A Novel Improved Binarized Normed Gradients Based Objectness Measure Through the Multi-feature Learning

Danfeng Zhao, Yuanyuan Hu, Zongliang Gan$^{(\boxtimes)}$, Changhong Chen, and Feng Liu

Jiangsu Provincial Key Lab of Image Processing and Image Communication, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
{13010609,13010602,ganzl,chench,liuf}@njupt.edu.cn

Abstract. In this paper, we propose a novel improved binarized normed gradients (BING) objectness method based on the multi-feature boosting learning. A series of difference of gaussians (DoG) of the images with given parameters are used for the feature extraction stage, since the image DoG filter can better describe objects border. In addition, in training phase, the classifier can adaptively combine the features from different scales and different frequency components. Moreover, since the norm of the feature gradients is a simple 64D feature in the proposed framework, the computational complexity of the algorithm is in the same level compared with the BING measure. Experiments on the challenging PASCAL VOC 07 dataset show that the proposed method can not only achieve higher detection rate and average score than some current related objectness measures, but also lead to a very competitive accuracy of locating objects, even in some difficult cases.

Keywords: Objectness measure · Multi-feature · Boosting strategy

1 Introduction

Object detection, as one of the most important areas in computer vision, has made impressive progress in recent years. Most state-of-the-art detectors are based on sliding window way to find the best object positions over the image. However, a sliding window method needs to evaluate a huge amount of windows [1–4], so it is necessary to constrain the number of locations considered for reducing computation. In order to accelerate object detection, training an objectness measure has recently attracted much attention [5–7, 10].

Objectness is used to quantify how likely it is for an image window to contain an object of any class [7, 18]. The goal of an objectness measure is to separate generic objects from non-objects by proposing candidate windows or proposals containing any object in an image. Generally, we divide an object detection process into two steps: coarse detection and fine detection. And, an objectness measure usually acts as a coarse detection for subsequent class-specific object detectors based on location priors. In other words, it generates proposals of bounding boxes rather than present a full detection system. Meanwhile, some other efficient classifiers [11–13] have been proposed for a fine object detection by screening the candidate windows derived from objectness proposal methods.

Recently, Cheng et al. [10] put forward a fast and efficient binarized normed gradients, namely BING, for objectness estimation at 300fps. A remarkable work of the BING measure is to propose an accelerated feature version, which greatly improves computational efficiency. However, BING feature would lose a part of discriminative ability due to the approximation of the normed gradients feature. Therefore, an improvement on it is necessary for achieving a better performance. Motivated by this, we propose a boosting objectness framework to generate object proposals based on several types of features and a boosting strategy. In our framework, we preserve the speed advantage of BING and focus on improving the accuracy and efficiency by combining multiple features into boosting learning. Our features are derived from a series of difference of gaussians (DoG) of the images with given parameters. Finally, within a single image, a larger overlap with the ground-truth bounding boxes means a higher rank, then the windows at the top of the ranking list can be taken as our object proposals.

2   Overview of BING [10]

The BING is a simple, fast, and high quality objectness measure by using binarized normed gradients feature. Different from amorphous background stuff, objects are standalone things with a well-defined boundary and center [8], hence gradient supporting for image edge examination is an efficient feature to extract them from background. An unusual observation is that, when resizing the corresponding image windows into a small fixed size, generic objects with well-defined closed boundary share strong correlation in normed gradients space. It is because that the little variation between closed boundaries could be brought out in such abstracted view, even though there is a huge difference between objects in terms of shape, color, texture, illumination and so on.

NG feature, i.e. a 64D normed gradients of a window, indicates the value in a 8×8 region (chosen for computational reasons that will explained in Sect. 3.3) of its resized normed gradients maps. Although NG feature is very simple, it has many advantages. Firstly, NG feature is insensitive to changes of translation, scale and aspect ratio. Secondly, it is extremely efficient to be calculated and verified owing to the dense compact representation. Conventionally, getting an 8×8 NG feature over an arbitrary rectangle window needs to calculate a loop computing access to 64 positions. To avoid the loop operations, an accelerated version of NG feature, namely BING, is proposed

by using the model binary approximation [15, 20]. As a result, the objectness calculation of each image window could be tested using only a few bitwise AND and bit COUNT operations.

