Downlink Human Detection Under Occlusion Based on Convolutional Neural Network

Chenxiang Zhang[1]
[1]*Suzhou Industrial Park Institute of Services Outsourcing, Suzhou 215123 China*
**Corresponding Author.*

Abstract

As one of the hotspots in the field of target detection, pedestrian detection has very high value in the application fields such as driverless vehicle assistance system, intelligent monitoring system and service-oriented intelligent robot. The pedestrian occlusion studied in this paper can be divided into two types: human to human self occlusion and object to human occlusion. Aiming at the problems of missing detection, high false detection and low detection accuracy of small-size targets in real scene pedestrian detection methods, a pedestrian detection algorithm based on improved convolution neural depth network model is proposed in this paper. The algorithm improves the original SSD network model by extracting the lower level output feature map, and uses the abstract features of the output of different layers of convolutional neural network to detect pedestrian targets respectively. This method combines the multi-layer detection results and improves the detection performance of small target pedestrians. The experimental results show that the accuracy of the proposed algorithm is 93.8% on the INRIA test set, and the missed detection rate is as low as 7.49%.

Keywords*: Pedestrian detection, pedestrian occlusion, detection accuracy, convolutional neural network*

I. Introduction

There are many differences in the appearance and dress of pedestrians in real scenes. Their body parts have various postures and shapes, and are also vulnerable to factors such as occlusion and illumination [1]. On the one hand, because of these appearance, posture and environmental factors, the research of pedestrian detection has many difficulties. In terms of appearance differences, the appearance of pedestrians is different due to different perspectives and pedestrian postures. For example, people wear various types of clothes, scarves, hats, luggage and other items, which have a great impact on pedestrian detection [2]. In terms of illumination difference, different brightness will also aggravate the difficulty of detection. In addition, in the complex pedestrian detection application scene, when the color, texture, shape and appearance of the object are very similar to the human body, the detector will be confused and can not accurately distinguish the target [3]. On the other hand, for the occlusion problem, in many application scenarios, there are often buildings blocking pedestrians or people blocking people.

Usually, the detector can only obtain part of the human body information, which brings serious challenges to the detection algorithm. In recent years, in order to improve the effect of pedestrian detection, researchers have applied convolutional neural network commonly used in the field of target detection to the field of pedestrian detection and made great progress [4]. However, in practical application, due to the occurrence of various pedestrian occlusion, the performance of pedestrian detection model is greatly reduced. The above factors make pedestrian occlusion one of the difficulties of pedestrian detection, and it is also the research object of modeling and experiment in this paper.

II. Problem statement

The problem to be solved in pedestrian occlusion detection is to mark the spatial position of all pedestrians in video

frames or pictures with rectangular boxes. Its detection method is similar to face detection, and it is also one of the typical problems of target detection. Some researchers use the methods proposed in target detection to detect pedestrians directly. However, these methods are difficult to obtain the best performance. The main reason is that pedestrians often gather together, which is easy to be blocked by pedestrians and other objects in reality. Because the pedestrian target is easy to be occluded, the model can not obtain complete information, that is, the features obtained in the feature extraction stage of target detection will be relatively reduced, resulting in some missed detection. The above shows that it is challenging and meaningful to deal with the occlusion problem in pedestrian detection. As shown in Figure 1, pedestrian occlusion in real scenes is divided into two types: intra class occlusion and inter class occlusion.

