Journal of Multimedia Information System VOL. 6, NO. 1, March 2019 (pp. 1-6): ISSN 2383-7632(Online)
https://doi.org/10.33851/JMIS.2019.6.1.1

Fast Extraction of Pedestrian Candidate Windows Based on BING Algorithm

Jiexian Zeng[1,2], Qi Fang[1], Zhe Wu[1], Xiang Fu[1], Lu Leng[1*]

Abstract

In the field of industrial applications, the real-time performance of the target detection problem is very important. The most serious time consumption in the pedestrian detection process is the extraction phase of the candidate window. To accelerate the speed, in this paper, a fast extraction of pedestrian candidate window based on the BING (Binarized Normed Gradients) algorithm replaces the traditional sliding window scanning. The BING features are extracted with the positive and negative samples and input into the two-stage SVM (Support Vector Machine) classifier for training. The obtained BING template may include a pedestrian candidate window. The trained template is loaded during detection, and the extracted candidate windows are input into the classifier. The experimental results show that the proposed method can extract fewer candidate window and has a higher recall rate with more rapid speed than the traditional sliding window detection method, so the method improves the detection speed while maintaining the detection accuracy. In addition, the real-time requirement is satisfied.

Key Words: Pedestrian detection, Pedestrian candidate windows, BING algorithm, Sliding window.

I. INTRODUCTION

The traditional sliding window algorithm is simple and widely used for candidate window selection. The sliding window algorithm was proposed in [1], and then became a popular method [2].

There are two popular multi-scale sliding window algorithms. Multi-scale detection windows are used in one algorithm, while the images are multi-scale resized in the other one. However, there is usually a large overlap between the adjacent sliding windows.

Although the sliding window algorithm has been widely used in various computer vision systems, it has two significant drawbacks. First, the number of candidate windows is very redundant, which degrades the real-time performance. An intuitive way to reduce the number of candidate windows is to increase the sliding step length of the window, but this may miss some positive pedestrian detection. Second, some non-pedestrian background areas, such as the sky and some complex background windows, are also judged as pedestrians by the classifier, which causes false detection.

The Caltech Pedestrian Dataset consists of approximately 10 hours of 640x480 30Hz video taken from a vehicle driving through regular traffic in an urban environment. About 250,000 frames (in 137 approximately minute long segments) with a total of 350,000 bounding boxes and 2300 unique pedestrians were annotated [3]. The annotation includes temporal correspondence between bounding boxes and detailed occlusion labels.

II. RELATED WORKS

Since the candidate window redundancy leads to low detection efficiency, a general target detection method is proposed to pre-select areas with a high recall rate, low computational complexity, high quality and a short time period [4]. With the gradual deepening of research in recent years, the scholars have proposed many general target detection methods [5], in which selective search is a classic method [6].

Selective search was proposed by J.R.R. Uijlings, which combines exhaustive search and image segmentation, and applies hierarchical clustering to the merging of regions [7]. The method first divides the image into several small regions, and then merges the regions belonging to the identical target to localize all the targets [8]. Compared with the traditional single strategy, selective search combines

Manuscript received March 4, 2019; Revised March 13, 2019; Accepted March 15, 2019. (ID No. JMIS-19M-03-005)
Corresponding Author (*): Lu Leng, Nanchang Hangkong University, 696 Fenghe South Avenue, Nanchang City, 330063, P.R. China, 0086-791-86453251, leng@nchu.edu.cn.
[1]School of Software, Nanchang Hangkong University, Nanchang 330063, China, zengjx58@163.com, 646767305@qq.com, 648178544@qq.com, fxfb163@163.com, drluleng@gmail.com.
[2]Science and Technology College, Nanchang Hangkong University, Gongqingcheng 332020, China

multiple strategies to enhance the robustness. Additionally, compared with the exhaustive search, the time consuming is greatly reduced due to the remarkably reduced search space. Because of its superior universal target detection performance [9], selective search became popular in many state-of-the-art object detection methods and is used for the extraction phase of the target candidate window [10, 11].

