Feature Extraction with Apparent to Semantic Channels for Object Detection

Lei Zhao*, Jia Su, Zhiping Shi, Yong Guan
Information Engineering College, Capital Normal University, Beijing, China.

* Corresponding author. Tel.: +86 17501135266; email: zhaolei@cnu.edu.cn
Manuscript submitted September 16, 2020; accepted December 13, 2020.
doi: 10.17706/jsw.16.4.157-166

Abstract: This paper focuses on using traditional image processing algorithms with some apparent-to-semantic features to improve the detection accuracy. Based on the optimization of Faster R-CNN algorithm, a mainstream framework in current object detection scenario, the multi-channel features are achieved by combining traditional image semantic feature algorithms (like Integral Channel Feature (ICF), Histograms of Gradient (HOG), Local Binary Pattern (LBF), etc.) and advanced semantic feature algorithms (like segmentation, heatmap, etc.). In order to realize the joint training of the original image and the above feature extraction algorithms, a unique network for increasing the accuracy of object detection and minimizing system weight called Multi-Channel Feature Network (MCFN) is proposed. The function of MCFN is to provide a multi-channel interface, which is not limited to the RGB component of a single picture, nor to the number of input channels. The experimental result shows the relationship between the number of additional channels, performance of model and accuracy. Compared with the basic Faster R-CNN structure, this result is based on the case of two additional channels. And the universal Mean Average Precision (mAP) can be improved by 2%-3%. When the number of extra channels is increased, the accuracy will not increase linearly. In fact, system performance starts to fluctuate in a range after the number of additional channels reaches six.

Key words: Feature, channels, faster R-CNN, semantic.

1. Introduction

In the field of target detection, Faster R-CNN [1] algorithm has excellent operation accuracy because of the low coupling degree of its internal modules, so it can optimize the structure of various application scenarios. This is also different from YOLO series [2] and SSD series [3], because the latter pay more attention to model integration and lightweight.

Since the convolution neural network (CNN [4], [5]) has been widely used in 2015, Faster R-CNN is combined with excellent convolutional networks such as residual network [6], and the introduction of Region Proposal Network (RPN) Local modules make the whole system be divided into three layers. The bottom layer is the feature extraction stage. The neural network algorithms used include ResNet 101, VGG-16 and so on. The second layer is the Region Proposal Network (RPN), This is also the most obvious difference between Fast R-CNN [7] and R-CNN [8]. The top layer roughly retains the framework of the previous generation, including the bounding box regression and the softmax layer. The Faster R-CNN algorithm still needs the optimization of object detection because the average accuracy of some well-known datasets (such as VOC 0712) are still less than 80% based on the analysis of experimental data. Therefore, this paper is dedicated

to algorithm optimization from the change of the structure of the input end of the algorithm. This structural optimization is based on the input of different image channels, and different channels can provide different semantic information, so that the bounding box can be framed more accurately in the RPN structure.

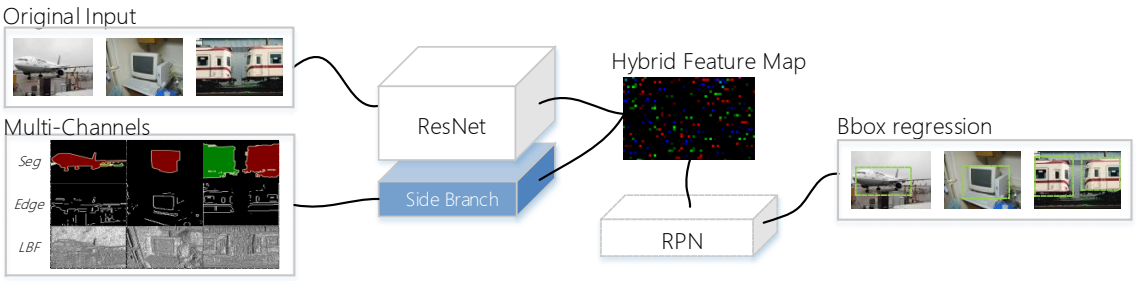


Fig. 1. The detection process of this paper.

