DEPTH-AWARE LAYERED EDGE FOR OBJECT PROPOSAL

Jing Liu[1], *Tongwei Ren*[1,*], *Bing-Kun Bao*[1,2], *Jia Bei*[1]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, China
jingliu@smail.nju.edu.cn, rentw@nju.edu.cn, bingkun.bao@ia.ac.cn, beijia@nju.edu.cn

ABSTRACT

Object proposal, typically served as preprocessing of various multimedia applications, aims to detect the bounding boxes of possible objects in an image. In this paper, we propose a novel object proposal method for RGB-D images based on layered edges, which can effectively eliminate the influence of the mixture of edges from objects and background and improve the accuracy of proposals. Firstly, we detect the sparse edges and correct depth on super-pixel representation. Then, we use depth-adaptive sliding windows in sampling of depth distribution and measure the objectness of each candidate box in multiple depth layers. Finally, the candidate boxes are ranked according to the integrated scores of all the depth layers, and the final proposals are generated. The experimental results show that the proposed method can outperform the state-of-the-art methods on the largest RGB-D image dataset for object proposal.

Index Terms— Object proposal, layered edge, depth correction, window scoring

1. INTRODUCTION

Object proposal aims to detect bounding boxes of class-independent objects in an image, which has been widely used as the fundamental for various multimedia applications such as object detection, tracking and retrieval [1–3]. To serve as an effective preprocessing, object proposal is required to generate bounding boxes with high accuracy for most existing objects within a given image by providing considerably small number of proposals [4, 5]. Moreover, it should also be efficient enough so as to well facilitate the subsequent processing [6].

Currently, the existing object proposal methods can be roughly categorized into two major paradigms, window scoring based methods and grouping based methods [6]. Window scoring based methods typically use sliding windows to sample candidate boxes, and measure the objectness, i.e., the likelihood to enclose an object, of each candidate box [7–9]. Grouping based methods usually over-segment the given image into super-pixels or regions, and group the segments according to their similarities. Compared to grouping based methods, window scoring based methods usually have higher efficiency, which are more suitable to serve as preprocessing for other applications, but they easily suffer the problem in proposal accuracy, i.e., they cannot provide acceptable proposals under high intersection over union (IoU) requirement [6].

Various features have been explored to improve the boundary box accuracy in object proposal, such as with color [7], saliency [10] and edge [8, 9]. Among these features, the effectiveness of edge has been proved in window scoring based object proposal methods [8, 9], for it can indicate object boundaries which play an important role in object estimation of human visual system [11]. However, current edge-based object proposal methods cannot discriminate the edges from objects and background within the candidate boxes, which may lead to inaccuracy in objectness measurement. Fig. 1 shows an example of the drawback of the mixture in edges from objects and background. When measured by a representative edge-based method, EdgeBoxes [8], a candidate box (green) obtains higher score than the ground-truth bounding box (red) due to the influence of the edges in background (Fig. 1(b)).

In this paper, we propose a novel object proposal method for RGB-D images based on layered edges. When viewing a scene, human vision system quickly moves fixation points in different depths by changing the diopter of eyes through lens adjustment, instead of completely capturing all the information in the scene [11]. Inspired by this characteristic, our method counts the edges in different depth ranges, scores each candidate box in each layer separately and combine the scores to measure the objectness of the candidate box (Fig. 1(c) and (d)). Fig. 2 shows an overview of the proposed method. We first use structured edge detector [12]

This work is supported by National Science Foundation of China (61321491, 61572503, 61202320), Beijing Natural Science Foundation (4152053), Natural Science Foundation of Jiangsu Province (BK20130588), Research Fund of the State Key Laboratory for Novel Software Technology at Nanjing University (ZZKT2016B09), National Undergraduate Innovation Project (G1410284074), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