The selective search algorithm consists of two models. The fast model generates approximately 2000 windows on an image. The recall rate is 98%, and the maximum average best overlap (MABO) reaches 0.804. The quality model produces about 10,000 windows with a recall rate of 99.1% and an MABO of 0.879 [12]. It is worth mentioning that the average speed of the algorithm processing is far from the real-time requirement for the fast extraction of the object candidate window in object detection fields [13]. In addition, the dimension of the selective search is too high.

The BING (Binarized Normed Gradients) algorithm [14] has received extensive attention from industry scholars because of its superior comprehensive detection performance. BING algorithm not only achieves similar detection accuracy to those of the selective search algorithm and the objectness algorithm on the pascal voc2007 dataset, but also improves the detection speed by three orders of magnitude. In only 3 ms, it can extract 1000 candidate windows that may be objects. Additionally, the recall rate is about 96%. Therefore, it is significant to improve the extraction of pedestrian candidate window based on the BING algorithm.

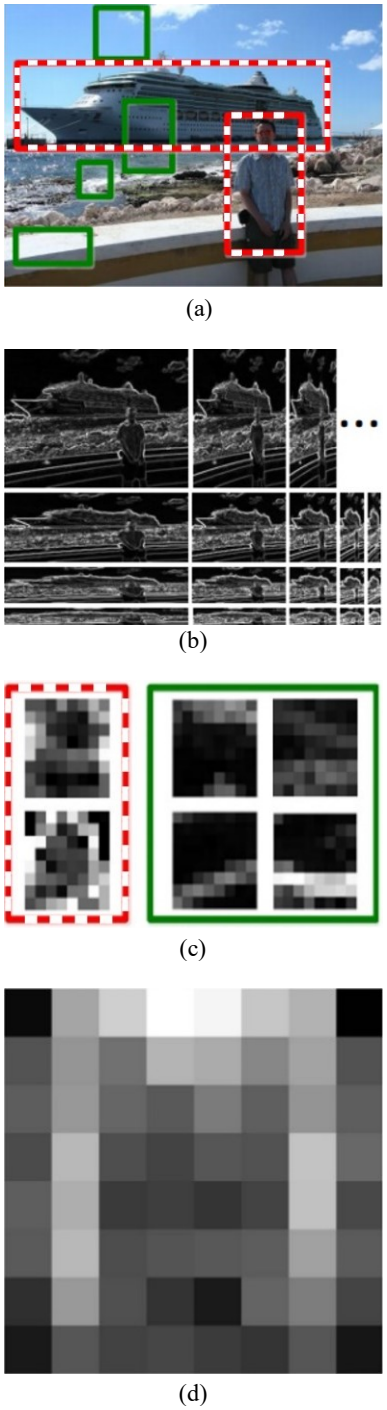
A general object is considered as the object that is not related to a category. The BING algorithm replaces the traditional sliding window scanning method in the field of object detection, and extracts as many candidate windows containing all objects as possible within milliseconds. The BING algorithm is computationally efficient because it uses simple gradient magnitude features and a linear SVM (Support Vector Machine) classifier. Under a fixed-size window, the gradient magnitudes of the object and the background are significantly different. The gradient distribution of the object is cluttered, while the gradient distribution of the background is uniform. The main reason for the difference of gradient distribution is that the objects usually have fully defined closed boundaries and centers [15, 16].

In Fig. 1(a), the red rectangles with dashed lines represent the general objects, which are a ship and a person. The green rectangles represent the random background portion. As shown in Fig. 1(c), after extracting the normed gradient (NG) features of all the rectangular frames, the distribution pattern of the normalized gradient features, which are extracted by the red rectangle frame with dashed lines, and the distribution pattern of the normalized gradient features, which are extracted by the green rectangular frame, are significantly different. The gradient features in the red boxes are more cluttered, while the gradient features in the green boxes are more evenly distributed.

The reasons why the BING algorithm is such highly efficient are:

(1) The original image is scaled to 36 different scales. Although some original information is lost, the structural outline of the objects remains intact. Therefore, the matching with an "8 × 8" template does not degrade the detection effect.

