EasyChair Preprint

№ 14826

Face Detection Algorithm in Classroom Scene Based on Deep Learning

Yi Zhang and Chongwen Wang

September 12, 2024

Face detection algorithm in classroom scene based on deep learning

Abstract As informatization progresses in education, the application of artificial intelligence technology in the field of education is increasing day by day. Although some breakthroughs have been made, the application in specific scenes (such as classroom scenes) faces many difficulties, such as small target detection, severe occlusion, and so on. We propose a target detection algorithm based on video data for the classroom scene, which combines the optical flow information to improve the accuracy and alleviate the impact of occlusion. We also put forward the counting method suitable for classroom scenes, which can be used as the evaluation standard of attendance rate, head-up rate, and other indicators in classroom quality evaluation.

Keywords target detection, optical flow method, video analysis, face detection

1 Introduction

With the development of educational informatization, deep learning and other technologies in education are also gradually increasing. For example, the target detection algorithm can be used to evaluate the probabilities of attendance, coming late, early leaving, and other indicators in teaching quality evaluation. However, the mainstream object detection algorithms such as YOLO, SSD, and faster RCNN have poor performance in college classroom scenes. We hope to propose a face detection algorithm suitable for classroom scenes and video data, especially for small targets and severe occlusion.

At present, target detection algorithms have made significant progress in accuracy and real-time performance. However, the image target detection network is usually detected after extracting keyframes when processing video data. The subsequent tasks such as tracking and motion prediction are also based on the bounding boxes. This method has good real-time performance, but some semantic information hidden between frames is ignored. In addition, it is difficult to deal with the occlusion problem, which is mainly limited by datasets and the setting of parameters. People usually adjust the threshold of the NMS [1] algorithm to balance the undetected error and residual error or use specific algorithms such as GAN [2] and composite network to alleviate the occlusion problem.

Although the deep learning method has been widely used in target detection, we still need to study many problems in video target detection. For example, this paper aims at the large classroom in colleges and universities, which has many people and small targets. There is a problem of severe occlusion and small target detection in this scene. Therefore, we combine optical flow information, design a video target detection algorithm, apply it to practical problems. We prove the algorithm can accurately detect students in classroom scenes.

2 Related works

2.1 Overview of target detection

With the development of deep learning, there has been a lot of progress and many algorithms with good performance in target detection. We will introduce target detection in detail from two aspects of image target detection and video target detection.

2.1.1 Image target detection

Image target detection algorithms are mainly divided into region proposal and end-to-end algorithms.

The algorithm based on region proposal, also known as the two-stage algorithm, needs to select the candidate region before extracting the feature and judging the category. Compared with the end-to-end algorithm, the detection speed is slower, but the detection accuracy is higher. The standard algorithms are fast RCNN and faster RCNN [3]. D2det [4] introduces dense regression to solve precise positioning and classification and predict multiple thick frame offsets of object proposals. This method is based on the standard faster RCNN framework. The traditional box offset regression of faster RCNN is replaced by the proposed dense local regression in this method.

End-to-end target detection algorithms mainly include YOLO [5], SSD, etc., also known as single-stage target detection algorithms. Instead of region nomination, they regard target detection as a complete algorithm and directly regress the image features extracted by a convolution neural network to obtain the position and category of target objects. Because the complex and time-consuming region nomination process is abandoned, the detection speed has been dramatically improved. However, because it sacrifices the detection accuracy, it has a poor effect on the problems of many targets, high overlap rate, and small target detection. Yolof [6] only uses the one-level feature layer for detection and introduces two key components: expansion encoder and uniform matching, which improves performance. Many experiments on the COCO benchmark show that the detection speed of Yolof is 13% faster than that of YOLOv4, and the accuracy is the same.

2.1.2 Video target detection

Video target detection is mainly divided into deep-learning-based methods and detection and tracking-based methods.

Video detection algorithms based on deep learning, such as DFF [7], due to the highly redundant characteristics of video and image, applying the detection model of the static image directly to video target detection will produce many problems, such as image blur, significant size change and so on. Scholars hope to use the motion information of the target in the video for detection, so they propose the concept of keyframe and carry out end-to-end training, which ensures the accuracy of detection and improves the defect of approximate but inaccurate features. Ba Griffin [8] et al.

proposed uncalibrated motion and detection-based depth estimation, derived analysis models and corresponding solutions, and developed a recursive neural network (RNN) to predict depth according to motion and boundary box to improve the general applicability in different fields.

