Pedestrian Detection Method Based on Deep Convolution Neural Network

SHEN Mengmeng1, WANG Yong1, MA Jiaqi1, LI Chuanguo1, HE Liangbo1, Gaurav Barnawal1, SHAN Wei1,2*
1 College of Physics and Electronic Information, Huaibei Normal University, Huaibei, Anhui, 235000, China
2 School of Computer and Electronic Information /School of Artificial Intelligence, Nanjing Normal University, Nanjing, Jiangsu, 210046, China
*Corresponding author’s e-mail: shanwei@chnu.edu.cn

Abstract. Compared with the traditional pedestrian detection technology, the pedestrian detection technology based on deep learning has achieved overwhelming advantages. However, due to the large scale of deep convolution network, the demand for dedicated processor limits the popularization of pedestrian detection system. To solve these problems, this paper proposes a deep convolution network model with moderate network scale, which improves the universality of the detection model on the premise of ensuring the detection accuracy. Based on the low dimensional shallow convolution neural network, the optimal network structure is found from three aspects of network layer number, sensor field size and characteristic graph, and the final network parameters are determined by guiding experimental evaluation. The pedestrian image to be detected is input into the above network model, and the pedestrian detection result is obtained. The pedestrian detection results on several popular pedestrian databases such as Daimler, MIT and INRIA show that the network structure designed in this paper not only has moderate network size, but also has good detection performance. Cross experiments also verify the robustness and generalization of the algorithm. The work of this article is funded by the undergraduate training program for innovation and entrepreneurship (202010373050, 202010373051, 202010373046).

1. Introduction

Traditional pedestrian detection mainly uses statistical classification method, first divides the image into regions of interest (ROIs)[1], then extracts the features of ROIs, and then feeds them into the classifier[2]. In the early days, the combination of gradient direction histogram (HOG) and support vector machine (SVM) has been widely used, and good results have been achieved in practice[3]. However, the detection effect of traditional methods depends on the quality of design features. Researchers are required to have considerable professional knowledge and experience in artificial design features, and the real-time performance and robustness are poor.

Hinton et al.[4] first introduced the convolutional neural network (CNN) for deep learning in the Journal Science in 2006. The most important feature is that it is no longer designed by human but inputs raw data. Through a large number of training samples, the built neural network with multiple hidden layers is trained, and the network automatically learns the required features[5,6]. Scholars attempt to use the in-depth learning network in the field of pedestrian detection[7]. Among them, Zhang Hui et al.[8] used Fast-RCNN model for reference, joined K-means clustering algorithm to extract initial candidate regions, and used Caffe framework to implement convolution neural network model. However, the
whole convolution layer network has a large number of layers, a large amount of computation is easy to overfit, and the real-time performance is low. Hao Xuzheng et al.[9] with YOLO (you only look once) convolution neural network based on 50-layer pre-activated depth residual network detection, to reduce the computational burden of high dimensions, they concatenates a fixed-size 1 after 50-layer deep residual network 1×1×1024 convolution layer. Although the above method achieves a good pedestrian detection rate, the network is overloaded, the convolution layer parameters are fixed, and the capture effect is not good in complex motion environment.

In order to solve the problems of the deep learning network mentioned above, this paper uses LENET-5 network[10] with fewer layers and faster convergence to detect pedestrians in visible images. The LeNet-5 network contains only two convolution layers, two pooling layers and one output layer. By analyzing the network construction process, the number of experiments is reduced, and the network parameters are selected and optimized through instructive experiments.

2. Deep Convolution Neural Network

2.1. LeNet-5 Convolution Neural Network

The LeNet-5 network consists of two convolution layers, two pooling layers, and one fully connected output layer. Since the input sample size of the network is 32×32, while the size of the visible pedestrian sample is 48×96, different image matching field sizes are different. The size of the field determines the size of the convolution template. The traditional LeNet-5 network can no longer adapt to the larger size of visible pedestrian sample classification, so it is necessary to adjust the size of the convolution template in the network and determine the number of feature maps through experiments.