(2) The gradient feature contains a small amount of data, describing the contour information of an object. The BING algorithm further simplifies the image data, discards the last four bits of the 8-bit data, and replaces the first four bits with its own data. This process of data reduction reduces subsequent bit operations by half the amount of shift operations.



(a) (b) (c) (d)

Fig. 1. Gradient distribution patterns of objects and background. (a) source image, (b) normed gradients map, (c) 8 × 8 NG feature, (d) learn model.

(3) From the computer hardware perspective, all the operations of shifting image pixels into an alignment operation greatly accelerate the calculation process.

III. PEDESTRIAN OBJECT EXTRACTION ALGORITHM BASED ON AN IMPROVED BING ALGORITHM

In the original BING paper, the training set is pascalvoc2007. The input image is resized to 36 different scales to detect objects of various sizes.

To better detect various pedestrian objects in daily street scenes, this paper proposes a pedestrian object extraction algorithm based on the improved BING algorithm.

The Caltech pedestrian dataset is selected as the training set. The object detection template in BING is set to the "8 × 16" size for the contour feature of the pedestrian. The pedestrian detection scale is set to a fixed 1:2 form. The specific detection sizes are set to "20 × 40", "40 × 80", "60 × 120", "80 × 160", "100 × 200", "120 × 240", "140 × 280", and "160 × 320". The Caltech datasets set00~set05 are the training set and set06~set10 are the test set. The pedestrians in the dataset are divided into three sizes, then the pedestrians at a close distance have more than 80 pixels, the pedestrians at a medium distance have 30-80 pixels, and pedestrians at a long distance have less than 30 pixels. Each frame of the 30 frames is used, and the training samples are 4250 images. Fig. 2 shows an example of some training samples in the Caltech dataset.

Fig. 2. Caltech Pedestrian Dataset Examples.

An improved BING template training process for pedestrian detection is as follows:

(1) Preparation stage for true positive and false negative sets

4250 images of the Caltech training dataset are used in this step. The images are resized to 8 different sizes. An "8 × 16" sized box is extracted for each sized pixel. The resized image with different scales are shown in Fig. 3.

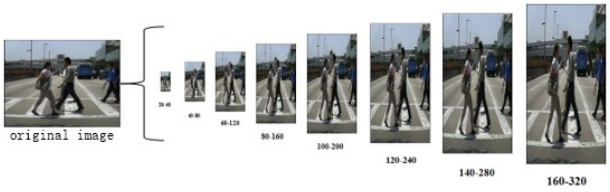


Fig. 3. Training image resized to 8 different scale.

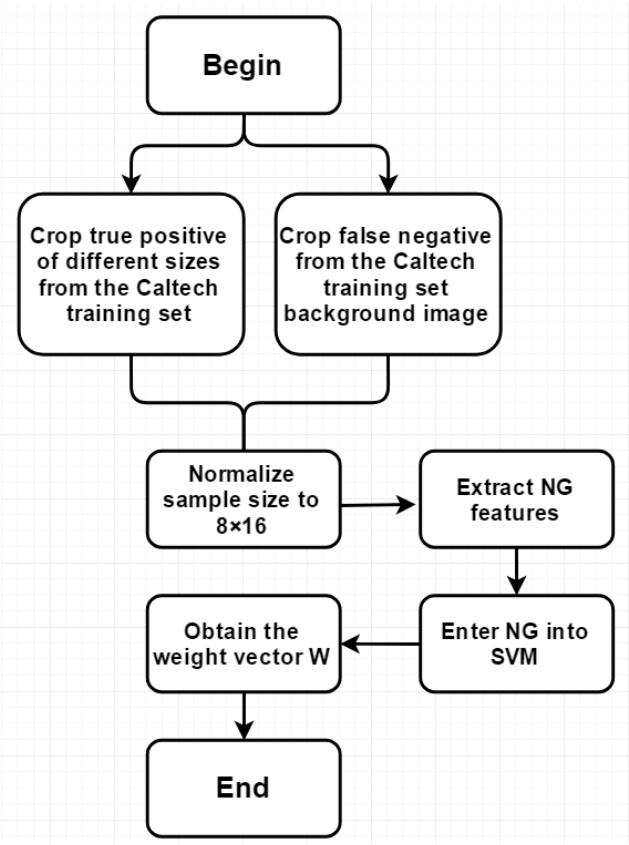
(2) First-level SVM training

The true positive and false negative of all scales are resized to an "8 × 16" size, and the BING features are extracted for linear SVM training.

