Object Proposals Using SVM-based Integrated Model

Wenjing Geng, Shuzhen Li, Tongwei Ren and Gangshan Wu
State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
Email: jenngeng@gmail.com, szli@smail.nju.edu.cn, {rentw, gswu}@nju.edu.cn

***Abstract*—Utilizing object proposals as a preprocessing procedure has been shown its significance in many multimedia computing tasks. Most state-of-the-art methods devoted to finding a generic objectness measure for rating the possibilities of the initial sliding windows with or without objects. In fact, the object criteria vary from one objectness measure to another, which leads to the definite bottleneck for the single method. By observing the performance of the state-of-the-art in the large dataset, an integrated objectness model is proposed in this paper by accumulating the advantages from selected state-of-the-art techniques. First, the initial bounding boxes are generated by the strategy as same as the method with the highest object detection rate and slowest intersection over union drop. Second, these candidate boxes are re-scored based on each method's objectness system. Then, a score feature is obtained for each bounding box. A support vector machines (SVM) is utilized to train a general model on the training set constructed from a series of score vectors and the probabilistic scores for the testing boxes are predicted according to the learned model. The final proposals are ranked on account of the predicted scores. The evaluation on the challenging PASCAL VOC 2007 dataset shows that the proposed method has dominant concentration with better performance compared to the single state-of-the-art method.**

I. Introduction

Enormous multimedia computing tasks, such as face recognition [1], [2], object detection [3], [4], start from generating millions of sliding windows. In a word, the sliding window approaches have dominated many multimedia and vision tasks for several years. However, some researches [5], [6] indicated that using a small number of object bounding boxes may somewhat improve detection accuracy because spurious false positives are reduced as much as possible. The technique of producing a small set of candidate windows that probably contain objects, called object proposals [7], [8], [9], [10], avoids handling with tremendous amount of bounding boxes over the traditional sliding window object detection paradigm. Generally, object proposal methods can be divided into segment-based proposals [11], [12] and the window-based proposals [13], [14], mainly according to the returned proposal types. However, due to the computational complexity, it is unpopular for the segment-based proposal methods serving as a pretreatment process in most detection tasks. Therefore, we mainly focus on generating bounding box proposals for their applicability, efficiency and convenience as a pre-filtering process for reducing the number of initial sliding windows.

Many existing methods, such as Objectness [5], Rahtu [13], Bing [15] and Edgebox [16] work on assigning a probabilistic objectness score to every sliding window. The key procedure of getting the objectness is designing a specific mechanism for combining one or more low-level features, such as saliency, color contrast, superpixels, image edges and gradients. Though a few encouraging achievements have been achieved, it is inevitable to improperly discriminate real objects in complicated scenes. It is because that objects under complex scenes always can not be distinguished far from being enough based on few image cues. Fig. 1 shows this kind of intrinsic deficiency, which we called inconsistency, caused by different objectness emphases. The light black lines mark the example of inconsistency existing in the proposal results generated by different methods. It is obviously shown that relying on different combinations of image features, some methods may lose efficiency upon objects with unclear color contrast or indistinct superpixels segments, while other methods may not tackle the situation with poor edges or gradients. Fig. 1 explicitly indicates that returning the same number of proposals

Fig. 1. The intrinsic deficiency existing in the proposal results by making a comprehensive overview of the proposals from single methods, Objectness [5], Rahtu [13], Bing [15] and Edgebox70 [16]. The solid rectangles indicate the annotated ground truth and the dashed rectangles illustrate the proposal boxes by each method. Red means the omitted bounding boxes and Green is the hit proposals. The light black lines mark the examples of proposal inconsistency among the different methods.