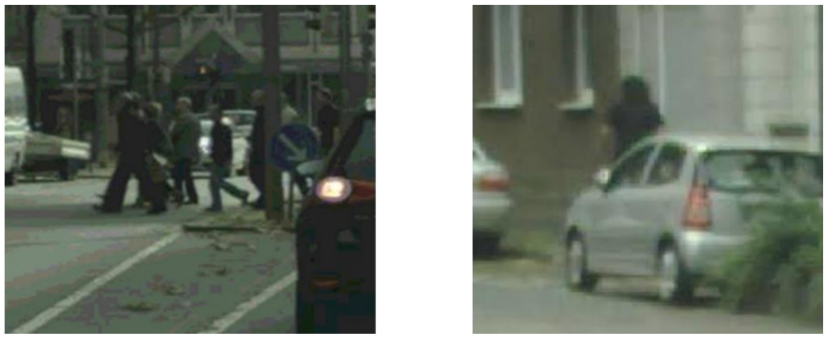

Fig 1: The occlusion type of pedestrian

1) Intra class occlusion: intra class occlusion mainly refers to the occlusion between pedestrians. Pedestrian occlusion usually occurs in the crowd, which shows that the body parts of pedestrians have a certain overlap rate. General detectors are easy to detect complete pedestrian targets, but for occluded pedestrians, due to the incomplete extracted features, they are easy to be ignored, resulting in false detection and missed detection.

2) Inter class occlusion: inter class occlusion mainly refers to the occlusion of objects to people. Generally, objects refer to non pedestrian parts, such as cars, street lamps and other common objects. When the car blocks the lower part of the pedestrian, the features learned by the detector are disturbed, which includes both the upper body features of the pedestrian and some car features, so it is impossible to confirm the pedestrian target. Therefore, it is easy to miss detection when the detector detects seriously blocked pedestrians. The goal of this paper is to use the full convolution pedestrian detection model to deal with different pedestrian occlusion patterns. For intra class occlusion, this paper mainly reduces the impact of other pedestrians on the target pedestrians, and adopts the method of introducing the rejection loss function into the pedestrian intra class occlusion detection model based on full convolution neural network; For inter class occlusion, this paper introduces the semantic attention module into the full convolution network. The model includes two parts, namely, semantic segmentation part and detection part. The semantic segmentation part promotes the latter detection module to extract the features of the visible part of the occluded pedestrian by strengthening the response to the pedestrian part.

III. Detection principle and algorithm

Machine learning based method is the mainstream of pedestrian detection algorithm at this stage. It mainly uses two modules: feature extraction and classification task to detect pedestrians in image or video frames. The feature extraction part mainly obtains the pedestrian appearance features and depth semantic feature information such as edge features, color features and texture features, and then applies these features to train the classifier to distinguish the background and pedestrian targets efficiently and output the classification results. Finally, the input test image is detected.

With the rise of big data, deep learning methods have produced a large number of research results in the fields of

speech recognition, pedestrian detection, face recognition, natural language processing and so on. Convolutional neural network is an end-to-end learning model. Compared with shallow learning structure, its deep learning structure is deeper and more powerful, and can learn richer and deeper features. The output of any layer of the network layer can be used as input data for feature extraction, which is the most prominent aspect of deep learning. The extracted features can also be further extracted and matched, and finally the classifier can be used to effectively complete target classification and recognition. The quality of feature extraction plays a key role in the operation effect of classifier. Traditional pedestrian detection uses the characteristics of manual design, and has achieved good results in some aspects, but its generalization is very limited. The convolutional neural network has better discrimination and generalization by using the features learned from the network layer.

A large number of studies show that this end-to-end learning mode is more conducive to achieve the best state of image recognition. As shown in Figure 2, the common convolutional neural network mainly combines the convolution layer, pooling layer and full connection layer into a basic structure. Among them, the input of CNN is some original data, such as RGB image, original audio data, etc. The output is the confidence that the image belongs to each category. The function of convolution layer is to convolute the input data to obtain feature map, that is, to extract features. The main function of the pooling layer is to carry out down sampling to reduce the number of network parameters by removing the unimportant parts in the feature map. The role of the full connection layer is to realize classification. The bottom layer of convolution neural network is to operate the receptive field of the local image, and then use convolution operation, aggregation operation and nonlinear variable mapping to obtain the classification probability.