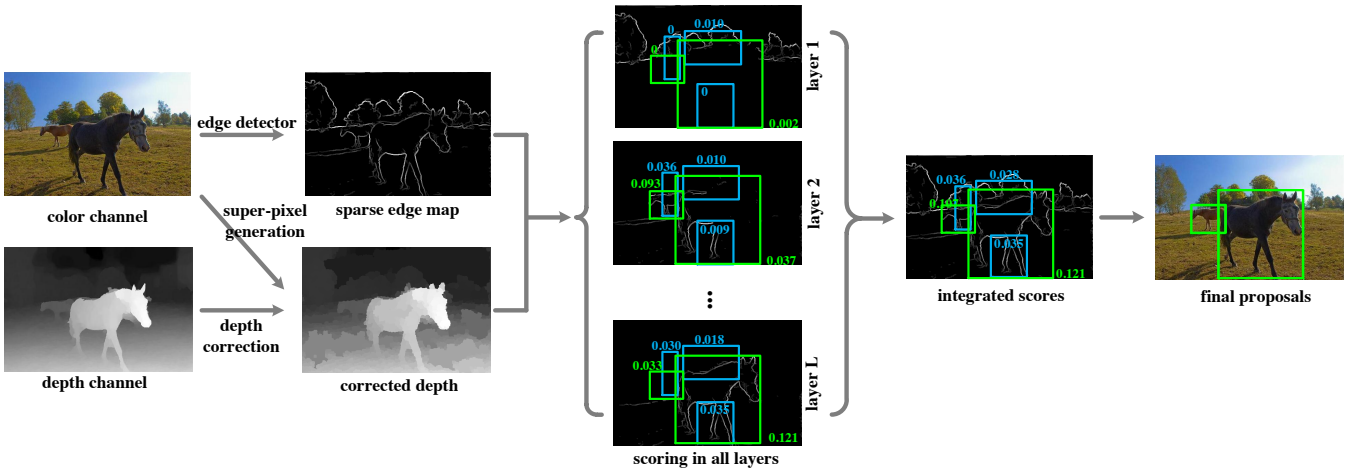

Fig. 2. An overview of the proposed method.

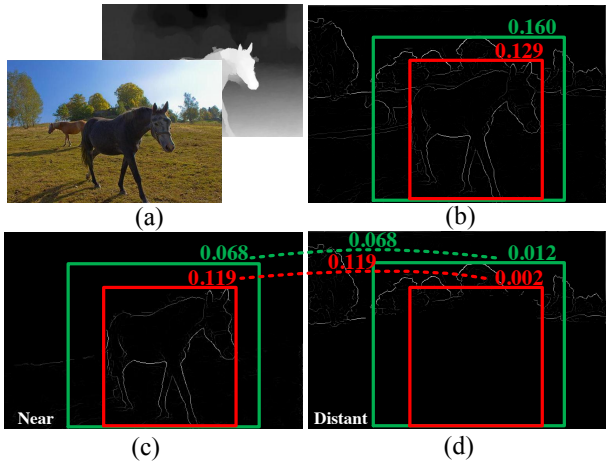

Fig. 1. An example of the drawback in mixing the edges from objects and background. (a) RGB-D image. (b) A candidate box (red) and the corresponding ground truth bounding box (green) with the scores measured by EdgeBoxes [8]. (c) and (d) The candidate box (red) and the corresponding ground truth bounding box (green) with the scores measured by our method in different depth layers.

to obtain the sparse edge map with No-Maximal Suppression performed. Then super-pixel is introduced to assist in correcting the depth maps. According to corrected depth map, we effectively divide the edges into several layers and perform window scoring in individual layers by calculating the magnitude and weight of edges wholly included in each box. Finally, the ultimate score of each box is determined by the scores inherent in layers and the weight computed by the depth of each layer. In this way, the edges in objects and background can be discriminated and the objectness measurement will be not misled. The method is validated on a public RGB-D image dataset for object proposal [13]. The experimental results show that it outperforms the state-of-the-art methods, especially for high accuracy requirement.

The rest of the paper is organized as follows. In Section 2, we briefly review the related work to the proposed method. In Section 3 and 4, we present the details of our method, and validate its performance by comparing with the state-of-the-art object proposal methods on *NJU1500* dataset. Finally, we conclude the paper in Section 5.

