Research of Pedestrian Detection Methods with Anchor Frame Based on Deep Learning

Tao Yan*

School of Computer Science and Engineering, Jishou University, Jishou, Hunan, 416000, China

Abstract: Pedestrian detection technology has been one of the hotspots for target detection tasks. Deep learning-based theories and techniques perform exceptionally well in the pedestrian detection domain, and several general-purpose target detectors are continuously utilized in the target detection domain. Deep learning-based pedestrian detection techniques are examined in this research, and anchor frame-based pedestrian detection techniques are separated, contrasted, and examined in light of the anchor frames. The occlusion problem and the scale change problem are one of the main causes of omission and false detection problems in practical applications of pedestrian detection. This paper first presents the conventional pedestrian detection techniques, then concentrates on the R-CNN detector in the anchor frame-based two-stage pedestrian detection algorithm. When addressing the scale change problem, the enhanced Faster R-CNN algorithm, which is based on R-CNN, greatly minimizes redundant computation and enhances recognition accuracy. The YOLOv3 model in the YOLO model exhibits notable modifications in the overall architecture of pedestrian identification in the single-stage pedestrian recognition method using anchor frames, greatly improving its capacity to handle scale changes and occlusion issues.

1 Introduction

The pedestrian detection task is a sub-task of the target detection task. It refers to the use of computer vision technology to accurately identify and locate pedestrian targets from images. This technology plays a vital role in real application scenarios such as intelligent transport systems, intelligent security, and so on.

Traditional pedestrian detection methods mainly rely on manually designed features for target characterization. Its theoretical foundation is relatively mature and simple to implement, but it also has more obvious limitations such as low recognition accuracy, poor robustness, and poor adaptability. Therefore, it is difficult for traditional pedestrian detection methods to meet the practical application needs of high accuracy and high efficiency required for today's pedestrian detection tasks.

In recent years, pedestrian identification methods utilizing deep learning have increasingly gained traction, owing to the rapid evolution of deep learning technologies. Convolutional neural network (CNN)-based deep learning algorithms have the ability to

* Corresponding author: taoyan@ldy.edu.rs

learn more expressive and generalized features by constructing intricate network architectures and exploring the intricate relationships within the data in depth. These deep learning-based pedestrian detection techniques achieve performance improvements in multiple aspects such as feature extraction, model optimisation, and classification decision-making, which in turn improves the accuracy and efficiency of pedestrian detection. In the absence of occlusion and with a simple scene, researchers have successfully integrated general-purpose target detection techniques such as a region-based CNN (Faster R-CNN), a two-stage detector based on anchor frames, and a single-shot multi-box detector (SSD), a single-stage detector based on anchor frames (SSD) and other general-purpose target detection models have been applied to pedestrian detection tasks with remarkable results.

Nevertheless, pedestrian detection still confronts several difficulties in real-world application scenarios, which are mostly represented in the two primary areas listed below:1) Occlusion problem. In intelligent monitoring, unmanned driving, and other scenarios, it is common for pedestrian targets to be occluded by various objects in real-world scenarios. Occlusion between pedestrian targets and scene objects, e.g., occlusion formed between pedestrians and trees, railings, moving vehicles, etc. Occlusion between people. The above two occlusion scenarios destroy the overall structural features of pedestrians, making it difficult for a general-purpose data detector to learn uniform features) Small Scale Problems. In the pedestrian detection task, pedestrian targets far away from the camera are small in size, and small-scale targets contain less valid information, making it difficult to achieve high-precision detection, and the fuzzy contour information they exhibit makes it particularly difficult to design unified feature processing strategies for targets of different scales.

Focusing on the above issues and challenges, the benefits and drawbacks of two-stage and single-stage anchored frame-based pedestrian detection algorithms in various contexts are thoroughly covered in this study. The purpose of this study is to guide for choosing pedestrian detection methods.

2 Traditional pedestrian detection methods

Before the extensive utilization of deep learning techniques in the domain of pedestrian detection, traditional pedestrian detection algorithms such as those relying on hand-designed features for target characterization and classifiers to achieve the detection of pedestrian targets showed both certain advantages and some limitations.

