Improved BING Method and Its Application in Object Detection

J.C. Cheng
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering
Tsinghua University
China

Y.L. Li
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering
Tsinghua University
China

S.J. Wang
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Electronic Engineering
Tsinghua University
China

***Abstract*-In this paper, we proposed an improved saliency computing method based on BING method. We observed that the undetected objects of BING method have something in common-that is, most of them are occluded or truncated. Therefore, we improved BING method by: Firstly, make a new training set with undetected objects (by BING method) and truncated ground truth of the original training set. Then, train an assistant filter on this new training set. The assistant filter supplements BING method by detecting objects which BING method misses successfully. The experimental results show that the detection rate with improved BING method is increased from 97.2% to 98.1% for 2000 proposals, and that our method, with training of an assistant filter, is better than original BING method at finding incomplete objects.**

Keywords-region proposal; saliency computing; BING; object detection; SVM

I. INTRODUCTION

With the popularity of network, video and image data accumulates rapidly. While human have easy access to massive visual data, dealing with the data manually costs them unaffordable time and resources. In this case, people need computer to process the increasingly large image data for them. For such reason, computer image processing technology develops rapidly at present.

Object detection technology is one of the most commonly used image processing technologies. Being the research hotspot in recent years, a variety of algorithms for object detection have come up, such as histograms of oriented gradients (HOG) for human detection [3], Object detection with discriminatively trained part-based models [6], object detection based on CNN [11, 19], and etc. Most of these algorithms use sliding window technique for window proposing, leading to high false-alarm probability and low detection speed. People need these algorithms to detect images as fast and accurate as human being. This means that current number of window proposed in the detection process is difficult to meet the needs. The research on saliency [20, 21, 22] springs up in response to such instant need.

The word "saliency" is used to depict the nature of an object which makes it particular. Researchers mainly studied two kinds of saliency: human saliency and object saliency. Human saliency is based on human visual characters; it describes where people tend to look at, which point people's attention is focused on, and etc. Object saliency is based on object itself; it describes where the object is placed and how it is different from background.

While these two kinds of saliency are in common with each other to a great extent for that people are usually interested in objects when dealing with visual information, there's some difference in certain cases. For example, in an advertising image, people are likely to pay more attention to ad words rather than objects in the image.

In the field of object detection, object saliency is more emphasized. Studying on object saliency can help researchers discover the essence of objects, and find the way to distinguish objects from background quickly and precisely. Theoretically, using object saliency algorithm can pick out a window collection composed of only a few proposals with all objects in the image covered. If this situation can be reached, it will definitely be a milestone in the progress of object detection.

Current researches on object saliency mostly lie in optimization of saliency definition, quantitative saliency computing method, fast saliency calculation, and etc. For instance, Alexe [2] defines five cues of object saliency. Siva [4] finds out a way to compute saliency based on frequency of image patches. Itti [5] builds a rapid saliency-based visual model.

Among many saliency computing methods, recently presented BING method has the fastest speed (300 fps) and the highest detection rate (96.2% with 1000 proposals).

Although existing saliency methods have achieved impressive results, there's still a long way to go compared with the ultimate goal-covering all objects in a few number

of proposals. There are three major difficulties which hold researchers back:

1. How to find out the intrinsic property for objects of all kinds by the perspective of machine rather than human;

2. How to distinguish the required objects from massive background information (like trees, buildings, and etc.), most of which can be regarded as object in broad sense;

3. How to reduce the time cost of computation and selection.

In this paper, we mainly study to improve saliency algorithm for BING method, and its application in object detection. We first analysis the proposed regions obtained by BING method and find out the problems which influence region performance. Then we propose an assistant filter to pick out occluded and truncated salient objects which are missed by BING method. Finally we apply the improved algorithm in object detection to speed up the object detection process.

II. RELATED WORK

This paper is related to works in two fields-object saliency and object detection.