The representative work of the video detection algorithm based on detection and tracking is TCNN [9], which can solve the problem of temporal and spatial consistency of targets in the video. Its core idea is to combine the material information of the video sequence learned by the tracking algorithm with the spatial information retained by the target detection algorithm. To improve the performance of the video target detection algorithm, TCNN links the detection frames extracted from each frame together in chronological order and then uses a neural network to classify and score, and finally completes the tracking. However, the limitation of TCNN is that it relies too much on the performance of the tracking algorithm, takes a long time, and does not make use of the motion information of the video. MW Ashraf [10] and others use the region nomination method and optical flow information to detect small video targets using the attention mechanism at the pixel level. Still, the approximate optical flow can not accurately track the matched targets. Roam [11] uses LSTM to solve the problem of target tracking, which is composed of a tracking module and model assistance update module, which can update the tracking model online.

2.2 Optical flow method

Hron Schunck proposed the gradient-based optical flow algorithm. Starting from the assumption of global smoothing, assuming that the optical flow field inside the moving object is the same, the constraint equation of the optical flow field inside the thing is obtained. This algorithm designs a total error parameter. When this parameter reaches the minimum value, it is considered that the calculated optical flow field is optimal, and the iterative formula of the optical flow field is obtained. Hron Schunck's algorithm can measure the roughness of the optical flow field. Jiang [12] et al. learned optical flow motion from sparse correspondence, which significantly reduced video memory use and maintained high accuracy.

With the development of deep learning technology, many methods of predicting optical flow using neural networks have appeared in recent years. These methods usually use multiplicative interaction to simulate the relationship between images. Differences and optical flow can then be inferred from potential variables. Flownet [13] inputs two frames of adjacent images, first extracts the respective characteristic pictures of the two images, then expands them to the size of the original image through the expansion layer, superimposes the two images according to the channel direction, and obtains the final optical flow prediction through a series of convolution layers and inverse convolution layers. Mask flownet [14] is a relatively new improvement, and its effect on the Kitti data set is also excellent. It is mainly aimed at the problem of some occluded areas in the background after the foreground moves. A subtask is designed to predict both the optical flow and the occluded to solve this problem. Yang G [15] et al. learned optical flow from unlabeled image sequences under enhanced self-supervised. The model uses unsupervised optical flow estimation to obtain the categories of reliable supervision information from transformations, which significantly improves the accuracy and has high compatibility and generalization ability.

3 PROPOSED METHOD

3.1 Model overview

Aiming at the problem of face detection in the classroom environment, we propose a two-stage detection network combined with optical flow named FFnet （Flow Face network）, as shown in Fig 1. The ideas to improve the detection accuracy of small targets in video data are as follows: (1) based on faster RCNN processing image RGB three-channel information, add two-dimensional optical flow information to enrich inter-frame semantics. (2) Refer to soft NMS algorithm and improve SEQ NMS algorithm to make it suitable for complex scenes with high target overlap in the classroom. (3) Sort algorithm is used to track to judge whether new students appear and leave the classroom.

Firstly, we use the optical flow prediction algorithm based on deep learning -a flownet network to extract the optical flow information and save it in a two-dimensional matrix with the same shape as the picture. The optical flow information matrix of the previous frame and the RGB image information of the current frame are spliced according to the channel dimension. The optical flow information matrix of the first frame is **0**. This paper proposes a target detection network FFnet suitable for a classroom scene based on optical flow information and a mainstream target detection network. The network architecture is shown in Figure 1.

In this paper, the dense optical flow information of video frames is extracted by flownet. The optical flow of all points on the image is estimated based on the displacement of pixels on the two frames. The image registration at the pixel level can be carried out, and its effect is better than sparse optical flow. The reason is that a group of points (such as corners) need to be specified to track the sparse light flow. The optical flow vector is not dense enough, but its calculation is faster than that of dense optical flow. After extracting the optical flow information, the optical flow information and image RGB information are spliced according to the dimension and sent to the following network for feature extraction, detection, and classification. The specific architecture of the network is described in detail below.

3.2 Network design

The design of the detection network of FFnet refers to the architecture of the classical algorithm faster RCNN. As shown in Figure 2, faster RCNN is mainly composed of four parts. Firstly, a set of convolution layers and pooling layers combined with the ReLU activation function are used to extract the features of pictures. The extracted features are sent to the RPN network to generate candidate regions. Softmax is used to judge whether the anchor belongs to the foreground or background, and then the bounding box regression algorithm is used to obtain the accurate candidate boxes. The generated candidate boxes are sent to the ROI pooling layer. After synthesizing the feature map and

candidate frame information, the candidate boxes feature map is extracted and sent to the full connection layer. The full connection layer does classification work, and the final accurate position is obtained by bounding box regression again.

