# Pedestrian detection algorithm based on Faster RCNN

YuLin Chen[1]

[1] Guangdong Provincial Key Laboratory of Optical Information Materials and Technology & Institute of Electronic Paper Displays, South China Academy of Advanced Optoelectronics, South China Normal University, Guangzhou 510006, P. R. China
452263754@qq.com

**Abstract.** Pedestrians are the most frequent targets in video surveillance and vehicle camera shooting, and the safety of pedestrians is the most concerned issue of social and public safety. In order to avoid accidents and traffic accidents caused by dense personnel, real-time detection of pedestrians on the street is particularly important. For pedestrian detection in various actual scenes, accuracy and real-time have always been the key indicators. Aiming at the difficulty of target detection in video, this paper adopts fast r-cnn algorithm for pedestrian detection, solves the problem of real-time pedestrian detection in video, and improves the accuracy of detection.

**Keywords:** Pedestrian detection, Deep Learning, Faster RCNN.

## 1 Introduce

In contemporary society, with the progress of science and technology and the rapid development of computer technology and hardware level, the intelligent analysis of video and image becomes more and more important. Pedestrian, as the most frequent target in video surveillance and mobile camera shooting, is the most concerned target of social public security. In particular, pedestrian detection technology plays a particularly prominent role in preventing accidents caused by mass gatherings and avoiding traffic accidents. In the assembly meeting, in 2014, an inter annual event resulted in a stampede accident due to the overcrowding of personnel. In 2014, the number of people controlling the scenic spots in the whole country was also consumed by manpower and material resources during the period of prevention and control in 2020. In terms of traffic accidents, pedestrians and bicycles account for more than 60% of fatal accidents. In the world, the number of deaths caused by traffic accidents in China is also in the forefront every year. A vision system that can detect pedestrians in real time can give early warning after the number of people reaches the predetermined threshold, so as to reduce various accidents caused by dense personnel; It can also make the car intelligent and actively pay attention to the relevant information of pedestrians in the process of street driving. In case of emergency, help the driver to respond quickly and reduce the occurrence of accidents[1].

EAI Urb-IoT 2021, December 20-21, Shenzhen, People's Republic of China

DOI 10.4108/eai.20-12-2021.2315018

The final form of pedestrian detection is to frame the pedestrian in the video by using a color rectangle. However, in the video, different pedestrians' dress, height, weight and other factors will lead to great differences in appearance between targets. Then there is the problem of target size[2]. The distance between the pedestrian and the camera will directly affect the size of the pedestrian in the video taken by the equipment. With the pedestrian from far to near, the corresponding target size in the video will also have a dynamic transformation from small to large, as well as the occlusion between pedestrians when they move. In addition, some application fields of pedestrian detection determine the real-time requirements of pedestrian detection algorithm. At present, most traditional pedestrian detection algorithms can not meet the requirements of the actual production environment in terms of real-time and overlapping pedestrian detection. Based on this, it is very necessary to introduce deep learning technology into the field of pedestrian detection. Firstly, in recent years, with the common development of computer software and hardware, target detection based on deep learning has made great progress[3]. Compared with traditional target detection algorithms, the speed and accuracy have qualitative changes. These advances also indicate the possibility of real-time pedestrian detection using deep learning technology. Secondly, the in-depth study of pedestrian detection algorithm based on deep learning is of great significance to the follow-up pedestrian tracking and pedestrian behavior recognition of pedestrian detection algorithm. The implementation of the complete framework of pedestrian detection technology based on deep learning requires the preprocessing of random data expansion and anchor box clustering before training; It is also necessary to extract pedestrian features with better robustness; At the same time, it is necessary to achieve accurate pedestrian positioning and reduce the offset; Finally, better post-processing is needed to stabilize the non maximum suppression operation and improve the recall rate[4].

## 2 Model method

Compared with voice and image, text is more complex and abstract. Human beings can have an overall understanding of the text content after reading the text according to their own understanding ability. However, the semantics in natural language is difficult to be directly understood by computers. Therefore, the text content must be expressed as forms that computers can understand and process, such as 0 and 1. Text representation model is to use numerical or symbolic vectors that can be expressed by computer to represent abstract and complex natural text. In order to better represent the text, it is necessary to extract the most representative features from the text data[5]. These features should have obvious statistical laws, which can reflect the text distribution in the feature space and minimize the computational complexity of text mapping to the feature space.

### 2.1 Convolutional neural network

CNN is a deep learning network widely used by the mainstream. It has a good performance in both supervised learning and unsupervised learning. It can often be seen in some international picture recognition competitions[6]. The data to be processed is first input from the input layer, then the features are extracted through the convolution layer, then the features are non-linear mapped in the excitation layer, then the features are clipped in the dimension in the pooling layer, finally the features are classified by the full connection layer, and finally the results are output.

