurrence and use it to label all the pixels in the segment.

The semantic segmentation method lets us split each bottom-up segmentation into sets of segments according to their class. The process of object proposal generation of Section III-A can now be applied to each set of segments independently. The effect is that, ideally, only segments that belong to one class are used in the candidate generation process at each time. E.g., when we generate candidates for *furniture*, segments that belong to other categories, e.g. *floor*, are not considered. Note that we do not require a one-to-one correspondence between actual object classes and the classes of the semantic segmentation algorithm; any partitioning that separates objects is likely to produce an improvement.

IV. EVALUATION

We evaluate our approach for generic object candidate generation on the NYU Depth Dataset V2[3] [5]. The dataset contains 1449 frames with dense ground truth as well as labels for the different object instances.

A. SEMANTIC SEGMENTATION

First, we evaluate our method for semantic segmentation (Section III-B). We use the NYU V2 dataset, using four classes as defined by Silberman et al. [5]: *floor*, *structure*, which denotes a permanency such as walls and ceilings, *furniture*, and *prop*, denoting movable objects. The dataset contains 795 training and 654 testing samples. The training set is used to learn the network parameters as explained in Section III-B and the testing examples are used to compute the average-per-class and pixelwise labeling accuracy. A more detailed evaluation of this method can be found in [4]. Table I shows our semantic segmentation results for the NYU V2 dataset, which are competitive with other recently proposed approaches. The results of our method for some frames of the dataset are shown in the bottom row of Fig. 5.

TABLE I
SEMANTIC SEGMENTATION ON NYU DEPTH V2

	Four-class accuracy (%)	
Method	Class average	Pixel average
Couprie et al. [29]	64.5	63.5
Khan et al. [30]	69.2	65.6
Stückler et al. [31]	70.9	67.0
Müller and Behnke [32]	72.3	71.9
Wolf et al. [33]	72.6	74.1
Eigen and Fergus [17]	*79.1*	**80.6**
Husain et al. [4] (Ours)	**79.2**	*78.0*

B. OBJECT CANDIDATES

We evaluate two variants of our proposed method (Section III-C): One using the *furniture* and *prop* classes, denoted as *Ours (2 classes)* in the plots, and the other one using the *structure*, *furniture* and *prop* labels, denoted as *Ours (3 classes)* in the plots. The reason is that the semantic segmentation method sometimes labels pieces of furniture or objects as walls, but one might want to recall those objects at the cost of some precision. We compare the two variants to our baseline (called Saliency Proposals in the plots, explained in Section III-A) and to several other state-of-the-art methods in object discovery: The depth-CPMC proposals of [34][4] , the Objectness measure of Alexe et al. [15], the Selective Search method of Uijlings et al. [14] and the EdgeBoxes method of Zitnick and Dollár [35]. The results of the last three methods were generated using the code that the respective authors provide online.

Ground truth: We generated ground truth for the *furniture* and *prop* labels, which correspond to the actual objects in the images. The dense labelling of the frames also contains an instance identifier to distinguish the different instances of the objects. We omitted classes unlikely to contain objects, i.e. *structure* and *floor*. An example of the ground truth that we used is shown in Fig. 3.

Metrics: Our evaluation protocol follows the procedure used by [2, 3]. We consider a candidate to be correct if the intersection-over-union (IoU) ratio with respect to the ground truth is greater than 0.5 [36]. This ratio can be computed in terms of pixel-precise object masks, but since some of the methods provide bounding boxes (EdgeBoxes and Objectness), we compute it for bounding boxes for all the methods. For a

[3] Available at http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

[4] We used the full set of proposals provided online: http://www.cs.toronto.edu/~fidler/projects/scenes3D.html