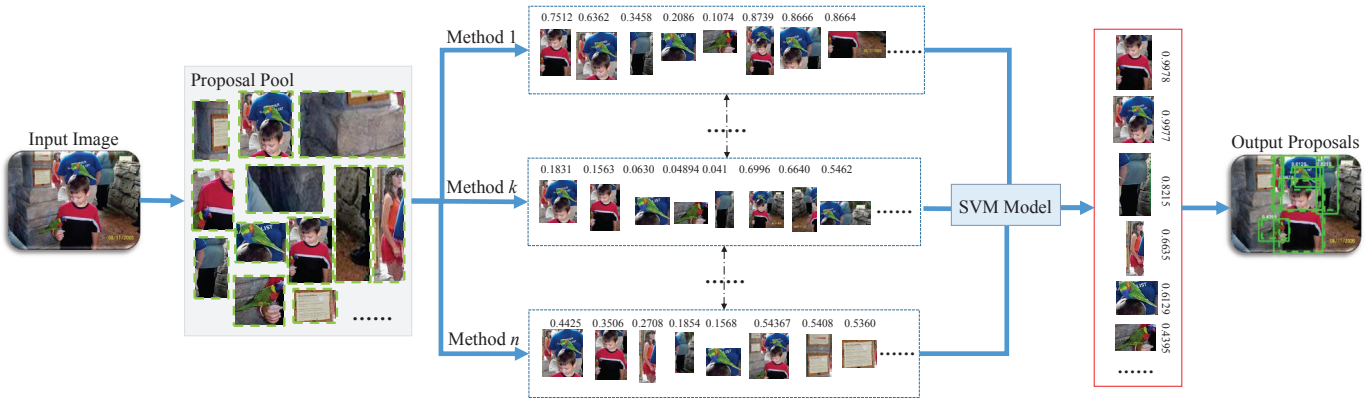

Fig. 2. An overview of the proposed method. Given an input image, the proposal pool is firstly built by candidate boxes generation. Then the score feature is extracted by considering the selected state-of-the-art methods. Finally, the output proposals are produced by ranking as the new objectness obtained by the learnt SVM model.

generated by different methods may cover distinct annotated bounding boxes, i.e., there are poor consistencies in proposals among different methods, which partly degrade the credibility of each method. This issue is mainly caused by the distinct score systems designed by different methods. Due to the scoring rules are based on different low-level features and each feature has its scope application, the proposal inconsistency is distinctly inevitable if only one proposal method is utilized.

In order to eliminate the inconsistency among different proposal methods as much as possible and keep the computing efficiency, an integrated model based on the support vector machines (SVM) is learned according to the score vectors constructed from the state-of-the-art techniques, which embraces general features contributing to generating favorable proposal results. The overview of the proposed approach is presented in Fig. 2. First, the initial bounding boxes are generated by the strategy with the method of comparable higher detection rate (DR) and slower intersection over union (IoU) drop. Second, the initial bounding boxes are re-scored by the selected state-of-the-art methods. Then the score vectors are constructed for each box and randomly selected as training set to learn a SVM model. The new objectness scores are predicted by the probabilistic output of the learnt SVM model. Finally, the proposals generated by the proposed method are ranked by the integrated scores. Experiments are performed on the challenging PASCAL VOC 2007 dataset [17], including training set, testing set and validation set. Compared to the state-of-the-arts, the proposed method shows the improved proposal performance while with the comparable efficiency.

The contribution of this paper can be briefly stated as follows. First, a straightforward framework is proposed aiming at taking advantages of the state-of-the-art proposal methods with little effort. Second, the score vectors are constructed to improve the detection rate on the specific number of proposals in an inexpensive way. Third, a simple machine learning method is adopted to learn a general integrated model to make best use of the state-of-the-arts and form a better and efficient paradigm for generating object proposals.

II. Related Work

In this section, a brief summary of the existing proposal methods and the works related to the proposed framework are provided. More details about the proposal methods can be found in the recent survey papers [18].

Object proposals are firstly proposed by Alexe et al. in [5], aiming at distinguishing windows containing an object from background windows and reducing the number of true negative sliding windows. Due to the applicability and efficiency in object detection [19], it has attracted much more attention from a growing group of researchers. Alexe et al. [6] presented a generic objectness measure to quantify how likely it is for an image window to contain an object of any class. They explored the image cues of multi-scale saliency, color contrast, edge density and superpixels straddling and found that the combinations of multi-scale saliency, color contrast and superpixels straddling achieve the best performance in their scoring system. However, it would fail to localize objects in the images with multiple salient objects or without clear salient objects. Rahtu et al. [13] proposed a method for generating the candidate windows and built an effective linear feature combination to score the windows. The features that they used include superpixel boundary integral, boundary edge distribution and window symmetry, i.e. gradients, as well as the superpixel straddling feature from [5]. This method gives up considering saliency and color contrast while achieving an improved results compared to [5]. Because of relying too much on the superpixels segmentation, it would be caught into confusion when facing the situation that there are many superpixels from different objects in the bounding box. [20], [21] and [14] separately adopted the gradient feature, saliency cues and superpixels straddling as same as [5] which achieve good performance in their application scopes. But mainly relying on the single image cue has not eliminated the deficiencies in the previous methods. Cheng et al. [15] raised a very fast framework to filter the bounding boxes at 300fps by merely relying on the norm of gradients collected from different image scales. This work is motivated by the fact that objects are

