Infrared Pedestrian Detection Based on Attention Mechanism

Yaling Zhu, Jungang Yang*, Xinpu Deng, Chao Xiao and Wei An

College of Electronic Science and Technology, National University of Defense Technology, Changsha, Hunan, 410073, China

*Corresponding author’s e-mail: yangjungang14159@163.com

Abstract. Pedestrian detection is one of the key technologies in computer vision, and plays an important role in surveillance and automatic driving. Compared with visible cameras, infrared cameras are more suitable for all-weather and all-day work. Recently, a number of methods have been proposed for infrared pedestrian detection, but cannot achieve a satisfactory performance in the case of small pedestrians. In this paper, we propose an improved RefineDet algorithm to solve the aforementioned problem. First, the aspect ratio in our method is modified to the range of an average person. Second, an attention mechanism is introduced to address the small spatial size of pedestrian. In addition, we develop a new dataset which includes small pedestrian for performance evaluation. Experiments demonstrated that our method can achieve a superior performance as compared to SSD and RefinDet methods.

1. Introduction

Pedestrian detection is one of the key technologies in computer vision, and is widely used in driving assistance and surveillance. In general, pedestrians are detected from optical images. However, it is infeasible to use optical images without sufficient illumination. As compared with optical images, infrared images are more suitable for all-weather and all-day pedestrian detection because it is not sensitive to the change of environment.

In recent years, researchers have put a lot of work into pedestrian detection. Existing pedestrian detection algorithms consists of two branches. One is traditional machine learning method, for example HOG [2], which is the first landmark achievement of pedestrian detection. At the same time, combining with the learning-based method such as support vector machine (SVM) [3], a better detection accuracy is achieved on HOG. In 2008, the deformable parts model (DPM) [4] algorithm proposed by Felzenszwalb is one of the best algorithms in the traditional learning mode. The core of the algorithm is the improved hog feature combined with the distance relationship of the component model. The other branch of existing pedestrian detection algorithms is the latest emerging deep learning pedestrian detection method such as [5-7]. The deep learning methods can be mainly divided into two categories: two-stage object detection and one-stage object detection. The main representative models of two-stage object detection are as follows: Fast RCNN [8], Faster RCNN [9], Mask RCNN [10], feature pyramid network (FPN) [11]. For one-stage detection methods, such as YOLOv1 [12], YOLOv2 [13], YOLOv3 [14], SSD [15] and DSSD [16], these algorithms are faster than the two-stage algorithms but the accuracy is lower.

The above algorithms can achieve a good performance on pedestrian detection on optical images. Compared with visible cameras, infrared cameras are more suitable for all-weather and all-day work. Infrared images are not sensitive to the changes of environment due to its dependence on infrared rays scattered by objects. In recent years, with the continuous development of electronic and computer
technology, low-cost infrared camera equipment appears. Low-cost infrared camera greatly promotes the application of infrared image processing technology in various fields. Thus, research has been performed on human detection at night using infrared camera. For example, Lee et al [17] proposed a multiple camera-based method. There is also significant progress in single camera-based methods [18]-[20] with the appearance of several infrared pedestrian detectors.

However, for thermal images, there are some difficulties that might reduce the detection accuracy such as the low resolution, contour blur and large noise. Pedestrian detection in infrared images is a challenging task due to these difficulties. Especially when a person occupies a small region in an infrared image, it is more challenging to detect. There were few works and dataset in literature, which made the research in community more difficult.

To solve these problems, we propose a novel infrared pedestrian detection network. There are several contributions in this paper. First, the aspect ratio of the RefineDet [1] algorithm is modified to a range of an average person aspect ratio. Second, the attention mechanism and RefineDet [1] algorithm are combined. Finally, we develop an infrared pedestrian dataset which can fill the blank in the area of infrared pedestrian detection.

2. Related Work
2.1. Pedestrian detection
Traditionally, there are many hand-crafted features and machine learning algorithms used to detect pedestrian such as ICF [21], ACF [22] and LDCF [23]. Since multi-scale sliding windows are used to traverse the whole image in the test phase, so the amount of computation increases dramatically and the speed is slow. Recently, deep learning algorithms [24-26] have been more popular because of their stronger feature extraction ability and higher generalization ability. The R-CNN framework [27] designed by Girshick opened the door for deep learning from target classification to target detection, which led to the upsurge of target detection algorithm with deep learning. Since then, Fast R-CNN [9], Faster R-CNN [9], YOLOv1 [12], YOLOv2 [13], YOLOv3 [14], SSD [15] and RefineDet [1] have been proposed one after another.

2.2. Attention mechanism
The attention mechanism [28-31] of deep learning is similar to the selective visual attention mechanism of human beings in essence. The core goal of it is to select the information which more critical to the current task goal from a large number of information. There are two kinds of attention mechanism, soft attention and hard attention. Soft attention pays more attention to channels or regions, and soft attention can be generated directly through the network, so it is a kind of deterministic attention. Moreover, the gradient of soft attention can be calculated by neural network. The weight of attention is obtained by propagating forward and backward feedback learning, which shows that the soft attention mechanism is differentiable. Hard attention is more inclined to the change of focus, that is to say, the mechanism of hard attention is a random prediction process, and every point in the graph may extend the attention. The training process of hard attention mechanism is usually completed by reinforcement learning, because hard attention is an indispensable attention.

