World Scientific Research Journal Volume 7 Issue 1, 2021
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

Research on Road Target Detection Method based on Feature Fusion

Lu Wang[1, 2]

[1]Department of Electrical Automation, Shanghai Maritime University, Shanghai, 201306, China

[2]Department of Image and Network Investigation, Railway Police College, Zhengzhou, 450053, China

Abstract

The increasingly complex road traffic environment has brought severe challenges to the safety of road traffic. In this paper, taking the accurate detection of road targets as the starting point, and combined with the current advanced convolution neural network technology, it analyzes and discusses the feature learning method based on feature fusion, and on this basis, it further proposes the target detection method based on feature fusion, and then it verifies the effectiveness of the proposed method through the simulation, thus, it provides a certain reference for road target detection in complex road traffic environment.

Keywords

Feature fusion; Road target detection; Multi-scale; Convolution neural network.

1. INTRODUCTION

With the further development of the economy and society, the number of cars is increasing, followed by complex road traffic conditions and frequent road traffic accidents. The complex road traffic conditions not only affect the travel, the living environment and living quality of urban residents, but also bring serious waste of resources, and also bring serious security risks to urban road traffic [1, 2, 3, 4, 5]. In the face of such a severe road traffic safety situation, how to alleviate the problem of road traffic safety through relevant technical means has become an important research direction. As the most basic research problem of road traffic safety, road target detection has become a research hotspot. The so-called road target detection is to find the interested road target in the scene image of road traffic[6,7,8]. Especially in the context of the current application of intelligent transportation, the complexity of road traffic scenes and the diversification of road targets bring certain challenges to road traffic safety, but also bring difficulties to the detection of road targets. Road target detection is not only an important issue in the research of intelligent road traffic, but also a problem involved in the field of intelligent public security.

Traditional target detection methods show unique advantages in image detection [9, 10, 11, 12, 13, 14, 15], and can be applied to road traffic target detection to a certain extent. However, due to the particularity and complexity of the target scale distribution in road traffic scene, the traditional manual feature target detection method and the classic deep learning target detection method still have some difficulties in the application of target detection in the actual road traffic scene, such as the inability to extract the effective features suitable for road multi-target detection, which affects the accuracy of target detection. Therefore, combined with the practical application needs of intelligent public security and intelligent traffic field for road

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

target detection, how to use more applicable target detection methods to extract more effective feature expression from complex road traffic scenes and detect road targets more accurately has become a research content to be solved in road target detection problem.

In this paper, aiming at the difficult problem of accurate target detection in road target detection, and taking the current advanced convolution neural network technology as the background, it studies and proposes a feature learning and target detection method based on feature fusion, so as to provide some technical reference for the road target detection in the current complex traffic environment.

2. FEATURE LEARNING METHOD BASED ON FEATURE FUSION

In the current field of visual image processing, especially in road traffic scene images, many scenes using the deep neural network method to extract and learn features on a single scale of the same size receptive field. This method has its certain success, but in the complex road traffic scenes with complex image context structure information, the single scale feature learning method also shows some limitations. Therefore, in order to further improve the performance of feature extraction for various kinds of complex multi-target in road traffic scenes, in this part, it proposes a road target feature learning method based on multi-scale feature fusion, which can improve the learning and expression ability of road target features, and also can improve the ability of network model to capture the information of image context structure. The road target feature learning model based on multi-scale feature fusion is shown in Figure 1.