stand-alone things with well-defined closed boundaries and centers [22]. Though efficient enough, the detection rate of this method has a significant drop when increasing the IoU overlap with the annotated bounding boxes. Zitnick et al. [16] found that only depending on the image edge feature can improve the detection rate compared to the state-of-the-arts. This method leverages both accuracy and efficiency very well and has the best performance even over the challenging overlap among the window-based proposals. However, this method would be easily misled when dealing with the bounding boxes with edges from multiple objects. There are many other segment-based proposals, such as [23], [11], [24] and [25], they adopted much more expensive features such as bag-of-words [26], SIFT descriptors. They also include some low-level features of edge and color, size, location, shape, contours with more complex procedures to achieve the accurate segment results. Among the segment-based methods, [25] can get the best detection rate in comparison with the other methods while need more computing time. In general, due to the complex of segment-based proposals and its less applicability as a pre-processing procedure, only window-based methods are explored in this paper to further improve the detection rate in an inexpensive way.

Besides, there are some other works trying to improve the detection rate among the current state-of-the-arts. Chen et al. [27] utilized superpixels straddling to refine the proposals by performing multi-thresholding expansion for each bounding box. By taking advantages of boundary-preserving superpixels, this method can be integrated into the existing models to generate object proposals with both high diversity and accurate localization. Therefore, the detection rate of this method keeps a slow drop when the overlap is increased. While the proposed method in this paper manages to improve the detection rate by re-scoring with an integrated model and the overall framework is easier to be reused. Pont-Tuset et al. [28] combined the proposals from different techniques to benefit the performance of the existing object proposals. In fact, they focused much about on exploring the performance of segment-based object proposals on the dataset of Microsoft Common Objects in Context (COCO) [29]. On the contrary, the proposed method mainly devotes to tackling the inconsistency among the current window-based object proposals and improving the detection rate in an acceptable IoU value.

III. SVM-BASED INTEGRATED MODEL

To address the solutions to the existing problems, Fig. 2 demonstrates the basic idea of the proposed method as well as the main procedures for generating the object proposals. It is obviously shown that the proposed approach aims at making best use of the current methods to further improve the detection rate with much higher IoU.

A. Candidate Boxes Generation

The intuitive intention of the proposed approach lies in yielding the improved detection rate with an applicable IoU by making best use of the state-of-the-arts as much as possible. Inspired by the concept of enhancing the advantages and avoiding the disadvantages, a comprehensive consideration about the existing works is made to find a better complementary mechanism. For most of the proposal methods are evaluated in the PSCAL VOC 2007 dataset, a thorough comparison should be performed on this large dataset with 9963 images to determine an efficient strategy for producing the initial windows. In fact, densely sampled windows could also be used as the initial windows, but reusing a more efficient strategy would further increase the computing efficiency. Three popular metrics which are introduced in Sec. IV, need to be calculated according to the proposal results on the entire dataset. Fig. 3 depicts the general comparisons among the four selected methods, showing that the Edgebox70 proposed in [16] has the acceptable detection rate even in a higher IoU and the DR drops slowly along with the change of IoU. For further improving the computing efficiency, N^I initial windows for image I generated by [16] are used as the candidate windows for it has achieved an applicable result leveraging both accuracy and efficiency. And the scores of the returned windows are calculated by Eq. (1).

$$h_b = \frac{\sum_i (1 - \max_T \prod_j^{|T|-1} a(t_j, t_{j+1}))m_i}{2(b_w + b_h)^\kappa}, \tag{1}$$

where $|T|$ is the length of an ordered path T of edge groups. $a(t_j, t_{j+1})$ is the affinity between the edge groups t_j and t_{j+1}. m_i is the sum of the magnitudes for all edges in the edge group. The detailed description about this scoring system is introduced in [16]. Although this method has to score many sliding windows, the computing time is very efficient to only half second for the adoption of using a structured edge detector proposed in [30] and reduced runtime by [31].

B. Score Features Extraction

After taking the image cues of some existing window-based proposal methods into consideration, it is found that most proposal methods prefer to use low level features. We roughly classified these cues into image saliency, color contrast, superpixel, image edge and gradient. And the usage of these cues is ticked as shown in Table III-B, which contributes to give a general impression of the usage distribution of each cue on different method. Tabulating each image cue by adding the ticks for every method visually shows that the combinations of the methods proposed in [16], [6], [13] and [15] contains the utmost image cues adopted by the most existing work. We assume that by combining the popular and efficient image cues together with a machine learning model, the inconsistency can be minimized by the feature complementary mechanism. Fortunately, these methods have published their source codes for further research study. Therefore, it is convenient to reuse the score system of each method.

