IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 9, 2020, accepted July 1, 2020, date of publication July 7, 2020, date of current version July 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007611

Two-Stream RGB-D Human Detection Algorithm Based on RFB Network

WENLI ZHANG, JIAQI WANG, XIANG GUO, KAIZHEN CHEN, AND NING WANG
Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
Corresponding author: Wenli Zhang (zhangwenli@bjut.edu.cn)

ABSTRACT In order to effectively combine RGB image features with depth image features for human detection, this paper proposes a two-stream RGB-D human detection algorithm based on RFB network. The proposed algorithm mainly contains three parts: RGB-stream, Depth-stream and Channel Weight Fusion (CWF) strategy. (1) The RGB-stream extracts RGB image features using RFB-Net as the backbone network. (2) By analyzing the results of depth features visualization, we build the Depth-stream, which can effectively extract the depth image features. (3) The improved CWF strategy can enhance the effectiveness of important channels in RGB-D fusion features and improve the capability of the network expression. The experimental results show that the proposed algorithm has a significant improvement compared with other algorithms on two common datasets.

INDEX TERMS RGB-D, human detection, fusion features, two-stream.

I. INTRODUCTION

Recently, the fields of smart building and intelligent security are developing rapidly, and the human detection has become a hot research topic in these fields.

In recent years, many researchers have conducted considerable work in using RGB images to detect human [2]–[8] and achieved good detection results. However, RGB images are easily affected by factors such as human occlusion, human attitude changes, illumination changes and complex background. In the complex and varied real scene, the detection accuracy based on the RGB images method may drop sharply.

With the popularity of depth cameras, depth images have attracted extensive attention in many fields [9]–[23]. Compared with RGB images, depth images are not affected by illumination changes, and easier to obtain object contours with low-noise. Li *et al.* [21] proposed an attention steered interweave fusion network (ASIF-Net) to detect salient objects. Han *et al.* [23] proposed a multiview CNN fusion model through a combination layer connecting the representation layers of multiple views to detect salient objects.

Some scholars have jointly used RGB images and depth images for human detection [9]–[19] to solve problems such as illumination changes and complex background. Among them, Zhou *et al.* [13] used two identical networks to process RGB images and depth images respectively, then compared the confidence scores of two network predictions and took the larger score as the final result. Eitel *et al.* [14] proposed a multi-modal RGB-D algorithm consisting of two separate CNN (Convolutional Neural Network) streams, both streams converge in the fully connected layer at the end of the network and directly utilizes fusion features for detection. Zhang *et al.* [15] proposed a multi-stream network for jointing human detection and head pose estimation method in RGB-D videos, using three identical network streams to process RGB data, depth data and optical flow data respectively, then three features are fused together for human detection and head pose orientation estimation. Many methods applying convolutional neural networks to RGB-D data simply combine RGB data and depth data into four-channel data or extract target features separately, and then fuse at the final fully connected layer.

Although the above human detection algorithms use RGB images and depth images for detection, there are two common problems: 1) Extracting RGB image features and depth image features by using two identical network structures. Because depth image lacks features such as texture and color of RGB image, when the depth image is processed complicated network, the target information depth feature maps will be lost. 2) Simple fusion of RGB features and depth features for human prediction. Because the response regions and values of each feature channel are different, when the simple RGB-D

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

fusion feature is used for human prediction directly, the detection result will be affected by the channels which response to non-target area. An excellent feature fusion strategy is essential for network tasks [21], [22], [24].

Therefore, how to effectively extract depth image features and utilize RGB-D fusion features are the keys to human detection by combining RGB images with depth images. Our goal is to deploy approached algorithm on lightweight edge devices. Consequently, we propose a two-stream RGB-D human detection algorithm based on RFB network [1]. The RFB network model uses SSD [25] as the backbone network, by embedding the RFB receptive field enhancement module, improving the detection accuracy on the basis of maintaining the reasoning speed. SSD is more streamlined and easier to modify than YOLO [26] and Faster RCNN [27]. It is more suitable for deployment to edge devices. Many lightweight detection algorithms [28], [29]are inspired by SSD.