A. *Object saliency*

Researchers started studying on saliency years ago. Early in 1998, Itti [5] proposed a classical Rapid Saliency Analysis Method-model of saliency-based visual attention for rapid scene analysis. This model is designed to simulate human vision system; it uses pixel-wise information of colors, intensity and orientations to form an image saliency map. Itti's model had presented earlier, it is simple, fast, and has multiple deformation algorithms. The shortcoming of this model is that the resulting saliency map is biased and of low resolution.

After Itti, numerous saliency computing methods has come out. Some of them are based on frequency [4, 10, 12, 13], like Siva's unsupervised Frequency -based Saliency Computing Method [4], which uses the negative correlation between saliency and image patch frequency to compute the image saliency map. The advantage of this method is that the saliency map is accurate and unsupervised. However, its assumption-object saliency is negatively correlated with frequency only-is too strong to be rigorous.

Some of them are based on self-defined cues [2, 14, 15], like Alexe's [2] Saliency Measurement Based on Objectness Cues, which defined object saliency, also called objectness, comprehensively. Proceeding from three distinctive characters of any object, Alexe gave out five cues related to object saliency, and combined them using naive Bayesian model. This method is rigorous, comprehensive, and has high detection rate but with slow speed.

The recently proposed BING Method [1] with fantastic result was drawn from a common sense-that is, every single object has a closed boundary. This method used binarized normed gradients feature and linear SVM model for saliency evaluation and window proposing. Compared with the previous ones, BING method is simple, fast, and achieves state-of-art detection rate.

B. *Object detection*

There are plenty object detection methods with various image features [8], most of them [3, 6, 7, 11] use sliding window algorithm for candidate window proposing and can be largely improved by saliency method. For instance, Dalal's Human Detection Method [3] on INRIA person dataset, which is classical in the field of pedestrian detection, used sliding window to propose candidates at different scales, and detected those candidates with SVM classifier [17,18] trained on HOG feature. The speed of this method is slow because sliding window algorithm presents too many candidates. Such problem just can be solved by reducing the number of candidate windows using object saliency.

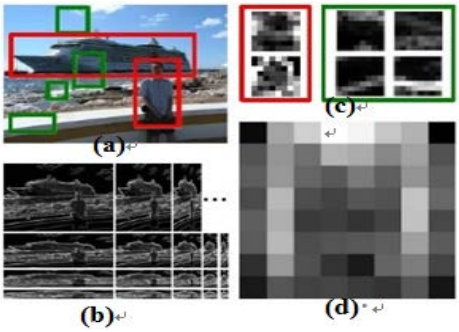

FIGURE I. ILLUSTRATION OF BING METHOD. (A)SOURCE IMAGE. (B)THE CORRESPONDING NORMED GRADIENTS, I.E. NG FEATURES. (C)OBJECT WINDOWS (RED) SHARE STRONG CORRELATIONS COMPARED WITH NON-OBJECT ONES (GREEN). (D) A SINGLE 64D LINEAR MODEL LEARNED FOR SELECTING OBJECT PROPOSALS BASED ON THEIR NG FEATURES

III. METHODOLOGY

A. *BING method*

BING method is inspired by the ability of human vision system which efficiently perceives objects before identifying them. As demonstrated in Fig.1 [1], the corresponding normed gradients (NG feature) of object windows share strong correlations which non-objects windows do not. Based on this observation, the method introduces a 64D binarized normed gradients feature (BING feature) which can efficiently capture the objectness of an image window. The general design idea of BING method is as follows:

Step I. Calculation of NG feature

This step is to get an image window's 64D NG feature. We firstly resize the window to 8×8 pixels and then calculate the resized normed gradients map. The 64D NG feature is obtained by lining up the 8×8 normed gradients map in a vector.

Step II. Learning objectness measurement with NG feature and two-stage cascaded SVM

The goal of this step is to get a linear model $w \in R^{64}$ which can score each image window with (1):

$$s_l =< w, g_l > \qquad (1)$$

where s_l is filter score, g_l is NG feature, $l = (i, x, y)$ is size and position of image window.

