Research on object detection technology for human detection

Yanling Niu* and Zhaohui Meng
School of computer and information technology, Hohai university, Nanjing, Jiangsu, 211100, China

*2649922019@qq.com

Abstract. There are many classical application scenarios in computer vision, and human detection in dense scenarios is a classical problem. Its goal is to find all the people in an image or each frame of a video, as well as their positions and sizes, and frame them in a rectangular box. In recent years, deep learning has been applied to the field of computer vision, laying a solid foundation for human detection based on deep learning. Aiming at the crowdhuman dataset, based on faster-RCNN and feature pyramid network (FPN), this paper studies the training effect. In order to detect small objects, a multi-scale feature fusion network (MFFN) was proposed for feature extraction. According to the serious occlusion of dataset, a double-branch structure was proposed to improve the detection accuracy. These two structures combine to form a new network, we call it MDN, namely multi-feature fusion and double-branch network. Experiments show that these methods are helpful to improve the detection accuracy.

1. Introduction

Given a picture, identifying the object of interest in the picture, and determining its position and category is the object detection. The traditional object detection method adopts the strategy based on sliding window, and the image features are artificially designed, resulting in poor detection effect, low accuracy and high computational complexity. With the continuous development of deep learning in recent years, the object detection algorithm based on deep learning has gradually replaced the traditional object detection algorithm. Common target detection tasks are mainly divided into two directions: two-stage algorithm such as Region-based Convolutional Network (R-CNN)[1] series, which uses an end-to-end approach to train CNNs and divides the proposal area into object categories or backgrounds. One-stage algorithm includes YOLO, SSD[2], etc. The main difference is that the two-stage approach requires first generating the region proposals and then fine-grained object detection. One-stage algorithm directly extracts features from the network to predict the classification and location of objects without generating regional proposals. R-CNN has two key points: one is to use region proposal and extract its features with CNN; the second is the use of pre-training model. SPP-Net[3] added the pyramid idea, that is, the pyramid pooling layer was added between the convolutional layer and the full connection layer in the feature extraction process, and after the extraction of the convolutional layer, the feature maps of different sizes were operated into the same size to ensure the effective input of the full connection layer. Calculate the shared feature map of the whole image, and then map the shared feature map to the corresponding feature vector according to the object proposal. On the basis of R-CNN, Fast-RCNN[4] adopts the SPP-Net method and shares the convolutional layer. Instead of each proposal box as input of CNN, a complete picture as input are obtained in the fifth convolutional layer, performance has improved significantly. Then, the job of finding proposal boxes was also given to the neural network, so a neural network, Region Proposal (RPN), was designed to extract proposal boxes in Faster-RCNN[1], which replaced the time-consuming selective search and
was the core of Faster-RCNN. Later, the Mask R-CNN[5] developed upgraded the ROI Pooling layer of Fast R-CNN to ROI Align layer, and added a branch FCN layer on the basis of boundary box recognition for semantic mask recognition. R-CNN series method based on region proposal is the most important branch of current object detection technology. Among them, the Faster-RCNN method is more classical and has a good effect. This paper is an improved research based on it.

2. Related Work

2.1. Faster-RCNN

Insert a RPN after the last convolutional layer. RPN trained to produce region proposals directly; No need for external region proposals! After RPN, using RoIPooling, upstream classifier and bbox regressor just like Fast R-CNN. RPN directly uses CNN to produce proposal regions, no longer using selective search, but using convolution network to extract features on the whole image, using RPN to slide on the last layer of the convolution layer, using the border regression mechanism and anchor mechanism of the network to obtain multiple proposal regions with different areas and different aspect ratios. See as figure 1.

2.2. Human Detection

In recent years, with the exploration of autonomous driving, intelligent robots and other fields, the detection of human body in dense scenes is an urgent problem to be solved. However, the shielding between human bodies and the shielding of objects from human bodies have a great impact on the detection accuracy. As shown in figure 2 and figure 3, the occlusion problem of human in real scene is very serious, the existence of small objects also affects the detection accuracy.

