Long-distance infrared video pedestrian detection using deep learning and background subtraction

Yaling Zhu¹, Jungang Yang*, Xiaokai Xie¹, Zhihui Wang² and Xinpu Deng¹

¹College of Electronic Science and Technology, National University of Defense Technology, Changsha, Hunan, 410073, China
²Air Force Early Warning Academy, Wuhan, 430019, China
zhuyaling_@mail.nudt.edu.cn
*Corresponding author’s e-mail: yangjungang14159@163.com

Abstract. Infrared video-based pedestrian detection plays an important role in surveillance and automatic driving. Compared with vision cameras, infrared cameras have the characteristics of all-day and all-weather availability. The current infrared video detection methods generally follow the traditional paradigm. However, when objects are at a long distance, it is challenging for existing methods to achieve accurate detection. Meanwhile, deep learning-based methods have been widely used in visible light object detection. In this paper, we introduce an algorithm which combines deep learning and background subtraction methods to achieve infrared pedestrian detection. Firstly, background subtraction method is performed to provide the inter-frame information for the deep learning module. Secondly, the RefineDet equipped with an attention module is used to improve the detection accuracy for small pedestrian. Moreover, we develop a novel infrared video dataset which includes long-distance pedestrian for performance evaluation. Experiments demonstrated that our method can achieve a superior performance as compared to RefinDet and SSD methods.

1. Introduction
Pedestrian detection is one of the key research technologies in computer vision. It has been widely applied in many fields such as surveillance, automatic driving and rescue. There are different sensors used in pedestrian detection, such as LIDAR sensors, thermal and vision cameras. Although visible images are more commonly used in object detection, it is hard to detect pedestrian in visible images at night or bad weather conditions. Compared with vision cameras[1], infrared cameras have the characteristics of all-day and all-weather availability and can record pedestrians at a distance. Some researchers refer to using thermal cameras to detect pedestrian. However, there are few researches about long-distance infrared video pedestrian detection. Therefore, in this paper, we discuss the problem of long-distance pedestrian detection in infrared videos.

Recently, with wider application of convolution neural network, deep learning-based target detection methods are rapidly developed. However, the state-of-art deep learning methods in object detection only utilize the context information, while ignoring the inter-frame information (i.e., temporal information) in videos. There are also some video object detection using deep learning methods such as [2-5]. They utilize both context information and inter-frame information in videos to achieve improved accuracy, but suffer from a heavy computational burden.

Long-distance infrared video pedestrian detection is a challenging task. There are some challenges
that lead to the low accuracy of current thermal pedestrian detection methods. Firstly, a person only occupies a small region in an infrared image when the camera is far away, resulting in difficulties for pedestrian detection. Secondly, the low resolution, occlusions and contour blur in thermal images also introduce difficulties to infrared pedestrian detection. Thirdly, the lack of the infrared video dataset makes the research in community more difficult.

To solve these problems, we propose a novel infrared video pedestrian detection network. Three contributions are made in this paper. Firstly, we use a background subtraction module to get the preprocessing images, which can provide the inter-frame information and for the deep learning module with a low computational complexity. Secondly, the deep learning module based on RefineDet [6] network uses an attention mechanism to improve the detection accuracy for small pedestrian. In addition, we develop a new dataset which includes long-distance pedestrian for performance evaluation. Experiments demonstrated that our method can achieve a superior performance as compared to RefinDet [6] and background difference methods.

2. Related work

2.1 video object detection
There are two branch of video detection methods. The first branch is traditional method [7-9]. There are three major traditional video detection methods: background difference-based methods [7], frame difference-based methods [8], and optical flow-based methods [9]. Background difference [7] is one of the most commonly used methods in video detection. It uses the difference between the current image and the background image to detect the moving area. Generally, it can provide the most complete feature data, but is especially sensitive to the changes of dynamic scene, such as illumination and interference of extraneous events. The frame difference [8] method obtains the contour of moving object through the difference operation of two adjacent frames in the video image sequence. It can be well applied to the case of multiple moving objects and camera moving. When abnormal object motion occurs in the monitoring scene, there will be obvious difference between frames. Two frames are subtracted to get the absolute value of the brightness difference between the two frames, and judge whether it is greater than the threshold to analyze the motion characteristics of the video or image sequence and determine whether there is object movement in the image sequence. Optical flow method [9] uses the change of pixels in the image sequence in time domain and the correlation between adjacent frames to find the corresponding relationship between the previous frame and the current frame, and calculates the motion information of objects between adjacent frames.