From the perspective of attention domain, attention mechanism can be divided into spatial domain, channel domain and mixed domain. The spatial domain transforms the spatial information in the original image into another space and retains the key information.

3. Approach
In this section, we introduce our approach for pedestrian detection in thermal images. There are three parts. First, we use the RefineDet [1] as a frame of our algorithm. Second, we modify the aspect ratio to an average person aspect ratio range. Third, we introduce attention mechanism to our algorithm.
3.1. Architecture

The overall network architecture is shown in Fig.1. Our algorithm is an improved algorithm based on RefineDet [1]. RefineDet [1] achieves better accuracy than two-stage methods and maintains comparable efficiency of one-stage methods. It consists of two modules named ARM (anchor refinement module) and ODM (object detection module). In addition, TCB (Transfer Connection Block) structure is designed, TCB maps the ARM features to ODM. On the other hand, TCB integrates the high-level semantic information with the low-level information.

![Fig.1 Architecture of our network](image)

Attention mechanism can select the information that is more critical to the current task goal from a large number of information, so it can be benefit for small oedestrian detection. In order to improve the detection accuracy of RefineDet [1] for small pedestrian, we introduce an attention block to RefineDet [1]. The attention block uses a spatial transformer. The architecture is as shown in (b). The input of this block is \( U \in \mathbb{R}^{H \times W \times C} \). H, W and C represent the length, width and channel number of the input feature respectively. Then the input enters two routes, one is localization net, and the other is sampling layer. The localization net generates a sampling signal, which is actually a transformation matrix. After multiplying with the original image, the transformed matrix \( V \in \mathbb{R}^{H \times W \times C} \) can be obtained.

Then a convolution operation is used to get the output \( W \in \mathbb{R}^{H \times W \times C} \). The architecture is as in Fig.2. The advantage of this block is that it can identify the key information of the signal in the upper layer. Such a block can be inserted into any layer, because the block can process the channel information and matrix information at the same time.

![Fig.2 Attention block](image)
3.2. Aspect ratio
The aspect ratio of RefinDet [1] are 1/3, 1/2, 1, 2, 3, but they do not match the size of the pedestrian. It is found that the size of pedestrians is generally about 0.41. Therefore, we modify the aspect ratio to about 0.41. We took some experiments, and the result showed this modification improved the detection accuracy by 1%.

3.3. Our dataset
The existing infrared pedestrian detection datasets are lack of small size pedestrians, so we develop a dataset with the pedestrian far from the camera. Below is a comparison of our dataset with other datasets.

We used an infrared camera to develop this dataset. The detector of this camera is 640*512, the pixel spacing is 17μm, and the type of lens are 35mm F1.0. Our dataset includes 1237 images, containing 2630 instances of pedestrians. These images are annotated by the VGG Image Annotator tool to generate the bounding boxes on pedestrians. Additionally, we select a set of 1049 images from our dataset to train the networks, and select the other 188 images to validate our detection networks. Finally, the size of images in our dataset is 720*480.

Below are images from our dataset and OSU Color and Thermal Database. We photographed pedestrians at a distance, so our dataset included more small size pedestrian.

![Fig.3 Our dataset (720*480) and OSU Color and Thermal Database (320*240)](image)

4. Experiments
We implement our algorithm in Pytorch. Experiments are conducted on our dataset. We set the learning rate to $10^{-4}$ for the first 2k iterations, and decay it to $10^{-5}$ for 2k-4k iterations, then decay it to $10^{-6}$ for training another 2k iterations, respectively. The batch size is set to 6, and the max iter is 6000. We compare our algorithm with SSD [15] and RefinDet [1] in table 1.

| Method   | mAP/% |
|----------|-------|
| SSD      | 57.4  |
| RefineDet| 59.4  |
| Our Model| 61.1  |

As can be seen from table 1, the detection accuracy of our algorithm in this paper is 1.7% higher than that of the RefinDet [1] network, and 3.7% higher than that of the SSD [15]. The performance of three method is as shown in Fig.3. Compared to SSD [15] and RefinDet [1], our model gets higher accuracy. As we can see from Fig.3, there is missing detection of small pedestrian in the SSD [15] and RefinDet [1] methods. In addition, our model gets higher confidence score to the same object.
5. Conclusion

The existing infrared pedestrian detection method is not effective for pedestrian detection with small target. This paper proposes an algorithm that introduces attention mechanism, which can make pedestrian achieve better results when a person occupies a small region. In addition, we modify the aspect ratio to the range of an average person in our algorithm, and we develop a dataset with the pedestrian far from the camera to validate our algorithm. It is found that the algorithm in this paper has a certain recognition ability for small targets, improves the problem of missing targets.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China under Grant 61401474