Thus, the objectness score can be defined as

$$o_l = v_i \cdot s_l + t_i \qquad (2)$$

where o_l, v_i, t_i, i are objectness score, separately learnt coefficient, bias and quantized size, respectively.

FIGURE II. MOST OF THE OBJECT MISSED BY BING METHOD (OBJECTS IN YELLOW WINDOW) ARE OCCLUDED OR TRUNCATED

The single model w for (1) is learned using linear SVM [9, 16], with NG features of the ground truth object windows and randomly sampled background windows used as positive and negative training samples respectively.

To learn v_i and t_i in (2) using a linear SVM, we use the selected (NMS, non -maximum suppression) proposals at size i as training samples and their filter scores (1) as 1D features, check the training image annotations and learn the optimal value of coefficient (v_i) and bias (t_i).

Step III. Binarization and speeding up

By binarizing the vectors in the algorithm, most of the calculation can be done using bitwise operation, which can largely accelerate the computing process. The learnt 64D filter w can be approximately expressed as:

$$w \approx \sum_{j=1}^{N_w} \beta_j a_j \qquad (3)$$

where N_w denotes the number of basis vectors, $a_j \in \{-1,1\}^{64}$ denotes a basis vector, and β_j denotes the corresponding coefficient.

There is:

$$< w, b > \approx \sum_{j=1}^{N_w} \beta_j a_j b$$

$$= \sum_{j=1}^{N_w} \beta_j (2 < a_j^+, b > - |b|) \qquad (4)$$

where b is a binarized feature, and $a_j^+ \in \{0,1\}^{64}$ s.t. $a_j = a_j^+ - \overline{a_j^+}$.

Thus, the 64D NG feature can be approximated by binarized normed gradients features as:

$$g_l = \sum_{k=1}^{N_g} 2^{8-k} b_{k,l} \qquad (5)$$

where N_g denotes the number of basis vectors, $b_{k,l}$ denotes binarized normed gradients feature.

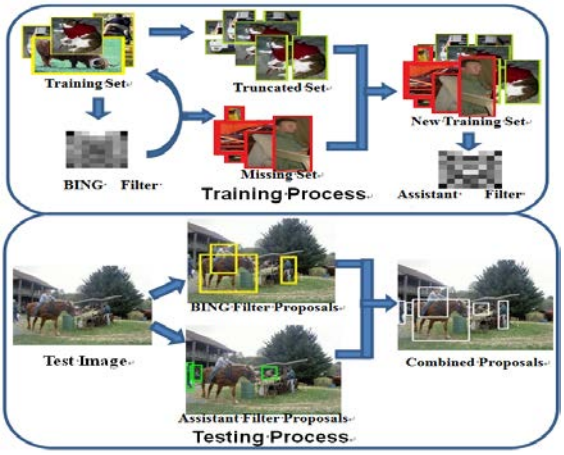

FIGURE III. TRAINING AND TESTING PROCESS OF IMPROVED BING METHOD

Therefore, the filterv score can be efficiently tested using:

$$s_l \approx \sum_{j=1}^{N_w} \beta_j \sum_{k=1}^{N_g} C_{j,k} \qquad (6)$$

where $C_{j,k} = 2^{8-k}(2 < a_j^+, b_{k,l} > - |b_{k,l}|)$, and the optimal $(N_w, N_g) = (2,4)$ according to experimental results.

BING method achieves amazing results. The first thousand window proposals contain 96% of objects. However, the latter part of the proposal set performs poor-detection rate (on single color space), it only rises 2% while the number of windows increases from 1000 to 5000 (Tab.1). The motivation of our method is intending to improve such situation.

B. Improved method

We find out that most of undetected objects by BING method are occluded or truncated (Fig.2). Based on this analysis, we propose an improved method: constructing an assistant filter which specializes in detecting occluded and truncated objects. The process of our improved method is shown in Fig.3.

