ADAPTIVE INTEGRATION OF DEPTH AND COLOR FOR OBJECTNESS ESTIMATION

Xiangyang Xu, Ling Ge, Tongwei Ren and Gangshan Wu

State Key Laboratory for Novel Software Technology
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing University, Nanjing 210023, China
xiangyang.xu@smail.nju.edu.cn, gelingnju@gmail.com, {rentw, gswu}@nju.edu.cn

ABSTRACT

The goal of objectness estimation is to predict a moderate number of proposals of all possible objects in a given image with high efficiency. Most existing works solve this problem solely in conventional 2D color images. In this paper, we demonstrate that the depth information could benefit the estimation as a complementary cue to color information. After detailed analysis of depth characteristics, we present an adaptively integrated description for generic objects, which could take full advantages of both depth and color. With the proposed objectness description, the ambiguous area, especially the highly textured regions in original color maps, can be effectively discriminated. Meanwhile, the object boundary areas could be further emphasized, which leads to a more powerful objectness description. To evaluate the performance of the proposed approach, we conduct the experiments on two challenging datasets. The experimental results show that our proposed objectness description is more powerful and effective than state-of-the-art alternatives.

***Index Terms*—** Objectness estimation, object proposal, depth map, generic object description

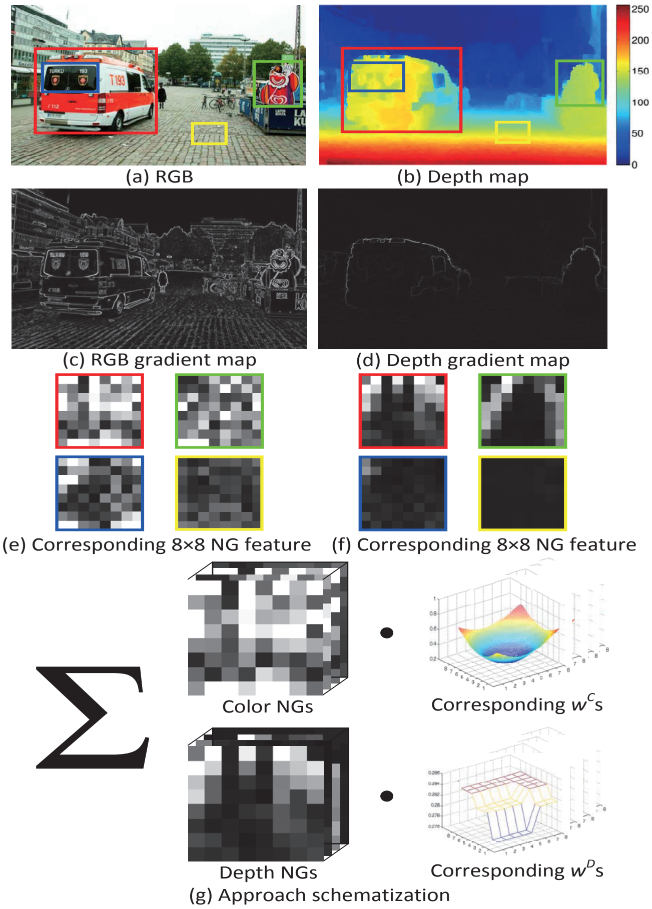
1. INTRODUCTION

Object detection, which aims to detect and localize objects in images, is widely embraced in many multimedia applications, including content analysis [1], image retrieval [2] and object-level editing [3]. Various efforts have been geared according to different processing paradigms, which can be roughly divided into two categories: exhausting sliding windows searching [4, 5] and objectness estimation [6, 7, 8]. Compared with sliding windows approaches, objectness estimation highly reduces the number of returned object proposals which substantially improves the subsequent object classifiers' efficiency. And the distractive false positives could be declined accordingly. Furthermore, the classification accuracy can be increased by enabling more complicated and discriminative classifiers owing to the reduced search space.