Image processing algorithms are gradually being replaced by neural networks in the field of object detection, but this does not mean that traditional algorithms have been abandoned, whether the traditional algorithm can be combined with neural network remains to be explored. Fig. 1 shows a brief process of Faster R-CNN with multi-channel. The multi-channels and side branch in this figure are the optimization goals of this article.

2. Proposal

2.1. Details of Apparent-to-Semantic Features

The extra image channels in Fig. 1 all belong to apparent-to-semantic features. For example, semantic segmentation maps, heatmap, HOG, ICF, etc. Apparent semantic features mean defining object categories from different perspectives. For example, semantic segmentation considers which category belongs to in the pixel level, while textures and edges tend to identify the detailed shape of each object. But we need to solve one problem that which kind of feature is effective and how it actually works to improve the R-CNN-based detectors. To answer this question, this paper explains from the following three aspects:

- Firstly, we integrate extra features as input channels into CNN-based detectors. To investigate apparent-to-semantic channels, extensive experiments are carried out on two datasets (PASCAL VOC 0712 [9] and COCO [10]), and inspired by the excellent work of J. Mao *et al*. [11], the feature extraction network is ResNet 101. Besides, the average recognition accuracy of benchmark is 77.61% [12] in this paper.
- Then, we experimentally analyzed both advantages and disadvantages of different channels of feature. Specifically, we quantify the improvement brought by different channel features and provide insight into the error sources. In Fig. 2, this figure shows the performance of additional channels in the KITTI dataset [13]. From the results, the accuracy of segmentation [8], [14], heat map [15], [16] and edge [17] is better than benchmark. Relatively speaking, the positive contribution rate of ICF [18], [19], HOG [20] and LBF [21] to the benchmark is less than 1%, and even the negative contribution rate appears. This is because of the low correlation, which cannot further improve the pixel level features.
- Moreover, a new framework called Multi Feature Channel Network (MFCN) is proposed to integrate two of extra channels. In the MFCN structure, extra channels will provide a real-time monitoring function that delimiting bounding box of Region of Interest (ROI).