Training process We first use BING method to train a model w, and use w to evaluate its training set. Compared with the annotation information of training set, we can get an undetected object set, which we call missing set. Then we form a truncated set by dividing the ground truth of training set into halves (each object can be divided into four: left, right, up, down half, respectively). Thus, a new training set for constructing assistant filter is obtained by combining missing set and truncated set. By training the assistant filter using the same method with BING, we get an additive model w^* for the assistant filter.

TABLE I. DETECTION RATE AS A FUNCTION OF THE NUMBER OF WINDOWS FOR BING AND IMPROVED METHOD. IMPROVED METHOD I TRAINS THE ASSISTANT FILTER ON MISSING SET; IMPROVED METHOD II TRAINS THE ASSISTANT FILTER ON MISSING SET AND TRUNCATED SET.

#WIN ($\times 10^3$)	1	2	3	4	5
BING method (%)	96.1	97.2	97.6	97.9	98.1
Improved method I (%)	96.3	97.6	98.1	98.8	99.1
Improved method II (%)	96.4	98.1	98.5	99.1	99.3

Testing process When proposing candidate windows, we use both filters to get candidate sets (w generates candidate set S1 and w^* generates candidate set S2), and combine their candidate sets together to form the final window set S. In this paper, the formulation of S is in accordance with a combination rule. The first 950 proposals are copied from S1, the latter part is composed half of S1 and half of S2. The rule can be expressed as:

$$\begin{cases} s_i = s1_i, i = 1,2,\dots 950 \\ s_{950+(2j-1)} = s2_j, j = 1,2,3,\dots \\ s_{950+2j} = s1_{950+j}, j = 1,2,3,\dots \end{cases} \quad (7)$$

where $S = \{s_1, s_2, s_3 \dots\}$, $S1 = \{s1_1, s1_2, \dots\}$, $S2 = \{s2_1, s2_2, \dots\}$, and all three candidate sets show window results sorted in descending order of scores.

IV. EXPERIMENTS

In this section we present the experimental results. Evaluation of our improved method compared with BING method is carried out on PASCAL VOC 2007 dataset, which consists of 4,952 images from 20 categories. The application of our method in human detection is realized on INRIA Person dataset, which has images with complex background, multiple postures and various illumination.

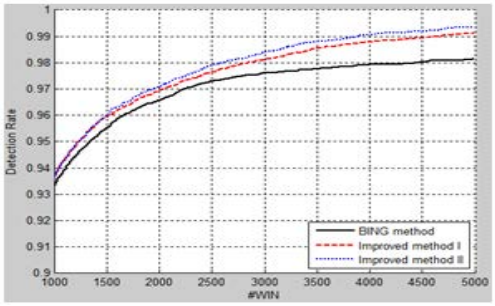

FIGURE IV. #WIN-DR CURVE- IMPROVED METHOD I: IMPROVED METHOD WITH ASSISTANT FILTER TRAINED ON MISSING SET; IMPROVED METHOD II: IMPROVED METHOD WITH ASSISTANT FILTER TRAINED ON MISSING SET AND TRUNCATED SET

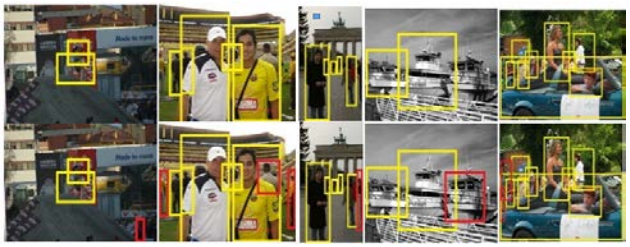

FIGURE V. WINDOW PROPOSALS BY BING VS BY IMPROVED METHOD-THE FIRST ROW SHOWS THE OBJECTS BING METHOD'S PROPOSALS COVERED (YELLOW WINDOWS), THE SECOND ROW SHOWS THE OBJECTS IMPROVED METHOD'S PROPOSALS COVERED (YELLOW AND RED WINDOWS). OUR IMPROVED METHOD IS BETTER AT DETECTING INCOMPLETE OBJECTS (RED WINDOWS).