Most of the existing objectness estimation methods work on the conventional 2D color images. Though many encouraging achievements [6, 7, 8] have been achieved, it is still very difficult to discriminate real objects from high textures, or to detect objects in complicated scenes. For instance, in Fig. 1(a), the ambulance has many inner distractions in the color map for its complex painting, which may split the ambulance into several parts during objectness estimation. In comparison, the depth cue provides a clean view of object structure, which is considerably powerful in predicting potential objects. As in Fig. 1(b) and (d), we could effortlessly infer that there are at least two objects. And the whole object body can be retained in this scene regardless of its color map. This is mainly owing to the obvious object boundaries, layered structures and cleanness of the object bodies in depth map. Moreover, depth cue has shown its superior effect in many recent applications, like salient object detection [9, 10, 11], image segmentation [12, 13] and activity recognition [14, 15]. Consequently, we consider to introduce the depth information into the objectness estimation task.

Meanwhile, it should be noted that depth is not perfect for general object description. First, the discriminative power of depth decays when the distance between the object and viewer increases. For example, the depth of trees is hard to read in Fig. 1(b) as it is too far from the viewer. On the other hand, it fails to detect the boundaries when the objects are in contact with background or each other, such as the boundaries of the van's wheels since they are in contact with the ground (Fig. 1(b) and (d)). Moreover, accurate depth map is still difficult to obtain with current techniques. The inaccuracy of depth map will inevitably bring in noises in object boundary description, such as the front of the van in depth map in Fig. 1(b).

Motivated by the above, we propose a novel adaptively integrated objectness description approach which takes full advantages of both the depth and color cues for objectness estimation. It is based on a recently developed promising objectness estimation method named BING [7]. In our proposed description, depth will contribute more when it has strong intensity, and the inner colored parts of objects (as Fig. 1(e) "blue" bounding box) are successfully suppressed. Furthermore, object boundaries will be emphasized by corresponding depth and color cues. Contrarily, as the



(a) RGB (b) Depth map

(c) RGB gradient map (d) Depth gradient map

(e) Corresponding 8×8 NG feature (f) Corresponding 8×8 NG feature

(g) Approach schematization

Fig. 1. First row shows a color image and its depth map (0 in depth map refers to infinity, 255 is nearest). Their corresponding gradient maps locate in the middle, which followed by some bounding boxes' 8×8 normed gradients (NG) features in color and depth space. The last row showcases the schematization of our proposed approach.

distance between object and viewer increases, the effect of the depth cue adaptively decays and the color will dominate the proposal prediction at places far away.

To evaluate the performance of the proposed approach, we build a large image dataset including more than 1000 stereo images with depth maps and manually labeled object bounding boxes. Besides, we also evaluate our method on a recently published RGBD dataset [16]. On these two challenging datasets, we compare the proposed method with a baseline which uses only depth as input and two state-of-the-art objectness estimation methods [7, 8]. The experimental results show that our proposed approach is superior to the existing methods. In summary, our major contributions include:

- We are the first to reveal the depth intrinsic characteristics in objectness estimation, which could greatly benefit predicting object proposals;
- We propose a novel generic object description approach by adaptively integrating depth and color information for objectness estimation, which outperforms the state-of-the-art methods;
- We build a stereo image dataset for objectness estimation consisting of a great diversity of images and manually labeled groundtruth[1], which can be used as a benchmark for further objectness estimation study.

2. RELATED WORK

The goal of our work is to introduce the depth cue to objectness estimation. In this section, we firstly outline the representative objectness estimation works briefly, which are performed on traditional 2D color images. After that, we will review the researches on depth-incorporated salient object detection, which are strongly related to our work.

Objectness estimation task aims to generate moderate generic-over-classes object proposals and is expected to cover all objects in an image [6, 7, 8]. According to the object distinctive characteristics, Alexe et al. [6] explored five window cues for measuring the objectness, such as multi-scale saliency, color contrast, edge density, superpixels straddling and window location and size. These cues are formulated in a Bayesian framework and each proposal is scored. But this framework costs much time to train and predict. With the similar manner, Cheng et al. [7] and Zitnick et al. [8] tried to assess each potential window with carefully defined "objectness" score in near real-time. Surprisingly, they all share a common idea that the object borders or edges in the image play a much more important role in objectness estimation and should be incorporated in this task. However, the edges of object inner parts are distractive in object judgement. Hence, in this paper, we adopt the depth map to address this problem.