To learn an objectness measure of image windows, the BING follows the general idea of a two-stage cascaded SVM [9]. The first stage of cascade model aims to learn a linear classifier to select a number of proposals in sliding window way. And in the second stage, a series of classifiers for a set of active quantized sizes are trained to rank the object proposals. Accordingly, the two-stage cascaded model enables us to incorporate variability in scale and aspect ratio of objects, which are treated separately. And different types of features could easily be incorporated in training phase.

Extensive experiments are carried out to evaluate BING measure on the challenging PASCAL VOC 2007 dataset. Compared to other popular alternatives [6, 7, 9], the BING can achieve better detection performance using a smaller set of proposals, is much simpler and 1000+ times faster, while being able to predict unseen categories.

3   The Proposed Framework

In this paper, we propose a boosting objectness framework (Fig. 1) for efficiently evaluating the possibility of an image window being an object, which is composed of several analogous two-stage cascaded models [9]. Any one cascaded model, as a sub one of proposed framework, is specialized for a certain type of objects feature. The first stage learns a model to find out local maximal scores of image windows, and the second one trains a set of classifiers for windows proposed by the first to achieve a final ranking list. Finally, we define the objectness score of an image window as a linear combination of all sub model scores. The usage of multiple improved gradient features and weighting score mechanism greatly improves accuracy and efficiency of our objectness estimation.
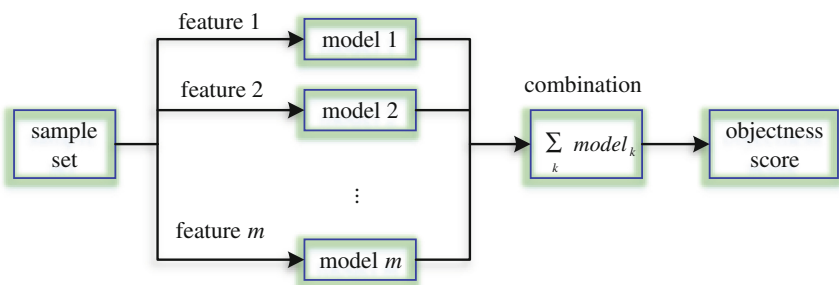

Fig. 1. The framework of our boosting method

In our proposed framework, a two-stage cascade model, named model k, is briefly explained in Fig. 2. Where, $k \in [1, 2, \ldots, m]$, m denotes the total number of sub models in our detection system. Difference between these sub models is that the feature matrix M is produced by difference of gaussians (DoG) with variable parameters (Sect. 3.1).

Our method involves the application of simple linear classifiers in training images. Firstly, we learn a simple SVM $w_k \in R^{64}$ using features of all bounding boxes annotated in the training set as positive samples. Secondly, we scan over various active quantized sizes (scales and aspect ratios), and each window within an image in model k is scored with the linear model w_k,

$$s_{kl} = <w_k, g_l> \tag{1}$$

$$l = (i, x, y) \tag{2}$$

Where, s_{kl}, w_k, g_l, l, i and (x, y) are the filter score, weight vector, feature, location, size and position of a window respectively. Using non-maximal suppression (NMS), we select a small set of proposals from each size i. Some sizes (e.g. 16×256) are less likely than others (e.g. 128×128) to contain an object instance. Thus we define the final objectness score (i.e. calibrated filter score) o_l as

$$o_l = \sum_{k=1}^{m} v_{ki} \cdot s_{kl} + t_{ki} \tag{3}$$

Where, $v_{ki}, t_{ki} \in R$ are learnt coefficient and a bias term for each quantized size i respectively in model k. Note that calibration using (3) is required to re-rank the small set of final proposals.
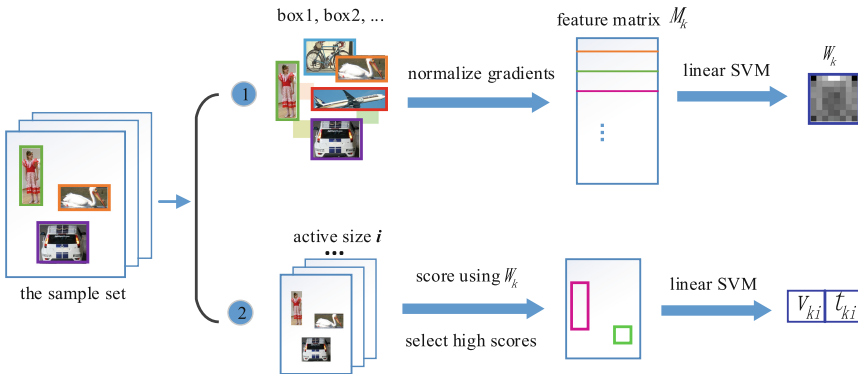