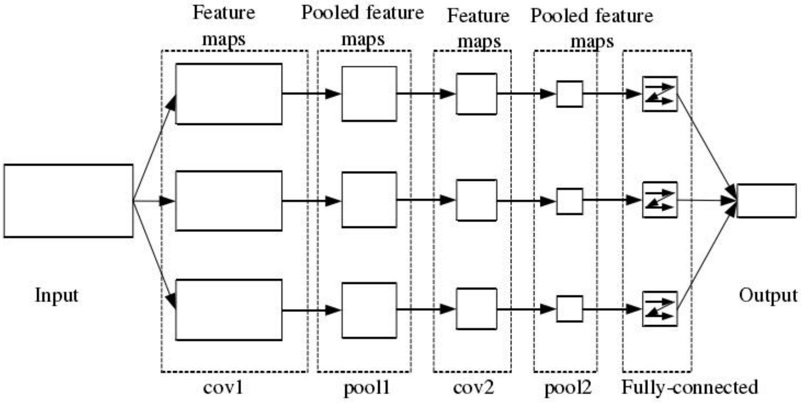

**Fig. 1.** The structure of convolutional neural network

The main function of pooling layer is to reduce the dimension of features and compress parameters and data to reduce the size of feature map. This can shorten the model training time, enhance the fault tolerance of the final model, and avoid the impact of over fitting to a certain extent. The main work of the full connection layer is to classify the previously extracted and processed features, and connect the upper neurons with each neuron in the lower layer in turn. In this layer, the feature map will be transformed into a two-dimensional feature vector and passed back through the excitation function. Convolutional neural network has the characteristics of simple training, local connection, weight sharing, down sampling and so on. Local connection is to connect each neuron with a small number of other neurons[7]. It is only used to learn local features, which can reduce many parameters; Weight sharing makes use of the similar features of the same target in the image, and a group of connected weights are shared by multiple groups, so that for the same target, multiple convolution kernels can extract the same features through weight sharing; Down sampling is mainly used to reduce the sampling in equal proportion according to the characteristics of different locations, so as to reduce the number of samples that are not particularly important, so as to further reduce the parameters.

### 2.2 Faster RCNN

Fast RCNN is widely used in the field of target detection. It inherits the part of feature learning and classification in fast r-cnn algorithm, then introduces RPN to replace SS algorithm to generate regional candidate box, and shares the features of the last convolution output between them, which greatly simplifies the calculation.

The introduction of RPN is the key to the wide application of fast r-cnn. It changes the selection search algorithm with huge amount of calculation, and allows the network to generate the target candidate box according to the feature map through the region recommendation algorithm. The advantage of this is to integrate the calculation of RPN into the training and testing process, which not only inhibits the generation of redundant candidate boxes, but also helps to improve the efficiency of training and testing, and opens a window for real-time target detection. In RPN network, the method of feature extraction according to recommendation frame has been optimized. Instead, nine anchors of different sizes are used as the detection frame. Firstly, the three sizes of 2128, 2256, 2512 are used as the benchmark, and then the detection frame of 9 sizes is obtained by zooming in and out in the ratio of 1:1, 1:2 and 2:1. When extracting features, input the whole picture, extract the features and output them, and then find the corresponding pixels on the original image according to the position of the feature points for regression and feature learning.

(1) Mapping ROI to the corresponding position of the feature map;

(2) The mapped area is divided into sections with the same size and number as the feature dimension according to the dimension of the feature set;

(3) Perform the maximum pool operation for each section divided in (2).

The above is the working steps of ROI pooling layer. After this operation, the characteristic map of the target to be detected will be output to the next layer in a fixed size, so that the whole network exists in an end-to-end structure.

Convolution neural network algorithm has specific requirements for the input image size, and the output image is also a specific size. This greatly limits the application of the algorithm. The ROI pooling layer is designed to break through this limitation. Generally speaking, the ROI pooling layer is a combination of ROI and pooling operations. ROI is the abbreviation of regions of interest, that is, regions of interest. In fact, ROI pooling layer is the pooling operation of regions of interest. For an image containing the target to be detected, the feature is extracted through the convolution layer, and then the proposed candidate area is provided through the RPN network.

Faster RCNN classifies by softmax function, and finally outputs the detection results after classification and regression.Fast r-cnn algorithm is a leader in the field of deep learning. Although it is not the best algorithm, from the perspective of maintaining the balance between detection speed and accuracy, fast r-cnn is the most appropriate choice for pedestrian detection in this scenario.