Depth-incorporated salient object detection has become an active research topic these years. Owing to the convenience of depth acquisition, depth information has been introduced into saliency or salient object analysis [17, 11, 18, 16]. In [17, 11], the authors investigated the matters of the stereopsis for salient object detection by leveraging stereo image pairs with implicit depth, which should be recovered by stereo matching, e.g., [19]. In [18, 16], RGBD data with explicit depth information, which directly read from depth cameras (such as Microsoft Kinect), was incorporated in saliency analysis and the authors demonstrated that the saliency models can be consistently improved by incorporating the depth priors. These salient object analysis works are dedicated to detect and segment the most salient object in each image, but this is not what objectness estimation desires, which tries to recall all objects from an image, not just the most salient one.

[1] Stereo objectness dataset: `http://mcg.nju.edu.cn/en/resource.html`

3. METHODOLOGY

Inspired by the observation that objects are stand-alone things with well-defined closed boundaries and centers such as vehicles, animals and so on [6], Cheng et al. [7] argued that the objects share strong correlation in the small normed gradient space, e.g., 8×8, as the "bounded" boxes shown in Fig. 1(e). Nevertheless, the "blue" van window, a part of the van, is a suspicious false positive object in color space. In comparison with color map, the "blue" bounding box can be inhibited with great confidence in depth map (Fig. 1(f)). On the other hand, it is hard to read the depth difference if the objects are in contact with backgrounds or others, such as the van chassis and the ground. Moreover, the discriminative power of depth decays as the distance to viewer increases. Consequently, we develop an adaptively integrated description of generic objects for further object proposal predicting and the approach's schematization is illustrated in Fig. 1(g).

3.1. Preliminaries

In [7], Cheng et al. proposed a surprisingly simple but very effective feature to describe objects, Normed Gradients (NG). Gradients g_x and g_y at each potential location are calculated along X and Y axes separately with the convolution mask $[-1, 0, 1]$, and then the 64-dimensional ($64D$) NG feature g_l is defined as:

$$\begin{aligned} g_l &= \min(|g_x| + |g_y|, 255), && (1) \\ l &= (i, x, y), && (2) \end{aligned}$$

where l is the window location, i is scale and (x, y) is position. These $64D$ NG features are fed into a two-stage cascade linear SVMs [20]. The first SVM is utilized to learn a generic object model m with groundtruth object bounding boxes as positive instances and randomly generated background windows as negative ones, respectively. According to PASCAL VOC criterion [21], if Intersection-over-Union (IoU) values between instances and groundtruth are not less than 0.5, the instances are treated as positives and vise versa. Due to the facts that different scales have different probability to cover an object, the second linear SVM is applied to learn the calibrated filter score at each window location l, which is referred to indicate how likely a window contains an object, i.e., objectness score o_l:

$$\begin{aligned} o_l &= v_i \cdot s_l + t_i, && (3) \\ s_l &= \langle m, g_l \rangle, && (4) \end{aligned}$$

where s_l is $1D$ filter score, $\langle \cdot, \cdot \rangle$ indicates vector dot-product and v_i and t_i are the learnt term. Remarkably, the approximate binarized model m and normed gradient bitmap $b_{k,l}$ are used to accelerate s_l calculation via bitwise operations, e.g., BITAND, BITCOUNT, etc. And specifically, the $64D$ m is approximated by a set of basis, $a_j \in \{-1, 1\}^{64}$ and $a_j = a_j^+ - \overline{a_j^+}$ ($a_j^+ \in \{0, 1\}^{64}$), so based on simple deduction, s_l can be rewrote as following:

$$\begin{aligned} s_l &= \langle m, g_l \rangle \approx \langle \sum_{j=1}^{N_m} \beta_j a_j, \sum_{k=1}^{N_g} 2^{8-k} b_{k,l} \rangle \\ &= \sum_{j=1}^{N_m} \beta_j \sum_{k=1}^{N_g} 2^{8-k} (2\langle a_j^+, b_{k,l} \rangle - |b_{k,l}|), \qquad (5) \end{aligned}$$

where N_m is the number of basis, N_g is the number of bitmaps and β_j is corresponding coefficient, and more details can be found in [22, 7].