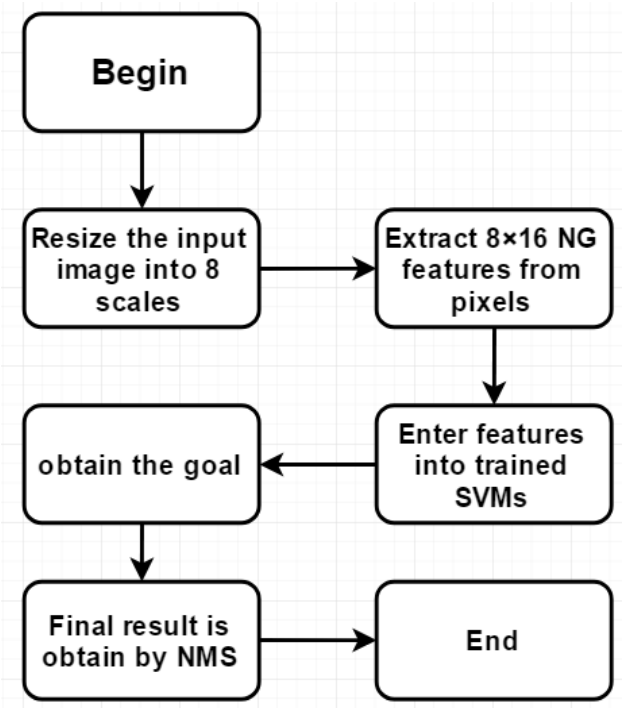
(3) Second-level SVM training

First, the BING template trained in the first-level is loaded. The training images are resized to 8 different sizes. The first-level BING template is used for general object detection at each size, and a small number of candidate windows are selected to form each specification using non-maximum suppression. Next, the retained windows of all scales are detected with the annotation information. The true positives have more than 50% of the intersection area, and the other ones are false negatives. The detection scores of true positive and false negative at different scales are taken as the features. Each SVM is trained once for each scale, that is, eight SVMs are trained. Then the final weight and offset are obtained.

The detection phase is divided into two steps. First the input image is resized to 8 different sizes, and the 8 × 16 sliding windows scan the 8 resized images. The first-level BING template is used for detection. A non-maximum suppression is used according to the score, and a partial detection window at each scale is retained. Then the remaining window is used to calculate the final scores, the scores are output from high to low. The overall processes of training and testing of the BING template are shown in Fig. 4.



(a)



(b)

Fig. 4. Flow chart of training and pedestrian detection using a BING template. (a) Training module, (b) Test module.

After training the pedestrian detection BING templates, they can be applied to extract the candidate window, and combined with any pedestrian classifier detection model. Fig. 5 shows the overall flow of the proposed algorithm. For an input image, first, the BING template is used to extract all candidate windows that may contain pedestrians. Then, these windows are input into the SVM (Support Vector Machine) classifier for classification to obtain the final test result.

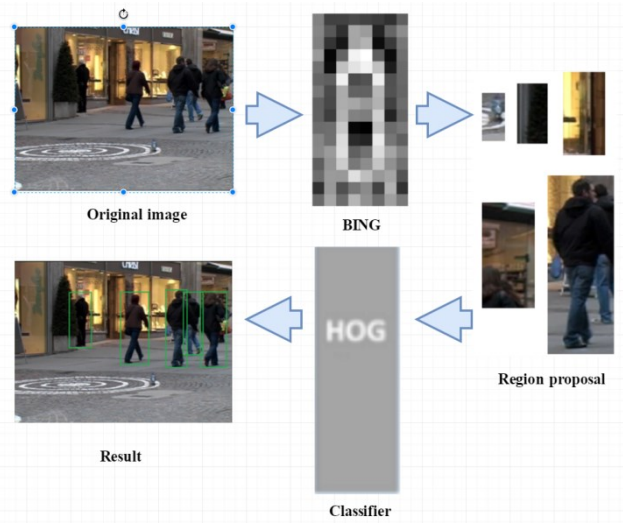


Fig. 5. Process combined with an improved BING algorithm for pedestrian detection.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experiments are performed on Caltech dataset to verify the advantages of the proposed method. The method is expected to significantly reduce the time cost for detection, and its accuracy is comparable to that of the HOG (Histogram of Oriented Gradient) algorithm.