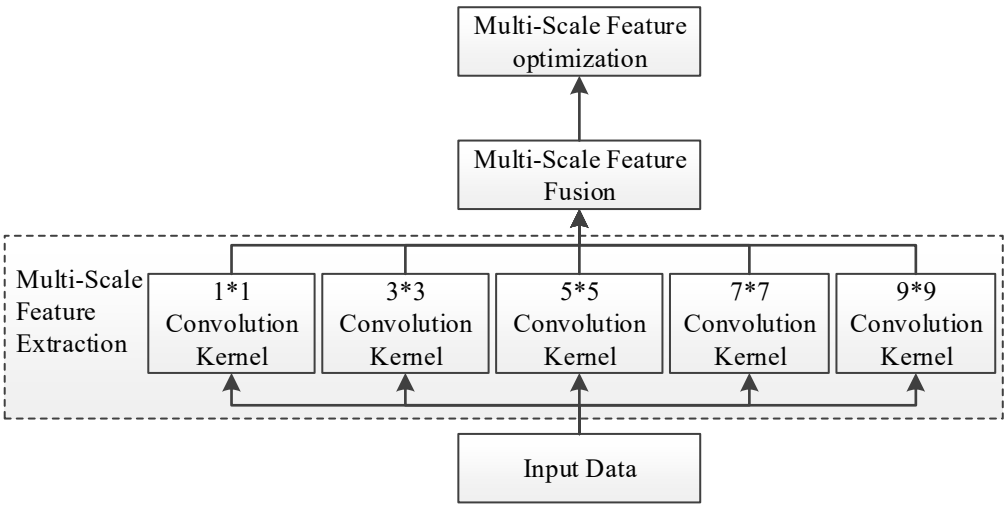

Figure 1. The road target feature learning model based on multi-scale feature fusion

It can be seen from the figure that the target feature learning model is mainly divided into three parts: multi-scale feature extraction module, multi-scale feature fusion module and multi-scale feature optimization module. The three main modules are introduced respectively as follows.

2.1. Multi-scale Feature Extraction

The feature extraction module is designed to solve the problem of insufficient expression ability of extracted features when extracting features with single scale from the input data. Different from the single scale feature extraction method which only uses one convolution kernel in feature extraction, the designed multi-scale feature extraction method here uses multiple convolution kernels with different scales to extract features in parallel simultaneously, which forms a parallel structure of multi-scale convolution kernels. In this parallel structure of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

multi-scale convolution kernels, each convolution kernel has a different size from other convolution kernels, which can be used to extract the features of the input image on the convolution kernel, while the convolution kernels with different sizes can simultaneously extract the features of the input image in parallel, so as to obtain the features of the input image with different scales. In the design of multi-scale feature extraction, different sizes of convolution kernels, such as 1×1 convolution kernel, 3×3 convolution kernel, 5×5 convolution kernel, 7×7 convolution kernel and 9×9 convolution kernel, can be used. Using the five convolution kernels at the same time, five different image features with different scales will be extracted. It should be noted that the 1×1 convolution kernel is designed to better retain the context details of the original input data in the shallow layer, so that after many complex convolution operations, the deep network features can still contain the context structure information. At the same time, in the design of multi-scale convolution kernel, it can design more abundant combination structure for convolution kernels with different sizes, and the feature extraction will be more abundant with the increase of convolution kernel scales. This will also increase the weight parameters of the whole network model, which will make the network over-fitting. Therefore, the selection and combination of convolution kernels with different scales should be adjusted according to the specific dataset size and network training requirements.

According to the design idea of convolution neural network, in the process of feature extraction of convolution neural network, after convolution operation of input data or features of the upper layer, nonlinear activation function is needed, so that the network model has good nonlinear performance. Therefore, for each convolution kernel operation with different scales, the nonlinear activation function should be used to nonlinearize the results. If the input image data of the whole network is X, and the network model contains multiple convolution layers, and the multi-scale convolution kernel in the operation process of each convolution layer is different, then the convolution operation expression with nonlinear activation for each convolution layer can be expressed as follows:

$$f_i(X) = \sigma_i(W_i * X + B_i) \tag{1}$$

Where i represents the i-th convolution kernel among all the multi-scale convolution kernels of one convolution layer, X is the input image, W_i and B_i are the weights and bias of the i-th convolution kernel, respectively, σ_i is the nonlinear activation. In convolutional neural networks, the nonlinear activation function often used is Relu, therefore, the expression of σ_i here is:

$$\sigma(x) = \begin{cases} x, & if\ x > 0 \\ 0, & if\ x \leq 0 \end{cases} \tag{2}$$

Which can also be simplified as follows:

$$\sigma(x) = max(0, x) \tag{3}$$

Where x is the convolution value input into the nonlinear activation function.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

2.2. Multi-scale Feature Fusion

After multi-scale feature extraction of the input image data, n convolution feature maps are obtained for each convolution layer, here, n is also the number of different convolution kernels in the convolution layer. Then, the n convolution feature maps will be fused and processed, that is, multi-scale feature fusion. When the n multi-scale convolution feature maps are fused, the feature fusion method is designed as: the different feature images obtained by different scale convolution kernels are superimposed. The number of channels of the superimposed feature map is equal to the total number of channels of different convolution kernels in the process of multi-scale feature extraction. The schematic diagram of multiscale feature fusion is shown in Figure 2.