3.2. Adaptively integrated description for objects

It is mentioned that depth information encodes the structure evidence but the discrimination power decays with the distance increasing. Besides, the gradients of object inner parts in color space should be prohibited in predicting object proposals. Therefore, we reformulate the object window gradient g_l:

$$g_l = \frac{1}{w_p} \sum_{c \in \{D, C\}} w^c \cdot g_l^c, \qquad (6)$$

where g_l^c and w^c ($c \in \{D, C\}$) are the gradient in depth and color space and their corresponding ***weight maps*** and normalizer $w_p = w^D + w^C$.

For the depth gradient map, if there is a strong intensity, it can be interpreted as a border of an object with great confidence. Hence, we formulate w^D using the Bayes' rule:

$$\begin{aligned} w^D &= P_h(p \in O | D) = \frac{P_h(p \in O, D)}{P_h(D)} \\ &= \frac{P_h(D | p \in O) P_h(p \in O)}{P_h(D)}. \qquad (7) \end{aligned}$$

In practice, for each depth map, its histogram distribution varies a lot responsible for that objects may occur in great range of depth, so we group the depth histograms into N clusters. And then the $P_h(p \in O) = \frac{\#(p \in \text{GT})}{\# p}$ (GT means object groundtruth, O refers to the object area and $h \in \{1, \cdots, N\}$) which indicates how likely the point p in the depth map belongs to the object. Likewise, the priori $P_h(D | p \in O)$ is counted according the object groundtruth bounding boxes in each cluster and depth distribution $P_h(D)$ is also cluster dependent. It should be noted that this depth prior w^D is evaluated on the training set. According to w^D, the importance of depth cue to the objectness estimation will be adaptively tuned during testing.

For the color gradient, the inner parts are possibly distractive and should be suppressed according to the depth prior, so a $2D$ Gaussian distribution function is a well-suited choice:

$$\begin{aligned} w^C &= 1 - \delta \cdot G(x, y, \sigma_x, \sigma_y) \\ &= 1 - \delta \cdot A \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right), \qquad (8) \end{aligned}$$

where σ_x and σ_y equal to half width and height of the object window respectively, A is a constant scalar and δ is an indicator function:

$$\delta = \begin{cases} 1, & \#(p > \text{GRA\_TH}) > 0.5\text{PER} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where the number of "strong" points p in the window's depth normed gradient map should be larger than half of the window's perimeter PER, and GRA_TH is a depth gradient threshold.

Our proposed objectness description has several advantages. First of all, when depth has strong intensity, w^D will contribute more and the inner parts of color normed gradient map (as "blue" bounding box in Fig. 1(c)) are successfully suppressed. Meanwhile, object bounding borders can be emphasized by corresponding depth and color normed gradient map. What's more, as the distance between object and viewer increases, w^D adaptively decays and the color normed gradient will dominate the proposal prediction at faraway places. Exceptionally, without depth prior δ, the normed gradient inner parts will always be suppressed by the Gaussian kernel even when the depth normed gradient map is less informative. And the "yellow" ground in Fig. 1(a) will result in a false positive proposal.

4. EXPERIMENTAL EVALUATION

To evaluate the proposed approach, we extensively conduct the experiments on two datasets, a self-built stereo image dataset and an RGBD image dataset [16]. By taking into account objectness estimation's efficiency requirement, we compare our approach with two state-of-the-art methods, BING [7] and EDGE [8], which are both near real-time in processing. We also treat BING-DEPTH as a baseline in comparison, which directly uses depth maps as input for BING. In all experiments, we adopt the authors' public source codes with suggested parameters in their papers. The detection rate (DR) with given number of windows (#WIN) (DR-#WIN) evaluation metric is utilized to evaluate the methods, which is defined as:

$$\text{DR-\#WIN} = \frac{\#(\text{IoU} \geq 0.5)@\#\text{WIN}}{\#\text{GT}}, \tag{10}$$

where IoU is the intersection-over-union score that is widely adopted to determine whether a proposal covers an object, and GT means object groundtruth bounding boxes.

4.1. Datasets and Experimental Settings

Due to lacking an image dataset with depth maps for objectness estimation, we collect over 1300 stereo images from three sources, daily photographs from a variety of outdoor and indoor places, sharing from Flickr[2] and the snapped frames from 3D videos, to keep the high diversity. Then we take a preprocessing on the collected stereo images, which includes rescaling, duplicates removing and stereo rectification. Since the depth information is implicitly encoded in stereo images, Sun et al.'s optical flow method [19] is employed for its accuracy, robustness and well edge-preserving in stereo matching. It is worth noting that the flow (disparity in our scenario) only occurs along the horizontal direction in calibrated stereo images, so we modify Sun's model to eliminate the vertical displacement. Apart from the self-built stereo objectness dataset, we also conduct experiments on another challenging dataset [16], which is collected by Microsoft Kinect and contains 1000 RGBD images.