![Figure 1. Structure of Faster-RCNN](image1)

![Figure 2. Small object scenario](image2)

![Figure 3. Occlusion scenario](image3)

A number of methods have been proposed for human detecting, some methods [6,7] use partial-based models to describe people in occlusion processing. These methods learn a series of partial
detectors and design mechanisms to fuse partial detection results to locate people in partial occlusion. Ouyanget al. [8] used multi-human detectors to assist single-human detectors to deal with partial occlusion, especially when people gather and block each other in real scenes. In [9,10], a group of human body occlusion patterns were found to learn the mixture of specific occlusion detectors. Mao et al. [11] proposed a novel network structure HyperLearner, suggesting that multi-task learning should aggregate additional features for human body detection. Brazil et al. [12] proposed a multi-stage autoregressive human detection system within RPN, in which each stage is trained with increasingly accurate labeling strategies. Zhou et al. [13] proposed joint learning of local detectors to reduce computing costs by using local correlation. Wang et al. [14] proposed a new bounding box regression loss Repulsion loss to detect human in the occluded environment. Liu et al. [15] introduced heat maps and simplified human detection into a direct, full-convolution central point and scale prediction task using the method without anchor box. Zhang et al. [16] divided people into five regions to improve the detection problem of shielding objects from the perspectives of ROI Pooling operation at the core of the two-stage detector and loss function. Aggregation Loss is designed to restrict the proposals to be closer to ground truth(GT) and to be as close as possible to proposals with the same object. The author predicted an occlusion score of 0 to 1 for the five local regions respectively, which was used to indicate whether the five regions were shielded or not. This part could be considered as a simple mask signal. Finally, the author multiplied the five scores and the features of the corresponding regions and added them together to obtain the final RoI feature. Brazil et al.[17] introduced semantic segmentation for human detection and achieved good results. In a word, deep learning promotes the development of human detection technology. Although there are a large number of human detection methods, it is still one of the key problems of human detector to carry out robust detection for each human in the occluded environment.

3. Proposed method

3.1. Feature Pyramid Network
Feature Pyramid Networks (FPN) was proposed by K. H et al. [2]. The specific structure is shown in figure 4. Originally to deal with the challenge of multi-scale objects detection, the detection of small objects has achieved good results. FPN is a feature extraction method based on pyramid aggregation. In addition to the bottom-up path, it also adds a down semantic information transfer path to realize the transfer of deep semantic information. The features of shallow graphics also have certain semantic information. Each layer can output corresponding detection results to realize the end-to-end multi-scale detection task.

3.2. Multi-scale feature fusion
Aiming at small targets in human detection, this paper proposes a multi-scale feature fusion network (MFFN). As shown in figure 5. FPN[2] and PANet[18] used side connection to carry out feature fusion, so that the fused features paid more attention to the adjacent resolution rather than other resolutions. In the process of information transmission, each fusion operation will make the semantic information of non-adjacent levels become diluted. MFFN firstly extracts the features \{C1, C2, C3, C4\} of four different scales from the image, and then rescale the features of these four scales to the features of the same scale: up sampling and down sampling (linear interpolation and pooling). Then sum it and take the average fusion operation to obtain the feature map after feature fusion, and then form the multi-scale feature map through convolution and deconvolution operation for subsequent detection.

3.3. Double-branch structure
People in complex environments have severe occlusion and some datasets have more abundant annotation information. How to use multi-part annotation information[19,20] to improve the accuracy has become an important research idea.
When both branches are considered positive, that is, $S^1 > 0.5$ and $S^2 > 0.5$, the probability of the two branches merging is higher. When one branch gives the positive example a low score (less than 0.5) and the other branch gives a high score (more than 0.5), the branch can increase the overall detection score. This will lead us to find complementarity between the two branches, thus reducing missed detection. If $\text{IOU}(P,F') \geq \alpha$ and $\text{IOU}(P,V') \geq \beta$, then the human proposal matches the human label well, where $\text{IOU}(P,F')$ is the intersection over union of the two regions P and F', and $\text{IOU}(P,V')$ is the proportion of the area of V' covered by P.