The second branch of video detection methods is deep learning. Currently, it has two main research directions. One is combination of detection and tracking, which is to track the detected target in the video, such as TCNN [10]. The other is using motion information such as Association LSTM [11]. These methods use the change of pixels in the time domain and the correlation between adjacent frames to find the corresponding relationship between the previous frame and the current frame, so as to calculate the motion information of objects between adjacent frames.

2.2 Attention mechanism
Attention mechanism has been widely adopted in many areas such as classification [12-13], segmentation[14-15], and image super-resolution [16-17]. An informal view of attention mechanism is that attention mechanism can make neural network have the ability to focus on its input (or feature) subset: select specific input. Attention can be applied to any type of input, regardless of its shape. In the case of limited computing power, attention mechanism is a resource allocation scheme to solve the problem of information overload, which allocates computing resources to more important tasks.

Attention mechanism is generally divided into two types. One is conscious attention from the top down, called focus attention mechanism. Focus attention mechanism refers to the attention focused on a certain object actively and consciously with a predetermined purpose and task. The other type is unconscious attention from the bottom up, which is called saliency-based attention. Saliency-based
attention is driven by external stimuli. It does not need active intervention, and has nothing to do with the task. If the stimulus information of an object is different from its surrounding information, an unconscious "winner-take-all" or gating mechanism can turn attention to the object. Whether the attention is intentional or unintentional, most human brain activities rely on attention, such as memorizing information, reading or thinking.

3. **Approach**

In this section, we introduce our infrared video pedestrian detection method. The above-mentioned traditional video detection methods cannot achieve a satisfactory performance. And the deep learning video detection methods cause large amount of transformation calculation. So, in this paper we propose an algorithm which combines deep learning and background subtraction methods. The architecture, aspect ratio and dataset of our method are introduced as follows.

3.1 **Architecture**

The architecture of our algorithm is shown in Fig.1. The network contains two modules: the background difference module and the deep learning module. The input is first preprocessed by the background difference module, then fed into the deep learning module.

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

The background difference module is combined with KNN background subtractor [18] and morphology open operation. Background subtractor is specially used for video analysis. For the new pixel value at a certain position of the image, compared with the historical information of the pixel value, if the difference between the pixel values is within the specified threshold value, the new pixel value is considered to be matched with the historical information. After all the historical information is compared, if the number of matching times exceeds the set threshold. Then the new pixel is classified as a potential background point. If the number of matched historical information that belongs to the background exceeds the set threshold, then the new pixel is classified as a background point. Finally, new pixels to historical information is saved according to certain rules. Morphology open operation takes corrosion and expansion in images. It has the function of eliminating small objects, separating thin objects and smoothing the boundaries of larger objects. The comparison of our method, mean-background-modeling and frame difference method is shown in Fig.2. The object is marked by red boxes. Our KNN background subtractor [18] achieves the best performance.
The deep learning module is an improved RefineDet [6] network which combined with attention mechanism. RefineDet [6] can achieve better accuracy than one-stage methods and maintains comparable efficiency of two-stage methods. This advantage can improve the detection performance of small targets in our long-distance task. It consists of three modules named ARM (Anchor Refinement Module), ODM (Object Detection Module) and TCB (Transfer Connection Block). The Arm module simulates the first step in two-step method, such as RPN in FasterR-CNN [19]. ODM module simulates the second step in the two-step method. ODM module does not use the time-consuming operation like ROI pooling, but directly connects through TCB to convert ARM features. In addition, attention mechanism can select the critical information from much information. In our task, the pedestrian in long distance occupies a small region, so they are easily to be ignored. The attention mechanism contributes to the detection of these pedestrians. The architecture of our attention block is shown in Fig.3.

### 3.2 Dataset
Large-scale datasets have been demonstrated to be contributive to deep learning-based algorithms in
various areas [20-22]. However, there are few datasets about thermal video pedestrian detection. In order to complete our task, we propose a dataset about long-distance infrared video pedestrian. The detector of the camera is 640*512, the pixel spacing is 17μm, and the type of lens are 35mm F1.0. The training set is a two-minutes video and 3208 images are annotated by the VGG Image Annotator tool. The test set has 320 annotations which is from another video. The size of all images in our dataset is 640*480.