2. RELATED WORK

2.1. Typical Object Proposal Paradigms

There are two major strategies of object proposal, including window scoring based methods and grouping based methods.

Window scoring based methods. Window scoring based methods initially generate a large number of candidate boxes through some sampling strategy, and then apply some measures to judge how likely a single box seems to exactly contain an object. Alexe *et al.* [7] first introduce the concept of objectness measurement and put forward a measurement computed by multiple appearance and geometry properties. Feng *et al.* [10] follow the sliding window strategy and propose a new measurement of saliency to score each location. Cheng *et al.* [9] train a linear classifier model over binarized normed gradient of edge features. Zitnick *et al.* [8] utilize structured edge detector to estimate object boundary and refine the initial boxes to improve the localisation. Window scoring based methods usually have high efficiency, but they are weak in providing the proposals with high IoU.

Grouping based methods. Grouping based methods is usually initialized with over-segmentation which generates quantities of segments like super-pixels. Carreira *et al.* [14] utilize several seeds for computing graph cuts in order to avoid initial segmentation and rank the resulting segments

by a large amount of features. Humayun [15] improve it by adopting edge detectors and applying multiple graph cuts. Uijlings *et al.* [16] propose Selective Search which greedily merge the super-pixels. Wang *et al.* [5] apply multi-branch hierarchical segmentation in Selective Search process and achieve improvement. Xiao *et al.* [4] also improve it through specifying the super-pixel merging in high-complexity scene. Long *et al.* [17] train a supervised model to greedily adjust the boxes after obtaining initial boxes through bottom-up merging. Arbelaez *et al.* [18] propose multiscale combinatorial grouping. Krähenbühl *et al.* [19] start from over-segmentation and use classifiers to place seeds in order to perform geodesic transform, so that object proposals are defined by level sets of each distance transform. Chen *et al.* [20] put forward a method to utilize multi-thresholding straddling expansion to adjust result bounding boxes generated by existing methods. Grouping methods can provide the proposals with high accuracy, for the boxes they generate fit to object boundaries well, but they usually suffer low efficiency problem.

2.2. Depth-assisted Object Proposal

In object proposal for RGB-D images, depth is considered as an effective cue to discriminate objects from background.

Xu *et al.* [21] firstly introduce depth cue into object proposal by adaptively integrating depth gradient map to RGB gradient map in BING [9]. Nevertheless, similar to BING, the method cannot provide acceptable proposals under the requirement of high IoU. Gupta *et al.* [22] incorporate depth into MCG framework and produce 2.5D Object Proposal. But its effectiveness of feature extraction depends on the consistency of depth camera parameters. Zheng *et al.* [23] presented a method which improve the 3D object category proposal generating pipeline introducing a shallow ConvNet layer for training to improve the accuracy. Its improvement for BING mainly focuses on loosing IoU threshold to 0.5, thus its cannot provide acceptable preprocessing results in detection or other real applications. Hence, the potential of depth cue has not been fully explored to efficiently generate proposals with high IoU.

3. METHODOLOGY

3.1. Sparse edge detection

We first generate the sparse edge map on color channel of RGB-D images by Structured Edge detector [12]. The pixel with magnitude not more than m_{thr} = 0.1 is regarded as not strong enough to be part of the boundary so that we remove them. As shown in Fig. 3(a), edges in different color means different edge groups which are the minimum units to judge whether it is wholly enclosed in a box. The edge group is formed by greedily merging the pixels on the edge until the sum of orientations of every two adjacent pixels larger than a threshold, which equals $\pi/2$ in our experiments.

3.2. Depth Correction

Due to the limitation of the existing depth estimation methods, inaccurate boundaries and noises often appear in depth channels of RGB-D images. The inaccuracy of depth channel will lead to the mistakes when assigning edges to different layers and further influence the objectness measurement of candidate boxes. So it is necessary to correct depth channel of RGB-D images before layering edges.