![Possible pedestrian detection network](image)

Figure 1 presents a possible pedestrian detection network model, where X and Y is the number of feature maps of layer 1 and 2 respectively. M×N, P×Q, m×n and p×q is the size of the feature map after convolution and pooling. Since the edge length of the signature graph must be an integer, p and q must both be non-zero integers. The next step is to explore the pending parameters in Figure 1.

2.2. Selection of Convolution Template Size

The size of the first layer convolution template that can be analyzed as shown in Table 1. Based on the size of the input image, we tentatively set the first layer convolution template C1 size to 7×7, 9×9, 11×11. Typically, the second convolution layer template C3 is less than or equal to the first convolution layer template C1, so the values are sorted in descending order starting with C1.

| group | width | height | C1   | M   | N   | P   | Q   | C3   | m   | n   | p   | q   | Possible to create network |
|-------|-------|--------|------|-----|-----|-----|-----|------|-----|-----|-----|-----|-----------------------------|
| 1     | 48    | 96     | 7×7  | 42  | 90  | 21  | 45  | 7×7  | 15  | 39  | 7.5 | 19.5 | No                           |

Figure 1 Possible pedestrian detection network

![Possible pedestrian detection network](image)
Table 1 shows that the size of C1 can only be 9×9 if p and q must both be non-zero integers with no image zeroing, as shown in Table 1 under group 4, 5, 6, 7. Considering that C1 is 9×9, P×Q is 20×44 and its size is close to 32×32. As the output size of C1 convolution layer decreases, the receptive field of corresponding C3 layer should also decrease. Excluding combination 4, we tentatively set the size of C3 of the second convolution layer as 3×3, 5×5 and 7×7, meaning the combination number in the table is 5, 6 and 7. Finally, the size of receptive field of C1 layer is 9×9, M×N is equal to 40×88, C3 layer tentative, next combined with the number of characteristic map X, Y, through the experiment to determine the size of C3 layer receptive field.

2.3. Selection of the number of characteristic graphs
The determination route of X and Y in Figure 1 is as follows: firstly, the number of characteristic graphs X(1~15) is changed successively in the single-layer convolution network (C1 and S2), and the value of X is determined at the minimum false detection rate, then the C1 layer and C3 layer are combined to form a two-layer network, and the number of characteristic graphs Y is changed successively, and the value of Y is determined at the minimum false detection rate. Because the receptive field of C1 layer is 9×9. The size of C3 layer is tentatively determined as 3×3, 5×5 and 7×7(Table 1). The following experiments are carried out for the three receptive field sizes of C3 layer.

3. Experiment
The experiment is divided into two parts, the first part is to determine the parameters, and the second part is to test the effect of the network after determining the parameters.

3.1. Determination of network parameters
This paper first selects the Daimler Stereo Pedestrian Detection Benchmark Dataset.

1) The selection of X
The single-layer convolutional neural network composed of C1 layer and S2 layer in Figure 1 is selected, and the receptive field size of C1 is 9×9, input image size 48×96, change the number of characteristic graphs X to 1~15 in turn, and take the false detection rate as measure performance.

\[
\text{False Positive (FP)} = \frac{ER}{TOTAL} 	imes 100\%
\]

In the formula (1), ER is the number of false samples, TOTAL is the total number of samples. The stronger the network extracts the sample characteristics, the lower the false detection rate. The experimental results are shown in Table 2. As shown in Table 2, when the size of the sensing field of C1 is 9×9, the number of signature graphs is too large or too small, the error detection rate of the network is high. When the X value is 8 in Table 2, the network error detection rate is the lowest and the result is the best, so the X value is 8.