Fig.4 is the comparison of images from our dataset and OSU Color and Thermal Database. We photographed at a distance, so our dataset included more small-size pedestrian. In addition, our dataset contains occluded pedestrian images.

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 maxiter is 20000. We compare our algorithm with RefinDet [6] and SSD [23] in table 1.

| Method     | mAP/% |
|------------|-------|
| SSD        | 37.4  |
| RefineDet  | 58.7  |
| Our Model  | 85.4  |

The results of SSD [23], RefinDet [6] and our algorithm is shown in Fig.5. The red box is
ground-truth bounding box, and the green box is predicted bounding box. Our method achieves the best performance. It can not only detect partially occluded pedestrians, but also has lower false alarm rate.

5. Conclusion
In this paper, a new thermal video pedestrian detection approach is introduced, and a new dataset is proposed to verify the effectiveness of the algorithm. The network contains two modules: the background difference module and deep learning module. The background subtraction method can provide the inter-frame information for the deep learning module. And the deep learning module combined the RfndDet [6] and attention mechanism, which can achieve a good performance in tiny person detection. It is found that our method has a certain recognition ability for small targets in thermal videos, and 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] Y. Wang, J. Yang, Y. Guo, C. Xiao, W. An, "Selective light field refocusing for camera arrays using bokeh rendering and superresolution", IEEE Signal Processing Letters, vol. 26, pp. 204-208, 2019.
[2] Nam H, Baek M, Han B, et al. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking[J]. arXiv: Computer Vision and Pattern Recognition, 2016.
[3] Yuan Y, Liang X, Wang X, et al. Temporal Dynamic Graph LSTM for Action-driven Video Object Detection[J]. In: International Conference on Computer Vision (ICCV). 2017.
[4] Zhu X, Xiong Y, Dai J, et al. Deep Feature Flow for Video Recognition[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[5] Xiao F, Lee Y J. Video Object Detection with an Aligned Spatial-Temporal Memory: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII[M]// Computer Vision – ECCV 2018.
[6] 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.
[7] Wren C R, Azarbayejani A, Darrell T, et al. Pfinder: real-time tracking of the human body[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(7): 780-785.
[8] Stander J, Mech R, Ostermann J, et al. Detection of moving cast shadows for object segmentation[J]. IEEE Transactions on Multimedia, 1999, 1(1): 65-76.
[9] Horn B K, Schunck B G. Determining Optical Flow[J]. Artificial Intelligence, 1980, 17(1): 185-203.
[10] K. Kang et al., "T-CNN: Tubelets With Convolutional Neural Networks for Object Detection From Videos," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896-2907, Oct. 2018, doi: 10.1109/TCSVT.2017.2736553.
[11] Lu Y, Lu C, Tang C, et al. Online Video Object Detection Using Association LSTM[C]. international conference on computer vision, 2017: 2363-2371.
[12] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
[13] Li X, Wang W, Hu X, et al. Selective kernel networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 510-519.
[14] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.
[15] Huang Z, Wang X, Huang L, et al. Cenet: Criss-cross attention for semantic
segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 603-612.

[16] Wang L, Wang Y, Liang Z, et al. Learning parallax attention for stereo image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 12250-12259.

[17] Ying X, Wang Y, Wang L, et al. A Stereo Attention Module for Stereo Image Super-Resolution[J]. IEEE Signal Processing Letters, 2020, 27: 496-500.

[18] Abeywickrama T , Cheema M A , Taniar D . k-Nearest Neighbors on Road Networks: A Journey in Experimentation and In-Memory Implementation[J]. Proceedings of the VLDB Endowment, 2016.

[19] 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.

[20] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3974-3983.

[21] Muller M, Bibi A, Giancola S, et al. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 300-317.

[22] Wang Y, Wang L, Yang J, et al. Flickr1024: A large-scale dataset for stereo image super-resolution[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019: 0-0.

[23] Liu W , Anguelov D , Erhan D , et al. (2016) SSD: Single Shot MultiBox Detector[C]. In: European Conference on Computer Vision. 21-37.