Fig. 2. One sub model of our boosting objectness framework

3.1 Features

As mentioned in Sect. 1, NG feature would lose a part of discriminative ability due to the approximation of the normed gradients feature [10]. Therefore, we propose a novel improved BING feature by combining DoG into the calculation of norm (i.e. magnitude) of the corresponding image gradients to help the search for objects. The purpose of the utilization of DoG is to increase the visibility of edges and other details present in source images, which would make our feature more discriminative.

Difference of gaussians is a feature enhancement algorithm that involves the subtraction of one blurred version of an original image from another, less blurred version of the original. When DoG is employed to different applications, its size ratio of kernel 2 to kernel 1 is usually different [19]. Especially, when utilized for image enhancement, the size ratio of kernel 2 to kernel 1 is typically set to 4:1 or 5:1 (Fig. 3). In this paper, we deal with an original image by operating difference of gaussians only on its Y channel before calculating its gradients. Splitting the original image into different channels is based on the fact that human eyes are more sensitive to luminance (Y) than chrominance (U, V).
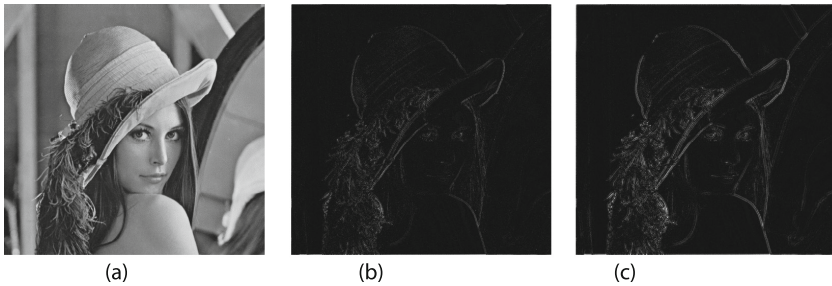

(a)   (b)   (c)

Fig. 3. An example of DoG with different size ratio of kernel 2 to kernel 1. (a) Original image. (b) Image using sigma1 = 0.5, sigma2 = 0.8. (c) Image using sigma1 = 0.4, sigma2 = 2.0.

Experimental results show that this combination of NG feature and DoG makes our measure more effective to detect objects at multi scales and aspect ratios, even when the testing images are with complex background stuff.

3.2 Boosting

Recent work [14] strongly suggests that we should explicitly learn different detectors for different scales, as the cascade model does. Moreover, in a cascade model, the performance of the follow-up classifier would be improved by adding various features into early stage training. Considering these, we designed a novel boosting objectness framework differing from previous methods [7, 9, 10]. In proposed framework, we integrate several cascade models through the multi-feature learning for improving positioning accuracy.

(1) Quantization Scheme
When objects in different class are in the same size, the possibility of its being an object to be detected is different. Based on this knowledge, we extend samples to multiple scales and aspect ratios. Intuitively, a bounding box is supposed to be scaled to some suitable sizes close to its original one. Thus, we adopt the following quantization scheme.

For ease of presentation of our quantization scheme, we provide some key parameters below:

- $t(x, y, w, h)$: a bounding box in an image;
- $s(x, y, w, h)$: the quantized box of a bounding box;
- S_t: the area of the bounding box t;
- $O(t; s)$: the overlap area between the bounding box t and s;
- $T \in [0, 1]$: the overlap threshold for discrimination;

Given a bounding box t, we resize it to N sizes $\{W_0, H_0\}$, in our experiments, $N = 36$, where $W_0, H_0 \in \{2^4, 2^5, 2^6, 2^7, 2^8, 2^9\}$), that is to say, the width and length of a quantized target window is normalized to a specified power of a fixed base 2 (see Sect. 3.3 for reasons). When the formula (4) is satisfied, we preserve this quantized window s for the training of first stage SVM classifier.