When each convolution layer uses convolution kernel to obtain the characteristic image of the next layer, and then uses activation function for nonlinear transformation, the expression and fitting ability of the network is improved. By abstracting and expressing the input data layer by layer, the high-level semantic information of the image can be extracted, and the learned features are more discriminative. As shown in Fig. 2, the image is input into the convolution neural network, passes through the convolution layer with three filters, performs convolution operation, and outputs three feature maps. Then, down sampling is carried out through the pool layer, and three characteristic diagrams with the size of 1 / 2 of the original size are output. Then input to the second convolution layer, there are 6 filters, perform convolution operation, and output 6 new characteristic graphs. After that, the second pool down sampling was carried out, and six characteristic diagrams with the size of 1 / 2 of the original size were obtained. Finally, the full connection layer is connected, the input characteristic graph is weighted and summed, and the softmax function is input for classification output.

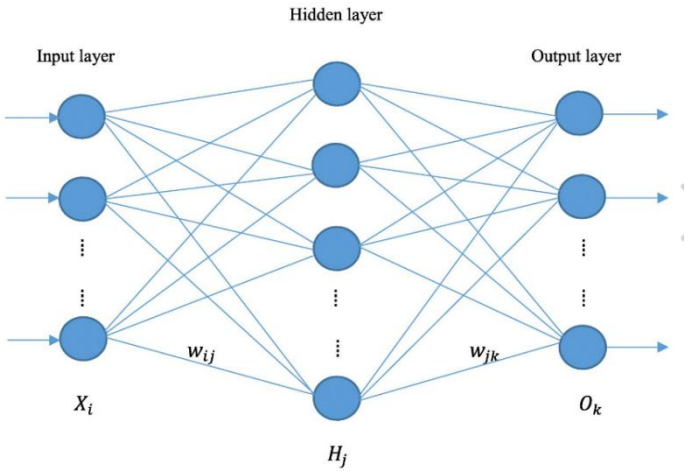

Fig 2: The topology of the convolution neural network

The basic algorithm is shown as the equation (1) [9-10]:

$$C^1 = C - C^0 , e^1 = e - e^0 ,$$
$$\eta^1 = \eta - \eta^0 , \rho_1 = \rho - \rho_0 \quad (1)$$

$$f(x,\omega) = f^0(x,\omega) + \int_V \mathcal{S}(x-x')(\mathrm{L}^1 F(\mathrm{y}')$$
$$+\rho_1\omega^2 \mathbf{g}(R)\mathrm{T}_1 f(\mathrm{y}')]S(\mathrm{y}')\mathrm{dy}' \quad (2)$$

The we get:

$$\frac{1}{\Gamma(1+\alpha)}\int_R \frac{f(t)}{(t-x)^\alpha}(dt)^\alpha$$
$$= \lim_{\varepsilon\to 0}[\frac{1}{\Gamma(1+\alpha)}\int_{-\infty}^{x-\varepsilon}\frac{f(t)}{(t-x)^\alpha}(dt)^\alpha +$$
$$\frac{1}{\Gamma(1+\alpha)}\int_{x+\varepsilon}^{\infty}\frac{f(t)}{(t-x)^\alpha}(dt)^\alpha] \quad (3)$$

(1) Convolution layer

The function of the convolution layer is to use the convolution kernel (i.e. filter) to extract the features of the input data, that is, the convolution kernel slides on the upper input layer to perform convolution operation. The convolution kernel parameters connected with the corresponding local area pixels are equivalent to the weight parameters set in the traditional neural network. The results on the convolution layer can be obtained by multiplying and summing the convolution kernel parameters with the corresponding pixel values (plus a bias parameter). As shown in Fig. 3, the input image is 32 × thirty-two × 3, 3 is the depth of the image (i.e. R, G, b). The convolution layer has two filters, each with a size of 5 × five × 3. The depth of the filter must be equal to the depth of the image. The input image passes through two filters and the output size is 28 × twenty-eight × Characteristic diagram of 2. When using convolution neural network to process image recognition task, deepening the level of convolution layer will extract deeper feature map, and the amount of calculation will also increase a lot. The following describes two important characteristics of the algorithm: local receptive field and weight sharing.