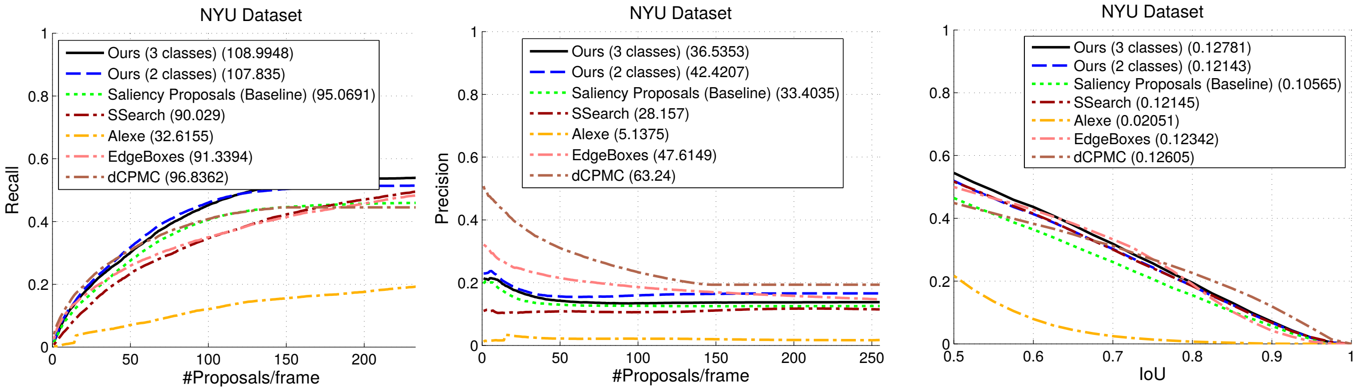

Fig. 4. Evaluation results on the NYU V2 Dataset. Left: recall over number of proposals. Middle: precision over number of proposals. Right: recall over intersection over union (IoU). The area under curve is given in parenthesis.

ground truth bounding box B_{gt} and a candidate bounding box B_p the IoU ratio is computed as

$$IoU(B_p, B_{gt}) = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})}.$$

A candidate B_p is considered correct if there is a ground truth object B_{gt} for which $IoU(B_p, B_{gt}) \geq 0.5$. Similarly, a ground truth object is matched if there is a sufficiently overlapping object candidate. Based on this, we can compute precision and recall values for every frame.

Results: We show the results of the evaluation in Fig. 4. In terms of recall (left), our proposed method (both with two and three classes) makes a significant improvement with respect to our internal baseline (the Saliency Proposals). By generating candidates for three classes (*structure*, *furniture* and *prop*) we obtain a small boost with respect to the two classes (*furniture* and *prop*): the semantic segmentation occasionally mislabels segments as *structure* when they actually belong to the *furniture* or *prop* categories; an example of this is the refrigerator in the rightmost image of Fig. 5, which is mostly labelled as *structure*. Regarding the competing methods, the Selective Search approach achieves similar recall values towards the end of the curve while Alexe et al. [15] obtains the worst results. The depth-CPMC [34] and EdgeBoxes [35] proposals obtain a recall similar to our baseline.

In terms of precision (centre of Fig. 4), depth-CPMC [34] obtains the best results (note, however, that no non-maxima-supression has been performed for them), followed by EdgeBoxes [35] and our two-classes method. Its curve is above our three-classes approach, which was predictable since in the three classes method we are additionally generating candidates for *structure*.

The quality of the object candidates is shown on the right of Fig. 4. In this plot, we show the recall as a function of the IoU ratio. We see that our three-classes method obtains a significant improvement with respect to our internal baseline and is above the other methods for the lower IoU values.

From these results, we can conclude that the use of semantic prior information to classify the bottom-up segments is beneficial in three ways: I) it leads to achieving a higher recall with fewer object candidates, II) more candidates correspond to actual objects, and III) the quality of the candidates themselves is higher. The reason is that segments from different categories are separated, and so the candidate generation process improves: e.g., if the algorithm starts generating an object candidate for a chair, it will not consider segments labelled as *structure*, which would make the candidate less precise or wrong. We show in Fig. 5 the successful candidates generated in five frames of the dataset.

V. CONCLUSION

We have proposed a method for object discovery that integrates a bottom-up mechanism for generating generic object candidates, and prior information in the form of semantic labels for improving their quality. The bottom-up component is based on saliency and segmentation, and is able to successfully locate objects. Based on a semantic segmentation method we improve the candidate generation process: Segments are now labelled according to different categories, and the bottom-up proposal generation process is performed for each category independently. Our evaluation on the NYU Depth V2 dataset shows that the proposed approach improves both recall as well as the quality of the object candidates.

REFERENCES

[1] F Husain, A Colome, B Dellen, G Alenya, and C Torras. "Realtime tracking and grasping of a moving object from range video". In: *Int. Conf. on Robotics and Automation (ICRA)*. 2014.

[2] E Horbert, G Martín García, S Frintrop, and B Leibe. "Sequence Level Object Candidates Based on Saliency for Generic Object Recognition on Mobile Systems". In: *Int. Conf. on Robotics and Automation (ICRA)* (2015).