$$\frac{O(s, t)}{S_s + S_t - O(s, t)} > T \tag{4}$$

In order to clearly illustrate the quantization scheme, we give a representation of two instances on the PASCAL VOC2007 training set in Fig. 4. As shown, it expands an object into a few surrounding sizes, which permits some translations of the images. Consequently, the scheme is robust not only for different kinds of objects, but also for those with different appearance in the same class.
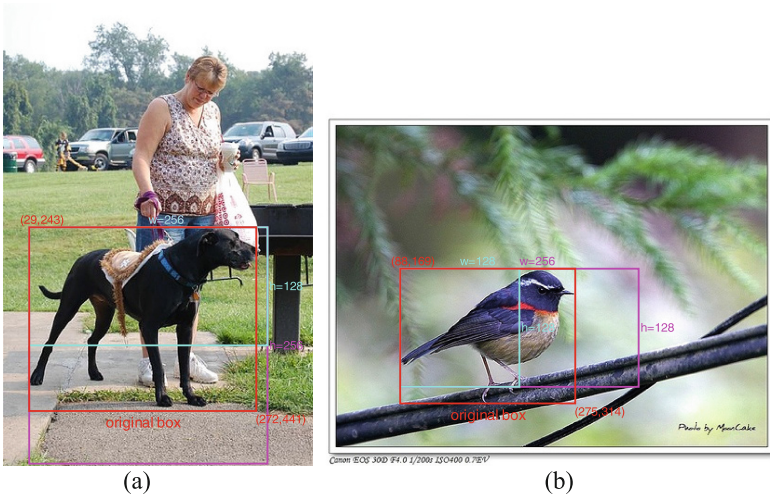

Fig. 4. Two instances of the quantization scheme on the PASCAL VOC2007 training set. The original bounding boxes are in red. (a) A dog with two quantized sizes (cyan, magenta). (b) A bird with two quantized sizes (cyan, magenta) (Color figure online).

(2) Two-Stage Cascade Model

Stage 1: The first stage of our cascade model aims to learn a linear SVM and select an *'active'* set of quantized scales and aspect ratios for the second stage, which is to rank

the proposals by learning a learnt coefficient and a bias term. Here, *'active'* means that this size are much likely than others to contain an object instance (e.g. 128×128 vs. 16×128). To learn a single model w using a linear SVM, improved NG features of all bounding boxes acquired by the above quantization scheme are used as positive training samples. While the negatives are sampled randomly from the background of every image (e.g. 50 per image). Furthermore, the overlap between a background window and any one of the true bounding boxes appended their quantized ones should be less than a certain threshold.

Stage 2: This stage is done by learning a classifier for each scale/aspect-ratio separately. At an active size i, every image of sample set is resized to it, and then a convolution is made by the weight vector w obtained in the first stage. Using non-maximal suppression (NMS), we select a small set of windows whose scores are at the top of ranking list. After converting them to their true locations in original images, we label every window by comparing it with annotated bounding boxes within a single image. When the overlap between a window and any one of bounding boxes exceeds the threshold, we label it +1, otherwise, label −1. Then, a linear SVM is applied to learn the learnt coefficient and bias term of this quantized size.

(3) Model Combination

Boosting is a machine learning algorithm which convert weak learners to strong ones. Most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. When weak learners are added, they are weighted in some way that is usually related to their accuracy. In our framework, we define the final objectness score as a linear combination of all sub-model scores (see Eq. (3)), and particularly, the weight of every model is all set to 1. The main reason for this setup is that there is no great difference between all sub models learned by different features (Sect. 4.1). This simplification by setting each weight to 1 avoids the tedious computation of weight coefficients and is a tradeoff between accuracy and speed of algorithm.

3.3   Speeding Up

One important performance index of an algorithm is time cost, especially for an objectness measure. As a low-level preprocessing stage, objectness not only improves computational efficiency by reducing the search space, but also allows the usage of strong classifiers during testing to improve accuracy for object detectors. Hence, it is better to accomplish the feature extraction and testing process with a lower computation complex.

Following [10, 16, 17], we accelerate our method by approximating gradients feature and the learnt model, which allows us to compute the response maps using only a few fast BITWISE and POPCNT SSE operators. Firstly, we approximate each weight vector $w_k = \sum_{j=1}^{N_w} \beta_j \alpha_j$ with a set of binary basis vectors following [10]. In order to compute the dot-product using only bit-wise operations, each $\alpha_j \in \{-1, 1\}^{64}$ is represented as a binary vector and its complement: $\alpha_j = \alpha_j^+ - \overline{\alpha_j^+}$, where $\alpha_j^+ \in \{0, 1\}^{64}$.