$$\boldsymbol{\sigma}_{ik}(\bar{k},\omega) = -\frac{1}{\eta_{11}^0}\frac{1}{\bar{k}^2} + \frac{1}{\rho_0\omega^2}\left(\frac{e_{15}^0}{\eta_{11}^0}\right)^2 \frac{\beta_\perp^2}{\bar{k}^2-\beta_\perp^2} , \gamma_i(\bar{k}_i,\omega) = \frac{1}{\rho_0\omega^2}\left(\frac{e_{15}^0}{\eta_{11}^0}\right)^2 \frac{\beta_\perp^2}{\bar{k}^2-\beta_\perp^2} m_i \quad (4)$$

In which,

$$\alpha^2 = \frac{\rho_0\omega^2}{C_{11}^0} ,$$

$$\alpha^2 = \frac{\rho_0\omega^2}{C_{66}^0} , \beta_\perp^2 = \frac{\rho_0\omega^2}{C'_{44}} ,$$

$$C'_{44} = C^0_{44} + \frac{(e^0_{15})^2}{\eta^0_{11}} \tag{5}$$

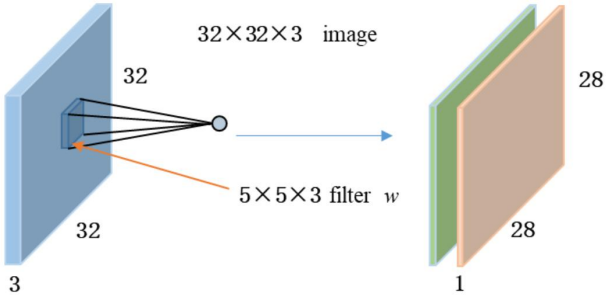

Fig 3: The diagram of convolution layer

1) Local receptive field: the filter of the convolution layer has fixed weight parameters, which is used to perform convolution calculation with the local area of the input data. After the data calculation of the current window is completed, the data window moves smoothly until all areas are convoluted. As shown in Figure 4, the size is 5 × five × The orange matrix of 1 represents the input data, and the size is 3 × three × The green matrix of 1 represents the filter. Convolute the filter with the area in the lower right corner, and the obtained value is 4. After all convolution operations, finally combine all local features to obtain the feature map of the blue part. This method is the local receptive field mechanism of CNN, which replaces the fully connected link mode in the traditional fully connected neural network, and greatly reduces the number of training parameters required by the network.

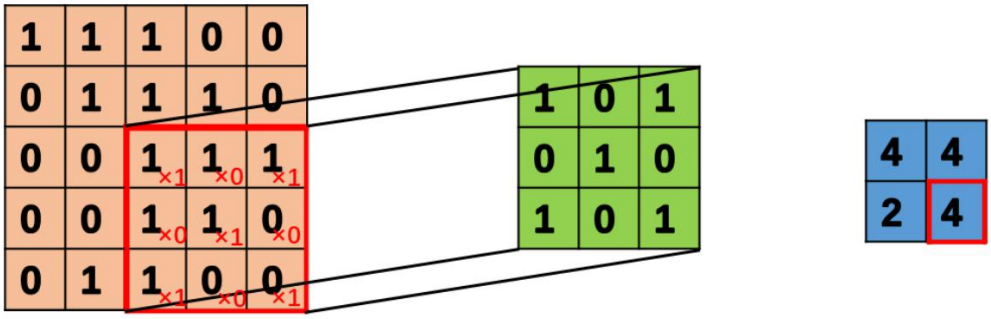

Fig 4: The diagram of local receptive field

2) Weight sharing: input a picture of a certain size, the filter first learns some features from the local region, and then uses the same method to extract the same features at different positions of the image. The parameters (i.e. weight) of the filter are fixed, which is equivalent to using the same convolution kernel to convolute the whole image at any position, and extract the features corresponding to the convolution kernel in the whole image. This weight is unchanged, that is, the weight sharing mechanism, which can greatly reduce the amount of calculation of the network.

(2) Pool layer

The function of the pool layer is to down sample the characteristic map output by the convolution layer. By reducing the spatial size of data and reducing the weighted parameters of neurons in CNN, the possibility of over fitting in the optimization process can be effectively reduced. Because the pooling layer will compress the input feature map to a certain extent, it can reduce the feature map and reduce the computational complexity, and the main features are extracted through compression. Random pooling, maximum pooling and average pooling are

commonly used. Because the maximum pool calculation is simpler and the effect is better, the maximum pool is commonly used in all practical applications. As shown in Figure 5, in 2 × The sample area of 2 takes the maximum value as the sample value of the last sampling.

