Multi-layer Feature Fusion Network with Atrous Convolution for Pedestrian Detection

You Li\textsuperscript{1,*}, Qingxuan Zhang\textsuperscript{1} and Yulei Zhang\textsuperscript{1}

Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology.

\textsuperscript{*}ly541911136@163.com

Abstract. In this paper, we present a simple but effective framework K-AFPN that incorporates feature pyramid method for small-size pedestrian detection, fully utilizing the lower-layer detail features and higher-layer semantic features. The method not only enhances the robustness of the features, but also improves the discrimination of the feature maps, achieving competitive accuracy. In addition, atrous convolution is used to optimize the network for high-resolution feature maps, avoiding information loss caused by frequent down or up sampling. On top of the backbone network, K-means algorithm is used to obtain optimal initial anchor base sizes, which reduces computational costs and improves location accuracy. Hence, our method pays more concentration on pedestrians, especially those of relatively small size. Comprehensive experimental results on two classic pedestrian benchmarks illustrate the effectiveness of the proposed approach.

1. Introduction

Pedestrian detection is one of the significant research tasks in computer vision, involving multiple technologies, such as computer graphics, artificial intelligence, image processing and deep learning. It plays a critical role in several real-world applications including robotics, automatic driving, human behaviour analysis, video surveillance and intelligent transportation \cite{1,2}.

Different from general object detection, pedestrian detection suffers from the interference caused by the complex backgrounds, various lighting conditions, different camera shooting angles and other factors. While many CNN based approaches have been proposed, pedestrian detection task remains challenging problems because that the appearance of pedestrians is susceptible to diversified appearances, walking postures, low resolution, occlusion issues and so on. So, it is a valuable and challenging subject in the field of computer vision.

We propose an effective pedestrian detection approach on the Faster R-CNN framework in this paper, which can overcome the limitations of current pedestrian detection methods, especially for the small-size pedestrian detection, and get rid of traditional hand-crafted features. At the same time, by optimizing the anchor scales, the detection miss rate is reduced and the generalization ability of the model is improved. We performed experiments on the Caltech and INRIA datasets respectively, and the results demonstrate that our method achieves competitive performance. The performance of small-scale objects is improved by 6% on the Caltech dataset, which is significantly better than the traditional methods and other Faster R-CNN based methods.

2. Related Works
Many efforts have been made to improve detection performance in recent years. Existing methods fall into two main categories: hand-crafted features based [3][4][5][6] and deep learning features based. With the establishment of various large-scale image datasets and the improvement of hardware computing capability, the application of convolutional neural network (CNN) makes computer vision tasks obtain state-of-the-art results in terms of accuracy [7][8][9][10]. The advantage of CNN is that it can autonomously learning deep semantic features from raw pixels. Thus, it is more robust for detecting objects from complex backgrounds. While deep learning features have improved object detection performance [11], there are still several shortcomings for pedestrian detection. In practical applications, such as autonomous driving and intelligent surveillance, pedestrians are of small sizes (e.g., 28*70 for Caltech). Small objects are turned to low-resolution feature maps after layer-by-layer down-sampling, and these features are not discriminative for detection [12].

Faster R-CNN [13][14][15] is a particularly successful approach based on CNN, which has leading accuracy on general object detection benchmarks. Therefore, we propose a multi-scale deep feature learning detector for pedestrian detection, which adopting Faster R-CNN as the basic framework, aiming at getting rid of hand-crafted features. Specifically, the detector is focus on pedestrian regions for its scale-awareness. We use the feature pyramid method to fuse multi-layers features in the stage of feature extraction, so that the lower-layer detail features and higher-layer semantic features can be fully utilized. The feature pyramid method enhances the robustness of the features, and improves the detection accuracy. Meanwhile, we adopt atrous convolution (i.e., dilated convolution) [16][17][18], avoiding continuous subsampling and increasing higher-layer feature map size. In addition, we also use K-means algorithm to obtain the optimal initial scales of anchors for fine-tuning the network parameters to improve the location accuracy and reduce the computational cost.

3. Approach
Our goal in introducing the K-AFPN method is to provide a better way to solve the small-scale objects detection problem in practical application. An overview of our approach is depicted in Figure 1. Given an input image, the network generates multiple convolutional feature layers with high resolution. The backbone of K-AFPN is a Resnet-101 network, which is fast and accurate for object recognition [19].

![Figure 1. Framework of K-AFPN.](image)

3.1 Feature Pyramid Module with Atrous Convolution
A multitude of previous object detection methods used only top-layer features for prediction, because the deeper features tend to encode more global and semantic information of objects that is robust
against appearance variations. Nevertheless, although the outputs of lower layers contain less semantic information, the features provide more precise localization. We use deep residual network (ResNet) as the backbone, which is an extremely deep convolutional neural network model [19]. It can avoid problems such as gradient disappearance and precision degradation caused by simple stacking of convolutional layers. The model is easier to be optimized and has excellent performance.