Most current researches usually use two identical branch networks to construct the two-stream network. It is generally believed that deeper network has stronger learning capabilities and could extract higher-level semantic information. However, because of the depth image has more object contours and more low-level information, the deeper learning will course the loss of important feature information in the depth image features, finally will affect the detection performance of the whole learning network.

Therefore, in this paper, an asymmetric two-stream network structure based on RGB-D fusion is proposed, which the RGB branch network has a deeper number of layers to learn more semantic features, and the depth branch network has a shallower number of layers to learn more lower-level features. Furthermore, because of the difference of the response area and the response value of each feature channel, the simple overlay fusion can not give full play to the advantages of RGB feature and depth feature. Therefore, we also propose a channel weight fusion strategy called as CWF, which is to increase the weight of the useful channel in the feature, suppress the weight of the useless channel, and improve the feature expression ability.

II. RELATED WORK

According to the type of image, we divided the methods of human detection into two categories for introduction: RGB human detection and RGB-D human detection.

A. RGB HUMAN DETECTION

Mu *et al.* [8] proposed local binary pattern (LBP) as regional features, and using two variants of LBP: Semantic-LBP and Fourier-LBP for human detection. Wang *et al.* [7] combined Histogram of Oriented Gradient (HOG) with LBP as features to effectively deal with partial occlusion in human detection systems. In recent years, with the development of deep learning, the algorithms in the field of object detection, such as SSD, YOLO, etc., have achieved state-of-the-art effects. Many scholars use the methods based on deep learning [2]–[7] for human detection, and the detection effect is significantly better than the traditional method. Ouyang and Wang [2] proposed a new deep learning network architecture by combining the HOG features extraction module. Wang *et al.* [3] proposed a new bounding box regression loss function, which improves the detection accuracy in crowded and occluded situations.

B. RGB-D HUMAN DETECTION

Spinello and Arras [9] designed Histogram of Oriented Depth (HOD), and combined with HOG and HOD in probability for human detection. Hu *et al.* [11] proposed a RGB-D + ViBe foreground extraction method, using three-dimensional moving and depth constraint methods to deal with human local occlusion problems. Liu *et al.* [12] determined the head vertices through depth images, then used the upper body locator and the Joint Histogram of Color and Height (JHCH) filter to encode the thick borders. Tian *et al.* [16] proposed Histogram of Multi-order Depth Template (HMDT) and Joint Histogram of Color and Distance (JHCD) methods. [13]–[15], [18] used deep learning methods for human detection. Among them, Zhou *et al.* [13] used two identical network structures to extract RGB features and depth features respectively, and chose the higher score as the final prediction result. The architecture proposed by Eitel *et al.* [14] consists of two independent CNN network streams, which processes RGB images and depth images respectively, then performs RGB-D features fusion at the fully connected layer and outputs the detection results. Zhang *et al.* [15] used three networks to extract RGB images, depth images and optical streams. Tian *et al.* [18] proposed a DMH (Depth map, Multiorder depth template, and Height difference map) representation method which can effectively capture the geometric structure information in the depth images. Fan *et al.* [19] proposed a simple baseline architecture, called Deep Depth-Depurator Network (D3Net), which consists of a depth depurator unit and a feature learning module, performing initial low-quality depth map filtering and cross-modal feature learning respectively. The human detection algorithms effect based on deep learning are superior to other methods.

III. METHOD

The overview of the network structure is shown in Section A. Section B describes the network structure construction based on visual analysis. Subsequently, the improved Channel Weight Fusion strategy is depicted in Section C.

A. OVERVIEW OF NETWORKS STRUCTURE

The proposed algorithm network structure is shown in Fig.1, which consists of three parts: 1) The Depth-stream is used to extract depth image features. 2) The RGB-stream is used to extract RGB image features. 3) The CWF strategy is used to improve weights of the important channels in RGB-D fusion features.