We adopt the abbreviations of Edgebox [16], Objectness [6], Rahtu [13] and Bing [15] for these methods in this paper. An effective solution to build a relationship among the state-of-the-arts is constructing score features f_b^I for each bounding

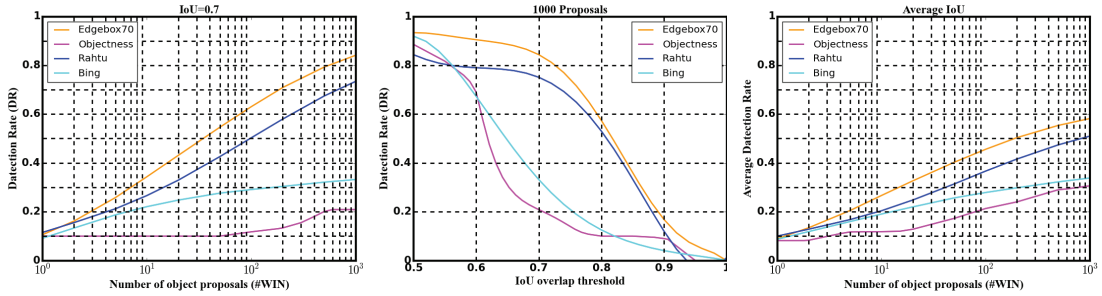

Fig. 3. The performance of the state-of-the-art methods on the entire PASCAL VOC 2007 dataset with 9963 images.

box b in an image I, defined in Eq. (2), where s_i indicates the score calculated by the evaluation systems designed by each method and n is the number of methods being used. Let s_1 represent the score calculated by Edgebox, s_2 by Objectness, s_3 by Rahtu and s_4 by Bing. For one candidate bounding box b in an image I, $s_1 = h_b$,$s_2 = p(obj|C)$, $s_3 = p(y, Y)$, $s_4 = o_l$. The detailed calculations are shown as Eq. (1), Eq. (3), Eq. (4) and Eq. (8). For further exploring, one could refer to the corresponding works.

$$f_b^I = \{s_1, s_2, ..., s_i, ..., s_n\}, \tag{2}$$

$$p(obj|C) = \frac{p(obj) \prod\limits_{cue \in C}^{cue \in C} p(cue|obj)}{\sum_c p(c) \prod\limits_{cue \in C}^{cue \in C} p(cue|c)}, \tag{3}$$

where $c \in \{obj, bg\}$, $cue \in C$ and $C = \{C|MS, CC, SS\}$.

$$p(y, Y) = F(w_1 BI(y), w_2 BE(y, Y), w_3 WS(y, Y)), \tag{4}$$

where y is one of the bounding boxes in the set of windows Y. $F(\cdot)$ represents a learnt function for the feature combinations and the weight vector w is learnt by Eq. (5). $BI(y)$, $BE(y, Y)$ and $WS(y, Y)$ separately represents superpixel boundary integral, boundary edge distribution and window symmetry, i.e., image gradients, defined in [13].

$$\min_{w,\zeta} \frac{1}{2} \parallel w \parallel^2 + \sum_i \zeta^i, \tag{5}$$

$$s.t. < w, \phi_{ij} > - < w, \phi_{ik} > \ \geq \ \Delta_{ik} - \Delta_{ij} - \zeta_{jk}^i, \tag{6}$$

$$\zeta_{jk}^i \geq 0 \quad \forall i, j, k \quad \zeta^i = \sum_{j,k} \zeta_{j,k}^i, \tag{7}$$

where ζ^i is a slack variable of image i, ϕ is the feature vector constructed by $BI(\cdot)$, $BE(\cdot)$ and $WS(\cdot)$. Δ_{ij} and Δ_{ik} are the corresponding loss of the j_{th} bounding box and the k_{th} box to the i_{th} window. The main schemes for getting the score calculated by [13] are listed from Eq. (4) to Eq. (7). Refer to [13] for more details.

$$o_l = v_q^i \cdot < w, g_l > + t_q^i, \tag{8}$$

where v_q^i and t_q^i separately denote the coefficient and the bias terms for each quantized size q of image i. $w \in R^{64}$ represents

TABLE I
THE IMAGE CUE DISTRIBUTION FOR SOME WINDOW-BASED PROPOSALS

Method \ Feature	saliency	color	superpixel	edge	gradient
Objectness [6]	√	√	√		
Rahtu [13]			√	√	√
Feng [21]	√				
Zhang [20]					√
RandSeeds [14]			√		
Bing [15]					√
Edgebox [16]				√	

the learnt linear model from different quantized window sizes of each image. And g_l is the learnt general objectness measure, NG feature, introduced in [15]. By inputting each bounding box $b \in N^I$ to different scoring system extracted from the current works, a score feature f_b^I for image I of one bounding box b is constructed for the following SVM-based learning.