A. Comparison with BING Method

The performance of our improved method compared with BING method is evaluated with curve measuring number of windows vs detection-rate (#WIN- DR curve). As shown in Fig.4, our method outperforms BING method. Table 1 gives the specific detection rate. Figure 5 illustrates that our improved method is better than BING at detecting occluded or truncated objects.

B. Human detection application

We replace the sliding window part in Dalal's human detection program [3] with our saliency method to propose candidate image windows on INRIA Person dataset. We successfully reduce the number of candidate windows by 18 times and speed up the whole process by 13 times (Tab.2). As the number of windows that may produce false alarm is decreased, there is an increase in detection accuracy as well (Fig.6).

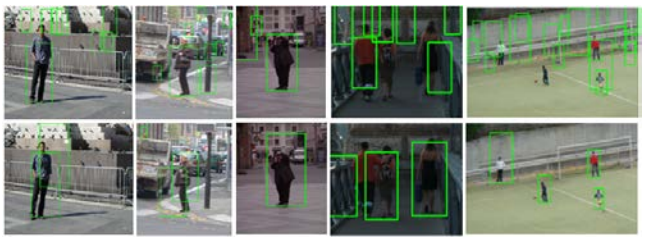

FIGURE VI. HUMAN DETECTION RESULTS WITH SLIDING WINDOW VS IMPROVED BING-THE FIRST AND SECOND LINE ARE RESULTS OF SAME IMAGES WITH SLIDING WINDOW AND IMPROVED BING METHOD RESPECTIVELY

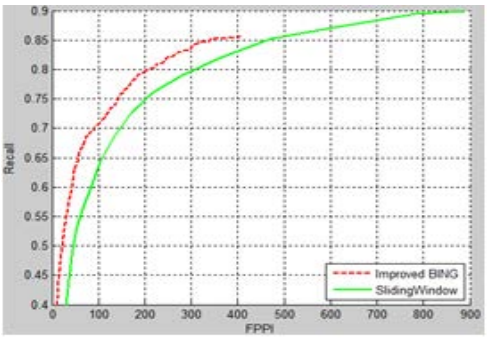

FIGURE VII. RECALL-FPPI CURVE- DETECTION RESULTS WITH BING METHOD (RED DASH LINE) HAS LESS FALSE POSITIVES, BUT ITS RECALL STOPS AT ABOUT 0.85. THIS IS BECAUSE THAT PROPOSALS BY BING DO NOT ALWAYS COVER OBJECTS ACCURATELY

TABLE II. AVERAGE NUMBER OF WINDOWS (AVENUM) AND AVERAGE DETECTION TIME (AVETIME) WITH DIFFERENT WINDOW PROPOSING METHODS.

Window Proposing Method	Improved BING	Sliding Window
AveNUM	1224	22945
AveTime(MS)	38.333	514.893

Figure 7 shows the Recall (detection rate) vs FPPI (False Positive Per Image) curve of both methods, it can be seen that our improved BING can achieve the same detection rate with less false positives. However, the highest detection rate with BING is limited, in that windows proposed by BING do not always cover objects accurately (e.g. BING may take a window covering a person without head as true positive, while the detection program regards it as a negative one).

In conclusion, the result shows that applying saliency to object detection can not only enhance the detecting speed but also increase its detection rate.

V. CONCLUSION AND FUTURE WORK

We present an improved region proposal method for saliency computing by constructing an assistant filter for

BING method. Our method achieves higher detection rate and is better at detecting occluded or truncated objects. We apply our saliency method to human detection program and find out that object saliency can help increase the speed as well as accuracy of detection.

However, our method does not reduce the number of windows proposed on the premise of unchanged accuracy. This means we still need to distinguish objects among thousands of window proposals. Another problem is that the windows proposed do not always cover objects accurately.