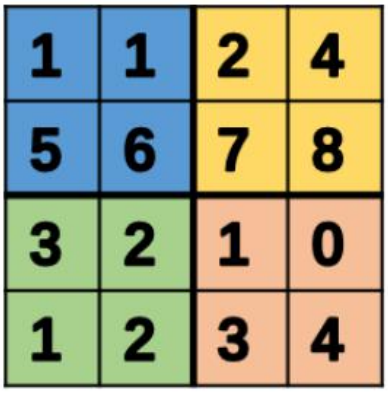

Fig 5: The diagram of max pooling

Rewrite again Eq. (4) as

$$\begin{aligned}\hat{f}_H^{\alpha}(x) &= \frac{1}{\Gamma(1+\alpha)}\int_{-\infty}^{\infty}\frac{f(t)}{(t-x)^{\alpha}}(dt)^{\alpha} \\ &= \frac{1}{\Gamma(1+\alpha)}\int_{-\infty}^{\infty} f(t)g(x-t)(dt)^{\alpha} \\ &= f(x)*g(x),\end{aligned} \tag{6}$$

(3) Full connection layer

The function of the full connection layer is to map the convoluted and pooled distributed features to the sample label, connect all the input features into a fixed length feature vector, and transmit the output value to the classifier (such as softmax classifier) for classification. For example, the convolution layer output of the previous layer of the full connection layer is w × H × 512, if the full connection layer is a vector with the size of 4096, the size of W can be used × H × five hundred and twelve × 4096 convolution kernel performs global convolution operation, which greatly reduces the influence of spatial structure characteristics on target detection and recognition. From Eq. (6), we can get:

$$\partial_j(C_{ijkl}\partial_k u_l + e_{kij}\partial_k\varphi) - \rho\ddot{u}_i = 0 \tag{7}$$

$$\partial_j(e_{ijkl}\partial_k u_l - \eta_{kij}\partial_k\varphi) = 0 \tag{8}$$

Compared with traditional methods, convolutional neural network has the advantage of automatically learning multi-level features, that is, the shallow level can obtain the boundary contour and other information of the target by extracting local features such as edge and texture; Abstract features can be learned at a deeper level, which is helpful to judge the category of objects in the image. FCN turns the full connection layer of traditional CNN into convolution layer.

IV. Experiment Results

In all experiments in this paper, vgg16 is used as the basic network structure, and the experimental training parameters in the detection module are set according to the standard recommended settings. For the semantic segmentation module, the first three convolution blocks (i.e. B3, B4, B5) have the same structure as vgg16, but the pooling layer is removed, and void convolution is applied to the last two convolution blocks (B4, B5). The hole step size of hole convolution is set to 2 and 4 respectively. B6 is 1 with sigmoid function as excitation function × 1, and the number of channels is 1, B7 is also set to 1 with sigmoid function as excitation function × 1.

The benchmark model includes many optimal methods for pedestrian detection, such as FasterRCNN [5], FasterRCNN + ATT Part [6], FasterRCNN + Reploss [7], Social Topology Line Localization (TLL) [8]. The detection principles of these pedestrian detection methods are briefly described below. The fasterRCNN method only uses the faster RCNN model as the pedestrian detection model for pedestrian occlusion detection [9]. FasterRCNN + ATT part [10] method improves the faster RCNN model for pedestrian occlusion detection by using the attention mechanism of channel features to activate body part response. The fasterRCNN + reploss method detects pedestrian occlusion by combining the rejection loss function with the faster RCNN model [11]. The somatic topology line localization method uses the head, bottom and body lines with the head pointing to the bottom to replace the commonly used pedestrian bounding box annotation for pedestrian occlusion detection.