Table 2 error detection rate under different X when the C1 layer is 9 × 9

| X   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| FP% | 4.31 | 2.98 | 1.85 | 1.70 | 1.55 | 2.10 | 1.70 | 1.55 | 1.90 | 2.71 | 4.17 | 4.01 | 3.87 | 3.90 | 5.00 |

3
2) C3 layer receptor field 3×3, selection of Y

Cascaded C1 layers and C3 layers form a two-layer convolution neural network. Set the number X of the feature diagrams in C1 to 8, and change the number Y of the feature diagrams in C3 in turn. Because C3 receptor field is 3×3(smaller than C1 9×9). In combination with Table 2, the error detection rate of corresponding network is higher when the number of characteristic graphs is too large or too small, so the range of Y values in this experiment is 3~9, and the experimental results are shown in Table 3. As shown in Table 3, when the values of Y are 5 and 8, the best results are obtained. Considering the speed of calculation, the value of Y is 5. Using the above steps, The best parameter of 3×3 size of C3 is obtained: 8 for X and 5 for Y can get the lowest false detection rate of 1.55%.

| Y  | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|----|-----|-----|-----|-----|-----|-----|-----|
| FP/%| 1.85 | 1.70 | 1.55 | 2.10 | 1.70 | 1.55 | 1.90 |

3) C3 size is 5×5, selection of Y

Setting the number X of the characteristic graph in C1 to 8 and changes the number Y(3~9) of the characteristic graph in C3. The experimental results show that the best parameter of 5×5 size of C3 is obtained: 8 for X and 5 for Y can get the lowest false detection rate of 1.35%.

4) C3 size is 7×7, selection of Y

Setting the number X of the characteristic graph in C1 to 8 and changes the number Y(3~9) of the characteristic graph in C3. The experimental results show that the best parameter of 7×7 size of C3 is obtained: 8 for X and 9 for Y can get the lowest false detection rate of 2.01%.

Comparing the experimental results of (1)~(4), it is finally determined that the C3 layer receptive field with the lowest rate of network misdetection is 5×5, Y is 5.

3.2. Network effect test

From the above discussion, we get a 5-layer deep convolution neural network: the first convolution layer convolution template size is 9×9, the number of feature maps X is 8; Second convolution layer template size of 5×5, the number of characteristic graphs Y is 5. In order to visually reflect the detection effect, the following performance indicators[11] are selected to evaluate the network.

\[
\text{Total False Positive (TFP)} = \frac{\text{ER}}{\text{TOTAL}} \times 100\% \quad (2)
\]

\[
\text{Detection Rate (DR)} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\% \quad (3)
\]

\[
\text{False Alarm Rate (FAR)} = \frac{\text{FN}}{\text{TP}+\text{FN}} \times 100\% \quad (4)
\]

In the formulas, ER is the total number of errors and TOTAL is the total number of samples. TP is the correct number of samples to be correctly classified, FP is the positive number of samples to be misclassified, and FN is the negative number of samples to be misclassified as positive.

To further verify the effectiveness of this method, we introduced MIT pedestrian database and INRIA pedestrian database. A large number of experiments have been done on the depth network proposed in this paper. Because there are only positive samples and no negative samples in the MIT pedestrian data, we can only calculate the detection rate but not the false alarm rate.

● Experiment 1: Using training set from INRIA database and test set from INRIA pedestrian database.
● Experiment 2: Using training set from Daimler database and MIT pedestrian database to test.
● Experiment 3: Using training set from Daimler database and test set from INRIA pedestrian database.
● Experiment 4: Using training set from INRIA database and test set from Daimler pedestrian database.
Experiment 5: Combining the Daimler database with the INRIA pedestrian database to form a new pedestrian database, assign the training and test sets, and do the training and test again. The results of the proposed algorithm and the traditional algorithm are compared as shown in Table 4.

