Detecting simple objects in RGB-D data

Anonymous CVPR submission

Paper ID ****

Abstract

In this paper we present an approach for detection of simple objects in RGB-D data. Object detection in clutterred indoors environments is an important perceptual capability of robotic systems required for object search and pick and deliver tasks. For long term autonomy robots should learn how objects look like and where they appear in an weakly supervised manner. In this work we exploit the depth information to provide evidence about occlusion boundaries and scale of the objects. The depth discontinuities along with image contours computed in the vicinity of the detection window boundary form an objectness *measure, which is used to train an SVM classifier. In the testing stage we exploit the knowledge of the actual size of the object to propose the scale of the detection window significantly pruning the number window candidates to be evaluated. We evaluate our approach for detecting simple objects on NYU RGB-D dataset, illustrate the effectiveness of our approach as well as difficulties with the standard evaluation methodologies.*

1. Introduction

With the advent of RGB-D cameras in recent years several approaches towards object detection, semantic segmentation and activity recognition as well as more general scene understanding have been developed [14, 12, 9]. The proposed approaches demonstrated the improved performance compared to purely image based methods thanks to availability of the depth data . Due to the range limitations of the sensor, most of the proposed methods focus on indoor environments. Different datasets have been proposed by researchers which are used for evaluation of respectively for semantic segmentation [16, 11], object detection and categorization [12] and localization [17]. These problems and class of environments commonly considered have are closely motivated by issues related to robot perception.

In order to enable long term robot autonomy and facilitate the more sophisticated robotics tasks, it is important that robots can localize objects at different scales in cluttered environments. In robotic setting the capability of generating hypotheses about presence of objects with particular aspect ratio and of particular size is of interest for tasks like object search, which precedes closer categorization, more detailed segmentation and manipulation. Hence considering this capability in the context of object search, it is also reasonable to assume that the actual size of the object to be located is known.

Figure 1. (a) Example scenes, with small simple objects and their bounding boxes from NYY RGB-D V1 dataset (b) Ground Truth labeling associated with the dataset, focuses typically on large regions. Many small objects are missed.

The goal of this paper is to advance the state of the art of detection of simple objects in cluttered RGB-D scenes. We consider simple objects where the apparent size of the object is possibly small and object's bounding box approximates well the extent of the object. Some related works approach this problem by means of semantic segmentation of the entire image, or use models of human attention to generate possible hypotheses about object location and size. In our approach we pursue the sliding window approach to object detection and make the following contributions: (i) We define an objectness measure computed over windows of both images and depth maps and use it to train a SVM classifier for scoring the windows as object or background; (ii) The classifier is trained on all bounding boxes regardless

of the object size and aspect ratio; (iii) In the detection stage we sample the actual object sizes to determine the scale of the window, significantly pruning the number of windows which need to be evaluated. The proposed approach is evaluated on a subset of scenes in NYU RGB-D V1 dataset, demonstrating the performance of the detection, compared to ground truth labeling.

2. Related work

The proposed work is related to several areas of research including semantic segmentation, object detection and saliency detection. While there is a large body of approaches which study these problems in the context of images only, we will focus here on the methods which exploit the depth information.

As mentioned before the nature of the datasets used to evaluate approaches to semantic segmentation and object detection and segmentation differ in their characteristics. The most important one is the scale at which objects appear in images. A successful approach to object detection in RGB-D data was proposed in the work on [13], where the objects are viewed in a table top setting at moderate scale. The authors formulated the object detection problem as an inference on a voxel grid, reconstructed from multiple frames of RGB-D data. The final inference is carried in MRF framework, where the data term accumulates evidence from the sliding window based detectors trained on different views of the objects. A variant of the HOG descriptor [12] was used for capturing the appearance and shape information of each view of an object and trained using SVMs. The outputs of multiple HOG detectors and multiple views were then combined to generate the score of the object presence at each 3D point. Additional features computed from the depth channel were used in the pairwise term of MRF model which further improved the object localization capability. The larger extent of the objects in the dataset [12] and sufficient number of training examples made the use of HOG detector feasible. Another related work on unsupervised object discovery [3] has shown promising results for closer range and small amount of clutter.