After the two-stage SVM training is completed, the linear SVM model learned with the BING features is shown in Fig. 6, which shows that the active white pixels are concentrated on the silhouette edge of a pedestrian. The SVM weights are very similar to the HOG feature weights learned with SVM.

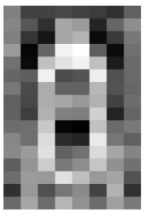


Fig. 6. Pedestrian detection BING template for training.

The detection effect is verified by adjusting the first-level BING template threshold to generate different numbers of candidate windows. The number of candidate windows decreases as the BING threshold increases.

Table 1 shows the detection time and missed detection rate at different BING thresholds. When the threshold is set in the range of [-0.05, 0.01], the missed detection rate remains unchanged, which are all 68%. The larger the BING threshold is, the faster the detection speed is. When the BING threshold is further increased, although the detection time is further decreased, the missed detection rate is greatly increased by discarding a large number of candidate windows containing pedestrians. Thus the optimal detection effect can be obtained when the BING threshold is 0.01. The detection speed of this algorithm is three times faster than that of the traditional Selective Search algorithm. And it has higher value in practical applications.

The Miss Rate formula is:

$$MR = FN/(FN+TP), \quad (1)$$

where FN is False Negative, TP is True positive. Time (s) indicates the time to process an image.

Table 1. Detection results at different BING thresholds.

BING threshold	Miss Rate	Ours Time (s)	Selective Search Time (s)
-0.05	68%	1	2.63
0	68%	0.88	2.16
0.005	68%	0.67	1.59
0.01	68%	0.32	1.04
10.02	75.6%	0.28	0.67
0.04	88.3%	0.12	0.29

V. CONCLUSIONS AND FUTURE WORKS

Firstly, the development history of the sliding window detection in the object detection field is introduced, and its shortcomings are summarized. Then, the widely used general object detection technologies are introduced, especially the selective search algorithm. Because the selective search algorithm has a serious time loss in the extraction of candidate windows, in this paper, improved BING algorithm is used to remarkably accelerate the speed, while the proposed method can achieve the similar detection effects to those of the selective search algorithm. The dedicated pedestrian dataset from Caltech is used to train the BING template, and the aspect ratio of the template is set to 1:2, which is "8×16" according to the appearance characteristics of a pedestrian. In addition, only the window with the aspect ratio of 1:2 is reserved during the detection phase, that is, the pedestrian objects at 8 different scales are detected. Finally, the pedestrian candidate windows extracted with the BING template are input to the SVM splitter for accurate classification. The experimental results show that the detection process time in this paper is only one-third of that in the original sliding window detection, while the detection accuracy does not degrade.

The features extracted by CNN (convolutional neural network) are commonly better than those of the traditional algorithms. We are going to combine the manual design candidate window with CNN to further improve the detection performance.

Acknowledgement

This work was supported partially by the National Natural Science Foundation of China (Grants No. 61763033, 61866028, 61662049, 61741312, 61881340421, 61663031, and 61866025), the Key Program Project of Research and Development (Jiangxi Provincial Department of Science and Technology) (20171ACE50024, 20161BBE50085), the Construction Project of Advantageous Science and Technology Innovation Team in Jiangxi Province (20165BCB19007), the Application Innovation Plan (Ministry of Public Security of P. R. China) (2017YYCXJXST048), and the Open Foundation of Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (ET201680245, TX201604002).

References

[1] Papageorgiou C, Poggio T., "A trainable system for object detection." *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15-33, Nov. 2000.