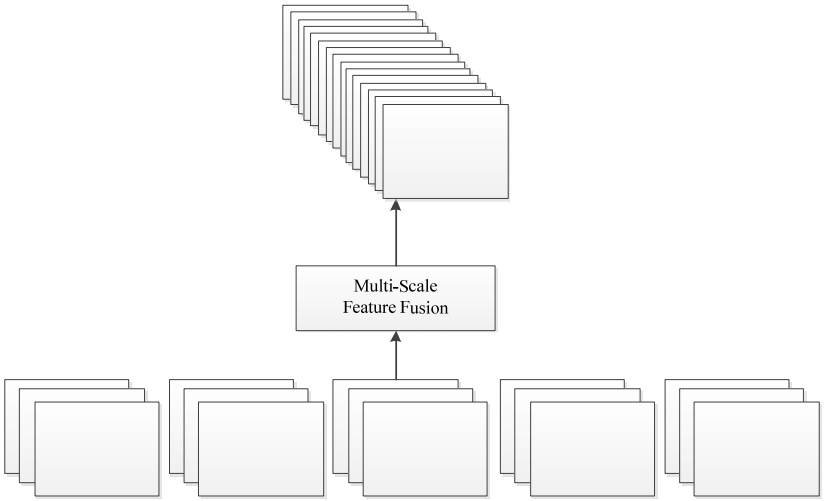

Figure 2. The schematic diagram of multiscale feature fusion

According to the principle shown in the above figure, for these n convolution feature maps obtained by convolution kernels of different scales, the expression of the above multi-scale feature fusion method can be expressed as follows:

$$f(X) = \sum_{i=1}^{n} f_i(X) = \sum_{i=1}^{n} \sigma_i(W_i * X + B_i) \quad (4)$$

Where i represents the i-th convolution kernel, namely, the i-th convolution feature maps.

2.3. Multi-scale Feature Optimization

In the process of multi-scale feature extraction for the input image, convolution kernels with different scales have a certain number of channels, and then they can extract rich features from the input image. However, when the convolution kernels of different scales have more channels and these convolution kernels are operated in parallel, the further feature fusion of multi-scale features will lead to more channels after fusion. These extracted features will also lead to the complexity of network scale and the improvement of computer complexity. Therefore, it is necessary to optimize the features and network. In convolutional neural networks, a common optimization method is pooling, however, pooling only reduces the resolution of the feature map itself, but does not change the number of channels of the feature map. Therefore, in order to reduce the number of channels of the feature map, other methods should be used. Here, one more effective method is to use 1×1 convolution kernel. In the specific design and implementation, the channel number of 1×1 convolution kernel is less than that of the fused feature, which can ensure that the channel number of feature maps after convolution operation

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

is relatively reduced, and the use of 1×1 convolution kernel can also ensure that all feature information in the fused feature map will not be lost. Thus, this method can effectively optimize the fused feature map. The optimized expression of feature map after multi-scale feature fusion can be expressed as follows:

$$F = \sigma(W * f(X) + B) \tag{5}$$

Where X is the input image, $f(X)$ is the output feature map after multi-scale feature fusion, W and B are the weight and bias for optimization, and σ is the Relu nonlinear activation function used after the convolution operation.

As can be seen from the above, the target feature learning model based on multi-scale feature fusion proposed in this section uses different convolution kernels to extract multi-scale features from the input image, and then the extracted multi-scale features are fused in parallel. After obtaining the multi-scale features of the image, the nonlinear mapping relationship between the input image and the output can also be obtained by using the nonlinear activation function. The optimization of multi-scale features can also improve the operation performance and time complexity of the whole network model, thus, it can be used to extract and express target features well. Meanwhile, this method can also be used for cascade design and improvement. The improved expression can be expressed as follows:

$$F^l = \sigma(W_{l_{n+1}} * \sum_{i=1}^{n} \sigma_{l_i}(W_{l_i} * F^{l-1} + B_{l_i}) + B_{l_{n+1}}) \tag{6}$$

The cascade learning method of target feature represented by this formula can extract the features of original image better, and thus can have better expression performance. However, the time performance of convolution calculation should be considered to find the most appropriate design model and related network parameters.

3. TARGET DETECTION METHOD BASED ON FEATURE FUSION

Based on the above learning method of the target features in the previous section, in this section, on the basis of the current classic convolution neural network, Faster-RCNN model [16], the road target features extracted by the proposed multi-scale road target feature learning method are applied to the Faster-RCNN model. By making full use of the efficient feature expression obtained by feature fusion, a more efficient target detection method and model is constructed to realize the road target detection more effectively.

