EasyChair Preprint

№ 8897

Lightweight Fusion Channel Attention Convolutional Neural Network for Helmet Recognition

Chang Xu, Jinyu Tian and Zhiqiang Zeng

October 3, 2022

Lightweight fusion channel attention convolutional neural network for helmet recognition

Chang Xu,Jinyu Tian,Zhiqiang Zeng*

Department of Intelligent manufacturing, WuYi University, JiangMen, 529030, China

**E-mail: zhiqiang.zeng@outlook.com*

Recently, in the context of complex production and construction environments, the detection of unsafe behavior becomes more and more necessary to ensure the safety of construction projects. In this paper, a multi-level pyramidal feature fusion network based on an attention mechanism is proposed for the detection and identification of helmets worn by personnel. To improve the detection speed and accuracy, the network uses a residual block structure design and introduces the ECAttention channel attention mechanism to achieve cross-channel interaction. By doing so, it significantly reduces the complexity of the model while maintaining a high level of performance. To verify the effectiveness of the proposed detection network, this study compares some outstanding detection methods, drawing on existing public datasets and images obtained from the Internet. The results show the proposed network's detection efficiency is higher, demonstrating the ability to achieve real-time high-precision detection of helmets worn at production sites.

Keywords: helmet detection;channel attention;multi-scale;feature fusion;

1. Introduction

During production, safety measures intended for the prevention of injury to operators are vital, as well the safety helmet, which is an important means to protect personnel. Typically, the environments in which they are necessary are intricate and hazardous, yet for various reasons, some operators ignore the importance of helmets. Therefore, there are still a large number of casualties caused by operators not wearing helmets as required [1]. Consequently, in such an environment, to better protect operators' safety, there is a need to detect and confirm when helmets are being worn. The traditional manual inspection is labor-intensive, error-prone, and cannot be monitored in real-time. Indeed, the incidences of workers not wearing helmets in accordance with safety regulations are many. To address this, the use of computer-generated vision to achieve real-time automatic detection

of personnel and determining whether or not they are wearing helmets is an effective way to solve the above mentioned problem.This method can be divided into two types of methods: traditional machine learning and deep learning.

In traditional machine learning algorithms, helmet detection is performed by isolating color and shape features. For example, Talaulikar et al.[2] used threshold processing to segment the safety helmet from the background, followed by extracting the shape features of the safety helmet and acting Principal Component Analysis (PCA), before finally constructing an MLP classifier to achieve detection of the safety helmet as worn by onsite workers. Chiverton et al.[3] proposed a new shadow detection method by using the reflective characteristics of the safety helmet and subtracting the background to extract the helmet's feature image, using an SVM classifier to detect it. The above techniques primarily rely on human feature extraction, which, due to the single feature and poor generalization ability, cannot effectively detect safety helmets in complex construction environments. As a result, traditional detection methods have limited ability to solve the detection problem in the context of wearing safety helmets.

Currently, the majority of deep learning-based means of object detection can be divided into two categories: two-stage and one-stage object detection algorithms. The two-stage network encompasses region extraction and classification, which has higher accuracy overall. However, the presence of a large number of operations in the proposed extraction region leads to a slow detection speed[4]. Conversely, the one-stage network uses advanced features of the image to directly predict location and category, which improves the speed, but compared to the two-stage network, its small object detection is less effective and the localization accuracy is often insufficient[4]. In the actual helmet detection scenario, the two-stage network approach cannot achieve the necessary real time detection. In order to meet the accuracy and speed requirements of helmet wearing recognition for operational scenes, this paper first proposes a backbone network (ConCaNet) that can effectively extract image features. Based on this, the proposed technique incorporates YOLOv3[5] Neck (neck structure), YOLOv3 Head (detection head), and said ConCaNet to construct a one-stage helmet detection model. The experimental results showed that the final algorithm mAP_{50} detailed in this paper outperforms some outstanding detection network algorithms. In addition, it has improved detection accuracy for small, medium, and large objects. The proposed ConCaNet is applied to the helmet detection algorithm with the following main contributions:

(i) A multi-route parallel residual structure is designed for the network to learn features more efficiently. The introduction of a channel attention in the structure block allows the network to reduce the number of parameters while being able to maintain the correlation between channels.
(ii) ConCaNet has an extra scale compared to YOLOv3, and feature fusion is done in the FPN structure to increase the recognition accuracy of tiny objects, thereby mitigating the circumstances in which YOLOv3 misses detection for small targets as may be applicable here.

The structure of this paper is as follows: in Section 2, the proposed EFFCA (Effective Feature Fusion Channel Attention) module, ConCaNet network, and helmet detection model are described; detailed experimental results and analysis of ConCaNet on helmet detection tasks are detailed in Section 3; and finally, conclusions are stated in Section 4.

2. Helmet Detection Algorithm & Model

In this section, the multi-channel attention interaction enhancement module (EFFCA) as designed in this paper in 2.1 is proposed, which can exchange information between multiple channels, and the EFFCA module improves the network feature's extraction capability. The helmet detection backbone network, ConCaNet, is introduced in 2.2. Finally, details related to the application to the task of detecting helmets when worn or not worn are presented in 2.3.

2.1. *Effective Feature Fusion Channel Attention (EFFCA) Module*

The EFFCA module contains the ECA[6] channel attention calculation operation, therefore, this section describes the module and the channel attention calculation process. In the ECA module, the input features first pass through the GAP (global average pooling) layer, followed by a convolutional layer with a convolutional kernel of size 1×1, a Sigmoid activation layer, before it is finally superimposed with the input features. The GAP can compress the feature maps on the channels into global features, thanks to which, EFFCA can learn the weight coefficients of each channel, as shown in Fig. 1 below. In the EFFCA module, the input features x will pass through the convolutional layers of convolutional kernels with sizes 1×1 and 3×3 in turn, followed by the ECA module to obtain

$Output_1$. The input features will also pass through the convolutional layers of convolutional kernel size 1×1 and BN (BatchNorm) operation, resulting in $Output_2$ and $Output_3$, respectively. Here the design of the 1×1 convolution can effectively integrate all channel information [7], and the BN layer can effectively prevent the neural network gradient from disappearing [8]. Finally, with input features x, the EFFCA module output results in $x + Output_1 + Output_2 + Output_3$.

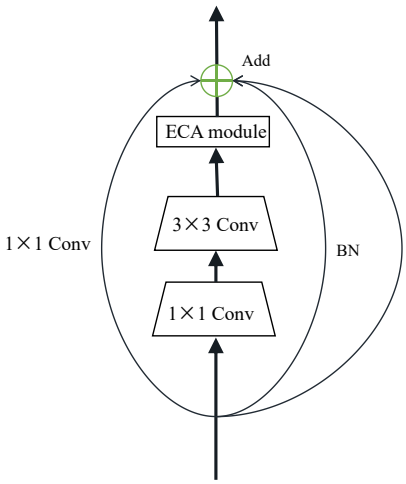

Fig. 1. EFFCA Block

Inspired by He et al. in ResNet [9], who used residual connections to train deep networks, a design for a four-way parallel structure in EFFCA is provided here. For a stacked block structure, it was assumed its learned features can be noted as $H_{(x)}$ when the input is x. In ConCaNet, the residual connected block represented in Fig. 2 is driven to learn features as $F_{(x)} = H_{(x)} - I_{(x)} - G_{(x)} - x$, where $I_{(x)}$ and $G_{(x)}$ represent the convolution and BN operations, respectively. Then, the original learned features may be expressed as $F_{(x)} + I_{(x)} + G_{(x)} + x$. This structure can make it easier for the network to learn features, and the structure in Fig. (1) allows the network to have better performance compared to those that learn the original features directly.