In the presented work we focus on the localization of simple objects in cluttered scene, such as the one depicted in Figure 1. Instead of striving to achieve complete semantic segmentation of these types of scenes as in [16, 9], we instead want to generate simple object hypotheses. The notion of a simple object here is the type of object whose shape can be well delimited by a bounding box. Our work is most closely related to work of [1] who considers the problem of detection of generic simple objects in an unsupervised setting. Authors in [1] use the computation of the boundary using both RGB and depth data, followed by a selection of salient points and boundary completion. This methods is very effective on closer range table top settings, where both depth discontinuities and support surfaces can be well estimated and the process of detection of image contours is more reliable. Their methods relies on a high quality contour detector [7], which is quite expensive to compute. While the produced contours are of high quality, the subsequent processing steps rely on more accurate depth estimates and supporting surfaces, which are harder to attain with varying viewpoints and far distance. With the change of scale of depth measurements, in many instances the depth measurements are missing and due to the common use of image in painting techniques the intensity and depth boundaries are not well aligned, making the countour based segmentations techniques very brittle.

Our approach is closely related and motivated by work of [2], who proposed a method for generic object detection in natural images. Authors in [2] pursue sliding window approach and learn how to classify generic backgrounds from object categories using cues characterizing the length of the contour close to the boundary sliding window, saliency measure and difference between color histograms in the outside and inside of the bounding box. The features are combined in Bayesian framework and greedy search over high scoring windows of all aspect rations and scales is proposed to select the top candidates. The approach performs well on the detection of isolated and often small number of objects in outdoors scenes (as tested in on PASCAL-VOC dataset). In indoors settings due to large amount of clutter color contrast feature is not so effective and the window scoring strategy along with greedy approach tends to selects windows of bigger size, missing smaller objects. In our work we also use the idea of presence of the contour close to object (window) boundary, but enhance the features by considering also the depth gradients, which are indicative of occluding contours. Instead of performing a greedy search, we use the depth information to select the scale of the window over which the score are computed, hence by passing the search over all possible aspect ratios and scales.

Another class of methods formulates the problem of object detection using over segmentation as initial representation and combines local evidence such as shape, appearance with pairwise interactions between regions in a MRF framework. [15, 11, 9]. The segmentation based approaches deal with imperfect segmentation by generating multiple segmentations and aggregating their results to form hypothesis about regions.

Biologically motivated approaches towards object detection use as starting point various saliency measures which are then enhanced using top-down information, or boosted using evidence from human attention models. In Itti [8] the problem of saliency object detection is studied jointly with the object search problem, where a model for combining bottom and top down cues is investigated. The idea of combining high level concepts and low level features to improve

current saliency models as well as to scale up current models to reach the human performance has been explored in the work of [5]. More recently the role of depth information in bottom up saliency models have been studied in [6] demonstrating that the availability of depth information affects human fixation. Authors propose to incorporate novel depth saliency priors to augment existing approaches which used only appearance information.

The goal of our work is to generate hypotheses about presence of objects in cluttered scenes. Examples of such scenes can be found in Figure 1. While there is large variety of objects and object classes, we are interested in detecting smaller objects which could possibly be manipulated. Note the scenes have large amount of clutter and large variety of objects appearing in them. For our experiments we use NYU RGB-D dataset [16], which as been introduced recently in the context of semantic segmentation.

3. Approach

In this section we describe the choice of the features and method for scoring of the candidate windows. Similarly as [2] our approach exploits the observation of the presence of the object boundary in the vicinity of the bounding box. This assumption is reasonable provided that the objects are relatively simple shapes, and majority of the true object boundary is close to the bounding box of the object (Figure 2).

Features We compute the gradient orientations in four blocks depicted in Figure 2c obtained by shrinking the bounding box boundary by θ_{bar} % of the size of the bounding box. The value $\theta_{bar} = 10\%$ has been used in current experiments. We also enlarge the windows of the ground truth bounding boxes by 5 pixels to mitigate some of the labeling errors. In each block we quantize the orientations from (0°- 360°) into 9 orientation bins. This is done both for intensity and depth channel yielding a $2 \times 4 \times 9 = 72$-dimensional feature. Prior to histogram computation, we normalized the gradients by a total energy in the bounding box. Average gradients of the ground truth windows, for a particular aspect ratio are visualized in Figure 3 along with average gradients of the windows used as negative examples.