The core of the detector proposed by VIOLA and JONES, specifically for face detection tasks, consists of Haar-like features, AdaBoost classifiers, and a cascade structure [1]. To address the discrepancy in their respective areas, the Haar feature often employs a multiplication operation to capture the difference between the aggregate pixel intensity in the black region and that in the white region. It is one of the commonly used features in these tasks due to its advantages of computational simplicity and robustness to illumination variations. And AdaBoost algorithm is an iterative algorithm that combines multiple weak classifiers into one strong classifier to improve classification efficiency. In the pedestrian detection task, the Haar-like features that best represent the pedestrian features are selected through continuous iteration to form a strong classifier, which can greatly improve the detection and classification efficiency. The Cascade structure will initially screen the images of each strong classifier, and only the images that satisfy certain conditions will be passed to the next classifiers for further processing. In this way, the cascade structure can gradually eliminate irrelevant image regions, thus speeding up detection and improving accuracy.

DALAL et al. introduced the algorithm known as Histogram of Oriented Gradients (HOG), which provides a novel approach. Each pixel point's gradient direction and intensity are determined by HOG, which then counts these gradient directions into the histogram to provide the image's HOG feature description, but in the case of occlusion of an object or an object with too large scale performance is poor [2]. By using a multi-component approach, FELZENSZWALB et al. suggested a Deformable Part Model (DPM) based on HOG [3]. Traditional pedestrian detection algorithms mostly use the sliding window strategy for screening one by one, which leads to inefficiency. The DPM algorithm has a high detection accuracy in the pedestrian detection task by decomposing the target into components such as head, torso, limbs, etc., and analysing the positional and deformation relationships of these components. In addition, manually designed feature extraction operators have limitations in terms of accuracy and robustness, especially in complex scenarios, where the detection accuracy decreases significantly. A new revolution in target detection has been brought about by the quick growth of deep learning. Through in-depth learning of image features, deep learning networks may learn feature extraction autonomously, replacing manually constructed feature extractors and greatly enhancing target identification performance as compared to older methods.

3 Two-stage pedestrian checking algorithm based on anchor frames

Pedestrian detection method based on anchor frame is a hot spot in many target detection methods. It can be primarily separated into two-stage and single-stage target detection algorithms. To determine whether pedestrians are present and where they located in the image, the two-stage pedestrian detection algorithm splits the pedestrian detection task into two steps: first, it creates candidate regions that might contain pedestrians; second, it fine-tunes the classification and position of these candidate regions. Algorithms such as R-CNN, Faster R-CNN, Cascade R-CNN, and others belong to the category of two-stage target detection methods that rely on anchor frames. To enhance Faster R-CNN, REN et al. added RPN and anchor frames [4]. This allows RPN to be trained from beginning to end to produce high-quality region suggestions, which Faster R-CNN then detects. The introduction of an anchoring frame improves the running speed of the model and performs well when trained and tested using single-scale images. VASCONCELOS et al. introduced Cascade R-CNN, a novel multi-stage architecture for high-quality target detection, which enhances the detector's performance by progressively increasing the Intersection over Union (IoU) thresholds, aiming to achieve an efficient solution [5].

3.1 R-CNN

The performance of target identification systems is greatly enhanced by GIRSHICK's suggested R-CNN, which uses pre-trained CNNS to generate rich feature representations from candidate regions [6]. This algorithm generates candidate regions using methods such as Selective Search. Then CNN takes these regions for feature extraction. Ultimately, classifiers like Support Vector Machine (SVM) are used to classify the features, and R-CNN further trains a bounding box regression model to increase the localization accuracy. The model refines the location of potential regions to better align with the actual position of the object, thereby facilitating the determination of whether a pedestrian is present.

In place of R-CNN's selection search network, DONG et al. suggested a candidate region acquisition approach based on the Aggregated Channel Feature (ACF) model, and

the workflow of the algorithm is illustrated in Fig. 1. The algorithm generates feature maps with rich information by aggregating multiple image features and generates candidate regions based on these feature maps, which improves the detection accuracy while discarding some useless candidate frames, thus improving the detection speed as well.