2.2. *ConCaNet*

The structure of the network model is shown in Fig. 2. The input image size is $224 \times 224 \times 3$, and the low-level features are first extracted by a two-layer convolution operation. The feature map then goes through an EFFCA convolution block to calculate the channel attention, and the number of output channels in the convolution block is equal to the number of input channels. After each EFFCA block, a convolution layer is passed to boost the number of feature channels.

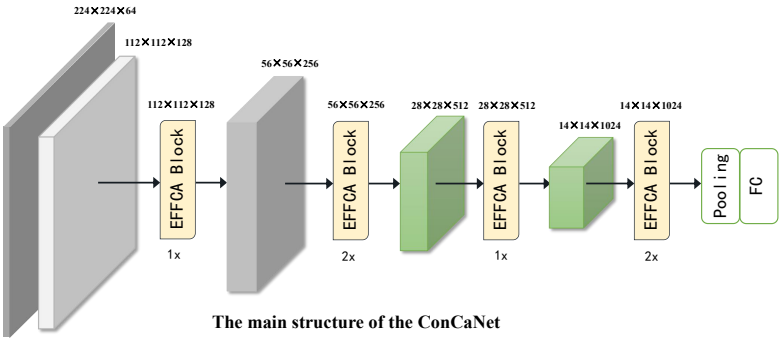

Fig. 2. ConCaNet Structure Diagram

2.3. *Design of a Network to Detect Helmets*

In object detection, the most basic network consists of a backbone network, a detection neck structure, and a detection head [10]. Briefly, the backbone network is used to extract image features, the neck structure is responsible for feature fusion, and the detection head eventually gives predicted values for the target location as well as the class to which it belongs. The structure of the helmet detection network is designed as shown in Fig. 3, where the input image size is $224 \times 224 \times 3$. The image is first fed into the feature extraction network ConCaNet, where four scale feature maps are used in the network before being passed into the neck structure for fusion to improve the detection accuracy for small, medium, and large objects. The neck structure fuses the four scale features and sends them into the detection head structure to obtain the object location information and the predicted value of the class.

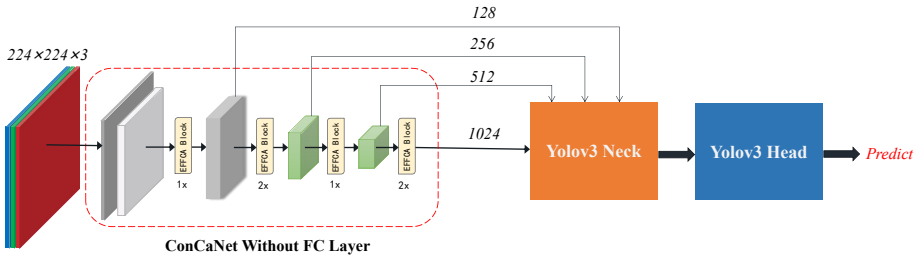

Fig. 3. Detection network structure design - the helmet detection network can predict the object class and location while generating candidate regions in a one-stage object detection algorithm.

3. EXPERIMENTAL

3.1. *Introduction to the Dataset*

The dataset used in this paper is obtained from the public safety helmet dataset of Baidu Paddle community and web crawler. The final COCO format dataset is built by using labelImg and other annotation tools. A total of 6000 photos make up the helmet dataset, which includes a training set, a validation set, and a test set. The training set accounts for 60%, the validation set for 20%, and the test set for 20%, all of which include data from tiny, middle, and large objects. Considering the application benefits and economic costs of the detection task, the detection objects are divided into two categories in this paper: those who wear helmets correctly are marked as Helmet category, and those who do not wear helmets are marked as Head. in addition, those who wear straw hats, bandanas, etc. as helmets in the dataset are marked as Head category, and those unused helmets are not marked in any way and are only treated as background.