In our method, we correct depth channel of an RGB-D image based on the super-pixel representation generated from its color channel. We first generate super-pixels with simple linear iterative clustering (SLIC) algorithm [24], and set the average depth of each super-pixel to all the pixels within it. Then, to each pixel in the detected sparse edges, its depth value is set as the nearest depth value of its eight neighboring pixels. In this way, the influence of inaccuract boundaries and noises in depth channel will be obviously eliminated. Fig. 3 shows an example of depth correction. Compared to the original depth channel (Fig. 3(a)), we can find that more boundary edges of the horses are located in correct depth after correction (Fig. 3(c)).

3.3. Depth-aware Layered Edges

Based on the corrected depth channel, we assign the sparse edges to multiple layers and independently measure the objectness of candidate boxes on each layer based on the corresponding edges.

In sparse edge assignment, we use an adaptive sliding window in depth distribution for sampling. The size of sliding window is adjusted from small to large when sampling from distant to near, for human vision system is more sensitive to the depth difference located in near than in distant. Assume the depth value is in the range of $[0, 1]$ (here 1 means the nearest and 0 means the most distant), we calculate the edge magnitude of pixel $p_{i,j}$ in layer l of all the L layers as follows:

$$e_{i,j}^{l} = \begin{cases} e_{i,j}, & \sigma(l-1) < d_{i,j} \leq w_0 + (\sigma+\rho)(l-1) \\ 0, & otherwise \end{cases} \quad (1)$$

where $e_{i,j}$ is the magnitude of pixel $p_{i,j}$ in sparse edge map; $d_{i,j}$ is the depth value of $p_{i,j}$; w_0 is the initial size of sliding window, i.e., window size in smallest depth value; σ is the step length of sliding; ρ is the addition of sliding window size. In our experiment, we use w_0 = 0.2, σ = 0.025 and ρ = 0.025. Totally, there are $L = \lceil \frac{1-w_0}{\sigma+\rho} + 1 \rceil$ layers generated for independent objectness measurement.

3.4. Proposals Ranking

As shown in [8], we sampling the candidate boxes and measure their objectness on each layer independently. We

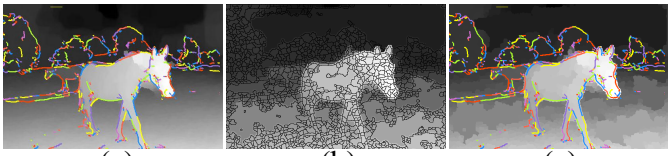

Fig. 3. Depth correction guided by super-pixels of color channel. (a) Intial depth channel of an RGB-D image with colored edge groups. (b) Depth correction on super-pixel representation. (c) Corrected depth with colored edge groups.

cluster the neighboring pixels of similar orientation to form edge groups and use edge groups as representation of edges as [8]. The affinity of two edge groups $g_{l,i}, g_{l,j}$ on layer l is defined as:

$$f(g_{l,i}, g_{l,j}) = |\cos(\alpha_{l,i} - \alpha_{l,i,j})\cos(\alpha_{l,j} - \alpha_{l,i,j})|, \quad (2)$$

where $\alpha_{l,i}$ and $\alpha_{l,j}$ are the mean orientation of edge group $g_{l,i}$ and $g_{l,j}$; $\alpha_{l,i,j}$ reflects the angle difference between $g_{l,i}$ and $g_{l,j}$, which is computed by difference of their mean position's orientation.

Then, we define the weight of ith edge group in a candidate box b_x on layer l as follows:

$$\omega^e_{x,l,i} = (1 - \max(\prod_{k}^{|Q_l|-1} f(q_{l,k}, q_{l,k+1}))), \quad (3)$$

where Q_l is a set of sequences of edge groups on layer l with length of $|Q_l|$, whose starting point is on candidate box b_x and ending point is on edge group $g_{l,i}$. In this way, all the edge groups out of candidate box b_x or overlap the boundary of b_x will be weighted to zero. And the path with the highest affinity between the ith edge group and an edge group that straddles the box will be found.