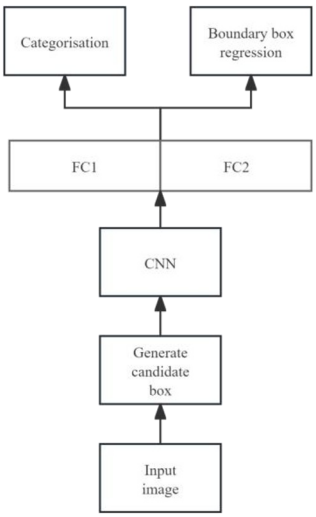

Fig. 1. Algorithm structure based on improved R-CNN [7]

3.2 Faster R-CNN

Given the prevalence of a large number of overlapping regions of candidate frames within an image, this results in a lot of computational redundancy when extracting image features. The Fast R-CNN approach, which GIRSHICK creatively suggests as a solution to this issue, first applies a convolution operation to the entire image before projecting the generated candidate frames onto the feature map. In contrast to the R-CNN method, which must convolve roughly 2000 candidate frames one at a time, Fast R-CNN drastically cuts down on computational redundancy.

Nevertheless, the Fast R-CNN method still generates candidate area frames using a selection search approach, which requires a significant amount of time. For further optimization, GIRSHICK then proposed the Faster R-CNN algorithm. It aims to detect objects in images faster and more accurately. It cleverly combines two stages, candidate region generation, and object detection, in a single network framework, enabling end-to-end training and optimization. A key element of Faster R-CNN is the Region Proposal Network (RPN), produces several anchor frames at every site, and these regions have different sizes and scales to cover a variety of objects that may appear in the image. Faster R-CNN achieves highly efficient target detection by introducing the RPN and Fast R-CNN detector heads, as shown in Fig. 2. It can decrease unnecessary computation, swiftly and accurately detect different objects in complicated images, and offer robust support for computer vision research and applications.

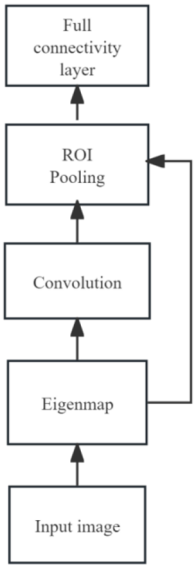

Fig. 2. Algorithm structure based on Faster R-CNN [8]

3.3 Cascade R-CNN

In the field of pedestrian detection, the Cascade R-CNN method increases object identification accuracy. This is accomplished gradually and phasedly by several cascade detectors while tackling challenging recognition tasks. Every detector in Cascade R-CNN is trained with a distinct IOU threshold. The model can learn more precise target features by gradually raising the IOU threshold, which is a gauge of the extent of overlap between the predicted and real frames. However, with the increasing IOU threshold, the detector also faces the risks of too few candidate frames and overfitting. Therefore, increasing the intersection and merger ratio thresholds will instead lead to a degradation of the detection performance. Cai and Vasconcelos introduced a cascade-based (setting multiple intersection and merger ratio thresholds) detection method.

BRAZIL G et al. proposed the AR-Ped framework based on Cascade R-CNN [9]. The AR-Ped framework consists of multiple cascaded detectors, each of which performs further processing on the candidate region output from the previous level of detector. Within the AR-Ped framework, a Region Proposal Network (RPN) employs a deep convolutional neural network alongside two concurrent fully connected layers to propose potential pedestrian regions and refine the positioning of bounding boxes.

4 Single-stage pedestrian checking algorithm based on anchor frames

The core of the single-stage pedestrian detection algorithm is to use anchor boxes as candidate regions to complete the classification and localization tasks of pedestrians in a single stage. It uses anchor boxes to forecast the classification results and the bounding box's return position without the need to create candidate boxes. As a result, it has a shorter detection time but a lower detection accuracy than the two-stage target detection method. The single-stage target detection algorithms based on anchor frames are SSD, YOLOv3, and so on. LIU's proposed SSD uses Non-maximum suppression (NMS) to choose the

bounding box above the threshold after predicting the bounding box's offset and the target object's classification probability value [10]. The results show that the SSD algorithm obtained 72.1% mAP (mean Average Precision) on Titan X at a rate of 58 frames when fed with a PASCAL VOC 2007 test image of 300×300 size. This was much faster than the best R-CNN at the time. The YOLOv3 algorithm proposed by Redmon et al. further improves the performance of the algorithm by improving the feature extraction network, introducing multi-scale prediction, optimizing the bounding box prediction, improving the loss function, and employing multi-label classification [11].

4.1 SSD-based pedestrian detection algorithm

The core of the SSD-based pedestrian detection algorithm lies in its prediction using multi-scale feature maps from CNNs. Unlike traditional target detection methods, SSD does not require an additional RPN to generate candidate regions, but instead directly sets a series of predefined anchor boxes on the feature map and performs classification and bounding box regression on these anchor boxes.