Object size We exploit in our approach the availability of the depth data in order to properly model the expected scale. The distributions of object sizes as well as aspect ratios are learned from the training data and we use this prior knowledge to speed up the process of windows sampling in testing stage. In our experiment less than 10,000 candidate windows are generated for the entire image at full resolution of 480×640. We firstly discretize the aspect ratios available in the training data into 10 bins. For any pixel (x, y) in the image, the corresponding point in the world coordinate (X, Y) can be obtained as $X = \frac{f}{Z}x$, and

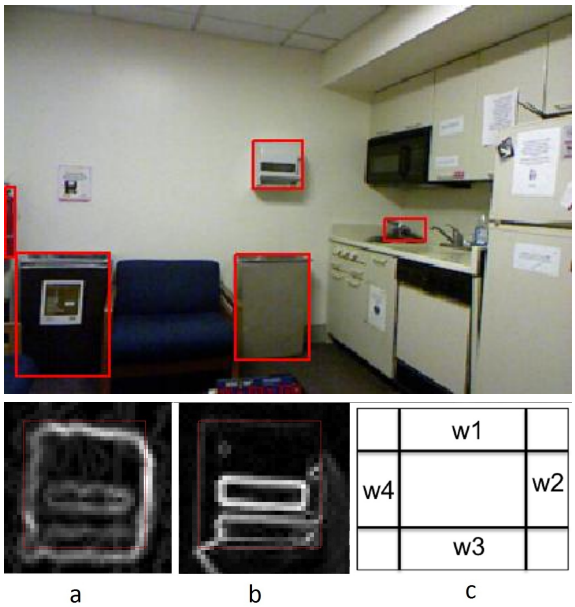

Figure 2. Examples of objects and their bounding boxes and close-up of the orientation energy for both intensity and depth channel. a) b) orientation energy for paper towel dispenser, where the image gradients and depth gradients complement each other well; c) an example of an object the strong orientation energy in the vicinity of the boundary occurs only at few locations.

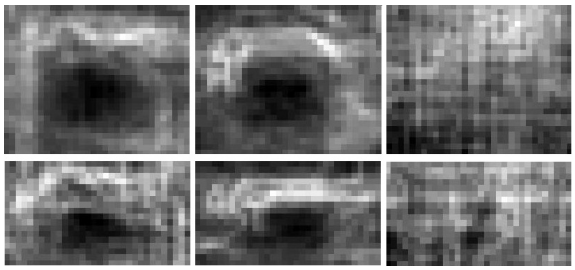

Figure 3. First and second columns are average depth gradients of positive examples from kitchen and bathroom datasets respectively. The last columns are the gradients computed over negative examples. The two rows visualize the averages for two different aspect ratios.

$Y = \frac{f}{Z}y$ using the median depth value Z in the bounding box. So for any two image points (x_1, y_1), (x_2, y_2) we have $\delta x = x_1 - x_2 = \frac{Z}{f}(X_1 - X_2) = \delta X$. This means the scale of an object at some distance can be determined by its aspect ratio and depth. For each aspect ration bin all possible object sizes are found by agglomerative clustering. One example of generated windows at some positions are shown in Figure 4, from which we can notice the effectiveness of our approach.

4. Experimental Setup

We carry out our experiment on NYU RGB-D V1 [16] dataset, which contains 7 different scene classes which in

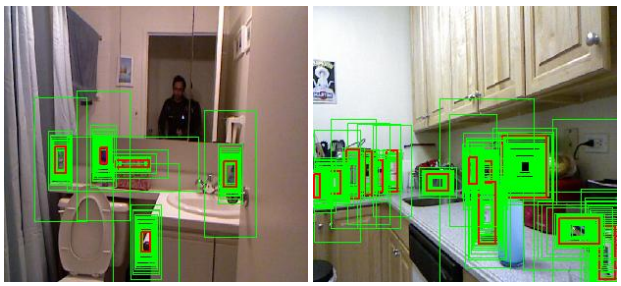

Figure 4. Candidate windows in some locations, the red box is the ground truth of an object while those green ones are proposed by our approach. Left is a bathroom scene and right is a kitchen scene

total has 64 scenes and 108617 frames. In the reported results, we only focus on the bathroom and kitchen scenes, which contain many simple objects (e.g. containers) . By filtering out those frames whose scene class has been wrongly assigned, we get 70 frames of 6 scenes for bathroom, and 276 frames of 10 different kitchen scenes.