And we calculate the score of b_x on layer l as follows:

$$s_{l,x} = \frac{\sum_i \omega^e_{x,l,i}\hat{m}_{l,i}}{2(w_x + h_x)^\eta} - \frac{\sum_{p_y \in b^{ct}_x} m_{l,y}}{2(w_x/2 + h_x/2)^\eta}, \quad (4)$$

where $\hat{m}_{l,i}$ represents the sum of magnitude of edge group $g_{l,i}$; $m_{l,y}$ is the edge magnitude of pixel p_y on layer l; b^{ct}_x is the box central in b_x with both half height and half width of b_x; η is a parameter to balance the naturally large sum of magnitude in larger boxes, which equals 1.5 in our experiments.

Finally, we integrate the scores of each candidate box b_x on all the layers. Considering that the distant content in an RGB-D image usually has low possibility to contain an object, we define a depth weight to bring in depth priority:

$$\omega^d_l = \nu + \frac{1-\nu}{L-1}(l-1), \quad (5)$$

where ν is a parameter of the minimum weight for the most distant layer, which equals 0.1 in our experiments. The overall score of candidate box b_x is calculated as:

$$s_x = \max(\omega^d_l s_{l,x}), l \in \{1, 2, \ldots, L\}. \quad (6)$$

We use the same refinement strategy as [8] to re-sample and score the candidate boxes around the locations with high scores. In addition, we perform non-maximal suppression to reduce the remaining number of boxes to improve the ranking of diverse boxes.

4. EXPERIMENTS

4.1. Dataset and Experiment Settings

We validate the proposed method on *NJU1500* dataset, which contains 1,500 stereo images with the corresponding depth maps generated by optical flow method [25] and manually labelled ground truths of object locations. To the best of our knowledge, it is the largest RGB-D image dataset for object proposal. Moreover, *NJU1500* dataset is balance in object number distribution among images, i.e., the numbers of images including 2, 3, 4, 5, and 5+ (more than five) are same. And its average object number in each image is 4.22, which is higher than the average object number 3.02 of the widely used *PASCAL VOC 2007* dataset [26].

In our experiments, we set the required IoU value to 0.8 to emphasize the accuracy of proposals, which is beneficial to the following processing in real applications. We also utilize Average Recall (AR) [6] as a criterion to comprehensively evaluate the performance under different IoU requirements, which is calculated as:

$$AR(\#prop) = \int_{0.5}^{1} Recall(x, (\#prop)) \quad (7)$$

where $Recall(x, (\#prop))$ is the function of recall over IoU; $AR(\#prop)$ actually calculates the area between $[0.5, 1]$ under the Recall-IoU curve when proposal number is $\#prop$. Furthermore, the Average Recall versus Number of Proposal curve is also used in performance evaluation for it is balanced to different IoU settings.

All the experiments were carried out with Intel i5 2.8GHz CPU and 8GB memory. For all the other methods engaged in comparision, we use the default settings suggested by the authors.

4.2. Comparision with State-of-the-Art Methods

To demonstrate its effectiveness, we compare the proposed method to the typical window scoring based methods and grouping based methods with public source codes. The compared methods includes adaptive integration depth and color (AIDC) [21], binarized normed gradients (BING) [9], edge boxes (EB) [8], objectness (OBJ) [7], geodesic object

Table 1. Comparison of our method in running time with the state-of-the-art methods.

Method	Type	Language	Time (s)
AIDC [21]	window	C++	0.07
BING [9]	window	C++	0.06
EB [8]	window	C++ & Matlab	0.69
OBJ [7]	window	C++ & Matlab	4.13
GOP [19]	grouping	C++ & Matlab	7.25
MCG [18]	grouping	C++ & Matlab	60.12
RCNND [22]	grouping	C++ & Matlab	65.52
SS [16]	grouping	C++ & Matlab	6.39
MEB [20]	integration	C++ & Matlab	0.99
Ours	window	C++ & Matlab	4.54

proposal (GOP) [19], multiscale combinatorial grouping (MCG) [18], selective search (SS) [16], and expansion by multi-thresholding straddling of edge boxes (MEB) [20].