Generally speaking, the deeper the network is, the stronger the translational and rotational invariance it has, which is of positive significance for ensuring the robustness of the model. For detection problems, the object locating task requires models to have a good perception of position information, but excessive translational and rotational invariance will weaken this performance. For deeper full convolutional neural networks, an obvious drawback of the Faster-RCNN detection framework [15] is that the sensitivity of the detector to the object position information and the detection accuracy decreases.

Region proposal network (RPN), the bounding box proposal generation module of Faster R-CNN, is a sliding-window object detector. Two paratactic 1*1 convolutions for classification and regression are designed to be a head, which follows a 3*3 convolutional layer on top of the final feature map. Object classification and bounding box regression are defined towards the anchors of multiple hand-crafted scales and aspect ratios. The anchors are designed to cover targets of different sizes.

In general, the most intuitive solution to the drawback of Faster R-CNN is to move the position of the RPN to shallower layers. However, this method will significantly increase the amount of computation of the down-stream classifier Fast-RCNN, making the detection speed slower and the feature robustness lower. This problem is alleviated in our method, as it is not constrained to use features of a certain layer or several shallow layers like previous method. We adopt Feature Pyramid Network (FPN) [20] to improve RPN for extracting region proposals, and attach the head to each level for prediction on every convolutional block independently. And we obtain 3 areas of the anchors through K-means clustering on \{P2, P3, P4, P5, P6\} respectively. So, it is unnecessary to have multiple aspect ratio on each level. Instead, we set the aspect ratio to be \{1:1\}. Totally, there are still 15 anchors over the framework.

We fix the number of feature dimension in all the feature maps to be 256-channel according to [20], and distribute the training labels as follows: An anchor is considered as a positive label, which has the highest IoU ratio for a given ground-truth bounding box or an IoU over 0.7 with any ground-truth bounding box. While if it has IoU lower than 0.3 for all ground-truth boxes, it is considered as a negative label.

It is a critical issue in pedestrian detection that detecting small-size objects is inclined to bring about feeble contrast and motion blur, while FPN can effectively handle small targets. The reason is that the structure of FPN can be described as two parts: the bottom-up part is the forward propagation process of neural networks as usual, and the feature map becomes smaller and smaller through convolution calculations; the top-down process is to up-sample the deeper features that are more abstract and more semantic. The horizontal connection is used to change the number of channels of the shallower-layer feature maps. Then the processed deep features and the processed shallow features are merged, and independent prediction is performed at each layer. Therefore, features from shallower but higher-resolution layers can provide more accurate details, meanwhile the deep-layer features provide stronger representation. The feature map of each prediction layer combines features of different resolutions and semantic intensities, so that the features are enhanced to be more discriminative, ensuring that objects of different sizes can be detected, especially for small targets. This method does not change the size of the acquired features by image scaling like image pyramid, which greatly improves the detection performance without increasing the calculation amount of the original model too much.

However, information loss occurs during multiple subsampling and up-sampling operations, so the location information is still not completely refined. Hence, how to minimize the information loss in the process of acquiring a larger perceptive field is the key to making detection more accurate.
Standard convolution has several problems in certain scenes such as image semantic segmentation. Reducing calculation amount and increasing the perceptive field through pooling operation, and expanding the image to the original size by deconvolution (up-sampling) lead to information loss, especially spatial structure information. And another problem caused by these operations is that small object information cannot be reconstructed. Dilated convolution [16][17][18] can enlarge the perceptive field without pooling, avoiding the compression of the image resolution, and the data structure inside the image is retained, which can obtain a better performance than the standard convolution.

There is an additional dilation rate in the dilated convolution to control the size of the kernel size expansion. The larger the dilation rate parameter is, the larger the perceptive field of same convolution kernel size dilates. We use dilated convolution to increase higher-layer feature map size.

Assuming that the input signal is two-dimensional, for each position \( i \) on an input \( y \) and the filter \( \omega \), the calculation formula for applying the atrous convolution on the input feature map \( x \) is:

\[
y[i] = \sum_k x[i + r \cdot k] \omega[k]
\]

where the hole rate \( r \) is the sampling stride, that is, to insert \( r-1 \) zero points between two neighboring convolution kernel values in each spatial dimension of the filter. Standard convolution is a special form of \( r = 1 \) for atrous convolution. The atrous convolution can modify the perceptive field of the filter by changing the value of the hole rate \( r \).