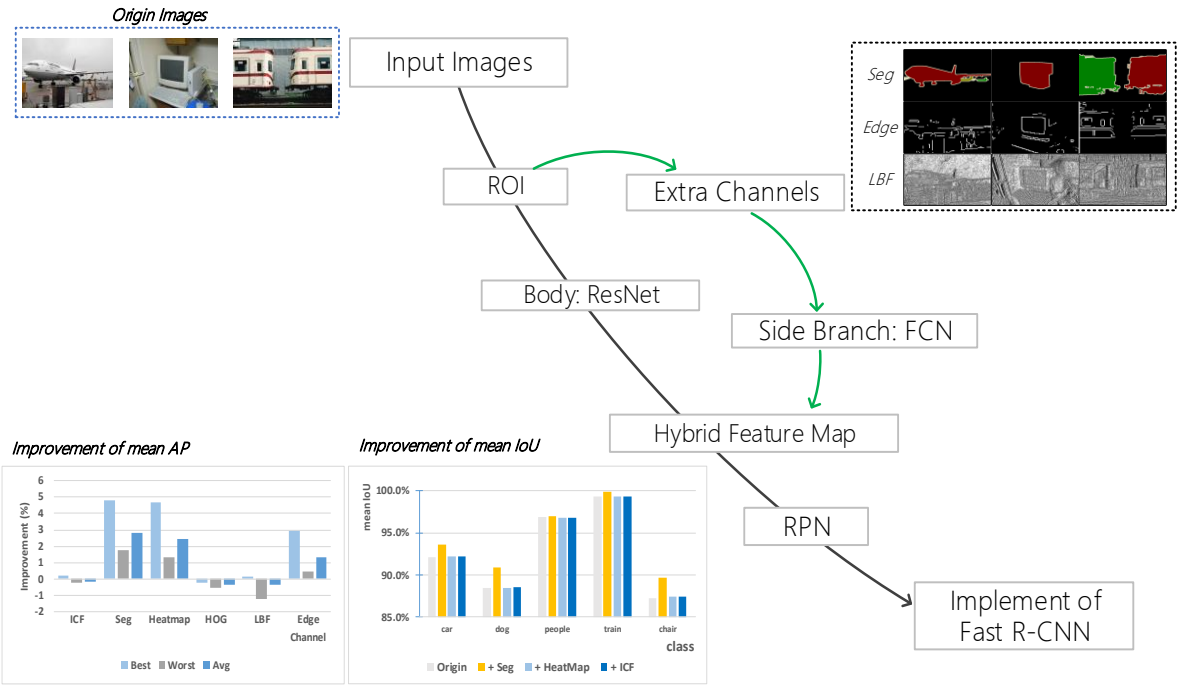


Fig. 2. Improvement with single extra channel.

2.2. Multi-channel Joint Learning

This paper optimizes based on Faster R-CNN and residual network proposed by Kaiming He, *et al*. Faster R-CNN is generally considered to have three layers of substructures.

In the substructure system of Faster R-CNN, what this paper considers is to design a multi-channel structure of the feature extraction stage. As shown in Fig. 3. This structure still retains the original picture input channel, and uses 1×1 and 3×3 convolution kernels for convolution (conv_1 to conv_4). The convolution layer designed for additional channels is a fully convolution network. Multi-channel images are pre-processed before the input, and images from multiple channels of the same image share the same information of location and category table.

Different image processing algorithm focuses on different point, such as texture and segmentation. Texture focuses more on the content of the category, and segmentation focuses on the interpretation of boundaries between different categories. This is also the problem that this paper needs to solve that through a large number of comparative experiments, which combination method is used to improve the accuracy of target detection, in other words, which kind of image processing methods can better be used with target detection.

Integrating channel features in the network can boost our detector working on images of both low resolution and high resolution. With these channel capabilities, we can close most of the gap between resolutions without introducing the need to enlarge the input image and drive the heavy computational costs of the latest technology. Compared with the basic Faster R-CNN, the computational cost of the brute force integration method is high. Therefore, we propose a multi-channel feature extraction algorithm. The framework is illustrated in Fig. 3. As shown, our system consists of four components: the body network of origin images, the multi-channel feature network (MCFN) of extra channels, the regional proposal network (RPN), and the fast R-CNN network of final detection tasks.

Body Network As shown in Figure 3, the body network on the left side means the original part that this paper referred from Faster R-CNN. The body network takes the raw image of shape $3 \times H \times W$, as its input, and outputs several activation maps. In this paper, the body network is a ResNet 101 network initialized with

the pre-trained weights. ResNet 101 contains five layers of convolution. The results obtained by the image through conv 1, conv 2_x, conv 3_x, and conv 4_x will become the input of RPN and ROI pooling, respectively, and the generated results will be passed to the next structure through conv 5_x and Average pooling.