The NYU dataset is typically used for evaluation of approaches for semantic segmentation. As a consequence many small objects are not labelled and in many cases the location of bounding boxes is not accurate and some bounding boxes are entirely missing. The labels are coarse (many objects are missing) and inaccurate (A frame with its labels is given in Figure 1). Secondly, the number of labeled objects is very small, which is insufficient for training. For the presented evaluation we firstly filter out the non-object labels and keep the remaining regions and their associated labels. In order to get larger number of training examples, we then sample around the ground truth bounding boxes to obtain more positive training examples. The negative examples are generated by uniformly sampling in the entire image (100,000) and filtering out those having a high overlapping with object windows (PASCAL score greater than 0.5). Finally, we get around 20,000 to 40,000 negative examples for one frame.

4.1. Training Stage

For each setting, 2/3 of the examples is randomly selected for training. Descriptors are computed on both positive and negative examples. We evaluate the performance of the proposed image descriptor, depth descriptors and concatenation of them. For classification we used SVM with intersection kernel. Because the number of negative and positive examples is unbalanced, for instance, there might be about 25,000 negative examples but only about 100 positive ones in a frame, so in this stage we experimented different ratio of the number of negative and positive examples. As expected, the more negative examples, the higher true negative rate. But the true positive rate decreases as a result although not dramatically. So in testing stage, we use the classifier learned with balanced number of positive and negative examples. In balanced case, we have 2041 positive examples and 2058 negative examples for kitchen setting; and for bathroom setting we have 1779 positive examples and 1758 negative examples.

bathroom	**TPR**	**TNR**	**PPV**	**NPV**
RGB	0.6206	0.8749	0.5708	0.8959
Depth	0.6257	0.8321	0.4996	0.8924
RGB-D	**0.6166**	**0.9125**	**0.6537**	**0.8988**
kitchen				
RGB	0.8423	0.8403	0.2407	0.9888
Depth	0.8162	0.8442	0.2395	0.9871
RGB-D	**0.856**	**0.8817**	**0.3031**	**0.9903**

Table 1. Testing results of different training models for bathroom and kitchen.

4.2. Testing Stage

Our experiments consists of two parts. Firstly, we evaluate our classifier on the object and non-object windows in the test data. Then the evaluation is performed on all frames, with the windows proposed by our algorithm.

Testing on known windows. For each frame in test data, we compute descriptors for both positive and negative windows, then we reported the true positive rate (TPR), true negative rate (TNR), positive prediction value (PPV) and nagetive prediction value(NPV) of our classifier. Also the precision/recall curve on testing data is reported. The results are shown in Figure 5 for bathroom and kitchen scenes. We can clearly notice that the performance is improved when combining depth data with RGB image. The quantitative results are given in Table 1.

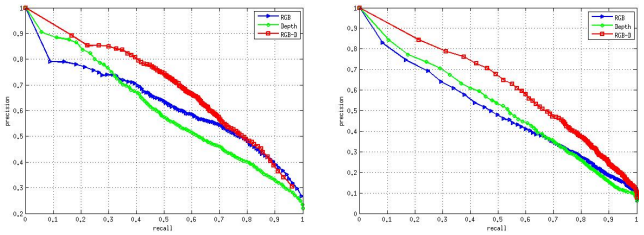

Figure 5. Precision/Recall curves for models trained on RGB only, depth only and both. Left is the result of bathroom scene, right is that of kitchen scene.

Testing on proposed windows In the detection stage traditional approaches [2] examine all possible window aspect ratios and all possible scales by generating increasingly complex scoring functions and greedily selecting the candidates in the subsequent steps. In our case we use the learned distribution of actual object sizes, to determine the apparent sizes of windows to be scored at selected locations. To further reduce the amount of locations visited we first oversegment the RGB image into superpixels. We have used two different oversegemtation strategies [10] and [18] on the

order of < 1000 small superxpixels. At the center of every superpixels the windows are generated according to the learned distribution of aspect ratios and scales. The results on 4 bathroom scenes and 10 kitchen scenes are shown in Figure 6, where the odd columns are ground truth and every even columns are our results. Our approach tends to detect all the small objects in the frame, although in some cases some objects are not labeled in the ground truth, an example is shown in right image of last row in Figure 6.