1) DEPTH-STREAM

Based on the characteristics of the depth image, we construct a network named Depth-stream, which can effectively extract

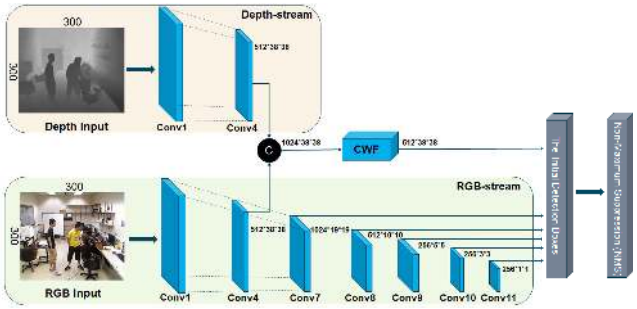


FIGURE 1. Algorithm network structure.

the depth image features. The Depth-stream consists of the first four convolutional blocks of VGG-16 network [30], Conv1-Conv4, with the 300*300 depth image as input, and output the feature map of 512 * 38 * 38. The feature map is more responsive to the human area and can separate the human information from the background image to reduce interference from background redundant information.

2) RGB-STREAM

The RGB-stream uses RFB-Net [1] as the backbone network, which is a multi-scale predictive one-stage network structure. RFB-Net uses SSD [25] network structure as the backbone network, and enhances the recognition ability and robustness of the features by embedding the receptive field module RFB, which makes the detector performance better.

3) CWF STRATEGY

The great channel weight fusion strategy can establish the connection between channels in RGB-D fusion features, assign different weights to each channel and enhance the peculiarity of RGB features and depth features. Therefore, this paper proposes the CWF strategy suitable for RGB-D fusion features, which can effectively improve the weights of important channels, realizing cross-channel interaction and information integration.

4) ALGORITHM FLOW

As shown in Fig.1, our network starts with two inputs (RGB image and depth image), and ends to six different scale feature maps to detect human of different sizes in the image. The last two parts of the network includes ''The initial detection boxes'' part and ''Non-Maximum Suppression (NMS)'' part. ''The initial detection boxes'' collects the detection boxes with confidence scores from the six feature maps which are from CWF, Conv7, Conv8. Conv9, Conv10 and Conv11. ''Non-Maximum Suppression (NMS)'' part sorts all detection boxes based on their scores. The detection box named M with the maximum score is selected and all other detection boxes with a significant overlap (using a pre-defined threshold) with M are suppressed. This process is recursively applied on the remaining boxes [31]. Finally, the final detection boxes will be output for further use.

The CWF block enables the network to learn the importance of each channel in RGB-D fusion features and improve the expression ability of fusion features. However, if the fusion features are directly input into the deep network to continue learning, the differences between RGB and depth images may lead to confusion, and the deep network cannot extract advanced semantic information. Therefore, only features after Conv4 in RGB-Stream are used to feed into Conv7 instead of features integrated after the CWF block.

The receptive field of the feature map generated by the shallow convolution layer is smaller, and the detection effect on the small-sized human in the image is better. At the same time, the RFB network detects the larger size target significantly better than the small size target, so we only perform RGB-D feature fusion on the feature map generated by the VGG-16 block4 convolution block. This approach can minimize the amount of computation and ensure the simplicity of the network structure and the speed of reasoning.

B. DEPTH-STREAM CONSTRUCTION BASED ON VISUAL ANALYSIS

Depth images highlight the edges, contours and shapes of the target without color or texture characteristics, thereby using the complex network to process depth images is likely to cause losing the target feature information. Therefore, excessively increasing the convolutional network hierarchy does not help extract effective depth image features.

The visualization method proposed by [32], [33] can observe the features in different convolutional layers. Among them, the guided backpropagation [33] method has the advantages of high resolution and outstanding details. Consequently, we use the visual method [33], utilizing the pre-trained VGG-16 as the backbone network, and visualizing the feature maps of Conv1, Conv2, Conv3, Conv4 and Conv5 respectively, as shown in Figure 2.