Atrous convolution can increase the perceptive field, but because the convolution kernel values are discontinuous, continuity of information may be lost, which has an impact on small objects detection. So, we only apply the atrous convolution to \( \text{Conv}4_x \) layer and \( \text{conv}5_x \) layer of ResNet101. The final feature map for ResNet101 is 32 times smaller than the size of the input image, so the output stride is 32. We set the step size of the last reduced resolution pooling layer or convolutional layer to 1 to avoid signal extraction, and the subsequent convolutional layers are replaced by atrous convolutional layers of \( r=2 \), which makes the final output stride remaining 8 and no longer increasing to 32, that can increase resolution and reduce stride. Thus, the network is able to extract denser features without having to learn any additional parameters. In addition, the deep feature map eliminates the up-sampling operation when merging, and only needs to change the dimension by 1*1 conv to merge with the upper feature map, which reduces information loss. In addition to the improvement of FPN with atrous convolution, we adopt the adaptations in the same fashion as in [20].

### 3.2 Anchor Clustering

The original RPN with FPN has anchors of hand-crafted aspect ratios and numbers. The parameters are adjusted to approach the ground truth with the iteration during training. Anchors of inappropriate aspect ratios are associated with few examples, so are noisy and harmful for detection accuracy and speed. We address this problem by adopting K-means algorithm.

Therefore, we conduct dimension clustering of ground truth boxes with K-means method. Taking \( K \) as the number of anchors, the sizes of the \( K \) clustering center boxes are set to be the dimensions of the anchors. The optimal number and dimensions of anchors can improve positioning accuracy and reduce unnecessary calculation. The traditional K-means uses the Euclidean distance function, which means that larger boxes will generate more errors than smaller ones. Hence, we use IOU so that the results are irrelevant to the size of the candidate boxes.

The distance function is set according to the method in [21]:

\[
d(box, truth) = 1 - IOU(box, truth)
\]

The objective function of the clustering operation is:

\[
S = \min \sum_{i=0}^{K} (1 - IOU(box - truth))
\]

where box is the candidate bounding box, truth is the ground truth box, and \( K \) is the number of anchors.
We clustered the Caltech dataset and the optimal dimensions are \[[9,22],[13,25],[12,35]], [[18,32],[17,42],[21,53]], [[34,37],[26,56],[23,70]], [[31,74],[36,97],[60,52]], [[53,129],[156,82],[100,244]]\), distributed on the five prediction layers according to the results of K-means algorithm.

4. Experiments

4.1Datasets
We evaluated the proposed approach on two benchmarks: Caltech [1][22] and INRIA [4]. Here we give a brief description of these datasets. The log-average miss rate is used to evaluate the detection performance and is calculated by averaging miss rate on false positive per-image (FPPI) in the range of \([10^{-2},10^0]\), denoted MR.

Acting as a representative benchmark for pedestrian detection, Caltech pedestrian is currently the most challenging and large-scale dataset. Over 70% of the annotated pedestrian instances have a height smaller than 100 pixels, including extremely tiny instances under 50 pixels. Pedestrians in the dataset are grouped into 3 subsets by their height in pixels: near (80 or more pixels), medium (between 30-80 pixels), and far (30 pixels or less). Assuming an urban speed of 55km/h, an 80-pixel person is just 1.5s away, while a 30-pixel person is 4s away. Thus, detection in the medium scales is essential for automotive applications. In the experiment, we extract 42782 images for training by sampling every 3rd frame from training sets, and evaluate the framework on a standard test set consisting of 4024 frames. And the performance is measured under the Medium subset [22].

The INRIA dataset is a classic pedestrian dataset. It has various backgrounds and high-quality annotations of pedestrians. The training set contains 2416 pedestrians, distributed in 614 positive training images. The test set contains 1126 pedestrians, distributed in 288 positive sample images [4].

4.2 Evaluation Metrics
To evaluate the detection performance, we plot miss rate against false positives per image (FPPI) (using log-log plots) by varying the threshold on detection confidence, according to the evaluation metrics defined by the Caltech Pedestrian method. We set the threshold to be 50 percent. When the IoU of a detected Bounding Box and a ground truth bounding box exceeds the threshold, then we consider that they are matched. The lower the MR-FPPI curve is, the better the performance of the pedestrian detection method will be [22].

4.3 Results and Analysis
For our experiments, the model was implemented using the TensorFlow machine learning library, and it was trained on a NVIDIA GeForce GTX 1080Ti GPU. For the feature extraction network, we used the ImageNet pre-trained ResNet-101 as initial parameters.