C. SVM Model Construction

SVM is a widely used technique in solving many multimedia computing tasks for its convenience and efficiency. Due to the distribution of the score feature constructed in this paper is unknown, we attempt to use different kernel function to learn a best model. For the introducing of Lagrange multipliers, a SVM optimization problem can be described as follows:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N_W} \sum_{j=1}^{N_W} \alpha_i \alpha_j y_i y_j K(f_i \cdot f_j) - \sum_{i=1}^{N_W} \alpha_i, \tag{9}$$

$$s.t. \sum_{i=1}^{N_W} y_i \alpha = 0, \ 0 \leq \alpha_i \leq C, \ i = 1, 2, ..., N_W, \tag{10}$$

where N_W is the selected number of samples from the N^I initial windows. $K(f_i \cdot f_j)$ is the kernel function, mainly including linear, polynomial, radial basis function (RBF) and sigmoid. Both f_i and f_j are the score vectors which are introduced in Sec. III-B for the bounding boxes. And C is a regulation parameter. The Lagrange multiplier is denoted as α_i, therefore the weight vector of the learnt model can be expressed as:

$$w = \sum_{i=1}^{N_W} \alpha_i y_i f_i. \tag{11}$$

For there are many excellent SVM tools, we adopt the implementation of LIBSVM [32] to learn the scoring model. For getting the continuous scores, the model is trained by the probability estimates, i.e., we adopt the SVM as a classification model while utilizing its probability output to assign a new score for a bounding box. We construct the training set firstly by calculating the VOC overlap score for every bounding box with Eq. (12) and each bounding box is labelled by Eq. (13).

$$s_{iou} = \frac{area(b_i \cap b_{gt})}{area(b_i \cup b_{gt})}, \tag{12}$$

$$l(i) = \begin{cases} 1, & s_{iou} > \tau \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Then the testing data matrix $F = [f_1, ..., f_n] \in R^{d \times n}$ with the corresponding label vector $L = [l_1, ..., l_n] \in R^{1 \times n}$ is built by randomly extracting m ($m \ll n$) boxes from the initial n windows for each image in the test set, including both positive and negative boxes which are scored by different methods presented in the previous section.

D. Predicted Score based Re-ranking

The objective of the proposed method is to learn a generic model contributing to grade the bounding box by taking the advanced state-of-the-art into consideration. In fact, the scoring model learned in Sec. III-C can be treated as a general objectness measure to assign a score for one bounding box. In the prediction stage, the first step for image object proposals is generating a series of bounding boxes by the initial window generation method. Then each bounding boxes are scored based on the different methods for acquiring a score feature. Let y_b denote y conditional on the event that b has the high possibility of being an object, so the classification setting can be simply described as follows:

$$y_b = \begin{cases} 1, & with\ probability\ p_b \\ 0, & with\ probability\ 1 - p_b, \end{cases} \tag{14}$$

where p_b indicates the probability of being an object for a bounding box and is used as the final score for each box.

For the training stage can be finished during the offline, the main computing time for the proposed method rests in the calculation for extracting the score features. In fact, the time consuming of the original methods is all within several seconds including the process of generating millions of initial windows. In the scope of the proposed method, we firstly adopt the same tactics for generating the initial windows which only takes around 1s. For acquiring the score features, only limited bounding boxes need to be input into different score system to get the corresponding scores. This procedure can be easily performed in parallel. Even without the parallel, the computing efficiency can also achieve several seconds. The final procedure of the proposed framework is ranking the proposals based on the new score p_b. It is noted that the proposed framework can be treated as a new scheme making best use of the state-of-the-art with little efforts but with comparable improvements. The integration of the existing work is definitely not limited to four, more diverse combinations can be tried to get further improved performance.