Fig.2(a) is the input depth image. Fig.2(b) and Fig.2(c) are respectively represent the feature maps of Conv1 and Conv2. There are more redundant background information in both images, which proves that the shallow network is difficult to learn target features. The background information in Fig.2(d) is effectively reduced, but the separation of foreground and background pixels is poor. Fig.2(e) shows the feature map of the Conv4, the response position pays more attention to the human area, effectively separates the human information from the background, and greatly reduces the interference of redundant information. The human characteristics in Fig.2(f) are significantly reduced and cannot provide effective assistance for human detection. The results of visual analysis prove that the feature map generated by Conv4 pays more attention to the human area, and can separate the human information from the background information to extract more discriminative semantic features. Therefore, we use the first four convolution blocks of VGG-16 to construct

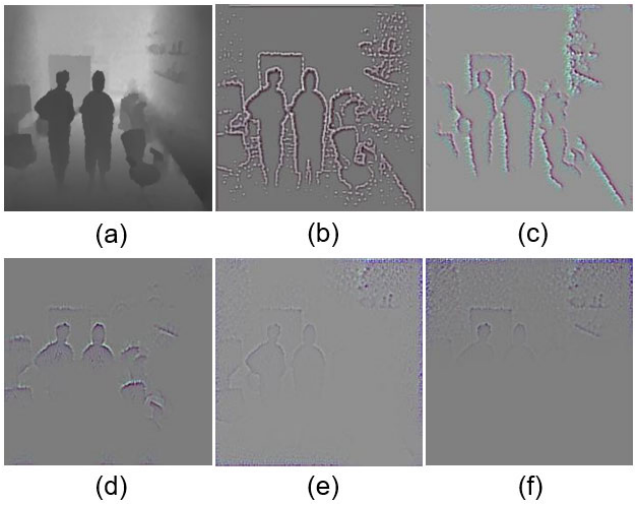


FIGURE 2. Feature maps visualization. (a) is the input depth image. (b), (c), (d), (e) and (f) are feature maps of Conv1, Conv2, Conv3, Conv4 and Conv5 respectively.

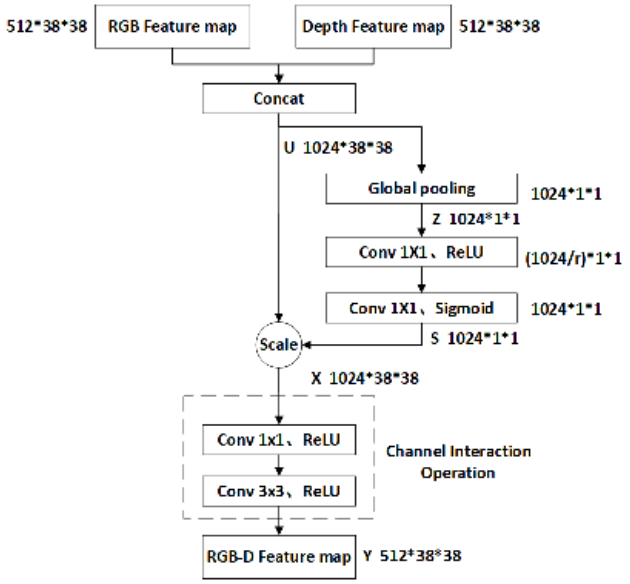


FIGURE 3. Channel Weight Fusion strategy.

the Depth-stream structure for extracting depth image features.