$$\text{IOU}(P,F') = \frac{\text{Area}(P \cap F')}{\text{Area}(P \cup F')} \quad (1)$$

$$\text{IOU}(P,V') = \frac{\text{Area}(P \cap V')}{\text{Area}(V')} \quad (2)$$

Let the training instance be represented as $X = (I,P,C,f',v')$. I represents the image that generates $P,C=0$ represents P does not contain human, $C=1$ represents P contains human. $f' = (f'^x,f'^y,f'^w,f'^h)$ and $v' = (v'^x,v'^y,v'^w,v'^h)$ are regression targets for whole-body and visible part estimation respectively[4]. $E = \{Xi = (Ii, Pi, Ci, f'i, v'i) | 1 \leq i \leq N\}$ is the definition of training.
instances, and N is the total number of training images. The model parameters of depth convolutional neural network are represented by M. The loss function of this structure is

\[ \text{L}(M, E) = L_{c_1}(M, E) + \lambda_1 L_p(M, E) + \lambda_2 L_{c_2}(M, E) + \lambda_3 L_v(M, E) \] (3)

where \( L_{c_1}(M, E) \) and \( L_p(M, E) \) represent the classification loss and regression box loss of the whole-body detection branch, while \( L_{c_2}(M, E) \) and \( L_v(M, E) \) represent the classification loss and regression box loss of the visible-part detection branch, respectively. The classification loss is calculated by cross entropy, and the regression loss is calculated by smoothing L1 loss [4].

3.4. Network architecture

MFFN was used to extract the features of the image and generate the feature map. The feature map is operated in the double-branch structure together with the proposal generated in RPN. This network structure is based on Faster-RCNN, using MFFN and double-branch structure, can produce better results, as shown in figure 7. The new network is called MDN, that is, MFFN and double-branch network.

4. Experiments

4.1. Dataset

The dataset for the experiment is called crowdhuman[21], there are 15,000 images for training and 4,370 images for testing. Some of the human instances in the image have a variety of occlusion conditions. Each human instance is annotated with a head boundary box, a human visible-part boundary box, and a human whole-body boundary box. Compared with some classical pedestrian detection datasets, Caltech, KITTI, CityPersons and COCOPersons, this dataset has higher density and more diverse scenes. See table 1 for a detailed comparison. If a single human instance is partially occluded, the annotator is required to complete the invisible-part and draw a complete whole-body boundary box. Cut out examples of each annotation from the image and give these cropped regions to the annotator to draw a visible-part boundary box. We further send the cropped regions to annotate a head bounding box.

4.2. Experimental details

In this paper, pytorch is used as the experimental framework, the computing platform is cuda10, GTX2080 graphics card is used for GPU acceleration, and it runs on Linux operating system. In the remaining experiments, we use \( \alpha = 0.5 \) and \( \beta = 0.5 \). Based on the Faster-RCNN network and FPN, resnet50 was used as the pre-training model to detect human. Stochastic Gradient Descent(SGD) was used to optimize the network on a GPU, with \( lr=0.0025 \), momentum factor of 0.9 and weight decay factor of 0.0001. Mini-batch was set as two images for each GPU, with a total of 600,000 iterations, which reduced the learning rate by 150,000 and 450,000 respectively. There are two evaluation indexes...
In this paper, average-precision (AP) and average-recall (AR). AP refers to the area enclosed by the precision-recall (p-r) curve, and generally the better the classifier, the higher the AP value. AP 50 means that the IOU threshold of the detector is greater than 0.5, and AP 75 means that the IOU threshold of the detector is greater than 0.75. Average-recall (AR) represents the maximum recall of a fixed number of 100 items detected per image, the average across categories and IOU ranges from 0.5 to 0.95.

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

TP(true positive): it is actually a positive example, and it is predicted to be a positive example; FP(false positive): the actual case is negative, while the prediction is positive. FN(false negative): the actual case is positive, the prediction is negative.