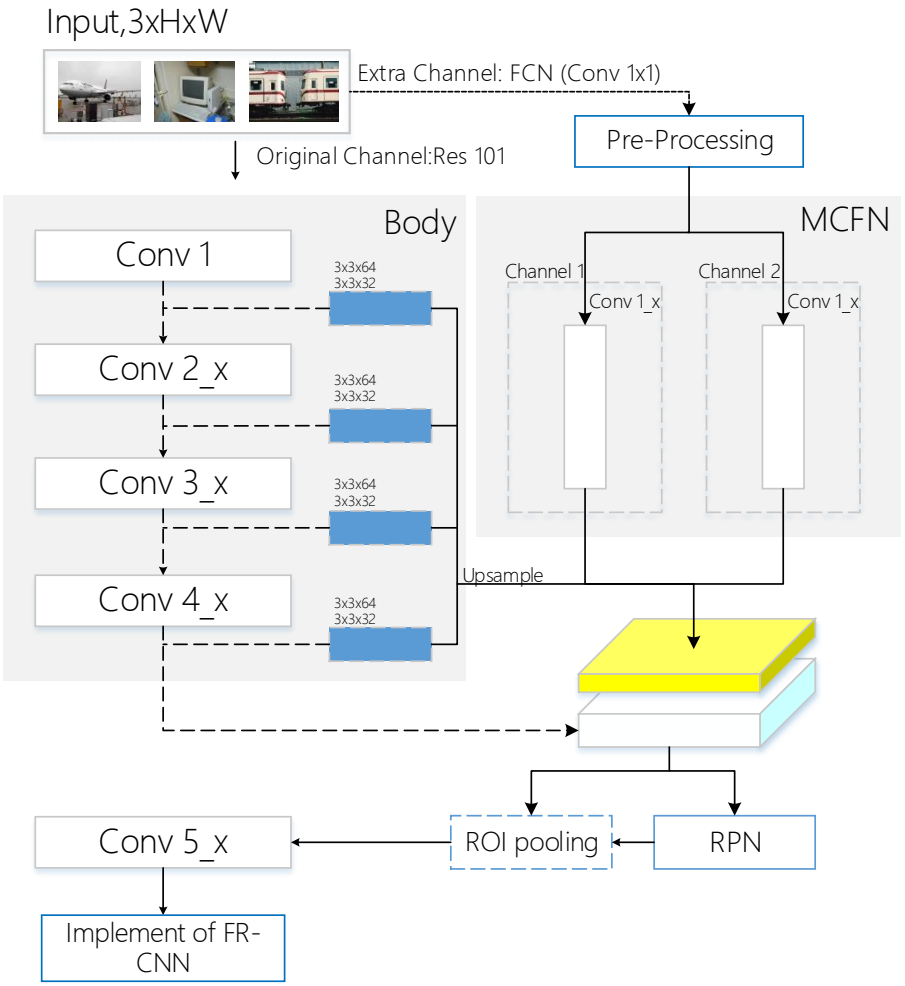


Fig. 3. Framework of this multi-channel network.

Multi-channel feature network (MCFN) The MCFN directly takes the aggregated activation maps to generate the predicted channel feature maps through fully convolution structure. The feature map obtained from the original image is summarized into MCFN after hierarchical up-sampling. MCFN is composed of multiple independent full convolution networks. The output after conv 4_x will be passed to the RPN and Fast R-CNN together with the output of CFN. The hybrid feature map acts as a significant role in the whole framework, which must be carefully trained. MCFN needs to be trained separately, considering model performance, single-threaded operations are required. Secondly, the body network (conv 1_x to conv 4_x) and MCFN are jointly trained to generate the hybrid feature map. Then, Fast R-CNN is trained separately with the same residual network. conv 5_x and the average pooling layer are added to perform bounding box regression.

Implement of RPN and Fast R-CNN As shown in Fig. 3, below the body network and MCFN, the implement of RPN and Fast R-CNN also refers the origin part from Faster R-CNN. We use the same structure for RPN and Fast R-CNN as proposed in [1]. RPN and Fast R-CNN now take both convolution activation map by MCFN in the ResNet network and the feature map from the body network as the inputs. The proposals generated by RPN are then fed into Fast R-CNN to perform final detection. In this process, faster r-cnn does not have the process of back propagation, and the process of updating the weight is estimated according to the relative

position of the bounding boxes and the optimization of the loss function.

In summary, our changes to MCFN are based on the structural optimization of the algorithm of Faster R-CNN, which is also the main idea that proposed in this article. And this article will conduct a comparative analysis through experimental data, show the results after adding different additional channels, and give an analysis of its reasons.

3. Experiment and Results

3.1. Single Extra Channel of Feature Extraction

The work of J. Mao *et al*. [11] is targeted at the application scenario of the pedestrian dataset KITTI, and an additional single-channel optimization experiment is designed on this basis. In this experiment, an additional single-channel experimental design is first performed on a dataset in a more general scenario. The purpose of this is to set a reference value and to obtain a preliminary expected range for different algorithms through experimental results.