C. CHANNEL WEIGHTS FUSION (CWF)

The RGB image and the depth image sources are different. Because of this, there are differences between RGB image features and depth image features. In Fig.1, 512-channel RGB features and 512-channel depth features are channel-connected to obtain the 1024-channel RGB-D features. Simple RGB-D fusion makes features more diverse, but there are three problems: 1) RGB features and depth image features are complementary and different. The two features are different in the contribution of the detection results. 2) Simple RGB-D fusion method cannot realize the information interaction between RGB features and depth feature channels. Consequently, the obtained RGB-D fusion features may be distributed disorderly and makes it impossible to complement each other and affect the performance of RGB-D human detection. 3) The number of feature channels is expanded from 512 to 1024, resulting in an increase in the calculation of network parameters, which increases the computational cost and affects the speed of algorithmic reasoning.

Hu *et al.* [34] proposed that SE establishes the relationship between feature channels, acquiring the importance of each channel through adaptive learning and enhancing the effectiveness of important channels. SE works well for RGB feature channels, but it does not apply to RGB-D fusion features. It is impossible to realize information interaction between RGB image feature channels and depth image feature channels and complement each other.

Therefore, this paper proposes the Channel Weight Fusion strategy suitable for RGB-D fusion features. It adds channel interaction operation based on SE, which can realize cross-channel interaction, reducing the dimension of fusion features, increasing nonlinear characteristics and improving the expressive ability of the network. The specific steps are shown in Fig.3.

Firstly, fusing 512-channel RGB features and 512-channel depth features into 1024-channel RGB-D feature by using channel-connection method. Then the Global pooling layer turns each two-dimensional channel of the RGB-D feature into the real number, and the output dimension matches the number of input feature channels. Each real number characterizes the global distribution of the feature channels, as shown in Equation (1), where z_c denotes the response of channel c in RGB-D feature U, u_c represents the channel c in RGB-D feature U, and H and W are the size of RGB-D feature U.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

Correlation between channels is established by two 1×1 convolutional layers, and the Sigmoid activation layer is able to obtain 1024 normalized weights between 0-1. Equation (2) describes Sigmoid gating function F_S, where σ denotes Sigmoid function, δ denotes the ReLU function, W_1 represents the parameters of the first convolutional layer, W_2 represents the parameters of the second convolutional layer and S is the normalized weights. After that, the Scale operation weights the normalized weights S onto each channel of the RGB-D feature U to obtain a 1024*38*38 channel weighted RGB-D feature X, Scale operation as shown in Equation (3).

$$S = F_S(Z, W_1, W_2) = \sigma(W_2 \cdot \delta(W_1 Z)) \tag{2}$$

$$x_c = F_{scale}(u_c, s_c) = s_c u_c \tag{3}$$

Finally, we add a channel interaction operation after Scale, aiming to enhance the weighted RGB-D features. As shown in the dotted line in Fig.3, it consists of a 1*1 convolutional layer and a 3*3 convolutional layer, and both convolutional

TABLE 1. Results of ablation experiment. Inference on single GPU GTX 1080TI.

Method	Precision	Recall	F1-score	FPPI	Miss Rate	Parameters	FPS
RFB	0.884	0.844	0.864	0.338	0.156	36.35M	67
RFB+VGG-16	0.785	0.839	0.811	0.701	0.161	58.16M	42
RFB+VGG-Conv4	0.814	0.856	0.834	0.6	0.144	48.72M	50
RFB+VGG-Conv4+SE	0.873	0.902	0.887	0.4	0.098	48.86M	50
RFB+VGG-Conv4+CWF(proposed)	**0.90**	**0.914**	**0.903**	**0.33**	**0.086**	51.74M	48

layers are activated by ReLU. The 1 * 1 convolution can perform linear combination operation on different channels to achieve cross-channel interaction. The 3 * 3 Convolution and ReLU activation can increase the nonlinear characteristics of the network. Formally,

$$Y = F_R(X, W_{1*1}, W_{3*3}) = \sigma(W_{3*3} \cdot \delta(W_{1*1}X)) \qquad (4)$$

where W_{1*1} represents the parameters of the 1*1 convolutional layer, W_{3*3} represents the parameters of the 3*3 convolutional layer and δ denotes the ReLU function. Through the above operations, CWF enhances the effectiveness of important channels in the RGB-D fusion feature and improves the expressive ability of the network.