The semantic attention map trained by the semantic segmentation module is visualized. In the hot spot map, the whole body of pedestrians and the visible part of blocked pedestrians have obvious response. Among them, the upper bodies of two pedestrians who were seriously blocked by cars still showed obvious response. This hotspot map shows that the semantic segmentation module can extract features from severely occluded pedestrians. Fig. 6 shows the effect before and after the introduction of exclusion loss and semantic attention network. The red boundary box represents the detection result, and the green boundary box represents the real box. The non optimized model can not detect pedestrians blocked by other non pedestrian objects, while the improved pedestrian detector significantly reduces the number of false and missed samples. In addition, the improved model can locate different pedestrians in the crowd, which shows that the method in this paper is effective for both inter class occlusion and intra class occlusion.

Fig 6: Experiment results

V. Conclusion

In this paper, a pedestrian occlusion detection scheme based on convolutional neural network is proposed. In order to deal with the occlusion between classes, a semantic segmentation module is introduced, through which the semantic attention map can be obtained by using the visible bounding box. And this module makes the later detection module more focused on extracting the features of the visible part of the occluded pedestrian. From the experimental results, the detection performance in terms of severe occlusion level has been significantly improved. This shows that the proposed method can effectively improve the performance of pedestrian occlusion detection in complex background.

Acknowledgements

The paper is supported by research results on the objects of “Cyan Engineering” supported by Jiangsu Universities in 2019 (No. JSQL201901).

References

[1] Lu Hongtao, Zhang Qinchuan. Review on the Application of Deep Convolution Neural Network in Computer Vision. Data Acquisition and Processing, 2016 (1): 1-17

[2] Tianyi Qin, Drivers drowsiness detection in embedded system, IEEE International Conference on Vehicular Electronics and Safety, 2007. ICVES

[3] Ian F. Akyildiz, David M. Gutierrez-Estevez, and Elias Chavarria Reyes. 2010. The evolution to 4G cellular systems: LTE-Advanced. Phys. Commun. 3, 4 (December 2010), 217-244.

[4] G. Araniti, V. Scordamaglia, M. Condoluci, A. Molinaro, A. Iera, ”Efficient Frequency Domain Packet scheduler for Point-to-Multipoint transmissions in LTE networks,” Communications (ICC), 2012 IEEE International Conference on , vol., no., pp.4405,4409, 10-15 June 2012.

[5] M. Condoluci, G. Araniti, A. Molinaro, A. Iera, J. Cosmas, ”On the impact of frequency selectivity on multicast subgroup formation in 4G networks,” Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on , vol., no., pp.1,6, 5-7 June 2013.

[6] A. Khandekar, N. Bhushan, Ji Tingfang, V. Vanghi, ”LTE-Advanced: Heterogeneous networks,” Wireless Conference (EW), 2010 European , vol., no., pp.978,982, 12-15 April 2010.

[7] Yan Chen, Shunqing Zhang, Shugong Xu, G.Y. Li, ”Fundamental tradeoffs on green wireless networks,” Communications Magazine, IEEE , vol.49, no.6, pp.30,37, June 2011.

[8] S. Frattasi, R.L. Olsen, M. De Sanctis, F. Fitzek, R. Prasad, ”Heterogeneous services and architectures for next-generation wireless networks,” Wireless Communication Systems, 2005. 2nd International Symposium on , vol., no., pp.213,217, 5-7 Sept. 2005.

[9] I. Bisio, M. Marchese, ”Power Saving Bandwidth Allocation over GEO Satellite Networks,” Communications Letters, IEEE , vol.16, no.5, pp.596,599, May 2012.

[10] J. Anand, A. S. Buttar and R. Kaur, "Handover Triggering Based Spectrum Handover Approach in Cognitive Radio Network: A Survey," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 830-835, DOI: 10.1109/ICECA.2018.8474549.

[11] N. K. Panigrahy and S. C. Ghosh, "Analysing the Effect of Soft Handover on Handover Performance Evaluation Metrics Under Load Condition," in IEEE Transactions on Vehicular Technology, vol. 67, no. 4, pp. 3612-3624, April 2018, DOI: 10.1109/TVT.2017.2786342.