We use pre-computed bit-count of w_k to achieve more efficiency in running-time. Secondly, a BING feature is represented as a single atomic variable (INT64 and BYTE), enabling efficient feature computation, see Fig. 5 for the details.
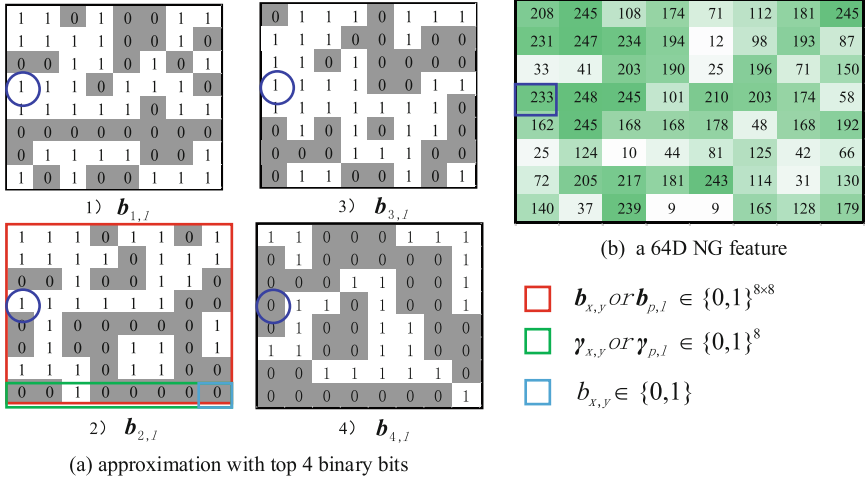

(a) approximation with top 4 binary bits

Fig. 5. Illustration of a binary approximation: (a) A BYTE value is approximated with top N_g (e.g. $N_g = 4$) binary bits. (b) The 64D NG feature g_l. Here is an example (in blue), Decimal: $233 \rightarrow$ Binary: **1110**1001$\rightarrow$ Top $N_g = 4$ bits: **1110**. In other words, $233 \approx \sum_{p=1}^{N_g} 2^{8-p} \boldsymbol{b}_{p,l} = 2^7 * \boldsymbol{1} + 2^6 * \boldsymbol{1} + 2^5 * \boldsymbol{1} + 2^4 * \boldsymbol{0} = 224$ (Color figure online).

The efficient BING feature calculation shares the cumulative nature with the famous integral image representation [21]. Instead of calculating a single scalar value over an arbitrary rectangle range, our filter score (3) of an image window corresponding to BING feature $\boldsymbol{b}_{p,l}$ can be efficiently tested using:

$$s_l = \sum_{j=1}^{N_w} \beta_j \sum_{p=1}^{N_g} 2^{8-p} \left(2 < \boldsymbol{\alpha}_j^+, \boldsymbol{b}_{p,l} > - |\boldsymbol{b}_{p,l}| \right) \tag{5}$$

4   Experiments

In PASCAL VOC2007, each image is annotated with ground-truth bounding boxes of objects from twenty categories (boat, bicycle, horse, etc.). Thus, it is very suitable for our evaluation, as we want to find all objects in the image, irrespective of their category. We evaluate on all 4952 images in the official test set where objects appear against heavily cluttered backgrounds and vary greatly in location, scale, appearance, viewpoint and illumination.

4.1 Parameter Setting

As demonstrated in paper [10], the BING outperforms many other related methods [6, 7, 9], and runs more than three orders of magnitude faster than most popular alternatives [6, 11] as well. In order to prove the effectiveness of our proposed method, we compare it with the BING measure. For a fair comparison, we train our cascade model with the suggested parameter setting in paper [10]. In experiments, 6 object categories (i.e. bird, car, cow, dog, and sheep) are used for training every two-stage cascade model, and the rest 14 categories (i.e. aeroplane, bicycle, boat, bottle, bus, chair, dining-table, horse, motorbike, person, potted-plant, sofa, train, and tvmonitor) are for testing. In addition, we sample about 1000 windows, which ensures covering most objects even in some very difficult cases (e.g. the image with undersized objects or objects occlusion), although 100 is roughly enough.

