Spatial and Channel-wise Attention in Multi-modality Feature Fusion for Multispectral Pedestrian Pedestrian Detection

Hao Wei , Jifeng Shen

School. of Electronic and Informatics Engineering , Jiangsu University

优胜奖

Abstract

Multispectral pedestrian detection with feature fusion of thermal and visible images has achieved great success in all-day scenarios, but few attentions have been paid to the quality of feature fusion. In this paper, we have proposed an improved attention aware dual-stream Faster R-CNN algorithm to ameliorate the feature quality. Quantitative analysis of three different feature fusion methods demonstrate that the performance of detector is highly related to the quality of feature fusion. Therefore, an improved multi-step spatial and channel-wise attention method is proposed to improve the feature fusion quality step by step based on the dual-stream Faster R-CNN framework. With the aid of attention mechanism, the model learning is optimized in a more smooth way, which can effectively enhance the response of fused feature map in each stage. Experiments based on the KAIST and CVC-14 dataset demonstrate that the proposed method has achieved better performance compare to the baseline method with a decrease of 4% and 8% in AMR respectively.

Proposed Method

1. Architecture

The dual branch backbone network is used to extract the thermal and visible features, and then the two features are fused, multiplied by the attention weight, and finally sent to the detection network to get the final result.The specific structure is shown in Figure 1.

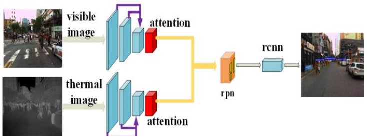


Fig.1 Architecture of our network

2. Implementation process

Firstly,we choose dual- stream VGG16 as the feature extraction layer, one branch extracts visible images features, the other branch extracts thermal image features. Secondly,we add a feature fusion layer to fuse visible and thermal features, thirdly, we add an attention mechanism layer , which is to improve the quality of features and filter useless features. Forthly, we cascade the RPN layer behind the feature layer to generate the region proposal. Finally, we add a fully-connected layer after RPN for final classification and positioning regression.

3. Fusion methods

We believe that different feature fusion methods schemes will affect the accuracy of the final detector. Because the expression ability of each layer's feature map is different.So we have designed three feature fusion schemes according to the different fusion positions of thermal and visible images feature, which are early-fusion, half- fusion, late-fusion.The experiment results show that the performance of the late fusion model is the best.The specific structure is shown in Figure 2.

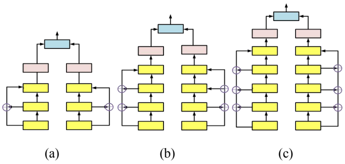


Fig.2 (a) Early fusion. (b) Half fusion. (c) Late fusion

4. Channel attention

The feature map after convolution contains a lot of redundant information while saving useful information. The purpose of channel attention is to filter useless information from the perspective of the channel. Specific implementation steps: first compress the feature map in the spatial dimension to obtain a one-dimensional vector. Our method uses max pooling, global average pooling, and full connection as compression methods for the feature maps on each channel. If only use max pooling , some feature-related information will be lost, so global average pooling is used to make up for this shortcoming. However, the above two pooling will slow down the convergence speed, so the fully connected layer is used as the third compression method. Next, the final attention weight will be obtained through the NIN layer, and NIN is to add the one-dimensional vectors obtained by the three compression methods element by element, then use 1 by 1 convolution to reduce to 16 times the original size.Finally ,the feature map is restored to the original size through the fully connected layer.The specific structure is shown in Figure 3.

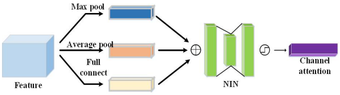


Fig. 3. The overview of channel attention module

6. Spatial attention

Similar to the channel attention, the network also needs to have a capability to understand that which parts of the feature map should have a higher response at the spatial level. Introduction to each part : firstly, we also use average pooling to make a spatial dimension compression on the input feature map. Then obtained feature map only has one channel; secondly, we make a scale transformation on the feature map by 3× 3 convolution and make it reduce to one-forth of its original size. Thirdly, we make it enlarge back original size by 3× 3 convolution, so that it can better fit the complex correlation in spatial.Finally, the channel and spatial attention module are connected in series to obtain the final hybrid-domain attention module.

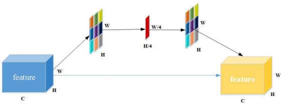


Fig. 4. The overview of spatial attention + module

Experiment

1.Dataset

KAIST：It is the multispectral pedestrian detection dataset which contains 7601 color-thermal image pairs for training and 2252pairs for testing.

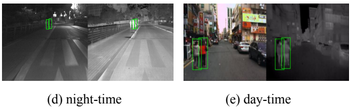
CVC-14: It's also multispectral pedestrian detection dataset.But we only use this dataset to test the generalization ability of the model.

2.Experiment result

TABLE Ⅰ. MR of Early-fusion, Half-fusion and Late-fusion

Methods	MR(%)
ACF+C+T	56.07
Fusion RPN+BDT	32.89
IATDNN+IAMSS	31.68
Halfway Fusion	23.78
Ours	19.44

Comparison of other methods.The method proposed in this paper is based on the improvement of Fast R-CNN, using RPN pre-defined anchor algorithm, fully connection network to fine-tune the region proposals generated by RPN.Finally,compared with ACF + C + T detector, the accuracy is improved by 37 percentage points.Our method combines the attention mechanism, which makes the network focus on the useful features and suppress the redundant features, so that the features extracted from the network can be more expressive. The specific results is shown in Figure 5.



(d) night-time (e) day-time

Fig. 5. Detection result of our method on the KAIST dataset

Conclusion

This paper focuses on the optimization method of multispectral pedestrian detection. The core algorithm is based on the improvement of Faster R-CNN. There are two specific improvements: first is features are obtained through by multi-scale information fusion; second is improved attention mechanism that we proposed is added to the multispectral pedestrian detection network. Finally, we also designed three feature fusion methods, namely early-fusion, half -fusion, and late-fusion. The final experiment proves that the late-fusion model has the highest accuracy. However, the detection speed of this detector model is not fast and it is still difficult to meet the requirements in actual scenarios. Therefore, the multispectral algorithm pedestrian detection needs to be further studied.

Reference

[1]Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1037–1045 2015

[2]Guan, Y. Cao, J. Yang, Y. Cao, and C.-L. Tisse. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. Applied optics, 57(18):pp108–116, 2018. 3, 4

[3]Jingjing, Z. Shaoting, W. Shu, and M. Dimitris. Multispectral deep neural networks for pedestrian detection. In British Machine ision Conference, pp 73.1–73.13, 2016. 1, 3

Contact us

Contact person：Hao Wei
Mobile phone：17826078206
E-mail：1594896411@qq.com