[3] G Martín García, E Potapova, T Werner, M Zillich, M Vincze, and S Frintrop. "Saliency-based Object Discovery on RGB-D Data with a Late-Fusion Approach". In: *Int. Conf. on Robotics and Automation (ICRA)* (2015).

[4] F Husain, H Schulz, B Dellen, C Torras, and S Behnke. "Combining Semantic and Geometric Features for Object-class Segmentation of Indoor Scenes". In: *IEEE Robotics and Automation Letters* (2016).

[5] N Silberman, D Hoiem, P Kohli, and R Fergus. "Indoor Segmentation and Support Inference from RGBD Images". In: *Europ. Conf. on Computer Vision (ECCV)*. 2012.

[6] J Hosang, R Benenson, P Dollár, and B Schiele. "What makes for effective detection proposals?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2015).

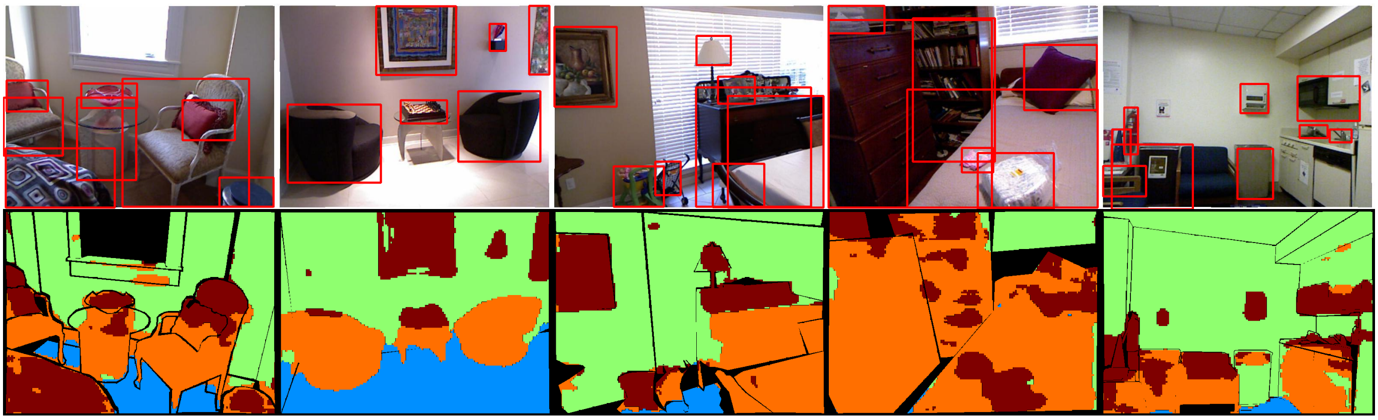

Fig. 5. Top row: example of the candidates that successfully retrieved objects using our proposed method in four frames of the NYU Depth V2 Dataset. Bottom row: the corresponding semantic segmentation obtained in those frames.

[7] RG Cinbis, J Verbeek, and C Schmid. “Segmentation driven object detection with Fisher vectors”. In: *Int. Conf. on Computer Vision (ICCV)*. 2013.

[8] R Girshick, J Donahue, T Darrell, and J Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Computer Vision and Pattern Recognition (CVPR), Conf. on*. 2014.

[9] K He, X Zhang, S Ren, and J Sun. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *Europ. Conf. on Computer Vision (ECCV)*. Springer, 2014.

[10] D Erhan, C Szegedy, A Toshev, and D Anguelov. “Scalable object detection using deep neural networks”. In: *Computer Vision and Pattern Recognition (CVPR), Conf. on*. 2014.

[11] G Kootstra and D Kragic. “Fast and bottom-up object detection, segmentation, and evaluation using Gestalt principles”. In: *Int. Conf. on Robotics and Automation (ICRA)*. 2011.

[12] AK Mishra and Y Aloimonos. “Visual Segmentation of Simple Objects for Robots”. In: *Robotics: Science and Systems (RSS)*. 2011.

[13] E Potapova, KM Varadarajan, A Richtsfeld, M Zillich, and M Vincze. “Attention-driven Object Detection and Segmentation of Cluttered Table Scenes using 2.5 D Symmetry”. In: *Int. Conf. on Robotics and Automation (ICRA)*. 2014.