By fusing the anchor boxes mechanism of the Faster R-CNN algorithm with the quick detection technique of the YOLO algorithm, LIU developed the SSD algorithm [12]. Since the SSD algorithm generates redundant anchor boxes on multiple feature maps, it is necessary to use non-maximal value suppression to remove redundant bounding boxes with high overlap. LI et al. suggested a multi-scale fusion pedestrian detection method that makes use of improved sparse connectivity inside the SSD framework to raise the detection accuracy of the SSD algorithm [13]. This method improves pedestrian identification performance by combining data from multiple scales. SSD is improved in the underlying network part by using convolutional kernels with different dimensions to limit the number of input signals and reducing the network dimensionality by adding a single-channel convolutional layer. It optimizes the feature extraction approach of the CNN in the SSD algorithm by different dimensional feature generation and feature complementary enhancement from the bottom to the top. Experimental validation of the SSD algorithm on PASCAL VOC and CUHK Occlusion image datasets has been carried out by LI et al.

According to the experimental findings, this approach outperforms the original algorithm in terms of accuracy, and its detection speed satisfies real-time requirements. Specifically, the detecting speed achieves 31 frames per second (fps), which satisfies real-time requirements and has a specific use case. However, LI notes that in situations where there is an occlusion in the crowd, the method does not considerably enhance pedestrian recognition.

4.2 YOLO pedestrian detection model

Addressing the challenge of reduced accuracy in pedestrian detection when objects are partially obscured, the YOLOv1 model, known for its efficiency and compactness, has gained increasing popularity in the domain of pedestrian detection. YOLOv1's network architecture is quite simple and primarily consists of a CNN. Two fully connected layers and twenty-four convolutional layers make up the network's more compact topology, which lowers computation and memory usage. The YOLO model breaks the framework of the traditional target detection methods and puts forward a new idea of considering target detection as a regression problem. However, the YOLOv1 model predicts a fixed number of bounding boxes per grid, and this coarse-grained prediction leads to inaccurate target localization and limits its ability to detect small targets. As a result, REDMON et al. enhanced YOLOv1 and suggested YOLOv2, which incorporates a BN layer after each convolutional layer to accelerate convergence and lower the chance of overfitting [14]. It

also employs Anchor Boxes to forecast the bounding box offsets and inherits the Anchor mechanism from Faster R-CNN.

The YOLOv3 has undergone greater modifications in its overall structure. It incorporates the current advanced design concepts of the detection framework, which further improves the detection accuracy while maintaining its speed advantage, especially the excellent detection ability for multi-scale targets. Furthermore, by extracting three distinct feature layers, YOLOv3's multi-scale feature fusion prediction approach successfully raises the algorithm's detection accuracy for small-scale targets. Finally, to ensure the prediction accuracy of each target with multi-target label classification ability, The original Softmax function is replaced with a new sigmoid loss function in YOLOv3.

The YOLOv3 detection model achieves the goal of improving detection accuracy while maintaining real-time performance by combining the Darknet-53 backbone network with an efficient detection network. Its network structure is shown in Fig. 3.

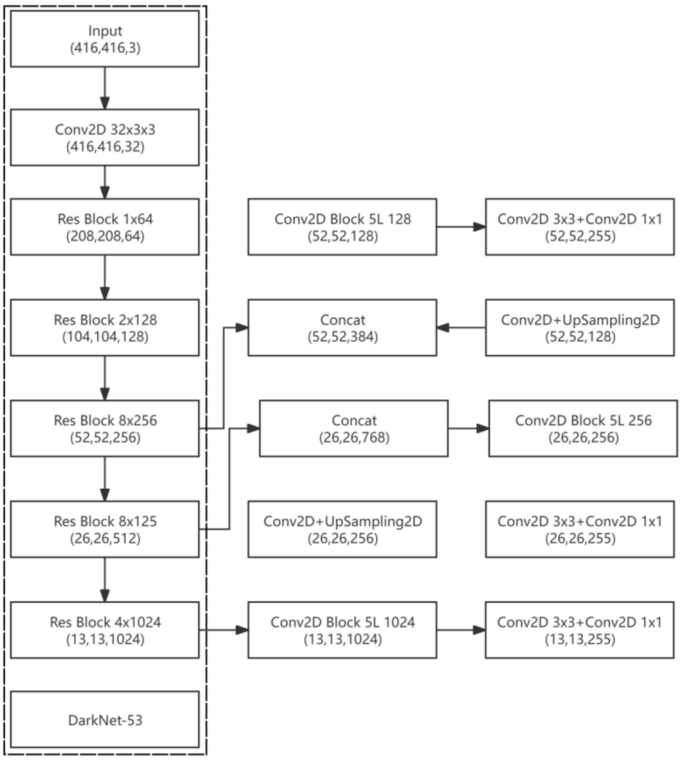

Fig. 3. Detection framework of YOLO v3 [15]