3.2. *Experimental settings and results*

In the helmet detection task, the training image is randomly cropped to a section measuring 224×224 pixels, random horizontal flipping and random cropping stitching are used to increase the amount of training data, and label smoothing[11] is carried out to avoid overfitting. The AdamW optimizer[12] was applied for optimization, where the momentum was set to 0.9, weight decay to 0.05, and batch size to 128. After five preparatory iterations, the warm-up learning rate was set to 1×10^{-6} while the initial learning rate was set to 1×10^{-4}, but because the cosine learning rate

strategy decreased, 311 iterations were trained on a single NVIDIA A100 GPU. The accuracy on the validation set was evaluated using images with a center crop of 224×224 pixels. For the helmet detection task, the algorithms performed to the standard as detailed in Table 1.

Table 1. Comparison of metrics of different algorithms

Model	#Params(M)	#FPS(img/s)	#FLOPs(G)	#mAP_50	#mAPs	#mAPm	#mAPl
Cascade_Rcnn_r50	68.93	21.72	**234.47**	83.7	35.0	60.0	63.7
YOLOF_r50	42.09	24.28	98.19	87.3	39.5	64.6	**70.9**
YOLOv3_d53	61.53	22.92	193.87	**89.1**	41.6	62.9	54.4
ConCaNet	**37.65**	**29.66**	137.02	88.2	**42.3**	**64.7**	70.3

'Params', 'FPS', and 'FLOPs' represent the number of parameters, the frame rate, and the number of floating-point calculations per second, respectively. The proposed detection algorithm as detailed here is the best according to four metrics: Params; FPS; mAPs; and mAPm. In the mAPl metric, it is superior by 28.3% and 15.9% when compared to Cascade Rcnn r50 and YOLOv3 d53, respectively. ConCaNet is effective in detecting helmet objects of all three sizes. In terms of FPS, the proposed algorithm demonstrated the fastest detection speed, showing it can be applied in practical environments since it can meet the demand for detecting worn helmets in terms of accuracy.

4. Conclude

The ConCaNet algorithm designed for the purpose of detecting helmets worn by operators, as proposed in this paper, can effectively learn data features while significantly reducing the number of necessary parameters, and improve the model detection performance using data enhancement, channel attention and multiple parallel residual mechanisms, and multi-scale feature fusion. This resulted in improved mAP50, which was 4.5% higher than in Cascade Rcnn. With a near match to the mAP50 of YOLOv3, the detection of small, medium, and large targets (relating to helmet size and proximity) was better with the ConCaNet as proposed in this paper. Future research will be conducted to investigate how the algorithm performs when subjected to adverse conditions including occlusion and bad weather and how this can be improved effectively.

References

1. S. Ajith, S. Chandrasekaran and V. A. Prabu, Safety and hazards management in construction sites - a review, *Journal of Technology* **35**, 175 (2020).
2. A. S. Talaulikar, S. Sanathanan and C. N. Modi, An enhanced approach for detecting helmet on motorcyclists using image processing and machine learning techniques (2019).
3. J. Chiverton, Helmet presence classification with motorcycle detection and tracking, *IET Intelligent Transport Systems* **6**, 259 (2012).
4. Y. Xiao, Z. Tian, J. Yu, Y. Zhang, S. Liu, S. Du and X. Lan, A review of object detection based on deep learning, *Multimedia Tools and Applications* **79**, 23729 (2020).
5. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
6. Q. Wang, B. Wu, P. Zhu, P. Li and Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
7. A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, Searching for mobilenetv3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.
8. W. Wang, X. Huang, J. Li, P. Zhang and X. Wang, Detecting covid-19 patients in x-ray images based on mai-nets, *International Journal of Computational Intelligence Systems* **14**, 1607 (2021).
9. K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
10. K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, Mmdetection: Open mmlab detection toolbox and benchmark, *arXiv preprint arXiv:1906.07155* (2019).
11. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
12. I. Loshchilov and F. Hutter, Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017).