REAL-TIME PEDESTRIAN DETECTION IN CROWDED SCENES USING DEEP OMEGA-SHAPE FEATURES

Yuting Xu, Xue Zhou, Pengfei Liu and Hongbing Xu*

School of Automation Engineering
University of Electronic Science and Technology of China, Chengdu, China
*Corresponding author: zhouxue@uestc.edu.cn

ABSTRACT

Region-based Fully ConvNet (R-FCN) designed for general object detection is difficult to be directly applied for pedestrian detection, due to being with large human pose and scale changes, and even with partial occlusion in surveillance scenarios. This paper presents a real time pedestrian detection method with partial occlusion handling, which builds on the framework of Region-based Fully ConvNet. We introduce a deep Omega-shape feature learning and multi-paths detection to make our detector being robust to human pose and scale changes. A novel predicted boxes fusion strategy is proposed to reduce the number of false negatives caused by partial occlusion in crowded environment. Our end-to-end approach achieved 95.35% mAP on the Caltech dataset and 97.43% on Bronze dataset at a test-time speed of 86ms second per image.

Index Terms— Pedestrian detection, Deep Omega-shape features, Non-maximum suppression with bootstrap, Region-based fully ConvNet

1. INTRODUCTION

Pedestrian detection has attracted more and more attention in recent years, as it is the foundation task of real-world applications such as automatic driving and intelligent surveillance. The state-of-art pedestrian detectors are hybrid methods which combine traditional, hand-crafted features and deep convolution features. These methods (e.g. [1, 2, 3, 4, 5, 6]) make progress in detection accuracy at the price of expensive time and memory costing, because of the structure of hybrid features and ensemble classifiers. They are not suitable for real-time applications on mobile platforms. On the other hand, region-based fully convolution network (R-FCN [7]) achieves 79.0% mean average precision (mAP) on PASCAL VOC [8] benchmark at the testing speed of 8 frames per second (FPS). Although these end-to-end methods (e.g. [7, 9, 10]) have achieved a good trade-off between accuracy and speed for general object detection, they have not presented satisfactory results on popular pedestrian detection datasets (e.g. the Caltech set [11]) because of lacking specialized optimization for pedestrian detection.

As is known to all that pedestrian detection is much harder than general object detection because of the various poses of people and different scales of changes. Moreover multiple people often occur in close proximity, which may cause different degrees of occlusion, making it particularly challenging to distinguish between nearby individuals. These reasons may explain the dissatisfactory performance directly using the aforementioned general object detection methods.

It is reasonable to fine-tune a convolutional network for pedestrian based on the general object detection framework. It can not only take the advantage of fast detection speed, but also solve the tough problems in pedestrian detection by adopting a set of optimized methods. RPN+BF [12] combines the region proposal network (RPN [13]) and boost forests (BF) with hard examples mining strategy to solve the the problem of different scales and various poses of people. Different from the above full-body human detectors which easily suffer from occlusions among individuals, some researches [14, 15] instead focus on the Omega-shape model (namely the head-shoulder part of human body), especially for top view surveillance scenes.

Similarly, this paper proposed a simple but effective method based on the aforementioned general object detection pipeline for real-time pedestrian detection. Our method selects R-FCN [7] as the baseline, because it can achieve a better trade-off between accuracy and speed [16]. This two-stage method consists of a RPN which aims to generate bounding boxes of the potential objects and position-sensitive score maps which are designed as the features for classification and location tasks.

The contributions of our work can be summarized as follows: (a) we choose the deep Omega-shape features as the representation model of people, making it possible for the R-FCN to learn robust features. (b) we introduce multi-paths detection and online hard examples mining to improve the detection performance in multiple scales and complex scenes. (c) we propose a non-maximum suppression with a bootstrap strategy which can pertinently solve the problem of partial occlusion in crowded scenes.

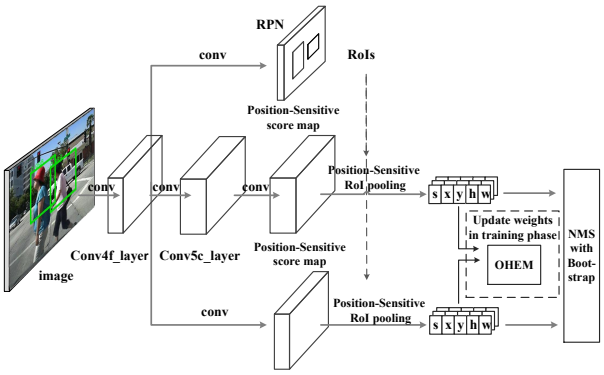

Fig. 1. The framework consists of the basic RPN, multi-path R-CNN and our rectified NMS. where (s, x, y, h, w) denotes the confidence score and centric location of detection boxes. OHEM is only used in training phase, making extracted features more discriminative.