Table 4 Comparison of the results between the algorithm and the traditional algorithm

| Supplementary experiment | The algorithm in this paper | HOG+SVM[3] |
|--------------------------|-----------------------------|------------|
|                          | TFP/% | DR/% | FAR/% | TFP/% | DR/% | FAR/% |
| Experiment 1             | 6.39  | 94.24 | 3.27  | 9.18  | 90.94 | 4.01  |
| Experiment 2             | 4.86  | 95.14 | —     | 9.30  | 90.70 | —     |
| Experiment 3             | 10.44 | 89.71 | 4.62  | 13.36 | 86.69 | 5.88  |
| Experiment 4             | 11.06 | 85.13 | 21.30 | 15.22 | 80.92 | 28.99 |
| Experiment 5             | 1.54  | 97.83 | 2.57  | 10.09 | 89.46 | 18.74 |

From Experiment 1, we can see that the network structure designed in this paper is reasonable to get better detection results. The cross experiment 2 indicates the total false detection rate of our network is 4.87% and the detection rate is 95.13%. Although the HOG+SVM method in reference [3] can achieve nearly 100% detection effect in MIT database, it uses MIT data set itself as training set and test set, and the effect is significantly reduced in cross experiment 2. Through cross experiments 2, 3 and 4, it can be proved that the performance of the traditional HOG+SVM method is significantly reduced when the training set and the test set are completely independent. On the contrary, the algorithm proposed in this paper still has good detection effect and all performance indicators are better than the traditional algorithm. Experiment 5 shows that the performance of the traditional HOG+SVM method is reduced when different data sets are overlapped, while the performance of this method is relatively stable. To sum up, this algorithm has higher robustness and better generalization ability.

4. Conclusion

Based on the deep learning convolution neural network, the pedestrian detection of visible light image is carried out by using 5-layer neural network. Through a large number of experiments, on the premise of defining the experimental steps, the parameters of the network are determined and optimized based on the best data. The results show that the detection rate of this network is higher, the false alarm rate is lower, and the performance of the network is better than other networks. The three pedestrian data sets from different countries also verify the robustness and generalization ability of the algorithm.

References

[1] J Liang, Huan M, Deng X, et al. Optimal Transmitter and Receiver Placement for Localizing 2D Interested-Region Target with Constrained Sensor Regions[J]. Signal Processing, 2021, 183(5): 108032.
[2] Saeidi M, Ahmadi A. High-performance and deep pedestrian detection based on estimation of different parts[J]. The Journal of Supercomputing, 2020(4): 2033-2068.
[3] Dalal N., Triggs B. Histograms of Oriented Gradients for Human Detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005, 886-893.
[4] Hinton G. E., Salakhutdinov R. R. Reducing the dimensionality of data with neural networks.[J]. Science, 2006, 313(5786): 504-507.
[5] ZENG X, OUYANG W, WANG X. Multi-stage Contextual Deep Learning for Pedestrian Detection[C]// Proceedings of International Conference on Computer Vision. 1-8 Dec. Sydney, NSW, Australia, IEEE, 2013: 14132232.
[6] TIAN Y, LUO P, WANG X, et al. Pedestrian Detection aided by Deep Learning Semantic Tasks[J] IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015(01): 5079-5087.
[7] Angelova A, Krizhevsky A, Vanhoucke V. Pedestrian detection with a Large-Field-Of-View deep network[J]. Proceedings-IEEE International Conference on Robotics and Automation, 2015, 2015: 704-711.
[8] ZHANG H, DU Y, NING S R, et al. Pedestrian detection method based on fast RCNN [J]. Transducer and Microsystem Technologies, 2019, 38 (02): 147-149 + 153 (Ch).

[9] HAO X Z, CHAI Z Y. An improved pedestrian detection method based on depth residual network [J]. Application Research of Computers, 2019, 36 (05): 1569-1572 + 1584 (Ch).

[10] LECUN Y., BOTTOU L., BENGIO Y., et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[11] FEI J C, RUI T, ZHOU Y, et al. Multi input convolution neural network based on gradient [J]. Opto-Electronic Engineering, 2015, 42 (03): 33-38 (Ch).