With the unavailability of object groundtruth bounding boxes, these two datasets cannot be straightly adopted in objectness estimation task. According to PASCAL VOC2007 annotation guidelines[3], five participants (three males and two females) are asked to draw the object bounding boxes for each image in these two datasets. If three out of five participants reach consensus, the object bounding boxes are averaged and the corresponding image are kept for subsequent evaluation. Finally, 1032 stereo images (about 2.98 objects per image) and 958 RGBD images (about 1.73 objects per image) are survived.

All the experiments are performed on a PC with two Intel Xeon E5540 CPUs and 12GB memory. The number of clusters N and depth gradient threshold GRA_TH are empirically set to 6 and 7, respectively. The constant scalar A is set to 0.75. We adopt 10 times *repeated random sub-sampling validation* to randomly spilt each dataset into two parts, in which 800 images are used as training set and the rest are treated as testing set. Finally, the average performance is recorded.

4.2. Experimental results and analysis

The quantitative performances in stereo objectness dataset and RGBD dataset are illustrated in Fig. 2 and Fig. 3, respectively. When truncated at small #WIN, e.g., not more than 100, the depth-only input method, BING-DEPTH, is comparable with color based methods. Top object proposals often locate near the camera where depth information comes with strong indications, so these objects can be well differentiated with depth maps. Yet despite all that, with the #WIN increasing, the color based methods catch up with the depth-only method and even perform better. We have made remarks on depth in Section 1 that depth map is a layered structure, and the larger depth, the less discriminant power. Therefore, the larger #WIN, the objects at larger distance can be returned by the color based methods, while the depth-only method becomes more and more incapable of. Nonetheless, by adaptively integrating the advantages

[2] https://www.flickr.com/

[3] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/guidelines.html

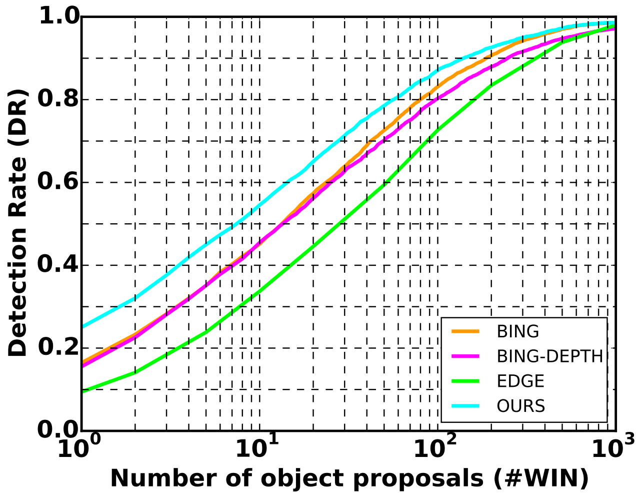


Fig. 2. Comparison of various approaches in stereo objectness dataset.

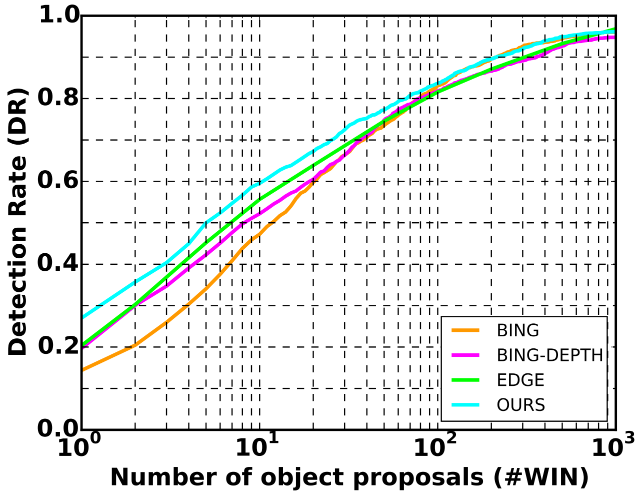


Fig. 3. Comparison of various approaches in RGBD dataset.

of depth and color, given small #WIN, the object borders are complementarily enhanced by depth and color gradient and the inner ambiguous parts of color gradient map are effectively suppressed by the Gaussian kernel with depth prior expressed in Equation (6) and (8). Although the depth's discrimination adaptively decays as #WIN increasing, the color cue still works. Therefore, the proposed adaptive integration approach performs superior to state-of-the-art competitors at any truncated level in stereo objectness dataset.