2. PROPOSED METHOD

Our approach inherits the advantages of R-FCN [7], which employs a region proposal network to generate candidate regions of interests (RoIs) and the position-sensitive score maps to classify and locate targets. Meanwhile, novel residual network (ResNet-50 [17]) is selected as the backbone of our framework for its fewer computations and parameters without accuracy decrease. Different from native R-FCN, We append another position-sensitive score map followed by the conv4f-layer and a fusion strategy of online hard examples mining in training phase to detect small objects. Fig. 1 illustrates the framework of our method.

2.1. Deep Omega-shape features

Most of the aforementioned methods use full-body human for detection, they are easily lead to unsuccessful detection when confronted with large pose variation and partial occlusions. Hence, different from full-body human detectors, a novel work [18] focused on the head part of human, and successfully proposed an effective pedestrian detection method in crowded scenes. It has been proved in our former research [19] that the Omega-shape model is a salient feature of head-shoulder region, especially for top view surveillance scenes. We took two different ways of annotation: the full-body annotation and the Omega-shape annotation on the same dataset. Then, we separately fine-tuned two models on the same dataset with different annotations. The different detection results are shown in Fig. 2. It is obvious that the deep Omega-shape features outperform full-body features for its lower probability of missing detection. In addition, Omega-shape annotation reduces the overlap rate of Ground-Truth boxes, making subsequent processing easier. The head-shoulder part of human body changes slightly despite the state of poses and angles, which makes its within-class distance smaller than full-body models.

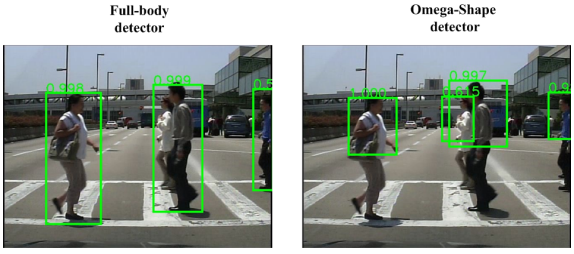

Fig. 2. The boxes denote the final detection results on Caltech dataset, with a confidence score above them. The left column uses a full-body annotation model, while the right column uses deep Omega-shape features.

2.2. Multi-paths detection with OHEM

Other than a good feature representation model, some detail tricks are requested after the analysis of real applications. Zhang et al. [12] found that RPN does have the ability to achieve competitive results on proposal quality, however the accuracy is degraded after feeding these proposals into the R-CNN [20] classifier. We argue that such unsatisfactory performance is attributed to two reasons as follows.

First, the responses of some small-size pedestrians disappear in the position-sensitive score map because of too large receptive field [21] in the last convolution layer (the conv5c-layer). Inspired by SSD [9] which merges multi-paths detection results, another detection path is added in our framework. A new position-sensitive score map followed by conv4f-layer together with the native one followed by conv5c-layer are both used to detect multi-scale pedestrians. The two paths with different receptive fields are complementary for each other.

Second, in pedestrian detection the false predictions are dominantly caused by confusions of hard background instances. So it is significant to adopt hard examples mining [22], which selects hard examples to perform back propagation during training. OHEM is nearly a cost-free hard examples mining because of all shared computations before position-sensitive RoI pooling. Our loss function defined on each RoI is a summation of the cross-entropy loss and the box regression loss from deferent detection paths:

$$L(s, t_{x,y,w,h}) = \sum_{i=1}^{2} \left(L^i_{cls}(s, c^*) + \lambda[c^* = 1] L^i_{reg}(t_{x,y,w,h}, t^*_{x,y,w,h}) \right) \quad (1)$$

where $c*$ is the RoI's ground-truth label, i is the number of detection paths (two detection paths in total), L_{cls} is the cross-entropy loss for classification, L_{reg} is the bounding box regression loss using $smooth_{L1}$ metric and t^* represents the ground truth box. $[c^* = 1]$ is an indicator which equals to 1 if the argument is true and 0 otherwise.