## 3 Experimental results and analysis

### 3.1 Dataset

In terms of model training, the Caltech public pedestrian data set is selected in this paper. The number of training iterations on the data set is set to 40000, in which the ratio of test set to training set is 1:4.

### 3.2 Model training and results

In this paper, experiments are carried out to verify the detection results of the algorithm from the following angles:
(1) Plenty of light

The light here refers to natural light, so this part of the experiment mainly selects daytime video frames for verification. When the light is sufficient, considering the influence of pedestrian posture, two different postures, normal walking and climbing over the fence, are selected for the experiment; Considering the influence of weather, three representative weather, sunny, rainy and snowy, were selected for the experiment.

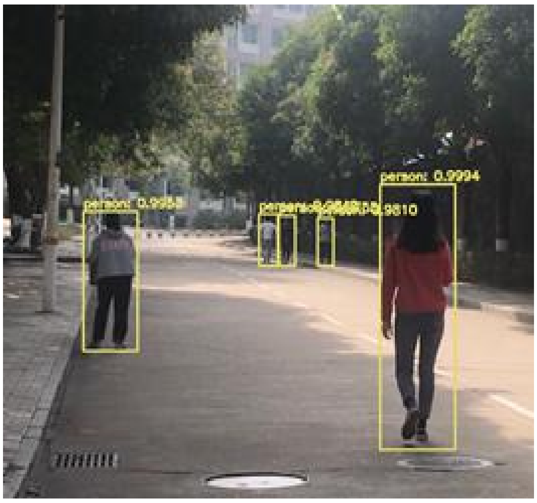

**Fig. 2.** Detection result

(2) Insufficient light
Aiming at the condition of insufficient light, the video at night and bad weather with low visibility are selected for the experiment. Because the application scenario of this method is mainly urban streets, the influence of artificial light source on the algorithm should be considered. Therefore, the selected video background contains colorful neon lights.

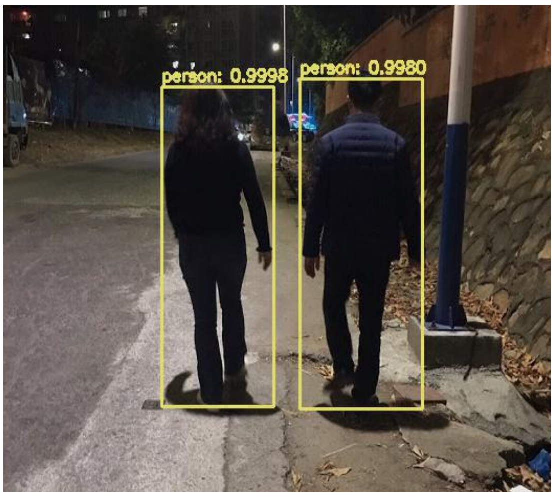

**Fig. 3.** Detection result

**Tab. 1.** Dectecion results

| | Accuracy | Error rate |
|---|---|---|
| Enough light | 90.36% | 10.48% |
| Insufficient light | 85.32% | 16.32% |

## 4 Conclusion

With the progress of science and technology and the increasing needs of human social life, people hope to make life more intelligent through science and technology to solve complex problems and tasks in life. The pedestrian detection algorithm used in this paper is tested in the daytime environment, under the conditions of complex background, blurred pedestrian image after shooting, and under the conditions of insufficient optical fiber and indoor environment at night, and the detection results are obtained.

## References

1. Gavrila D . Pedestrian Detection from a Moving Vehicle. Springer, Berlin, Heidelberg, 2000.
2. P Dollár, Appel R , Kienzle W . Crosstalk Cascades for Frame-Rate Pedestrian Detection. Springer Berlin Heidelberg, 2012.
3. F Xu, Fujimura K . Pedestrian detection and tracking with night vision. IEEE, 2003, 1:21-30.

4. Ouyang W , Wang X . Joint Deep Learning for Pedestrian Detection. IEEE International Conference on Computer Vision. IEEE, 2014.
5. Seemann E , Leibe B , Mikolajczyk K , et al. An Evaluation of Local Shape-Based Features for Pedestrian Detection. British Machine Vision Conference. DBLP, 2006.
6. Keller C G , Enzweiler M , Rohrbach M , et al. The Benefits of Dense Stereo for Pedestrian Detection. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(4):1096-1106.
7. Angelova A , Krizhevsky A , Vanhoucke V , et al. Real-Time Pedestrian Detection With Deep Network Cascades. British Machine Vision Conference. 2015.