YOLOv3's core delivers a more potent basic feature extraction network while including the ResNet residual network concept, Darknet-53. Compared with the previous network structure, Darknet-53 accelerates the detection process to a certain extent by adopting a multi-scale fusion prediction method with a total of 3 feature layers extracted, which partially improves the accuracy, with the maximum accuracy compared to the 74.1% of Darknet-53 to 77.2%, while the number of floating point operations is almost doubled compared to Darknet-19.

The Darknet-53 model architecture of the YOlOv3 model consists of 53 layers of convolutional operations and 23 residual connections (jump connections). Compared to the YOLOv2 model, the deepening of the network hierarchy allows for a better application of including residual modules, multiscale detection, and feature fusion processes. YOLOv3

uses FPNs for feature fusion. The FPNs extract features from an image at different scales and resolution features from an image and combine them into a feature pyramid. Fig. 3 shows the Input input module, using 416×416×3 input, for different scales of feature maps, up-sampling and splicing operations are performed in the Concat module for feature fusion. The Detection Head is the output part of YOLOv3, which is responsible for converting the feature maps output from the feature fusion network into the detection results, and three convolutional layers make up the output section. These layers are in charge of shrinking the feature maps, extracting the features, and predicting the coordinates of the bounding boxes, confidence scores, and category probabilities, respectively.

5 Conclusion

Ensuring the accuracy of pedestrian detection in areas such as intelligent transport, driverless, and smart cities is a problem that researchers have been exploring to solve. Based on the introduction of conventional pedestrian detection techniques, this study first focuses on the creation of deep learning-based anchor frame pedestrian recognition algorithms on the occlusion problem and size change problem. In other words, the occlusion issue and computational redundancy are partially resolved by the switch from R-CNN to Faster R-CNN. The advancements in accuracy and speed of general-purpose target recognition algorithms, such as YOLOv3, are the main focus of the YOLO series. The subject of pedestrian identification has significantly advanced thanks to deep learning-based approaches, but the occlusion and scale variation problems need to be solved urgently and are one of the current hot issues in the field. Although pedestrian detection techniques are currently widely employed in intelligent security, automated driving, and other industries, there is still a gap between the detection goals of high efficiency and high accuracy. Therefore, enhancing detection accuracy and efficiency also remains a viable area for future research.

References

1. P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), IEEE, (2001)
2. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, 1:886-893 (2005)
3. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, et al., Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1627-1645 (2009)
4. S. Ren, K. He, R. Girshick, et al., Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137-1149 (2017)
5. Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6154-6162 (2018)
6. R. Girshick, J. Donahue, T. Darrell, et al., Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580-587 (2014)

7. P. Dong, W. Wang, Better region proposals for pedestrian detection with R-CNN. 2016 Visual Communications and Image Processing (VCIP), IEEE, 1-4 (2016)
8. R. Girshick, Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 1440-1448 (2015)
9. G. Brazil, X. Liu, Pedestrian detection with autoregressive network phases. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7231-7240 (2019)
10. W. Liu, D. Anguelov, D. Erhan, et al., SSD: Single shot multibox detector. Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14, Springer International Publishing, 21-37 (2016)
11. J. Redmon, A. Farhadi, YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767, (2018)
12. W. Liu, D. Anguelov, D. Erhan, et al., SSD: Single shot multibox detector. Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14, Springer International Publishing, 21-37 (2016)
13. X. Li, X. Luo, H. Hao, et al., Pedestrian detection method based on multi-scale fusion inception-SSD model. 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), IEEE, 9:1549-1553 (2020)
14. J. Redmon, S. Divvala, R. Girshick, et al., You Only Look Once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779-788 (2016)
15. J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement. arXiv e-prints, (2018)