As show in Table 1 and Table 2. This idea is confirmed in the previous pedestrian detection, so this need to be carried out in a general target detection scenario. Synchronous experiments are a prerequisite for subsequent experiments. The reason is that the structure of the algorithm for pedestrians has not changed structurally after the abstraction of specific pedestrian detection scenes to general target detection. The recognition categories are more abundant, and the recognition image data set has also changed. The two tables show together that it is effective and feasible to add an additional channel without changing the overall experimental structure. By using two different datasets, the obvious gained results could be detected easily.

The reason for their effectiveness is that the additional channels can provide enough boundary information or semantic information through layer-by-layer upsampling, which can more accurately frame the starting point during RPN and frame regression and the regression function will also improve the results.

Table 1. Experiment Result with Single Extra Channel in PASCAL VOC 0712 Dataset

Models	PASCAL VOC 0712 *			Improvement		
	Best	Worst	Avg	Best	Worst	Avg
Fr R-CNN	79.75	76.44	77.61	-	-	-
+ ICF	79.05	75.20	76.62	-0.67	-1.24	-0.99
+ Seg	83.22	78.62	80.59	+3.47	+2.18	+2.98
+ Heatmap	82.80	77.91	80.23	+3.05	+1.47	+2.62
+ LBP	80.01	75.71	77.49	+0.26	-0.73	-0.12
+ Edge	81.19	76.25	77.88	+1.44	-0.19	+0.27

In the two selected datasets, the experimental results of adding segmentation and heatmap channels outperform than other algorithms. It is also worth noting that not all channels will necessarily improve, such as adding LBP and ICF channels. This is because the loss function is related to the frame regression of the obtained feature map. For example, LBP pays more attention to the texture results of the image, but the design of the loss function pays more attention to the boundary, which results in no actual improvement in the results.

Table 2. Experiment Result with Single Extra Channel in COCO Dataset

Models	COCO *			Improvement		
	Best	Worst	Avg	Best	Worst	Avg
Fr R-CNN	37.00	29.20	34.10	-	-	-
+ ICF	36.09	28.20	33.14	-0.91	-1.00	-0.96

+ Seg	39.23	31.15	36.19	+2.23	+1.95	+2.09
+ Heatmap	38.56	30.53	35.55	+1.56	+1.33	+1.45
+ LBP	36.45	28.36	33.40	-0.55	-0.84	-0.70
+ Edge	37.78	29.53	34.66	+0.78	+0.33	+0.56

3.2. Multi-extra-Channel of Feature Extraction

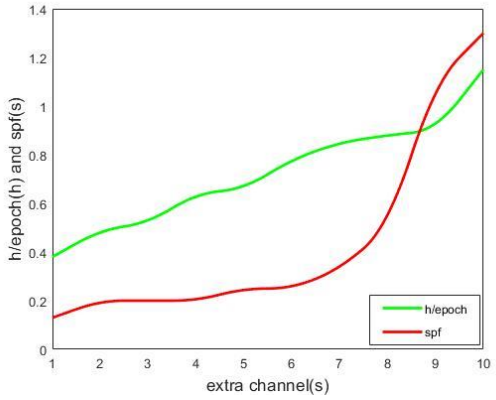
Designing a dual-channel algorithm on the basis of an additional single-channel experiment is the most important step. Here it involves the choice of the algorithm combination and the priority. As show in Table 3. This table explains the feasibility of the network structure optimization proposed in section of proposal. After comparing the experimental results of several sets of additional dual channels, it can be concluded that under the same experimental environment, the results can be improved by increasing the number of channels and selecting a better performing algorithm. From the comparison of (0, 1, 2, 3), we can see the different results of different combinations of segmentation and other algorithms. The comparison of (3, 4) is for experiments on feature selection algorithms. (1, 5) and (5, 6) also compare the experimental results of adding channels on the basis of ICF and Edge.