Fig. 4 shows the comparison result. As shown in Fig. 4(a), we can find that our method obviously outperforms the other methods under IoU = 0.8. To the second place method, MCG, our method is still nearly 0.2 higher in recall, even MCG is more than thirteen times slower than our method (Table 1). And RCNND does not outperform MCG here, for the inconsistent camera parameters of RGB-D images in our dataset seriously influence the effectiveness of feature extraction on depth. Moreover, as shown in Fig. 4(b) and (c), our method keeps the best performance on average recall when the number of proposals is larger than 1,500, and it outperforms other methods under IoU from 0.7 to 0.9.

Fig. 5 shows some examples of object proposal results generated by our method. The best bounding boxes to ground truths within the top 5,000 of each image are marked with green bounding boxes. It is found that almost all the objects are detected by our method.

We also validate the efficiency of the proposed method. As shown in Table 1, our method retains relatively high efficiency as well as achieve high accuracy in object proposal.

5. CONCLUSION

In this paper, we propose an effective object proposal method RGB-D images based on layered edges. To discriminate the edges from different objects and background, the sparse edge map detected on the color channel of RGB-D image is layered with adaptive sliding window according to the corrected depth channel. For each candidate box, its objectness is independently measured on all the layers and then integrated to generate the proposals. The experimental results show that our method obviously outperforms the state-of-the-art methods under high IoU requirement.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[2] P. Liang, C. Liao, X. Mei, and H. Ling, "Adaptive objectness for object tracking," *CoRR*, vol. abs/1501.00909, 2015.

[3] X. Xu, W. Geng, R. Ju, Y. Yang, T. Ren, and G. Wu, "Obsir: Object-based stereo image retrieval," in *ICME*, 2014.

[4] Y. Xiao, C. Lu, E. Tsougenis, Y. Lu, and C.-K. Tang, "Complexity-adaptive distance metric for object proposals generation," in *CVPR*, 2015.

[5] C. Wang, L. Zhao, S. Liang, L. Zhang, J. Jia, and Y. Wei, "Object proposal by multi-branch hierarchical segmentation," in *CVPR*, 2015.

[6] J. H. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals," 2015.

[7] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *CVPR*, 2010.

[8] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[9] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014.

[10] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *ICCV*, 2011.

[11] R. L. Solso, O. H. MacLin, and M. K. MacLin, *Cognitive Psychology*, 8 edition, 2007.

[12] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *ICCV*, 2013.

[13] J. Liu, T. Ren, and J. Bei, "Elastic edge boxes for object proposal on RGB-D images," in *MMM*, 2016.

[14] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *CVPR*, 2010, pp. 3241–3248.

[15] A. Humayun, F. Li, and J. M. Rehg, "RIGOR: Reusing inference in graph cuts for generating object regions," in *CVPR*, 2014, pp. 336–343.

[16] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," in *IJCV*, 2013, vol. 104, pp. 154–171.