IV. Experiments

A. Experimental Settings

The proposed approach is evaluated on the PSCAL VOC 2007 dataset [33] which contains 9963 images from 20 category, 2501 for training data, 4952 for testing data and 2510 for validation data. Unlike the tag annotations [34], the ground truth is annotated by labelling bounding boxes around the object-like parts. All experimental results are separately reported on the test set and the validation set for cross validation. The parameters are set as $\{N, n, \tau, m\} = \{10^4, 4, 0.7, 20\}$, which means that only $N = 10^4$ initial candidate windows are generated, the score features are constructed from $n = 4$ methods, the labeled data are constructed based on $\tau = 0.7$ and $m = 20$ bounding boxes of each image are selected as the training set. Besides, all the experiments are conducted on a machine with a 3.4GHz Intel i7-4770 CPU and 16GB memory. By taking the comprehensiveness into account, four state-of-the-art methods are compared with the proposed approach, including OBJECTNESS [5], RAHTU [13], BING [15] and EDGEBOX70 [16]. After multiple comparisons, we find that the linear kernel with default parameters got the best results. We also treat the average DR of all the methods as a baseline in comparison. All the scores are calculated on the similar initial windows generated by the introduced method in Sec. III-A.

The authors' public source codes with optimized parameters in their papers are adopted in all the experiments for efficiency and fairness. Three popular evaluation metrics are utilized to quantitatively evaluate the performance of the proposed method. They are the detection rate (DR) with given number of windows (#WIN) (DR-#WIN), DR with variational IoU threshold covered by ground truth annotations for a fixed number of proposals (DR-IOU), and the average detection rate (ADR), i.e., average recall (AR) [18] between 0.5 and 1 by averaging over the overlaps of the images' annotations with the closet matched proposals (ADR-#WIN). Let #GT represent the number of the annotative ground truth for one image, o be the IoU overlap, the DR-#WIN and ADR are separately calculated according to Eq. (15) and Eq. (16).

$$\text{DR-\#WIN} = \frac{\#(o > \epsilon)@\#\text{WIN}}{\#\text{GT}} \quad \epsilon \in \{x | 0.5 \leq x \leq 1\}, \tag{15}$$

$$ADR = 2 \int_{0.5}^{1} DR(o) \text{d}o, \tag{16}$$

where DR-#WIN is curved by a fixed IoU threshold ϵ between 0.5 and 1 with incremental number of windows, while DR-IOU is plotted based on the different IoU between 0.5 and 1 with a fixed number of windows. And the ADR is calculated according to the different DR on distinct IoU with changing number of proposals.

(a) Color image (b) Ground truth (c) Edgebox70 (d) Objectness (e) Rahtu (f) Bing (g) Ours

Fig. 4. The improvement of the proposed approach (g) on the inconsistency among the different methods (c-f). (a) is the original color images and (b) shows the annotated ground truth bounding boxes of the dataset.

B. Comparison with the State-of-the-Art

Generally, each method has distinct procedures to generate limited proposals. The key idea of the proposed method lies in utilizing these unique score systems to assign a score vector for the same bounding box so that the new objectness score can be predicted based on the simple machine learning method. The most noteworthy is that the score features are constructed based on different objectness system for the same candidate windows. But the problem is that some original methods only selectively calculate a part of the sliding windows and generate a few proposals. Hence, some modifications should be executed on the original methods. Therefore, it is necessary to evaluate the proposal performance generated by the original methods and the single performance on the same initial bounding boxes. Meanwhile, the average performance of the four popular state-of-the-art methods is calculated as the baseline to show the improvement of the proposed method.

Qualitative evaluation Fig. 4 shows that the improvement of the proposed approach compared with the four state-of-the-art methods by utilizing the same initial bounding boxes. The red solid rectangles represent the annotated ground truth which is omitted by the top 100 proposals. Fig. 4 (g) shows the hit proposals of our method. It is obviously shown that by integrating the advantages of the current state-of-the-art, the new proposals can benefit from the distinct objectness score. For example, there are three annotations of the image in the first row, but every the single method fails to cover all the annotations. While due to the usage of complementarity effect, the proposed method can cover all the ground truth with only 100 proposals. The other images can also express this kind of improvement clearly.