Table 3. Improvement of Double Channels in two Datasets

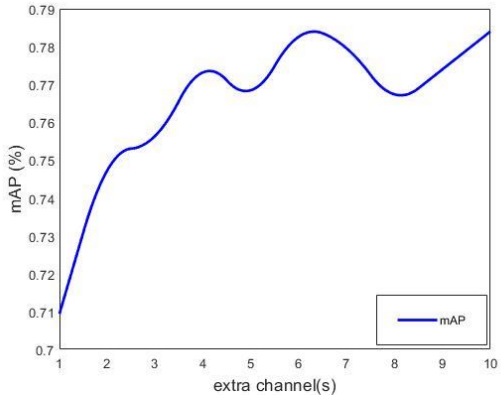
No.	Combinations	PASCAL VOC 0712			Improvement		
		Best	Worst	Avg	Best	Worst	Avg
0	Benchmark	79.75	76.44	77.61	-	-	-
1	Seg + ICF	83.06	78.28	80.19	+3.31	+1.84	+2.58
2	Seg + Edge	82.69	79.00	80.36	+2.94	+2.56	+2.75
3	Seg + Seg	83.62	80.04	81.35	+3.87	+3.60	+3.74
4	Seg + Heatmap	83.25	79.78	81.03	+3.50	+3.34	+3.42
5	Edge + ICF	81.07	75.88	77.99	+1.32	-0.56	+0.38
6	Edge + Heatmap	82.26	77.50	79.40	+2.51	+1.06	+1.79

Different algorithm combinations are not randomly selected. What needs to be considered is whether algorithm A and algorithm B can provide image semantic information in the target detection scenario. If it is just an RGB-level matrix, then it will only increase the depth of the convolution rather than provide accurate position information. In addition, can the two algorithms operate in parallel? Because the image information provided by the two algorithms is aggregated after upsampling by the network to form more layers of input features. Such feature maps are layer-by-layer and the matrix operations may focus on the same area at the same time, but get different classification results. This also leads to errors.

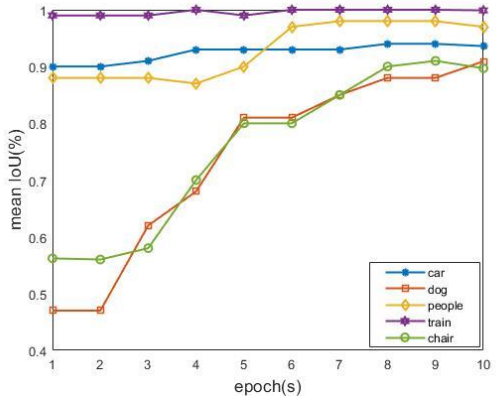
In addition, when the number of channels increases, the relationship between recognition accuracy and algorithm performance is also worth considering. The curve in Fig. 4. (a) represents the relationship between second per figure (spf), hours per epoch (h/epoch), and the number of channels after the number of channels has been increased to 10. Fig. 4. (b) shows the relationship between the number of channels and accuracy. From this experimental result, it can be seen that when the number of channels is continuously increased, the burden of model training will be increased, and the recognition rate will be reduced. The recognition accuracy obtained in this way is not worth doing.