Moreover, the difference of gaussians plays an important role in our objectness framework as well as boosting strategy. It is a crucial step for proposed method to determine a set of appropriate DoG parameters for achieving good performance. Therefore, we conduct a large number of experiments to pick up better parameters. As shown in Tables 1 and 2, we list respectively the detection rate and average score of windows in a single two-stage cascaded model by setting up several groups of DoG parameters, where the maximum values of rows and cols are in bold type.

Table 1. The detection rate in single two-stage cascaded model with different DoG parameters.

Sigma \ Size	(0.4,0.5)	(0.5,0.8)	(0.5,1.2)	(0.5,2.0)
7	**0.964181**	0.964173	0.962062	0.962788
9	**0.965142**	0.963439	0.961649	**0.964047**
11	0.962820	**0.966074**	**0.963893**	0.963183
13	0.962313	**0.964754**	0.962064	0.961325
15	0.964230	**0.964841**	0.962860	0.959824

Table 2. The average score in a two-stage cascaded model with different DoG parameters.

Sigma \ Size	(0.4,0.5)	(0.5,0.8)	(0.5,1.2)	(0.5,2.0)
7	**0.667807**	0.666295	0.664853	0.665763
9	**0.667396**	0.664670	0.665560	**0.665815**
11	**0.667648**	**0.667258**	**0.666077**	0.665779
13	**0.667608**	0.666715	0.665212	0.665585
15	**0.667799**	0.665961	0.665743	0.665583

As we can see from above tables, it gives rise to a little difference on the detection rate and average score of windows when the parameters (size and sigma of kernel 1,

kernel 2) are changed. In the boosting framework, we apply those parameters which can achieve better performance considering both the detection rate and average score.

4.2   Results

Following [6, 7, 9, 10], we evaluate the detection rate (DR) and average score (AS) of an image window given the number of proposals (#WIN) on PASCAL VOC 2007. As shown in Fig. 6, our method can achieve a higher DR (97.35 %), with three sub models using 1,000 proposals, than the BING measure (96.25 %), and the average score of an image window is greater than that of the BING by 1.3 % (Fig. 7), which fully testifies the effectiveness of our proposed algorithm.
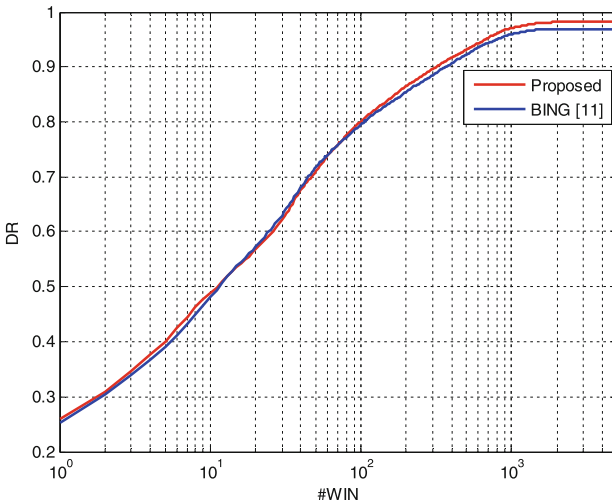

Fig. 6. Tradeoff between #WIN and DR. Our method achieves 97.35 % DR with three sub models using 1,000 proposals, and 96.25 % DR in paper [10].

As shown in Fig. 8, these examples clearly demonstrate that our proposed method can detect many difficult instances that are not captured by the BING measure. Especially, when partial occlusion exists between objects (e.g. a man riding a motor-cycle, two pedestrians walking side by side), it is usually hard for BING to detect all objects, while our method almost can. This is mainly caused by that BING is only based on a simple normed gradients feature, which can't extract complete contour information of occluded objects for evaluating them effectively.

In addition, we intuitively illustrate in Fig. 9 that our improved feature can locate objects more accurately, which means our boosting framework can give more rea-sonable proposals than the BING measure even when they are with the same detection rate. Besides, it shows that our objectness score defined by combining all the scores of sub models makes sense. In some cases, e.g. a few objects close to each other are similar in color and shape, the BING is very likely to mark off a bigger or smaller
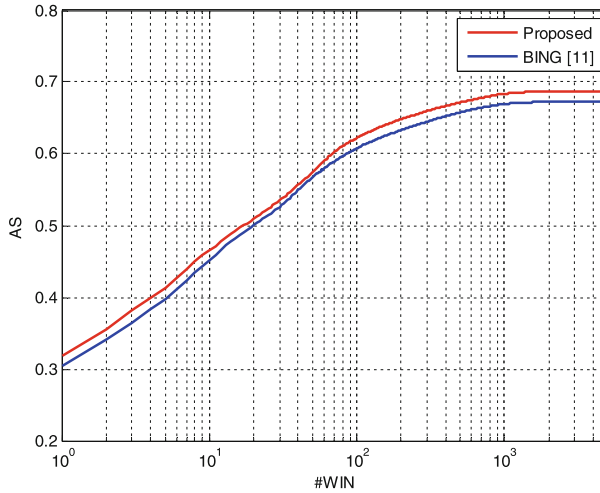
Fig. 7. Tradeoff between #WIN and AS (average score). Our method achieves 68.59 % with three sub models using 1,000 proposals, and 67.29 % in paper [10].