References

[1] S. Zhang, L. Wen, X. Bian, Z. Lei and S. Z. Li. (2018) Single-Shot Refinement Neural Network for Object Detection[C]. In: Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City. pp. 4203-4212.
[2] N. Dalal, B. Triggs. (2005) Histograms of Oriented Gradients for Human Detection[C]. In: Conference on Computer Vision and Pattern Recognition (CVPR), San Diego. pp. 886-893.
[3] Platt, John. (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines[J]. Advances in Kernel Methods-Support Vector Learning. 208.
[4] P. F. Felzenszwalb, R. B. Girshick, et al. (2010) Object Detection with Discriminatively Trained Part-Based Models[J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627-1645.
[5] Song H, Choi I K, et al. (2018) Vulnerable pedestrian detection and tracking using deep learning[C]. In: International Conference on Electronics,1-2.
[6] Tian Y, Luo P, Wang X, et al. (2015) Deep Learning Strong Parts for Pedestrian Detection[C]. In: International Conference on Computer Vision (ICCV). Santiago. pp. 1904-1912.
[7] Li J, Liang X, Shen S M, et al. (2018) Scale-aware Fast R-CNN for Pedestrian Detection[J]. IEEE Transactions on Multimedia, 20(4):985-996.
[8] Ross Girshick. (2015) Fast R-CNN[C], In: International Conference on Computer Vision (ICCV), Santiago. pp. 1440-1448
[9] Ren S, He K, et al. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region
Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6):1137-1149.
[10] Kaiming H , Georgia G, et al. (2018) Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-1.
[11] Lin T Y , Dollar P, et al. (2017) Feature Pyramid Networks for Object Detection[C]. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu. pp. 936-944.
[12] Redmon J , Divvala S, Girshick R, et al. (2016) You Only Look Once: Unified, Real-Time Object Detection[C]. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[13] Redmon, Joseph, et al. (2017) YOLO9000: Better, Faster, Stronger[C]. In: Computer Vision and Pattern Recognition (CVPR). Honolulu. pp. 6517-6525.
[14] Redmon, Joseph , et al. (2018) YOLOv3: An Incremental Improvement[C]. In: Computer Vision and Pattern Recognition (CVPR). Salt Lake City.
[15] Liu W , Anguelov D , Erhan D , et al. (2016) SSD: Single Shot MultiBox Detector[C]. In: European Conference on Computer Vision. 21-37.
[16] Fu C Y , Liu W , Ranga A , et al. (2017) DSSD: Deconvolutional Single Shot Detector[C]. In: Computer Vision and Pattern Recognition (CVPR). Hawaii.
[17] Ji, Hoon, Lee, (2015) Robust pedestrian detection by combining visible and thermal infrared cameras.[J]. Sensors. 15(S):10580-10615.
[18] Qi B , John V , Liu Z , et al. (2016) Pedestrian detection from thermal images: A sparse representation based approach[J]. Infrared Physics & Technology. 76:157-167.
[19] Ghose D , Desai S M , et al. (2019) Pedestrian Detection in Thermal Images using Saliency Maps[C]. In: Computer Vision and Pattern Recognition (CVPR). Long Beach. 0-0.
[20] Biswas S K , Milanfar P . (2017) Linear Support Tensor Machine With LSK Channels: Pedestrian Detection in Thermal Infrared Images[J]. IEEE Transactions on Image Processing. 26(9):4229-4242.
[21] Dollár, Piotr, et al. (2009) Integral Channel Features[C]. In: British Machine Vision Conference (BMVC). 1-11.
[22] Dollar, Piotr, et al. (2014) Fast Feature Pyramids for Object Detection[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions, 36:1532-1545.
[23] W. Nam, P. Dollar, and J. H. Han. (2014) Local decorrelation for improved pedestrian detection[C]. In: NIPS. 1:424-432.
[24] Zhang L , Lin L , et al. (2016) Is Faster R-CNN Doing Well for Pedestrian Detection?[C]. In: European Conference on Computer Vision. Amsterdam. 443-457.
[25] Lin C , Lu J , Wang G , et al. (2020) Graininess-Aware Deep Feature Learning for Pedestrian Detection[J]. IEEE Transactions on Image Processing. 29:3820-3834.
[26] Hou Y L , Song Y, et al. (2017) Multispectral pedestrian detection based on deep convolutional neural networks[C]. In: IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). Xiamen. pp. 1-4
[27] Girshick R , Donahue J , et al. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation [C]. In: Conference on Computer Vision and Pattern Recognition (CVPR). Columbus. 580-587.
[28] Wang X , Girshick R , et al. (2018) Non-local Neural Networks[C]. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City. 7794-7803.
[29] Hu J , Shen L , Sun G . (2019) Squeeze-and-Excitation Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1-1
[30] Wang L , Wang Y , et al. (2019) Learning Parallax Attention for Stereo Image Super-Resolution[C]. In: Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach. 12250-12259.
[31] Huang Z , Wang X , et al. (2018) CCNet: Criss-Cross Attention for Semantic Segmentation[C]. In: International Conference on Computer Vision. Seoul. 603-612.