However, in RGBD dataset, there are lots of "flat" objects, such as paintings, windows (see Fig. 4) and so on. For these objects, the depth normed gradient map cannot tell their boundaries. Then, the proposed approach's performance gain in this dataset is not as much as that in stereo objectness dataset. Another observation for RGBD dataset is that due to the *ambient infrared light* influence, the outdoor scenes' depths in the RGBD dataset are essentially wrong, so depth information in this dataset is not discriminating enough. However, even with these distractions, the proposed approach's performance is still comparable, especially at small truncated level.

As for the computational performance, the proposed method inherits the computation superiority of BING [7]. Therefore, our approach is more powerful and effective with the adaptively integrated description. Some qualitative results are demonstrated in Fig. 4. The positive proposals closest to the groundtruth are highlighted.

5. CONCLUSION AND FUTURE WORK

In this paper, by adaptively integrating depth and color cues, we propose a generic object description approach for objectness estimation. Based on the depth priors, object inner distractive regions can be effectively suppressed. Meanwhile, the object boundaries can be emphasized by the complementarily informative parts in depth and color gradient map. On the contrary, as the distance between object and viewer increases, the effect of the depth cue adaptively decays and the color will dominate the proposal prediction at places far away. Experimental results on two challenge datasets, stereo objectness dataset and RGBD dataset, show that the proposed approach outperforms state-of-the-art alternatives.

However, our method is based on the observation that bounded objects share strong correlation in the normed gradient space. It seems to be incapable of some special shaped objects, such as snakes and "T-shaped" objects. Moreover, the depth discrimination power decays for "flat" objects or those locate far away. Nonetheless, our proposed description can perform at least as good as color-only methods. Furthermore, we will investigate how to integrate depth into other stages of objectness estimation as opposed to the current generic object description stage in the future.

Acknowledgements. We would like to thank the anonymous reviewers for their valuable comments. This work is supported by the 863 Program of China (2012AA063304), National Science Foundation of China (61321491), Research Project of Excellent State Key Laboratory (61223003), and National Special Fund (2011ZX05035-004-004HZ).

6. REFERENCES

[1] Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang, "Object-based visual sentiment concept analysis and application," in *MM*. 2014, pp. 367–376, ACM.

[2] Yang Yang, Linjun Yang, Gangshan Wu, and Shipeng Li, "A bag-of-objects retrieval model for web image search," in *MM*. 2012, pp. 49–58, ACM.

[3] Jianru Xue, Le Wang, Nanning Zheng, and Gang Hua, "Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting," *PR*, vol. 46, no. 11, pp. 2874–2889, 2013.