As can be seen from table 2 of the experimental results, the method adopted in this paper has improved the human detection to some extent. Relatively speaking, the double-branch structure has a better improvement effect in complex environments.

| Caltech | KITTI | CityPersons | COCOPersons | CrowdHuman |
|---------|-------|-------------|-------------|------------|
| Images  | 42,784| 3,712       | 2,975       | 64,115     | 15,000     |
| Persons | 13,674| 2,322       | 19,238      | 257,252    | 339,565    |
| Ignoreregions | 50,363| 45          | 6,768       | 5,206      | 99,227     |
| person/image | 0.32 | 0.63        | 6.47        | 4.01       | 22.64      |
| Uniquepersons | 1,273|<2,322       | 19,238      | 257,252    | 339,565    |

5. Summary and outlook

In this paper, crowdhuman dataset is applied. The scene of this dataset is complex and challenging. On the basis of the traditional convolutional network, the experiment adopted the method of double-branch structure and multi-scale feature fusion, aiming at to improving the problem of crowd occlusion and small target detection. Experiments show that the methods proposed in this paper improve these problems. Human detection is still a hot issue in the future. How to further improve the detection accuracy and speed is still the focus of research.

Acknowledgment

I would like to express my sincere thanks to my teachers and classmates in the laboratory, whose help and encouragement enabled me to complete this paper.

References

[1] Ren, S., He, K., Girshick, R.,& Sun, J.( 2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149.

[2] Lin, T.,Dollár,P.,Girshick,R.,He,K.,Hariharan, B.,& Belongie, S.(2017)Feature Pyramid Networks for Object Detection.In:IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, pp. 936-944.
[3] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916.

[4] Girshick, R. (2015). Fast R-CNN. In: IEEE International Conference on Computer Vision, Santiago, pp. 1440-1448.

[5] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2020). Mask R-CNN. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386-397.

[6] Tian, Y., Luo, P., Wang, X., & Tang, X. (2015). Deep Learning Strong Parts for Pedestrian Detection. In: IEEE International Conference on Computer Vision, Santiago, pp. 1904-1912.

[7] Zhou, C., & Yuan, J. (2016). Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In: Lai SH., Lepetit V., Nishino K., Sato Y. (eds) Computer Vision – ACCV 2016. Springer, Cham. Lecture Notes in Computer Science, vol 10112.

[8] Ouyang, W., & Wang, X. (2013). Single-Pedestrian Detection Aided by Multi-pedestrian Detection. In: IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, pp. 3198-3205.

[9] Tang, S., Andriluka, M., & Schiele, B. (2012). Detection and tracking of occluded people. In: The British Machine Vision Conference. UK. pp. 1-11.

[10] Pepik, B., Stark, M., Gehler, P., & Schiele, B. (2013). Occlusion Patterns for Object Class Detection. In: IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, pp. 3286-3293.

[11] Mao, J., Xiao, T., Jiang, Y., & Cao, Z. (2017). What Can Help Pedestrian Detection? In: IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, pp. 6034-6043.

[12] Brazil, G., & Liu, X. (2019). Pedestrian Detection With Autoregressive Network Phases. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 7224-7233.

[13] Zhou, C., & Yuan, J. (2017). Multi-label Learning of Part Detectors for Heavily Occluded Pedestrian Detection. In: IEEE International Conference on Computer Vision, Venice, pp. 3506-3515.

[14] Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., & Shen, C. (2018). Repulsion Loss: Detecting Pedestrians in a Crowd. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Portland, OR, pp. 7774-7783.

[15] Liu, W., Liao, S., Ren, W., Hu, W., & Yu, Y. (2019). Center and Scale Prediction: A Box-free Approach for Object Detection. https://arxiv.org/abs/1904.02948?context=cs.

[16] Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd. https://arxiv.org/abs/1807.08407

[17] Brazil, G., Yin, X., & Liu, X. (2017). Illuminating Pedestrians via Simultaneous Detection and Segmentation. In: IEEE International Conference on Computer Vision, Venice, pp. 4960-4969.

[18] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 8759-8768.

[19] Zhou, C., & Yuan, J. (2018). Bi-box Regression for Pedestrian Detection and Occlusion Estimation. European Conference on Computer Vision. Springer, Cham.

[20] Zhang, K., Xiong, F., Sun, P., Hu, L., Li, B., & Yu, G. (2019). Double Anchor R-CNN for Human Detection in a Crowd. https://arxiv.org/abs/1909.09998.

[21] Shao, Z., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., et al. (2018). CrowdHuman: A Benchmark for Detecting Human in a Crowd. https://arxiv.org/abs/1805.0012.