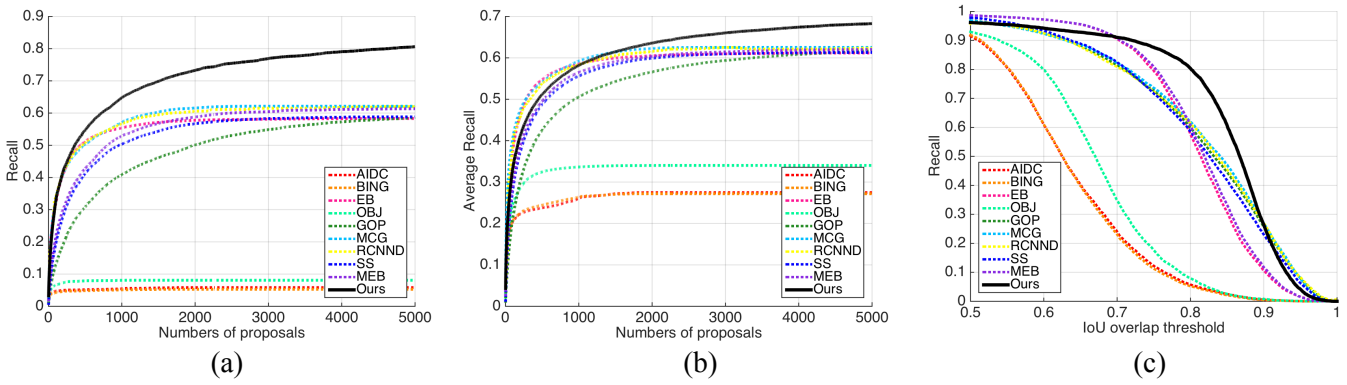

Fig. 4. Comparision with state-of-the-art methods. (a) Recall versus number of proposals curve (IoU = 0.8). (b) Average Recall versus number of proposals curve. (c) Recall versus IoU curve on the top 5,000 proposals.

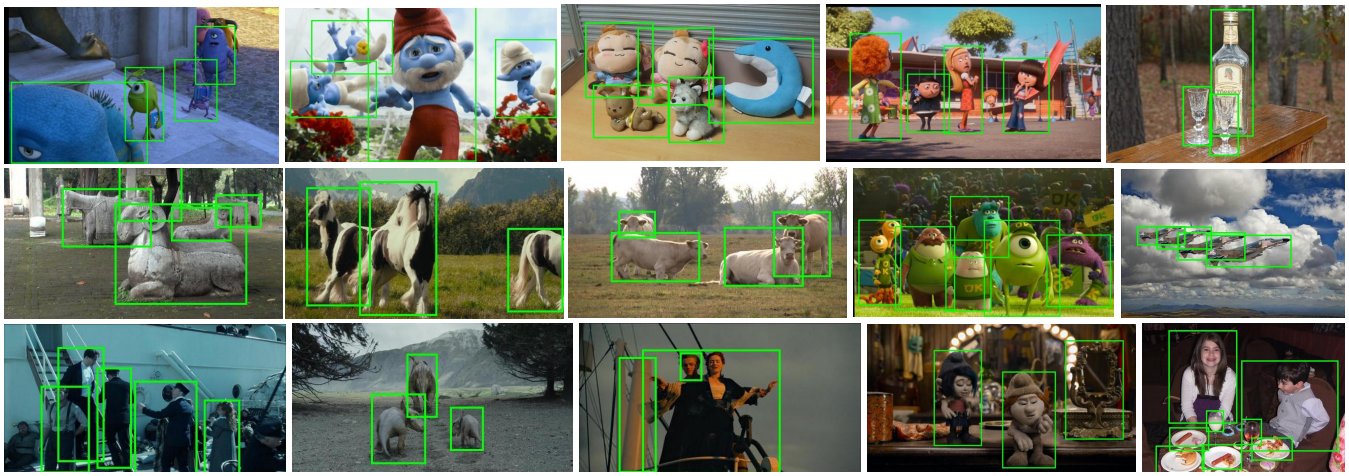

Fig. 5. Examples of object proposal generated by the proposed method. All the green bounding boxes are the best bounding boxes to ground truths within the top 5,000 proposals.

[17] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets re-localization," in *ACCV*, 2014.

[18] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014.

[19] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *ECCV*, 2014.

[20] X. Chen, H. Ma, X. Wang, and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," in *CVPR*, 2015.

[21] X. Xu, L. Ge, T. Ren, and G. Wu, "Adaptive integration of depth and color for objectness estimation," in *ICME*, 2015.

[22] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *ECCV*, 2014.

[23] S. Zheng, V. A. Prisacariu, M. Averkiou, M.-M. Cheng, N. J. Mitra, J. Shotton, P. Torr, and C. Rother, "Object proposals estimation in depth image using compact 3D shape manifolds," in *GCPR*, 2015.

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *TPMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[25] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *CVPR*, 2010.

[26] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.