Fig. 8. Comparison of the false negatives between proposed method and BING [10] on VOC2007 test images. The results of BING [10] are shown at the top row (red), and the corresponding ones of proposed method are demonstrated at the bottom row (green) (Color figure online).

rectangular window than the true bounding box, which will lead to an inaccurate detection for the follow-up class-specific object detectors. And yet, our method makes a great improvement on it, giving a series of exact proposals around the boundaries of objects. It is mainly because that difference of gaussians can increase the visibility of edges and other detail present in an image, and the usage of multiple normed gradients features makes it fetch more information from objects.

Further experiments on VOC2007 test images indicate that, with the increase of the number of sub models and proposals, our proposed measure can reach better performance, with a higher DR and fewer false negatives and positives, however, the run time will naturally increase linearly.
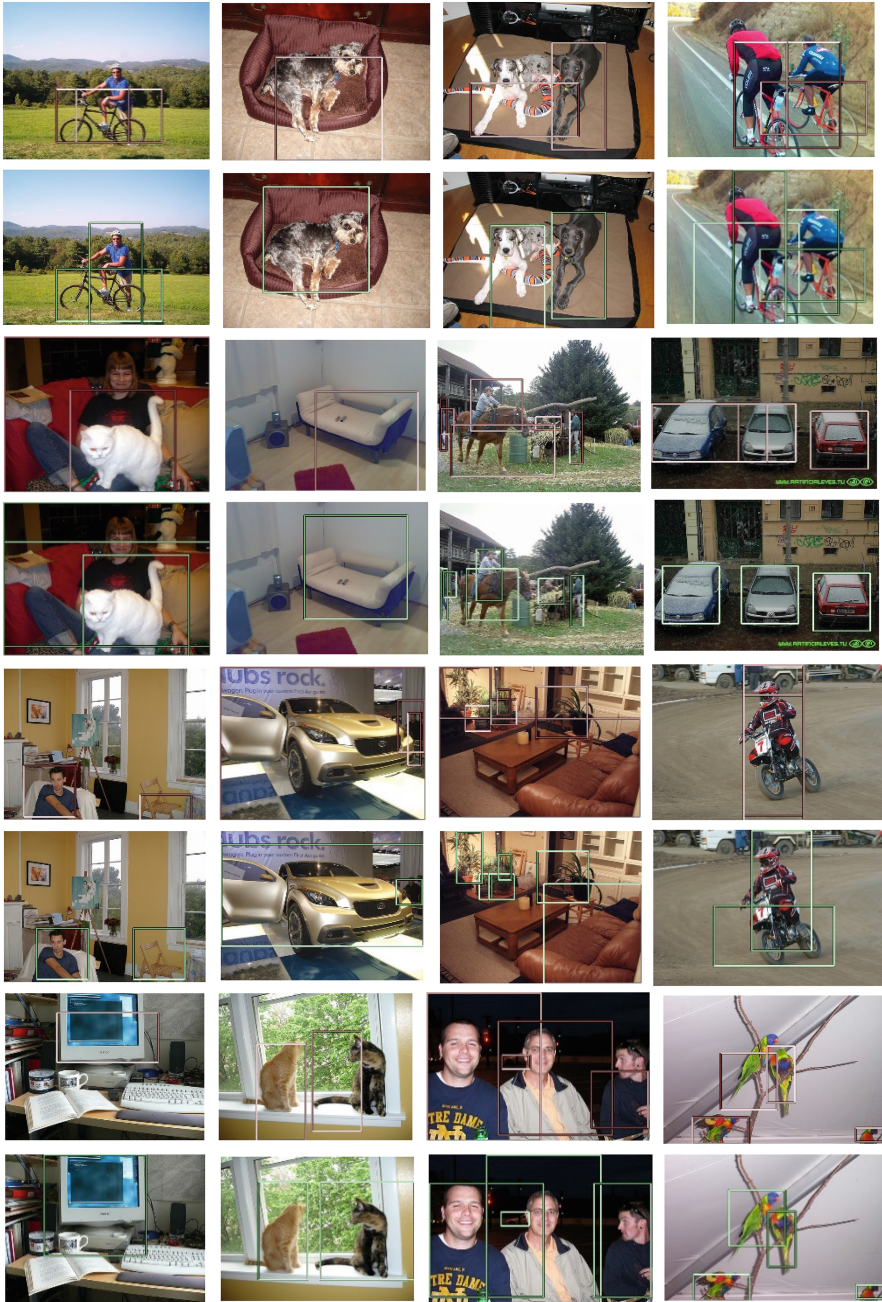
Fig. 9. Comparison of the false positives between proposed method and BING [10] on VOC2007 test images. The results of paper [10] are shown in odd rows (red), and the corresponding ones of proposed method are displayed in its next row (green) for a clear comparison (Color figure online).