2.3. Non-maximum suppression with bootstrap

In state-of-art detectors [7, 9, 13], non-maximum suppression (NMS [23]) is used to obtain the final set of detections as it

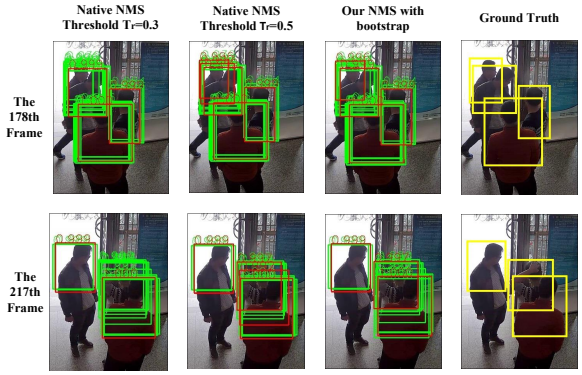

Fig. 3. The green boxes denote the detection boxes before NMS, and the red ones denote post-NMS predicated boxes. The native NMS usually causes either missing detection or duplicate detection. Our bootstrap strategy can deal with this problem, especially in the condition of partial occlusion.

significantly reduces the number of false positives. The native NMS method often leads to missing detection of neighbour pedestrians especially in crowded scenes. Different from the traditional greedy NMS, Soft-NMS [24] decreases the detection scores according to an increasing function of overlap instead of directly setting the scores to zero. This method does not work well enough in crowded scenes, in addition, changing scores of detection boxes is not beneficial to subsequent analysis. Hence, we propose a rectified method with a strategy of bootstrap. Overlap rates r and similarity metrics m of two boxes are used simultaneously to distinguish whether the two boxes are belong to the same person. We define a similarity metric for bootstrap strategy in Eq.(2).

$$\begin{aligned} m &= e^{-((1-2\lambda)*dxy+\lambda*dw+\lambda*dh)} \\ dxy &= \sqrt{\left(\frac{x^*-x}{w}\right)^2+\left(\frac{y^*-y}{h}\right)^2} \\ dw &= \|\frac{w^*}{w}-1\| \\ dh &= \|\frac{h^*}{h}-1\| \end{aligned} \tag{2}$$

where (x, y, w, h) denotes the box's center coordinates, its width and height, respectively. Variables (x^*, y^*, w^*, h^*) are for the predicted box with a maximum score in the sorted queue Q, and (x, y, w, h) for the other boxes in the sorted queue. $\|\cdot\|$ denotes the L1-norm. λ denotes a weight, here we set $\lambda = 0.3$. The final metric is a weighted sum of the three deviations which represent the similarity degree of two detection boxes. The processing scheme of our rectified NMS is shown in Algorithm 1.

We find that the similarity metric is a salient difference among nearby boxes because of the different figures and poses for different individuals. Moreover, our defined similarity metric is insensitive of parameters, it can perform well under a large scale of threshold T_m range from 0.2 to 2.0. We used the native NMS and our NMS with bootstrap respectively after the R-FCN detector, great improvements have been achieved in Fig. 3. Our NMS with a bootstrap strategy does not require any extra training and is almost as fast as the native version in terms of implementation, it can be easily integrated into any other object detection pipeline.

Algorithm 1 Non-maximum suppression with bootstrap strategy

Inputs :detection boxes X with scores S,
threshold of similarity metrics T_m

Output :results list L

1. select the detection boxes X whose score S is greater than 0.5 and sort them in a descending order based on scores S in a queue Q.
2. calculate the overlap rates r and similarity metrics m between the first one and the rest boxes in the sorted queue Q according to Eq.(2).
3. remove the boxes in sorted queue Q whose overlap r is greater than 0.3 to a new backup queue B.
4. search the boxes in the backup queue B whose similarity metric m is greater than a threshold T_m and meanwhile overlap r below 0.5 (namely, meet the condition of $(0.3 < r < 0.5)\&(m > T_m)$).
5. append the searched boxes in B to the bottom of sorted queue Q, and then remove the first box from the sorted queue Q to the output list L.
6. return the output list L if the sorted queue Q is empty otherwise repeat steps 2–5.

Table 1. Comparions of deep Omega-shape and full-body features on Bronze and Caltech sets. All methods are based on standard R-FCN with an optional strategy of OHEM.

Datasets	annotation	OHEM	mAP
Bronze	full-body		88.01%
Bronze	omega-shape		93.69%
Bronze	full-body	yes	88.58%
Bronze	omega-shape	yes	**94.01%**
Caltech	full-body		82.40%
Caltech	omega-shape		81.35%
Caltech	full-body	yes	93.05%
Caltech	omega-shape	yes	**93.43%**

3. EXPERIMENTS

3.1. Implementation Details

We comprehensively evaluated our method on two datasets: Caltech [11] and Bronze. The Caltech dataset is made up of approximate 250,000 images taken by in-vehicle cameras, with a resolution of 640*480 pixels. The Bronze dataset is our self-built dataset with images taken in an indoor surveillance scenario from the top view. It contains 2,600 pictures with 26,895 bounding boxes. Each image has a resolution of 960*540 pixels and ten instances on average. Two annotations according to full-body and human head-shoulder omega shape were labelled before training.