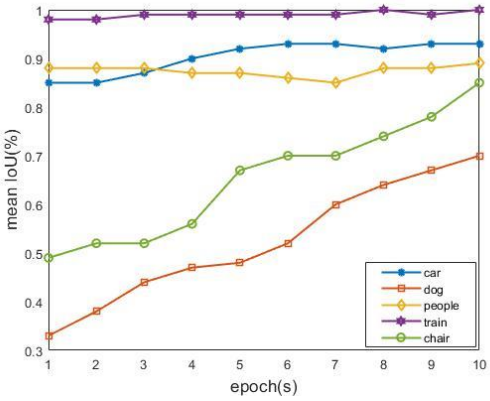
(a) h/epoch and spf



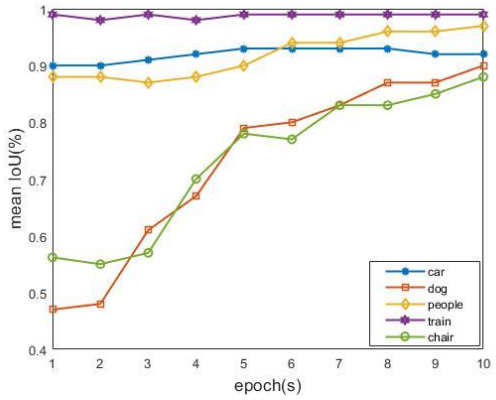
(b) mAP



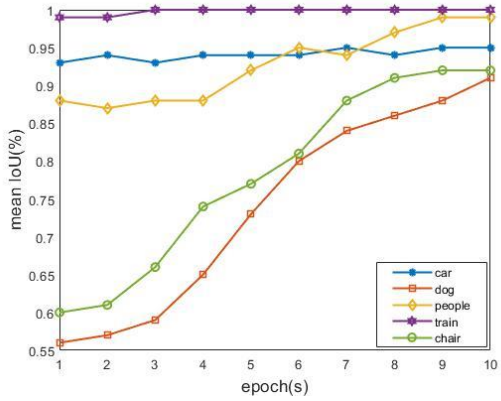
(c) $mIoU_1$: Seg



(d) $mIoU_2$: Heatmap



(e) $mIoU_3$: Seg + Heatmap



(f) $mIoU_4$: Seg + Seg

Fig. 4. Curves of some parameters.

The curves from Fig. 4. (c) to Fig. 4. (f) show the change of the mIoU value after adding different additional channel numbers. The value of mIoU generally reflects the gap of semantic information such as segmentation and heat map, and can intuitively reflect the contribution to the accuracy of object recognition in different situations. In (c), a single segmentation channel is added to the experiment, in (d), a single channel of heat map is added to the experiment, while in (e), two channels are added, segmentation and heat map. In (f), two channels are also added, both of which are segmentation graphs. The following is an overall evaluation of the mIoU experiment results. From the comparison between (c) and (d), The influence gap between single channels on mIoU is more obvious, and the segmentation map will have obvious advantages, because the segmentation map will generate stronger semantic information, and the boundary information included in the regression operation will be more. It can be seen from the two groups, {c, e} and {d, e}, on the basis of

single channel, it is worth trying to combine two single channels for multi-channel calculation. This includes the fusion calculation of multiple information and the redundancy, complementarity and rejection of calculation parameters. As can be seen from (e) and (f), the comparison experiment between different combinations of multiple channels is also an important work, which shows that it is necessary to choose some of these combinations that have more advantages and can explain the direction of the reason.

4. Conclusion

This paper combines traditional apparent algorithms and advanced semantic algorithms to implement the framework optimization of target detection algorithms based on Faster R-CNN.

The experimental results are discussed in two aspects. The first aspect proves that without changing the internal structure of the network, only changing the mode from the input can improve the detection accuracy. There will be a difference in improvement when using different image processing algorithms. For example, segmentation and heatmap are 3% better than ICF and LBP. This also reflects the advantages of semantic information in feature extraction.

The second aspect shows that by implementing the form of multi-channel additional side branches, the images obtained by these algorithms and the original image are taken as inputs from different dimensions, and the final hybrid feature map obtained by parameter screening has certain optimization effects. For example, the channels combining segmentation maps and heat maps have an average recognition accuracy improvement of 3.74% compared to traditional Faster R-CNN. Meanwhile, when the number of experimental channels increases, it does not necessarily result in a significant increase of accuracy. In contrast, the performance of the model will increase dramatically. This also reminds us that the algorithm performance is also a point to consider optimization.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Lei Zhao, Jia Su conducted the research; Lei Zhao, Zhiping Shi analyzed the data; Lei Zhao, Jia Su and Yong Guan wrote the paper; all authors had approved the final version.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61702348, 61772351, 61602326, 61602324, 61572076), National Key R \& D Plan (2017YFB1303000, 2017YFB1302800), the Project of the Beijing Municipal Science \& Technology Commission(LJ201607), Capacity Building for Sci-Tech Innovation-Fundamental Scientific Research Funds (025185305000), Youth Innovative Research Team of Capital Normal University, Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five–year Plan, Shanghai Key Lab of Digital Media Processing and Transmission (STCSM 18DZ2270700).