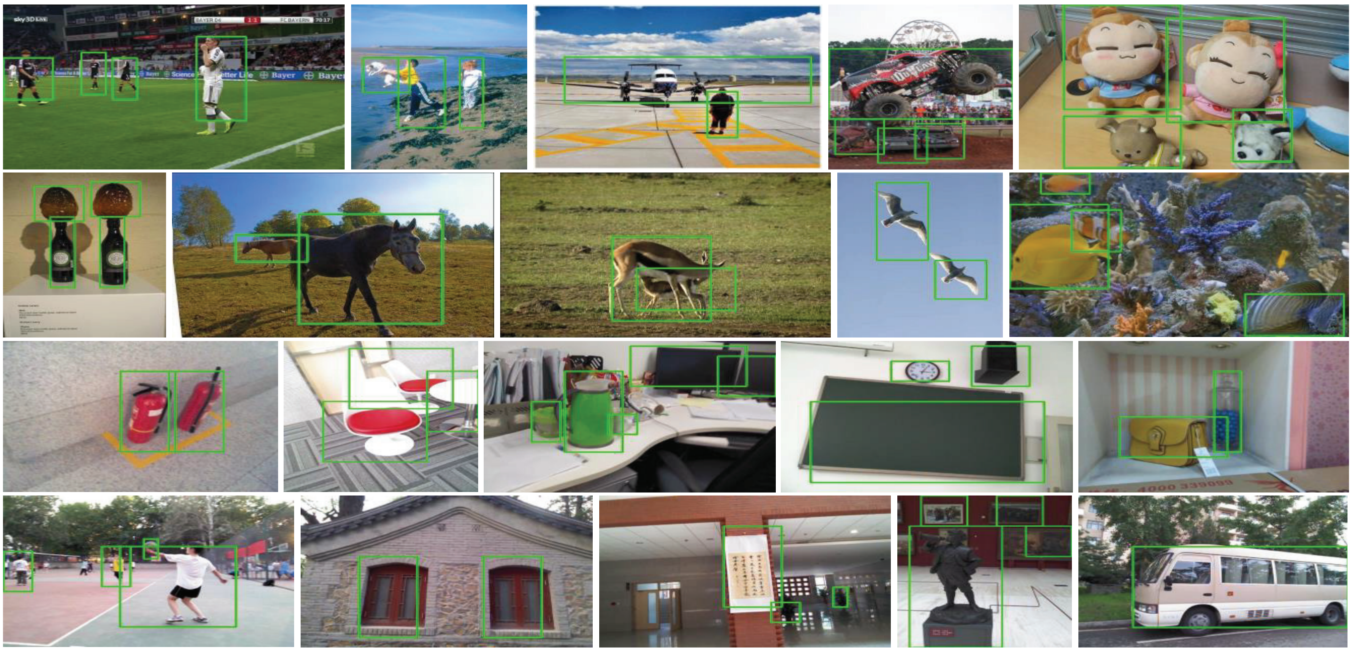


Fig. 4. Exemplars of some positive proposals closest to the groundtruth in the test images. The top two rows are from our stereo objectness dataset and bottom two are from RGBD dataset.

[4] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *CVPR*. 2005, pp. 886–893, IEEE.

[5] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[6] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, "Measuring the objectness of image windows," *TPAMI*, vol. 34, no. 11, pp. 2189–2202, 2012.

[7] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*. 2014, IEEE.

[8] C Lawrence Zitnick and Piotr Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*. 2014, pp. 391–405, Springer.

[9] Jie Feng, Yichen Wei, Litian Tao, Chao Zhang, and Jian Sun, "Salient object detection by composition," in *ICCV*. 2011, pp. 1028–1035, IEEE.

[10] Xiangyang Xu, Wenjing Geng, Ran Ju, Yang Yang, Tongwei Ren, and Gangshan Wu, "Obsir: Object-based stereo image retrieval," in *ICME*. 2014, pp. 1–6, IEEE.

[11] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu, "Depth saliency based on anisotropic center-surround difference," in *ICIP*. 2014, pp. 1115–1119, IEEE.

[12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*. 2012, pp. 746–760, Springer.

[13] Ran Ju, Xiangyang Xu, Yang Yang, and Gangshan Wu, "Stereo grabcut: Interactive and consistent object extraction for stereo images," in *PCM*. 2013, pp. 418–429, Springer.

[14] Omar Oreifej and Zicheng Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *CVPR*. 2013, pp. 716–723, IEEE.

[15] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*. 2011, pp. 1297–1304, IEEE.

[16] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji, "Rgbd salient object detection: A benchmark and algorithms," in *ECCV*. 2014, pp. 92–109, Springer.

[17] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu, "Leveraging stereopsis for saliency analysis," in *CVPR*. 2012, pp. 454–461, IEEE.

[18] Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*. 2012, pp. 101–115, Springer.

[19] Deqing Sun, Stefan Roth, and Michael J Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *IJCV*, vol. 106, no. 2, pp. 115–137, 2014.

[20] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "Liblinear: A library for large linear classification," *JMLR*, vol. 9, pp. 1871–1874, 2008.

[21] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[22] Sam Hare, Amir Saffari, and Philip HS Torr, "Efficient online structured output learning for keypoint-based object tracking," in *CVPR*. 2012, pp. 1894–1901, IEEE.