3.1. Faster-RCNN Target Detection Model

Based on the original Fast-RCNN target detection model [17], Faster-RCNN target detection model[16] uses the region proposal network (RPN) to replace the selective search method[18] in Fast-RCNN, that is, RPN is used to generate anchors with the size of $\left\{(w \times h), (\alpha w \times \alpha h), (w\gamma, \frac{h}{\gamma})\right\}$ for each pixel in the image, where w and h represent the width and height of the initial anchor, respectively, α and γ represent the frame scaling ratio of the initial anchor, and there are $\alpha \in (0,1]$, $\gamma > 0$. It can be seen that when there are n anchors and m anchor scaling ratios, a total of $n \times m$ anchors will be generated. It judges each generated anchor, and then those belonging to the background are filtered out, and those belonging to the target are further processed by ROI pooling and border regression processing. The anchor diagram of Faster-RCNN target detection model is shown in Figure 3.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

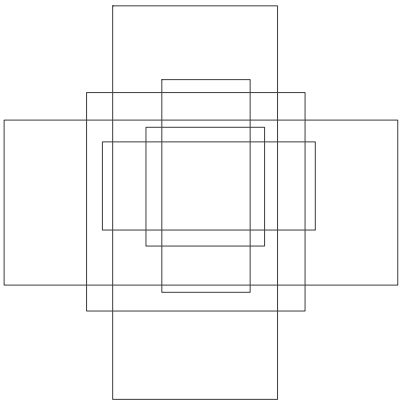

Figure 3. The anchor diagram of Faster-RCNN target detection model

3.2. Feature Fusion-based Target Detection

Based on the above Faster-RCNN target detection model, in the specific implementation of the designed target detection model, the output results of the target features obtained by the multi-scale target feature learning method are then applied to the RPN network of the Faster-RCNN model. The fused feature with better expression performance can obtain more accurate target detection performance. The proposed feature fusion-based target detection model is shown in Figure 4.

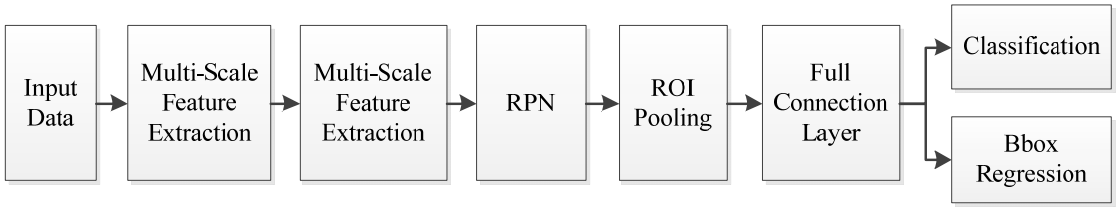

Figure 4. The target detection model based on feature fusion

4. EXPERIMENTAL SIMULATION AND ANALYSIS

In this paper, it uses the images provided by KITTI as dataset to verify the proposed method. KITTI has a relatively complete road traffic scene, and it is also one of the most frequently used road traffic dataset in the field of computer vision. The dataset contains real-life images of urban areas, roads, villages and so on, and each image contains many complex road targets such as different types of vehicles and pedestrians with different shapes. In the specific design of the experiment, the representative images are selected to form the training set and test set. Using the method proposed in this paper, it inputs the fused features into the Faster-RCNN network, and then analyzes and verifies the target detection performance of the method. On this basis, the Faster-RCNN method and the proposed method are evaluated respectively. Table 1 shows the statistical results of target detection performance under the two methods. In this paper, the mean average precision(mAP) and average detection time are used to evaluate the performance.

Table 1. Statistical results of road target detection performance under the two methods

Method	mAP(vehicles) (%)	mAP(pedestrians) (%)	average detection time (s)
Faster-RCNN	65.56	63.12	3.36
The proposed method	65.62	63.28	5.82

It can be seen from the table that when using the Faster-RCNN method, the mAP of vehicles and pedestrians in road traffic images is 65.56% and 63.12%, respectively. Compared with the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

original Faster-RCNN method, the proposed method has slightly improved the mAP of both the targets of vehicles and pedestrians, which shows that the proposed method can extract the expression features of road traffic targets more effectively, and obtain relatively better target detection accuracy through the target fused features with rich expression ability. However, in the process of target extraction and detection, the average detection time of the proposed method is significantly longer than that of Faster-RCNN method, which also reflects that the target detection method based on feature fusion proposed in this paper also takes some time in the specific implementation process, and this time consumption problem also increases the calculation cost of network model. This time consumption problem also needs to be optimized and improved in the proposed method, and it is also the next research direction of the research group.