References

[1] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

[2] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real- time object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Sharma, R., Biookaghazadeh, S., Li, B., & Zhao, M. (2018). Are existing knowledge transfer techniques effective for deep learning with edge devices. *IEEE International Conference on Edge Computing (EDGE)*.

[4] Erhan, D., Szegedy, C., Toshev, A., & Anguelov, D. (2014).1 Scalable object detection using deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2155-2162).

[5] C. S. *et al*. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1-9).

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770- 778).

[7] Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1440-1448).

[8] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for ac- curate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1476-1481).

[9] Vicente, S., Carreira, J., Agapito, L., & Batista, J. (2014). Reconstructing pascal voc. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 41-48).

[10] Pont-Tuset, J., & Gool, L. V. (2015). Boosting object proposals: From pascal to coco. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1546-1554).

[11] Mao, J., Xiao, T., Jiang, Y., & Cao, Z. (2017). What can help pedestrian detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6034-6043).

[12] Wu, N., & Haruyama, S. (2019). Real-time sound detection and regeneration based on optical ow algorithm of laser speckle images. *Proceedings of the 28th Wireless and Optical Communications Conference (WOCC)* (pp. 1-4).

[13] Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*.

[14] Sande, K. E. A. V. D., Uijlings, J. R. R., Gevers, T., & Smeulders, A. W. M. (2011). Segmentation as selective search for object recognition. *Proceedings of the International Conference on Computer Vision (ICCV)* (pp. 1879-1886).

[15] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (pp. 834-848).

[16] Costea, A. D., & Nedevschi, S. (2016). Semantic channels for fast pedestrian detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 2360-2368).

[17] Sharma, R., Biookaghazadeh, S., Li, B., & Zhao, M. (2018). Are existing knowledge transfer techniques effective for deep learning with edge devices. *Proceedings of the IEEE International Conference on Edge Computing (EDGE)* (pp. 42-49).

[18] Zhang, H., & Zhao, L. (2013). Integral channel features for particle filter based object tracking. *Proceedings of the 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (ICISC)* (pp. 190-193).

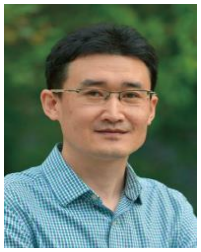
[19] Gong, L., Hong, W., & Wang, J. (2018). Pedestrian detection algorithm based on integral channel features. *Proceedings of the Chinese Control and Decision Conference (CCDC)* (pp. 941-946).

[20] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 886-893).

[21] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3431-3440).

Lei Zhao was born in Taizhou city, Jiangsu province, China in 1997. He received the B.S. degree in computer science and information engineering from Jiangsu University, Jiangsu, China, in 2014. He is currently pursuing the M.S. degree in information engineering at Capital Normal University, Beijing, China.

Jia Su received the B.E. degree in telecommunications engineering School of Xidian University, China, in 2006; and received her M.E. degree both in Graduate School of Information, Production and Systems, Waseda University and School of Microelectronics in Xidian University in 2008 and 2009, respectively. She received the Ph.D. degree in electronic information from the Waseda University, in 2012. She became a member (M) of IEEE in 2019. She is currently a professor with Capital Normal University, Beijing, China. Her research interests include video compression and computer vision.



Zhiping Shi received the B.Sc. and M.Sc. degrees in electronic engineering from Zhejiang University, China, in 1985 and 1988 respectively, and the Ph.D. in computer engineering from Nanyang Technological University, Singapore. Currently, he is an assistant professor with Nanyang Technological University, Singapore.

Yong Guan received the Ph.D. degree in computer science from the China University of Mining and Technology, Beijing, China, in 2004. He is currently a professor with Capital Normal University, Beijing. His current research interests include formal verification, PHM for power, and embedded system design. He is a member of the Chinese Institute of Electronics Embedded Expert Committee and the Beijing Institute of Electronics Professional Education Committee, and Standing Council Member of the Beijing Society for Information Technology in Agriculture.