IV. EXPERIMENT

In this chapter, we will introduce the datasets, the experimental environment, and comparison with other algorithmic experimental results.

A. DATASETS AND EXPERIMENTAL ENVIRONMENT

We conduct experiments in two public datasets: the Office Dataset and the Mobile Platform Dataset [35]. The Office dataset contains 17 videos, and the people perform different postures such as standing, walking, and sitting. The Mobile Platform dataset consists of 18 videos, with an approximately horizontal view. We select 691 pairs of RGB-D images from the Office Dataset, 2209 pairs of RGB-D images from the Mobile Platform dataset for experiments. Two datasets are captured with the Kinect depth camera. Each image has a resolution of 640*480.

Data preprocessing: Since the Kinect camera is affected by the mirror object and the shooting distance is limited, the acquired depth image will generate a void area. Therefore, in our experiments, we use the hole repair method [36] to repair 2900 depth images, and label the head and shoulders of the person in each image.

Training data and parameters: We divide 2900 pairs of images according to the ratio of 6:4, 1740 pairs of images to train, and 1160 pairs of images to test. Our training strategy follows RFB training methods, including data expansion, hard negative sample mining, using the smooth L1 loss function for position regression and the Softmax loss function for human classification. The optimization method is SGD with 0.9 momentum, 0.0005 weight decay and 0.001 initial learning rate. The training time is about 12 hours with one NVIDIA GTX1080TI.

B. DISCUSSION

We use the Miss Rate-FPPI indicator to evaluate the effect of the model, where Miss Rate represents the rate of missed detection targets to the ground truth, and FPPI represents false positive per image. If the detected IOU (Intersection over Union) overlap rate of the bounding box and the labeled bounding box is greater than 0.5, the detected bounding box is considered correct.

1) ABLATION EXPERIMENTS

We use five methods to perform experiments on the dataset of the Clothing Store dataset and the Office dataset, as shown in Table.1. RFB: utilizing RFB-Net for human detection based on RGB images. RFB + VGG-16: the RFB-Net is used as RGB-stream to extract RGB images features, and the total VGG-16 network is used as Depth-stream to extract depth images features. RFB + VGG-Conv4: the difference with RFB + VGG-16 is that Depth-stream is constructed by the first four convolution blocks of the VGG-16. RFB + VGG-Conv4 + SE: utilizing SE to weight channels of RGB-D features. RFB + VGG-Conv4 + CWF: the proposed algorithm. The results have shown that our proposed algorithm achieves the best experimental performance on the basis of guaranteeing the 48 FPS real-time reasoning speed.

Firstly, it is obvious that all of the VGG-Conv4 networks combining the RGB and Depth streams, are more effective and lighter than the RFB+VGG-16 network, which proves the complicated network is not help for extracting depth image features. Especially, the proposed RFB+VGG-Conv4+CWF algorithm not only has fewer parameters and faster inference speed, but also the F1-score is improved by 9.2% than the RFB+VGG-16 network. Secondly, compared with the RFB+VGG-Conv4 and RFB+VGG-Conv4+SE network, the proposed algorithm can achieve better detection performance with almost the same of parameters and inference speed. At the same time, it demonstrates that the CWF strategy is more suitable for RGB-D fusion features. Finally, since the proposed algorithm contains RGB-stream and Depth-stream, compared with the RFB network which only has RGB-stream, its network parameters and inference time have a little increased. However, the proposed algorithm still significantly improves the detection performance: the precision increased by 1.6%, the recall rate increased by 7%, F1-score increased by 3.9%, while maintaining effective real-time reasoning speed 48 FPS.