[2] Dalal N, Triggs B., "Histograms of Oriented Gradients for Human Detection." *Computer Vision and Pattern Recognition (CVPR)*, San Diego, pp. 886–893, Jun. 2005.

[3] Dollar P, Wojek C., "Pedestrian detection: A benchmark." *Computer Vision and Pattern Recognition (CVPR)*, Miami, pp. 304-311, Jun. 2009.

[4] Alexe B, Deselaers T, Ferrari V., "Measuring the objectness of image windows." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189-2202, 2012.

[5] Endres I, Hoiem D., "Category-independent object proposals with diverse ranking." *IEEE transactions on pattern analysis and machine intelligence,* vol. 36, no. 2, pp. 222-234, 2012.

[6] Alexe B, Deselaers T, Ferrari V., "What is an object?" in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE*, San Francisco, pp. 73-80, Jun. 2010.

[7] Endres I, Hoiem D., "Category independent object proposals," in *European Conference on Computer Vision*. Springer Berlin Heidelberg, pp. 575-588, 2010.

[8] Van de Sande K E A, Uijlings J R R, Gevers T., "Segmentation as selective search for object recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE*, Barcelona pp. 1879-1886, Nov. 2011.

[9] Zhang Z, Warrell J, Torr P H S., "Proposal generation for object detection using cascaded ranking svms," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE,* Barcelona pp. 1497-1504, Nov. 2011.

[10] Uijlings J R R, van de Sande K E A, Gevers T., "Selective search for object recognition," in *International journal of computer vision*, vol. 104, no. 2, pp. 154-171, Sep. 2013.

[11] Girshick R., "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision,* Santiago, pp. 1440-1448, Dec. 2015.

[12] Girshick R, Donahue J, Darrell T., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Columbus, pp. 580-587, Jun. 2014.

[13] Felzenszwalb P F, Huttenlocher D P., "Efficient graph-based image segmentation." *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167-181, Jun. 2004.

[14] Cheng M M, Zhang Z, Lin W Y, et al., "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proceedings of the IEEE conference on computer vision and pattern recognition.* Columbus, pp. 3286-3293, Jun. 2014.

[15] Forsyth D A, Malik J, Fleck M M., "Finding pictures of objects in large collections of images," in *International Workshop on Object Representation in Computer Vision*. Springer Berlin Heidelberg, pp. 335-360, 1996.

[16] Heitz G, Koller D., "Learning spatial context: Using stuff to find things," in *European conference on computer vision*. Springer Berlin Heidelberg, pp 30-43, 2008.

Authors

Jiexian Zeng received his master's degree in engineering from Northwestern Polytechnical University in 1997. Currently, he is a professor at Nanchang Hangkong University. He is mainly engaged in image processing, pattern recognition and computer vision research. He has published over 100 papers.

Qi Fang received a bachelor's degree from Yanjing Institute of Technology, Hebei province, China. He has obtained a master's degree from Nanchang Hangkong University.

Zhe Wu received his master degree from Nanchang Hangkong University, in 2017 and received the bachelor degree from Wuhan Donghu University, in 2014. His research interests include image processing, object detection, and semantic segmentation.

Xiang Fu received his Ph.D. degree from Xidian University in 2008. Currently, he is an associate professor at Nanchang Hangkong University. His research interests include computer vision, image processing and pattern recognition.

Lu Leng received his Ph.D. degree from Southwest Jiaotong University, Chengdu, P. R. China, in 2012. He performed his post-doctoral research at Yonsei University, Seoul, South Korea, and Nanjing University of Aeronautics and Astronautics, Nanjing, P. R. China. He was a visiting scholar at West Virginia University, USA. Currently, he is an associate professor at Nanchang Hangkong University.

He has published more than 60 international journal and conference papers and been granted several scholarships and funding projects for his academic research. He is the reviewer of several international journals and conferences. His research interests include image processing, biometric template protection, and biometric recognition.

Dr. Leng is a member of the Institute of Electrical and Electronics Engineers (IEEE), the Association for Computing Machinery (ACM), the China Society of Image and Graphics (CSIG), and the China Computer Federation (CCF).