A predicted box is considered as a positive example if it has an Intersection-over-Union (IoU) ratio greater than 0.5

Table 2. Different combinations of tricks on two datasets using only Omega-shape annotation.

Datasets	OHEM	Multi-paths	NMS with bootstrap	mAP
Bronze				93.69%
Bronze	yes			94.01%
Bronze	yes	yes		94.72%
Bronze	yes		yes	96.10%
Bronze	yes	yes	yes	**97.43%**
Caltech				81.35%
Caltech	yes			93.43%
Caltech	yes	yes		94.44%
Caltech	yes		yes	94.40%
Caltech	yes	yes	yes	**95.35%**

with one ground truth box, and otherwise negative. This article adopts the area under the Precision-Recall curve (AUC) as the mean average precision (mAP) of the detection algorithms. We fine-tuned RPN to generate 300 proposals in both training and testing phases, and then feed them into the R-CNN classifier. The first half samples which are considered as hard examples in a mini-batch are adopted to update the weights in back propagation. After the convolution neural network, we adopt our rectified non-maximum suppression with a similarity threshold $T_m = 0.8$ to fuse the predicted boxes. Other hyper-parameters of the network are the same as in R-FCN [7]. Our proposed method is fine-tuned with a learning rate of 0.001 for the first 40k mini-batches and 0.0001 for the second 30k mini-batches. We achieved a test-time speed of 86ms on single Nvidia 1080 GPU using the platform of Caffe. Code and datasets are made publicly available at: *http://github.com/xuyuting45/pedestrian-detection-in-crowded-scenes*.

3.2. Experiments on two datasets

We fine-tuned the standard R-FCN on both Caltech and Bronze datasets to analyze which kind of annotation method is better for pedestrian detection. The performances in the form of Precision-Recall curve are shown in Fig. 4(a) and the quantitative indexes are given in Table 1. On Bronze dataset, it is obvious that deep Omega-shape features are superior to full-body features at an increasing mAP of 5.4–5.7 percent. On Caltech dataset, deep Omega-shape features are a little inferior to full-body model when OHEM is not adopted. Because there are some small size of pedestrians in Caltech dataset (about 40*20 pixels). The head-shoulder region is half the size of full body, which makes the head-shoulder region a hard example to detect. However, deep Omega-shape features can achieve additional 0.4 percent of improvement than full-body features when OHEM is adopted. So we can draw a conclusion that deep Omega-shape features are more suitable for pedestrian detection in surveillance scene by employing OHEM.

We designed an experiment of different combinations of tricks on both Bronze and Caltech datasets with only Omega-shape annotation. The results are listed in Table 2. Our structure of multi-paths detection can achieve an increase of 0.7–1.3 percent of mAP on Bronze and 0.95–1 percent on Caltech. Meanwhile our rectified non-maximum suppression with a bootstrap strategy can obtain satisfactory 2.1–2.7 percent of improvements on Bronze and 0.9–1 percent on Caltech. On the other hand, OHEM is necessary for Caltech, which achieves more than 12 percent of promotion, Because OHEM can well solve the problem of confusion instances on a complex background.

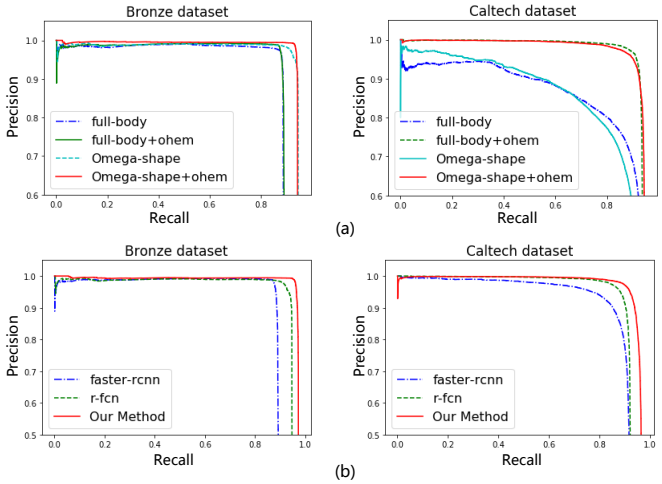

Fig. 4. (a) shows the comparisons of full-body and deep Omega-shape features on two datasets. (b) shows the improvements of our method compared with faster-rcnn and r-fcn.