In order to evaluate the accuracy of the proposed approach, 25% of boxes ground truth boxes have been correctly detected by our approach. For evaluation we use the PASCAL VOC (intersection/union) score of 0.5. Despite the apparent improvement while visually examining the results, there are several reasons for low values of the score. For small objects the PASCAL criterion of 0.5 is rather strict and it is often the case that the location of many ground truth bounding boxes have errors which exceed the score. Another side effect of the ground truth labeling is the fact that many objects are labelled as a group and many objects which we successfully detect are not labelled at all. We also suffer at certain locations from errors in misalignment of image and depth boundaries which are due to in painting algorithms used to fill the missing values. In the supplemental material [4] we present a comparison with the existing approaches for object detection [2] and [1] using the code made available by authors. We also present comparison our the sliding windows based methodology with bottom up saliency based methods such as the methods used in [2] to select initial windows. Since in most of these methods adopt the notion of saliency of local neighborhood, by measuring the difference from the surroundings, the presented examples clearly demonstrate the problem of these methods in cluttered environments.

5. Conclusions

We have presented a method for detecting simple objects in in cluttered scenes using RGB-D data. In order to overcome the brittleness of the boundary based methods (both depth and image), we propose to adopt a discriminative approach using intensity and depth gradient features computed in the vicinity of the bounding capturing the notion of closed boundary. We evaluate the feasibility of the objectness measure on the bounding boxes selected from the NYU RGB-D Dataset, which is typically used for evaluation of semantic segmentation and considers many of the small objects as part of the background. In the actual object detection stage, we presented a method for exploiting the available depth information for determining the apparent size of the objects and significantly pruning the number of window candidates which need to be evaluated. The presented approach shows promising results as well as point out many open problems with the current evaluation pipelines and ground truth datasets. Furher improvements can be achieved by incorporating additional features and other types of contextual informations present in inodores environments.

References

[1] A. S. Ajay K. Mishra and Y. Aloimonos. Segmenting simple objects using rgb-d. In *IEEE Conference on Robotics and Automation*, 2012. 2, 5

[2] B. Alexe, T. Deselaers, and V. Ferrarii. What is an object ? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2, 3, 4, 5

[3] M. H. Alvaro Collet, Siddhartha Srinivasa. Structure discovery in multi-modal data: a region-based approach. In *IEEE Conference on Robotics and Automation*, 2011. 2

[4] Anonymous. Supplementary material. In *CVPR*, 2013. 5

[5] A. Borji and L. Itti. Boosting bottom-up and top-down visual features for saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3

[6] T. V. N. e. a. C. Lang. Depth matters: Influence of depth cues on visual saliency. In *Proc. of European Computer Vision Conference*, 2012. 3

[7] a. D.Martin, C.Fowlkes. Learning to detect naturalimage boundaries using local brightness, color and texture cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004. 2

[8] L. Elazary and L. Itti. A bayesian model for efficient visual search and recognition. *Vision Research*, 50:1338–1352, 2010. 2

[9] R.-D. S. L. Features and Algorithms. Xiaofeng ren, liefeng bo, and dieter fox. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2

[10] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *Int. Journal on Computer Vision*, 59(2):167–181, 2004. 4

[11] H. S. Koppula, A. Anand, T. Joachims, and A. Saxena. Semantic labeling of 3d point clouds for indoor scenes. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 244–252. 2011. 1, 2

[12] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *IEEE International Conference on on Robotics and Automation*, 2011. 1, 2

[13] K. Lai, L. Bo, X. Ren, and D. Fox. Detection based object labelling in 3d scenes. In *IEEE Conference on Robotics and Automation*, 2012. 2

[14] P. K. N. Silberman, D. Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. of European Computer Vision Conference*, 2011. 1

[15] I. S. S. of Objects via multiple segmentation. T. malisiewicz and a. efros. In *Proc. of British Machine Vision Conference*, 2007. 2