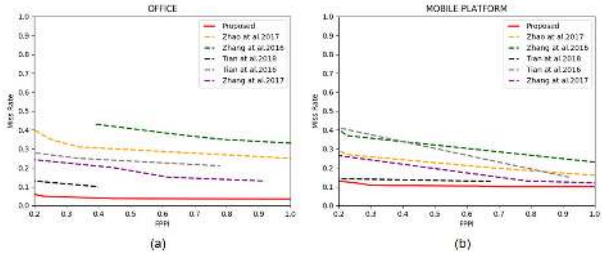


FIGURE 4. Human detection results on (a) Office Datasets and (b) Mobile Platform Datasets. The abscissa FPPI indicates false positive per image, and the ordinate Miss Rate indicates the rate of missed detection targets to the ground truth.

FIGURE 5. Human detection results display. (a) Office Datasets and (b) Mobile Platform Datasets.

2) COMPARISON WITH OTHER ALGORITHMS

We compare the proposed method with Eitel *et al.* [14], Zhang *et al.* [15], Tian *et al.* [18], Zhang *et al.* [37] and Zhao *et al.* [38]. Among them, Zhang *et al.* [37] uses the depth value and the real head size to obtain the head region, but the coarse extraction method has a large positioning error, and it is difficult to accurately locate the person coordinates. They [14], [18], [38] search for the location of the extreme points of the head contour and relies heavily on the accuracy of the original depth data. Its [15] multi-stream network is highly complex, and simple features fusion is performed only at the end of the network, which increases the risk of features redundancy. The experimental results are shown in Fig.4. In the experimental results of two datasets, the detection accuracy of the proposed algorithm is better than other algorithms. The results show that the proposed algorithm can extract effective depth image features and enhance the effectiveness of important channels in RGB-D fusion features. Fig.5 shows several visualized detection results of proposed algorithm on the two datasets.

V. CONCLUSION

This paper proposes a two-stream RGB-D human detection algorithm based on RFB network. By analyzing the feature map of each layer, the Depth-stream is constructed by shallow convolution layers, which can extract the more effective depth image features than complicated network. The improved CWF strategy is suitable for RGB-D features, which enables cross-channel interaction and improve the expressive ability of the network. After experimental analysis, the proposed algorithm can enhance the effectiveness of important channels in RGB-D fusion features and improve human detection accuracy. In the future, we will design a lighter network and a more efficient RGB-D fusion strategy to further improve the speed and accuracy of the algorithm.

REFERENCES

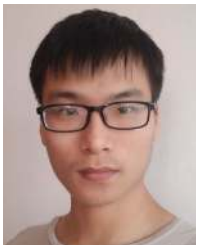
[1] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 385–400.
[2] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2056–2063.
[3] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7774–7783.
[4] S. Wu, W. Wu, S. Lei, S. Lin, R. Li, Z. Yu, and H.-S. Wong, "Semi-supervised human detection via region proposal networks aided by verification," *IEEE Trans. Image Process.*, vol. 29, pp. 1562–1574, 2020.
[5] W. Liu, S. Liao, and W. Hu, "Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding," *IEEE Trans. Image Process.*, vol. 29, pp. 1413–1425, 2020.
[6] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5187–5196.
[7] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 32–39.
[8] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[9] L. Spinello and K. O. Arras, "People detection in RGB-D data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 3838–3843.
[10] B. Li, H. Jin, Q. Zhang, W. Xia, and H. Li, "Indoor human detection using RGB-D images," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Aug. 2016, pp. 1354–1360.
[11] T. Hu, H. Zhang, X. Zhu, J. Clunis, and G. Yang, "Depth sensor based human detection for indoor surveillance," *Future Gener. Comput. Syst.*, vol. 88, pp. 540–551, Nov. 2018.
[12] J. Liu, G. Zhang, Y. Liu, L. Tian, and Y. Q. Chen, "An ultra-fast human detection method for color-depth camera," *IEEE Trans. Ind. Electron.*, vol. 31, pp. 177–185, Aug. 2015.
[13] K. Zhou, A. Paiement, and M. Mirmehdi, "Detecting humans in RGB-D data with CNNs," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 306–309.
[14] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 681–687.
[15] G. Zhang, J. Liu, H. Li, Y. Q. Chen, and L. S. Davis, "Joint human detection and head pose estimation via multistream networks for RGB-D videos," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1666–1670, Nov. 2017.
[16] L. Tian, G. Zhang, M. Li, J. Liu, and Y. Q. Chen, "Reliably detecting humans in crowded and dynamic environments using RGB-D camera," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.
[17] J. Liu, Y. Liu, G. Zhang, P. Zhu, and Y. Q. Chen, "Detecting and tracking people in real time with RGB-D camera," *IEEE Trans. Ind. Electron.*, vol. 53, pp. 16–23, Feb. 2015.
[18] L. Tian, M. Li, Y. Hao, J. Liu, G. Zhang, and Y. Q. Chen, "Robust 3-D human detection in complex environments with a depth camera," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2249–2261, Sep. 2018.
[19] D. Fan, Z. Lin, Z. Zhang, M. Zhu, and M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, Jun. 2020, doi: 10.1109/TNNLS.2020.2996406.
[20] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, and F. Xu, "3D room layout estimation from a single RGB image," *IEEE Trans. Multimedia*, early access, Jan. 17, 2020, doi: 10.1109/TMM.2020.2967645.
[21] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, early access, Feb. 13, 2020, doi: 10.1109/TCYB.2020.2969255.