In future work, we will focus on reducing number of windows, and the way to make proposed windows cover objects more accurately.

ACKNOWLEDGEMENTS

This work was supported by the National High Technology Research and Development Program of China (863 program) under Grant No. 2012AA011004 and the National Science and Technology Support Program under Grant No. 2013BAK02B04.

REFERENCES

[1] Cheng M M, Zhang Z, Lin W Y, et al. BING: Binarized normed gradients for objectness estimation at 300fps, *IEEE CVPR*, 2014.

[2] Alexe B, Deselaers T, Ferrari V. Measuring the objectness of image windows. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34.11, pp. 2189-2202, 2012.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Computer Vision and Pattern Recognition, 2005. CVPR 2005. Computer Society Conference on IEEE Vol.1, pp. 886-893, 2005.

[4] Siva, Parthipan, et al. "Looking beyond the image: Unsupervised learning for object saliency and detection." Computer Vision and Pattern Recognition (CVPR), IEEE Conference, 2013.

[5] Itti, Laurent, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis." IEEE Transactions on pattern analysis and machine intelligence 20.11, pp.1254-1259, 1998.

[6] Felzenszwalb, Pedro F., et al. "Object detection with discriminatively trained part-based models." Pattern Analysis and Machine Intelligence, IEEE Transactions on 32.9, pp. 1627-1645, 2010.

[7] Yali Li, Shengjin Wang, Qi Tian, Xiaoqing Ding, Learning Cascaded Shared-Boost Classifiers for Part-based Object Detection, IEEE Transactions on Image Processing,Volume:23, No.4, pp.1858-1871, 2014.

[8] Yali Li, Shengjin Wang, Qi Tian, Xiaoqing Ding, A survey of recent advances in visual feature detection, Neurocomputing, Volume 149, Part B, pp. 736–751, 2015.

[9] Chen P H, Lin C J, Schölkopf B. A tutorial on v machines[J]. Applied Stochastic Models in Business and Industry, 21(2), pp. 111-136, 2005.

[10] R. Achanta, S. Hemami, F. Estrada and S. Süsstrunk, Frequency-tuned Salient Region Detection, IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 1597 - 1604, 2009

[11] Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097-1105, 2012.

[12] O. Boiman and M. Irani. Detecting irregularities in images and in video. IJCV, 74(1), pp. 17~31 2007.

[13] Breitenstein M D, Grabner H, Van Gool L. Hunting nessie-real-time abnormality detection from webcams[C]. Computer Vision Workshops (ICCV Workshops), IEEE, pp.1243-1250, 2009.

[14] W. Einhauser and P. Konig. “Does luminance-contrast contribute to saliency map for overt visual attention?” European Journal of Neuroscience, 5(17), pp. 1089–1097, 2003.

[15] Valenti R, Sebe N, Gevers T. Image saliency by isocentric curvedness and color[C]. Computer Vision, 12th International Conference on IEEE, pp. 2185-2192, 2009

[16] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. The Journal of Machine Learning Research, 9, pp. 1871-1874, 2008.

[17] Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2.2, pp. 121-167, 1998.

[18] Sanchez, A., and V. David. "Advanced support vector machines and kernel methods." Neurocomputing 55.1, pp. 5-20, 2003.

[19] Bengio, Yoshua. "Learning deep architectures for AI." Foundations and trends® in Machine Learning 2.1, pp. 1-127, 2009.

[20] Canny, John. "A computational approach to edge detection." Pattern Analysis and Machine Intelligence, IEEE Transactions on 6, pp. 679-698, 1986.

[21] Uijlings, J. R. R., et al. "Selective search for object recognition." International journal of computer vision 104.2, pp. 154-171, 2013.

[22] Van de Sande K E A, Uijlings J R R, Gevers T, et al. Segmentation as selective search for object recognition[C] Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, pp. 1879-1886, 2011.