Quantitative evaluation The quantitative performances in both test set and validation set in PASCAL VOC dataset are illustrated in the first and second row of Fig. 5 respectively. The first column depicts the DR-#WIN, the second column shows the DR-IOU and the third column is the ADR-#WIN. It is clearly shown that the proposed approach outperforms the original state-of-the-art methods, which are shown as long dashed lines. According to the learnt objectness model, the true positive proposals can get higher scores than from relying on single score system, shown as the solid lines. The DR-#WIN curve is drawn based on the 0.8 IoU, illustrating that the proposed method improves the detection rate compared to the state-of-the-art even in such a challenging overlap value. In fact, the proposed method can achieve good performance under the popular IoU between 0.5 and 1, such as 100 proposals shown in the second column of Fig. 5. It is worth noted that for benefiting from the state-of-the-art methods, we preserve the limited bounding boxes generated by the method with highest DR and slowest DR-IOU drop and re-score these bounding boxes according to the unique scoring system presented by the different methods. Considering both accuracy and efficiency, only 10^4 windows are generated as the initial candidate boxes in the whole experiments, which is far less than the number of initial windows compared to the state-of-the-art methods. Nonetheless, the proposed integrated model has achieved the improved detection rate, and not only outperforms the original methods, but also being superior to the methods with single score system. The average detection rate also presents the superiority of the proposed approach under the changing IoU and number of proposals. The integrated model is learnt in

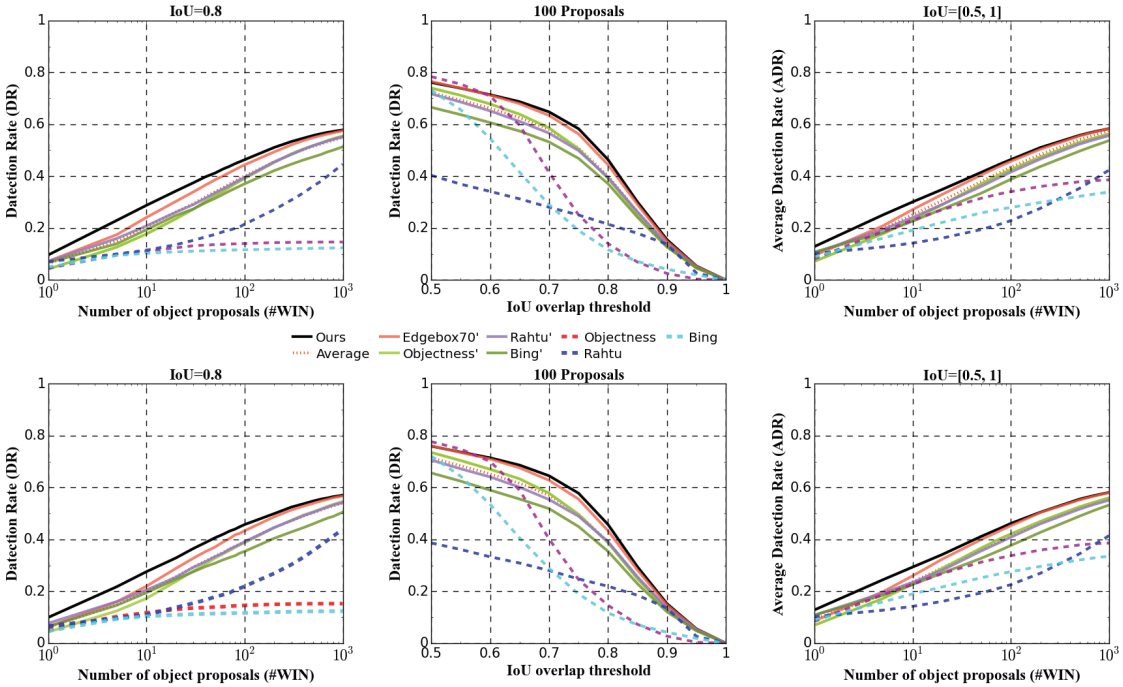

Fig. 5. The comparisons of the detection rate on the PASCAL VOC 2007 test set (the first row) and validation set (the second row).

the training set, while performs better both in the test set and validation set. Therefore, the proposed method has a good generalization ability. And the computing efficiency for predicting an objectness score for a bounding box is obviously efficient. Meanwhile, in view of the most research results in the field of object detection, 0.8 overlap is highly competent for fulfilling most tasks.

C. Discussion

More sample results are presented in Fig. 6. It could be shown that the proposed method could cover much more ground truth with only top 100 proposals on highly overlap with the annotated bounding boxes. And Fig. 7 are some results showing that the proposed approach may also omit some ground truth windows. Such failures can be explained in three aspects. First, some annotations are too small or far away from the salient objects. And some annotations contain the incomplete objects. These factors make it hard to localize every kinds of annotations accurately. Second, the proposed model are built on making best use of the state-of-the-art methods, it would lose efficiency when most of the methods could not present good objectness. Third, we only adopt 10^4 windows as the initial candidate boxes, which is much less than the number of sliding windows. Hence, the diversity of the windows may be degraded in some way. But the little loose in the diversity can bring highly efficiency.