[22] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2019.
[23] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.
[24] C. Yan, B. Gong, Y. Wei, and Y. Gao, "Deep multi-view enhancement hashing for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, Feb. 24, 2020, doi: 10.1109/TPAMI.2020.2975798.
[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[26] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," Apr. 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[28] Y. He *et al.*, "LFFD: A light and fast face detector for edge devices," Apr. 2019, *arXiv:1904.10633*. [Online]. Available: https://arxiv.org/abs/1904.10633
[29] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3FD: Single shot scale-invariant face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 192–201.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[31] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5561–5569.
[32] J. T. Springenberg *et al.*, "Striving for simplicity: The all convolutional net," Dec. 2014, *arXiv:1412.6806*. [Online]. Available: https://arxiv.org/abs/1412.6806
[33] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.
[34] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[35] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1577–1591, Jul. 2013.
[36] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 175–185.
[37] G. Zhang, L. Tian, Y. Liu, J. Liu, X. A. Liu, Y. Liu, and Y. Q. Chen, "Robust real-time human perception with depth camera," in *Proc. 22nd Eur. Conf. Artif. Intell.*, Aug. 2016, pp. 304–310.
[38] J. Zhao, G. Zhang, L. Tian, and Y. Q. Chen, "Real-time human detection with depth camera via a physical radius-depth detector and a CNN descriptor," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1536–1541.

WENLI ZHANG received the M.S. and Ph.D. degrees from The University of Tokyo, Japan. She worked at the Tokyo Research Center, Panasonic Electric Works Company Ltd., as a Senior Researcher. She is currently working as a Professor with the Faculty of Information Technology, Beijing University of Technology. She has published more than 30 international academic articles and declared nearly 30 national invention patents, including five international invention patents and ten authorized patents. Her main research interests include image processing and pattern recognition.

JIAQI WANG received the B.E. degree, in 2018. He is currently pursuing the master's degree with the Beijing University of Technology, Beijing, China. He has published two international academic articles and declared two national invention patents. His main research interests include object detection and pattern recognition.

XIANG GUO received the B.E. degree, in 2017. He is currently pursuing the master's degree with the Beijing University of Technology, Beijing, China. He has published five international academic articles and declared five national invention patents. His main research interests include object detection and pattern recognition.



KAIZHEN CHEN received the B.E. degree, in 2019. He is currently pursuing the master's degree with the Beijing University of Technology, Beijing, China. His main research interests include object detection and pattern recognition.

NING WANG received the B.E. degree, in 2019. She is currently pursuing the master's degree with the Beijing University of Technology, Beijing, China. Her main research interests include object detection and pattern recognition.