To train the network, the standard stochastic gradient descent with momentum was employed for optimization, where the initial learning rate, momentum, and weight decay were set to 0.001, 0.9, and 0.0005. By default, an IoU threshold of 0.5 was used for determining “true positives”. The NMS was adopted both region proposal stage and classification stage to avoid redundant detections.
Figure 2 indicates that the ROC plot of miss rate against FPPI for the available top performing approaches based on Faster RCNN and several other outstanding methods reported on Caltech medium and reasonable subsets. In medium scale case, as shown in Figure 2 (a), our K-AFPN achieves 35% miss rate, which significantly outperforms the other detectors. Using K-means to initialize the anchor scale instead of hand-crafted settings improve the performance by 4%. As depicted in Figure 2 (b), our detector is slightly better than the previous best methods on Caltech reasonable subset. Although the performance improvement is not very remarkable on the reasonable subset, since our approach is mainly aim at small targets, but it still acquires the state-of-the-art performance with 9% miss rate.

For INRIA, we trained our model with 614 positive images by excluding negative images and evaluated on the test set. Figure 3 shows the comparison with previous methods on INRIA test dataset in terms of miss rate, which illustrates that our method gets a compelling result with 9% miss rate, better than the competitive approaches by 2%. The performance reveals that our approach, taking full advantage of multi-level features, has strong generalization ability, which yields state-of-the-art results even if the training set is very limited.

Figure 3. Comparison results of K-AFPN with state-of-the-art methods on INRIA dataset (lower is better).

5. Conclusion
A K-AFPN method for pedestrian detection is proposed in this paper, which merges detail information on shallower layers into higher-layer features to gain more discriminative features for detection, while most existing pedestrian detection methods only consider low resolution features extracted from the higher-layer feature maps. Atrous convolution is used to optimize the network for high-resolution feature maps. In addition, we use k-means to obtain appropriate anchor number and dimensions for pedestrian. Hence, our method pays more concentration on pedestrians, especially those of relatively small size. Experimental results on the two widely used pedestrian benchmarks demonstrate the advantages on detection robustness and efficiency of the proposed method. As for the future work, we will focus on the overall performance, and occlusion handling will be involved.

References

[1] Dollár P, Wojek C, Schiele B and Perona P. Pedestrian detection: a benchmark. In Computer Vision and Pattern Recognition (CVPR), 2009, pp. 304–311.

[2] Benenson R, Omran M, Hosang J, et al. Ten years of pedestrian detection, what have we learned? In European Conference on Computer Vision (ECCV), 2014, pp. 613-627.

[3] Dollár P, Tu Z W, et al. Integral channel features. In British Machine Vision Conference, 2009.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition (CVPR), 2005, pp. 886-893.

[5] Zhang S, Bauckhage C, Cremers A B. Informed haar-like features improve pedestrian detection. In Computer Vision and Pattern Recognition (CVPR), 2014, pp. 947–954.

[6] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. In Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1-8.

[7] Tian Y L, Ping L, Wang X G, et al. Pedestrian detection aided by deep learning semantic tasks. In Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5079-5087.

[8] Zhang S, Benenson R, Omran M, et al. How far are we from solving pedestrian detection? In Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2-10.

[9] Li C, Wang X, Liu W. Neural features for pedestrian detection. In Neurocomputing, vol. 238, no. 2017, pp. 420-432, 2017.

[10] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation. In IEEE Trans on Pattern Analysis & Machine Intelligence (TPAMI), vol. 38, no. 1, pp. 142, 2016.

[11] Krizhevsky A, Sutskever I. Imagenet classification with deep convolutional neural networks. In the Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097-1105.

[12] Zhang L, Lin L, Liang X, He K. Is faster r-cnn doing well for pedestrian detection? Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 443-457.

[13] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580-587.

[14] Girshick R. Fast R-CNN. In International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448.

[15] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. In IEEE Trans on Pattern Analysis and Machine Intelligence (TPAMI), vol. 39, no. 6, pp. 1137-1149, 2017.

[16] Yu F, Koltun V. Multi-scale context aggregation by dilated convolution. In International Conference on Learning Representations (ICLR), 2015, arXiv preprint arXiv: 1511.07122.

[17] Chen L C, Papandreou G, et al. Rethinking atrous convolution for semantic image segmentation. In Computer Vision and Pattern Recognition (CVPR), 2017, arXiv preprint arXiv: 1706.05587.

[18] Li Y H, Zhang X F, Chen D M. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In Computer Vision and Pattern Recognition, 2018, pp. 1091-1100.

[19] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[20] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In Computer Vision and Pattern Recognition (CVPR), 2017, pp. 936-944.

[21] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In IEEE Conference on Computer Vision & Pattern Recognition (CVPR), 2017, pp. 6517-6525.

[22] Dollár P, Wojek C, Schiele B, et al. Pedestrian detection: an evaluation of the state of the art. In IEEE Trans. Pattern Anal. Mach. Intell (TPAMI), vol. 34, no. 4, pp. 743, 2012.