[14] J Uijlings, K van de Sande, T Gevers, and A Smeulders. “Selective Search for Object Recognition”. In: *Int. Journal of Computer Vision (IJCV)* (2013).

[15] B Alexe, T Deselaers, and V Ferrari. “Measuring the Objectness of Image Windows”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 34.11 (2012).

[16] S Manén, M Guillaumin, and L Van Gool. “Prime Object Proposals with Randomized Prim’s Algorithm”. In: *Int. Conf. on Computer Vision (ICCV)*. 2013.

[17] D Eigen and R Fergus. “Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture”. In: *CoRR* abs/1411.4734 (2014).

[18] J Long, E Shelhamer, and T Darrell. “Fully Convolutional Networks for Semantic Segmentation”. In: *Computer Vision and Pattern Recognition (CVPR), Conf. on*. 2015.

[19] C Wu, I Lenz, and A Saxena. “Hierarchical Semantic Labeling for Task-Relevant RGB-D Perception”. In: *Robotics: Science and Systems (RSS)*. 2014.

[20] A Hermans, G Floros, and B Leibe. “Dense 3D semantic mapping of indoor scenes from RGB-D images”. In: *Int. Conf. on Robotics and Automation (ICRA)*. 2014.

[21] S Frintrop, T Werner, and G Martín García. “Traditional Saliency Reloaded: A Good Old Model in New Shape”. In: *Computer Vision and Pattern Recognition (CVPR), Conf. on*. 2015.

[22] L Hurvich and D Jameson. “An opponent-process theory of color vision”. In: *Psychological review* 64.6 (1957).

[23] PF Felzenszwalb and DP Huttenlocher. “Efficient Graph-Based Image Segmentation”. In: *Int. Journal of Computer Vision (IJCV)* (2004).

[24] A Richtsfeld, T Morwald, J Prankl, M Zillich, and M Vincze. “Segmentation of unknown objects in indoor environments”. In: *Intelligent Robots and Systems (IROS), Int. Conf. on*. 2012.

[25] MK Hu. “Visual pattern recognition by moment invariants”. In: *IRE Transactions on Information Theory* 8.2 (1962).

[26] S Gupta, R Girshick, P Arbelaez, and J Malik. “Learning Rich Features from RGB-D Images for Object Detection and Segmentation”. In: *Europ. Conf. on Computer Vision (ECCV)*. 2014.

[27] M Oquab, L Bottou, I Laptev, and J Sivic. “Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks”. In: *Computer Vision and Pattern Recognition (CVPR), Conf. on*. 2014.

[28] P Sermanet, D Eigen, X Zhang, M Mathieu, R Fergus, and Y LeCun. “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks”. In: *Int. Conf. on Learning Representations (ICLR)*. 2014.

[29] C Couprie, C Farabet, L Najman, and Y LeCun. “Indoor Semantic Segmentation using depth information”. In: *Int. Conf. on Learning Representations (ICLR)*. 2013.

[30] S Khan, M Bennamoun, F Sohel, and R Togneri. “Geometry Driven Semantic Labeling of Indoor Scenes”. In: *Europ. Conf. on Computer Vision (ECCV)*. Vol. 8689. 2014.

[31] J Stückler, B Waldvogel, H Schulz, and S Behnke. “Dense real-time mapping of object-class semantics from RGB-D video”. In: *Journal of Real-Time Image Processing* (2013).

[32] A Müller and S Behnke. “Learning depth-sensitive conditional random fields for semantic segmentation of RGB-D images”. In: *Int. Conf. on Robotics and Automation (ICRA)*. 2014.

[33] D Wolf, J Prankl, and M Vincze. “Fast semantic segmentation of 3D point clouds using a dense CRF with learned parameters”. In: *Int. Conf. on Robotics and Automation (ICRA)*. 2015.

[34] D Lin, S Fidler, and R Urtasun. “Holistic scene understanding for 3d object detection with rgbd cameras”. In: *Int. Conf. on Computer Vision (ICCV)*. 2013.

[35] CL Zitnick and P Dollár. “Edge boxes: Locating object proposals from edges”. In: *Europ. Conf. on Computer Vision (ECCV)*. Springer, 2014.

[36] M Everingham, L Van Gool, CK Williams, J Winn, and A Zisserman. “The pascal visual object classes (VOC) challenge”. In: *Int. Journal of Computer Vision (IJCV)* 88.2 (2010).