In addition, we made a comparison of our method with fine-tuned faster-rcnn [13] and r-fcn [7] on both Bronze and Caltech datasets. It is obvious in Fig. 4(b) that our method makes satisfactory promotion on the base of state-of-art algorithms, which achieves a mAP of 97.43 percent on Bronze and 95.35 percent on Caltech.

4. CONCLUSION

In this paper, we present a simple but effective method based on Region-based fully convolution network [7] for pedestrian detection. A successful transfer learning from state-of-art general object detection to pedestrian detection was accomplished by our work. We find that deep Omega-shape features are more effective than full-body representation model in surveillance. Not only the detection accuracy but also the testing speed are promoted by our three improvements.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61472063) and in part by the 2018 Fundamental Research Funds for the Central Universities.

5. REFERENCES

[1] Yonglong Tian, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Pedestrian detection aided by deep learning semantic tasks" in *CVPR*, pp. 5079-5087, 2015.

[2] Dollar Piotr, et al. "Fast feature pyramids for object detection" in *PAMI*, vol.36, no. 8, pp. 1532-1545, 2014.

[3] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele, "Filtered channel features for pedestrian detection" in *CVPR*, 2015.

[4] Dollar Piotr, et al. "Integral channel features" in *British Machine Vision Conference* , 2009.

[5] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele, "Taking a deeper look at pedestrians" in *CVPR*, pp. 4073-4082, 2015.

[6] Zhaowei Cai, Mohammad Saberian, and Nuno Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection" in *ICCV*, pp. 3361-3369, 2015.

[7] Jifeng Dai, Yi Li, Kaiming He, Jian Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks" in *Advances in neural information processing systems*, pp. 379-387, 2016.

[8] Everingham, Mark, et al. "The PASCAL Visual Object Classes (VOC) Challenge" in *IJCV*, vol.88, no. 2, pp. 303-338, 2010.

[9] Liu Wei, et al. "SSD: Single Shot MultiBox Detector" in *ECCV*, pp. 21-37, 2016.

[10] Joseph Redmon, Ali Farhadi, "YOLO9000: Better, Faster, Stronger" in *arXiv preprint arXiv:1612.08242*, 2016.

[11] Dollar Piotr, et al. "Pedestrian detection: An evaluation of the state of the art" in *TPAMI*, vol.34, no. 4, pp. 734-761, 2012.

[12] Liliang Zhang, Liang Lin, Xiaodan Liang, Kaiming He, "Is Faster R-CNN Doing Well for Pedestrian Detection?" in *ECCV*, pp. 443-457, 2016.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Network" in *NIPS*, pp. 91-99, 2015.

[14] Li, Min, et al. "Rapid and Robust Human Detection and Tracking based on Omega-Shape Features" in *IEEE International Conference on Image Processing*, pp. 2545-2548, 2009.

[15] Li, Min, et al. "Estimating the Number of People in Crowded Scenes by Mid-based Foreground Segmentation and Head-shoulder Detection" in *International Conference on Pattern Recognition*, pp. 1-4, 2008.

[16] Jonathan Huang, et al. "Speed/accuracy trade-offs for modern convolutional object detectors" in *CVPR*, 2017.

[17] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun, "Deep Residual Learning for Image Recognition" in *CVPR*, pp. 770-778, 2016.

[18] Stewart Russell, Mykhaylo Andriluka and Andrew Y. Ng, "End-to-end people detection in crowded scenes" in *CVPR*, pp. 2325-2333, 2016.

[19] Pengfei Liu, Xue Zhou, and Shibin Cai, "Omega-Shape Feature Learning for Robust Human Detection" in *CCPR*, pp. 290-303, 2016.

[20] Girshick Ross, "Fast R-CNN" in *arXiv preprint arXiv:1504.08083*, 2015.

[21] Karel Lenc and Andrea Vedaldi, "R-CNN minus R" in *arXiv preprint arXiv:1506.06981*, 2015.

[22] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, "Training region-based object detectors with online hard example mining" in *CVPR*, pp. 761-769, 2016.

[23] Zitnick, C. Lawrence and Dollar, Piotr, "Edge Boxes: Locating Object Proposals from Edges?" in *ECCV*, pp. 391-405, 2014.

[24] Navaneeth Bodla, et al. "Soft-NMS – Improving Object Detection With One Line of Code" in *ICCV*, pp. 5561–5569, 2017.