[16] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*, 2011. 1, 2, 3

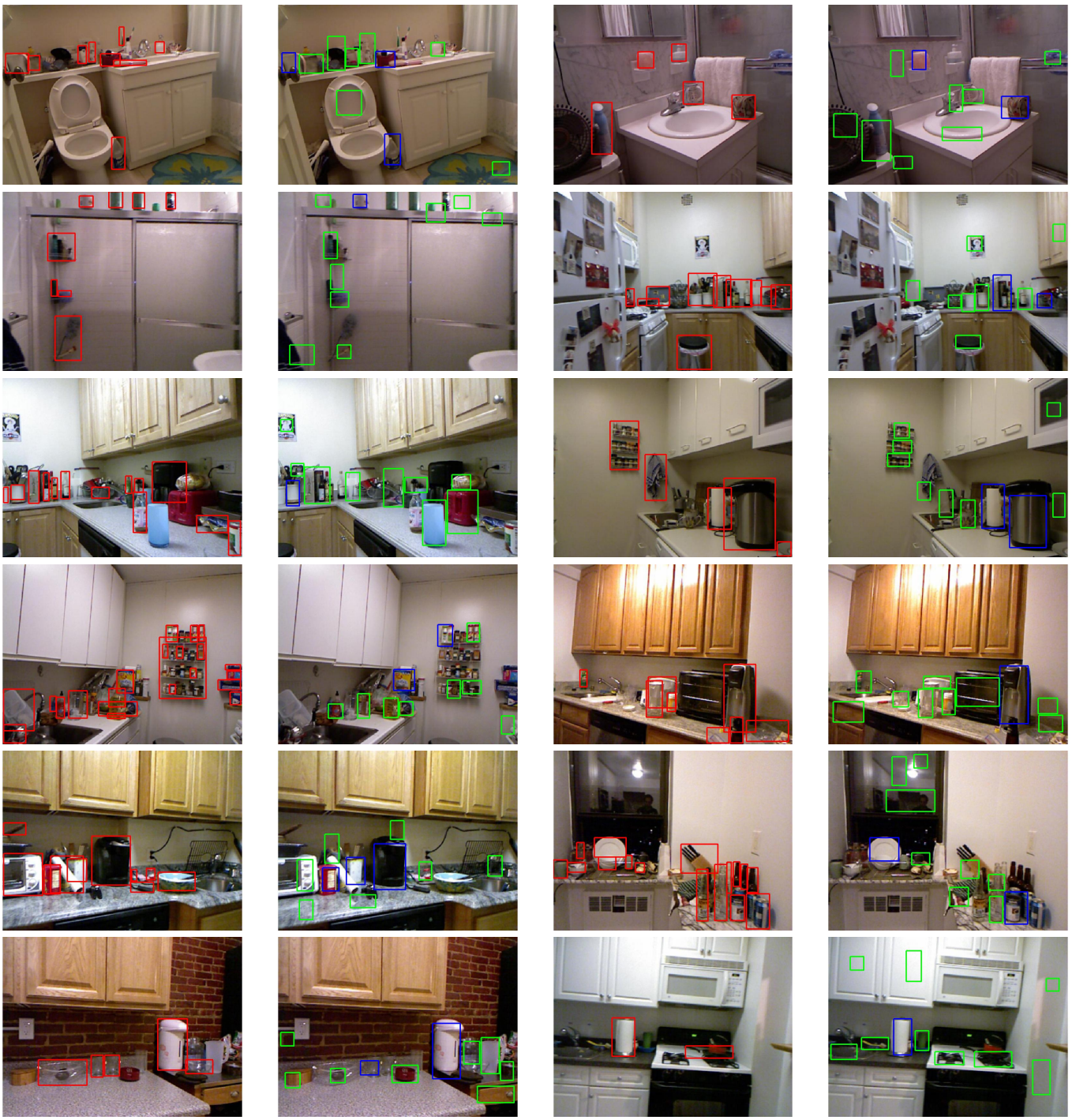

Figure 6. Results on different scenes, left column is the ground truth; right column is 10 detected objects with highest score, and those having pascal score greater than 0.5 are marked blue.

[17] J. Sturm, S. Magnenat, N. Engelhard, F. Pomerleau, F. Colas, W. Burgard, D. Cremers, and R. Siegwart. Towards a benchmark for rgb-d slam evaluation. In *Proc. of the RGB-D Workshop on Advanced Reasoning with Depth Cameras at Robotics: Science and Systems Conf. (RSS)*, Los Angeles, USA, June 2011. 1

[18] H. Wildenauer, B. Micusík, and M. Vincze. Efficient texture representation using multi-scale regions. In *ACCV (1)*, pages 65–74, 2007. 4