5. CONCLUSION

Due to the particularity and complexity of the target scale distribution in road traffic scene, the traditional manual feature target detection method and the classical deep learning target detection method have certain difficulties in the actual road traffic scene target detection application. Therefore, road target detection in complex environment has become an urgent research content. Aiming at this problem, in this paper, it takes the current advanced convolution neural network technology as the background, takes the feature fusion as the main research object, and on the basis of analyzing and discussing the feature learning method based on the feature fusion, further proposes the road target detection method based on the feature fusion, and through the construction of the model framework, the road target detection technology is deeply studied. Thus, it provides some ideas for the further study of related theories in the field of road target detection, and also provides technical support for safe road traffic environment.

ACKNOWLEDGEMENTS

This work is supported by Ministry of Public Security Technology Research Project of China (2019JSYJC25), Fundamental Research Funds for the Central Universities of China (2019TJJBKY012, 2020TJJBKY001), Henan Province Educational Commission Key Scientific Research Project of China (21B580005).

REFERENCES

[1] Prakash I, Neves O, Cumbe E, et al. The Financial Burden of Road Traffic Injuries in Mozambique: A Hospital-Related Cost-of-Illness Study of Maputo Central Hospital, World Journal of Surgery, 2019, 43(12):2959-2966.

[2] Garg K, Ramakrishnan N, Prakash A, et al. Rapid and Robust Background Modeling Technique for Low-Cost Road Traffic Surveillance Systems, IEEE Transactions on Intelligent Transportation Systems, 2019, PP(99):1-12.

[3] Zhao J W, Hua J, Liu Y T, et al. Research on Road Traffic Accident Information Collection Technology and Its Application, Journal of Chongqing University of Technology, 2019, 33(07):28-36.

[4] Nogal M, Honfi D. Assessment of road traffic resilience assuming stochastic user behavior, Reliability Engineering & System Safety, 2019, 185(5):72-83.

[5] Zhu Q. Research on Road Traffic Situation Awareness System Based on Image Big Data, IEEE Intelligent Systems, 2020, 35(1):18-26.

[6] Song J G, WU Y. Improved Road Target Detection Algorithm Based on YOLOv2 Model. Software Guide, 2019, 18(12):126-129.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202101_7(1).0025

[7] LI H G, LU C Y, QI L. Road Target Detection Based on Otsu Multi-Threshold Segmentation, Mechanical Engineering and Control Systems, 2016: 265-269.

[8] Zhang H Y, Qin H B. FMCW Radar Moving Road Target Detection System with GPRS Communication, Technology of IoT & AI, 2018, 1(01):31-34+44.

[9] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection, Computer Vision and Pattern Recognition, 2005 CVPR, San Diego, CA, USA. IEEE:886-993.

[10] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features, Computer Vision and Pattern Recognition, 2001 CVPR, Kauai, HI, USA, IEEE:1-11.

[11] Teoh S S, Braunl T. Symmetry-based monocular vehicle detection system, Machine Vision and Application, 2012,23(5):831-842.

[12] Sivaraman S, Trivedi M M. Active learning for on-road vehicle detection: a comparative studay, Springer-Verlag New York, Inc., 2014.

[13] Cheon M, Lee W, Yoon C, et al. Vision-Based Vehicle Detection System With Consideration of the Detection Location, IEEE Transaction on Intelligent Transportation System, 2012, 13(3):1243-1252.

[14] Sivaraman S, Trivedi M M. A General Active-Learning Framework for On-Road Vehicle Recognition and Tracking, IEEE Transactions on Intelligent Transportation System, 2016, 11(2):267-276.

[15] Lowe D G. Object Recognition from Local Scale-Invariant Feature, Computer Vision, 2002 The Proceedings of the Seventh IEEE International Conference, Kerkyra, Greece. IEEE,2:1150-1157.

[16] Ren S, He K, Girshick R & Sun J. Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017, 39(6): 1137-1149.

[17] Girshick R. Fast R-CNN, Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015:1440-1448.

[18] Sandekea V D, Uijlings J R, Gevers T, et al. Segmentation as selective search for object recognition, Proc of International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2011:1879-1886.