V. Conclusion

In this paper, a straightforward but ingenious framework is proposed for object proposals aiming at taking advantages of the state-of-the-arts. By utilizing an SVM learnt integrated model, the popular rich features are interacted with each other to form a complementary function for object proposals, which is helpful to improve the detection rate while with little efforts. Furthermore, the proposed method could be effortlessly extended by making the different combinations among the good object proposal methods. Experiments show that the proposed framework is propitious to enhance the advantages of the state-of-the-arts and minimize the disadvantages by adopting a complementarity effect superiority.

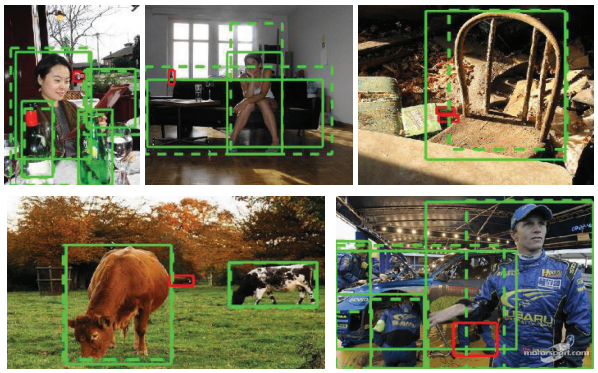

Fig. 7. The demonstration of omitted proposals for the proposed method.

Acknowledgements. This work is supported by the National Science Foundation of China under Grant No.61321491 and No.61202320. Research Project of Excellent State Key Laboratory (No.61223003), Natural Science Foundation of Jiangsu Province (BK20130588) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

Fig. 6. Illustration of the ground truth (the solid and green rectangle) and the true positive object proposals (the dashed and green rectangle) generated by the proposed integrated model for the PASCAL VOC 2007 test and validation images.

REFERENCES

[1] F. Song, X. Tan, S. Chen, and Z.-H. Zhou, "A literature survey on robust and efficient eye localization in real-life scenarios," *PR*, vol. 46, no. 12, pp. 3157–3173, 2013.

[2] Y. Zhao, Y. Liu, Y. Liu, S. Zhong, and K. A. Hua, "Face recognition from a single registered image for conference socializing," *ESWA*, vol. 42, no. 3, pp. 973–979, 2015.

[3] B.-K. Bao, G. Liu, R. Hong, S. Yan, and C. Xu, "General subspace learning with corrupted training data via graph embedding," *TIP*, vol. 22, no. 11, pp. 4380–4393, 2013.

[4] S. Yi, Z. He, X. You, and Y.-M. Cheung, "Single object tracking via robust combination of particle filter and sparse representation," *Signal Processing*, vol. 110, pp. 178–187, 2015.

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

[6] B. Alexe, T. Deselaers, and et al, "Measuring the objectness of image windows," *TPAMI*, vol. 34, no. 11, pp. 2189–2202, 2012.

[7] J. Carreira and C. Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," *TPAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.

[8] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

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

[10] J. Liu, T. Ren, and J. Bei, "Elastic edge boxes for object proposal on rgb-d images," in *MMM*. Springer, 2016.

[11] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized prim's algorithm," in *ICCV*. IEEE, 2013.

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

[13] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *ICCV*. IEEE, 2011.

[14] M. Van den Bergh, G. Roig, X. Boix, S. Manen, and L. Van Gool, "Online video seeds for temporal window objectness," in *ICCV*. IEEE, 2013.

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

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

[17] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2009," in *2th PASCAL Challenge Workshop*, 2009.

[18] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *arXiv preprint arXiv:1502.05082*, 2015.

[19] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *TPAMI*, vol. 36, no. 2, pp. 222–234, 2014.

[20] Z. Zhang, J. Warrell, and P. H. Torr, "Proposal generation for object detection using cascaded ranking svms," in *CVPR*. IEEE, 2011.

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

[22] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *ECCV*. Springer, 2008.

[23] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *ICCV*. IEEE, 2011.

[24] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *CVPR*. IEEE, 2014.

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

[26] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision*. IEEE, 2003, pp. 1470–1477.

[27] X. C. H. M. X. Wang and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," 2015.

[28] J. Pont-Tuset and L. Van Gool, "Boosting object proposals: From pascal to coco," in *Computer Vision*. IEEE, 2015.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.

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

[31] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *TPAMI*, vol. 36, no. 8, pp. 1532–1545, 2014.

[32] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *TIST*, vol. 2, no. 3, p. 27, 2011.

[33] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[34] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *TMM*, vol. 14, no. 3, pp. 883–895, 2012.