5   Conclusions

In this paper, we have presented an objectness measure trained to distinguish object windows from background ones, getting a small set of proposals to cover nearly all potential object regions. And the main character about our framework is very flexible. First, we incorporate different types of features over various active quantized sizes (scales and aspect ratios). And second, the number of two-stage cascade models in the whole system, where a cascade model is used as one sub model to generate object detection proposals, can be adjusted according to different time and accuracy requirements. Evaluation results on the challenging PASCAL VOC2007 test images demonstrate that our proposed method is a superior in detection rate, average score of a window, false negatives and false positives compared with the BING measure. Moreover, since we adopt an improved binarized gradients feature in the proposed framework, the computational complexity of the algorithm is in the same level compared with the BING. Similar with other proposal generations, our method can be easily applied to a complete object detection pipeline, which can greatly reducing the number of windows evaluated by class-specific detectors.

References

1. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
3. Harzallah, H., Jurie, F., Schmid, C.: Combining efficient object localization and image classification. In: IEEE ICCV, pp. 237–244 (2009)
4. Zhu, L., Chen, Y., Yuille, A., Freeman, W.: Latent hierarchical structural learning for object detection. In: CVPR (2010)
5. Endres, I., Hoiem, D.: Category independent object proposals. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 575–588. Springer, Heidelberg (2010)
6. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV **104**, 154–157 (2013)
7. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 2189–2202 (2012)
8. Heitz, G., Koller, D.: Learning spatial context: using stuff to find things. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 30–43. Springer, Heidelberg (2008)
9. Zhang, Z., Warrell, J., Torr, P.H.: proposal generation for object detection using cascaded ranking SVMs. In: IEEE CVPR, pp. 1497–1504 (2011)
10. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: BING: binarized normed gradients for objectness estimation at 300fps. In: IEEE CVPR (2014)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich: feature hierarchies for accurate object detection and semantic segmentation. In: IEEE CVPR, (2014)

12. Dean, T., Ruzon, M., Segal, M., et. al: Fast, accurate detection of 100,000 object classes on a single machine. In: CVPR (2013)
13. Wang, X., Yang, M., Zhu, S., Lin, Y.: Regionlets for generic object detection. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 221–231 (2014)
14. Park, D., Ramanan, D., Fowlkes, C.: Multi-resolution models for object detection. In: ECCV 2010, pp. 241–254(2010)
15. Hare, S., Saffari, A., Torr, P.H.: Efficient online structured output learning for key point-based object tracking. In: IEEE CVPR, pp 1894–1901(2012)
16. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: CVPR (2007)
17. Carreira, J., Sminchisescu, C.: CPMC: automatic object segmentation using constrained parametric min-cuts. IEEE TPAMI **34**(7), 1312–1328 (2012)
18. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR (2010)
19. Marr, D., Hildreth, E.: Theory of edge detection. Proc. Roy. Soc. Lond. Ser. B Biol. Sci. **207** (1167), 187–217 (1980)
20. Zheng, S., Sturgess, P., Torr, P.H.S.: Approximate structured output learning for constrained local models with application to real-time facial feature detection and tracking on low-power devices. In: IEEE FG (2013)
21. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV **57**, 137–154 (2004)