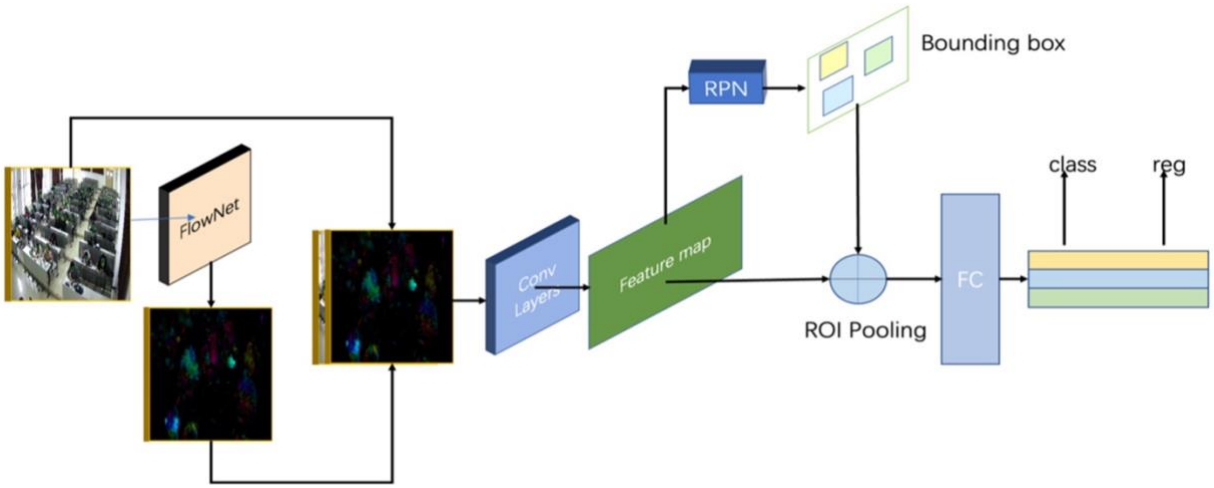

Fig. 1 The architecture of the FFnet.

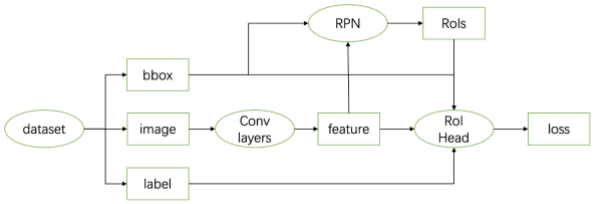

Fig. 2 The architecture of the Faster RCNN.

3.2.1 Feature extraction

Because we use additional two-dimensional optical flow information, the input layer dimension of the feature extraction network is set to 5 * w * H. The setting of convolution layers refers to vgg16 network architecture which is shown in Figure 3. The convolution layer and ReLU layer do not change the input and output size, while the pooling layer changes the output length and width to 1 / 2 of the input length and width. To sum up, after the entire convolution layers, the matrix with the input shape of W * h becomes (w / 16) * (H / 16). Due to many vgg16 parameters, the feature extraction network selects five convolution layers, five ReLU activation layers, and four pooling layers to save the amount of calculation, as shown in Figure 4.

3.2.2 Region Proposal Network

RPN (region proposal network) is mainly used to generate candidate regions. The proposed structure replaces the previous method of violently enumerating and selecting candidate boxes and significantly reduces extraction time consumption. The head of the RPN network is used to generate basic anchors. In this paper, three aspect ratios (0.7, 1.0, 1.3) and three scaling scales (2, 4, 8) are set. Nine anchors with different lengths and widths can be obtained by combination. The figure below shows the anchor corresponding to one position.

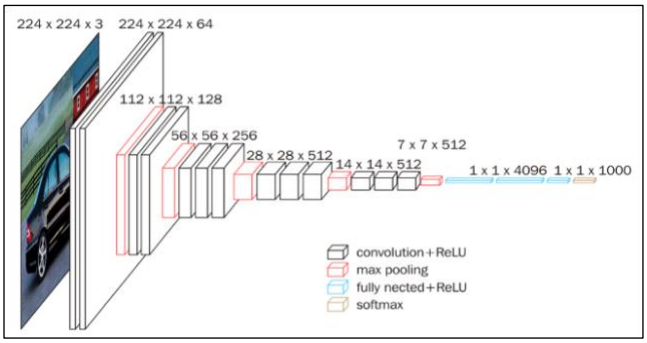

Fig. 3 The architecture of the VGG16.

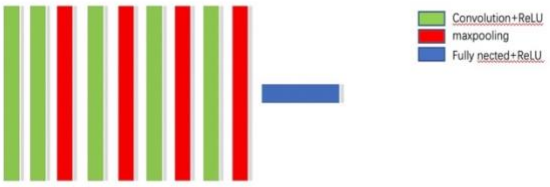

Fig. 4 Feature extraction architecture of the FFnet.

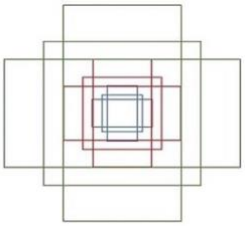

Fig. 5 Anchor of different sizes.

According to the generated basic anchor, boundary boxes of different sizes are developed with the pixel as the center for each pixel on the feature map. Finally, the anchor is screened. Firstly, the coordinate offset output by the positioning layer is applied to all generated anchors, and then all anchors are scored and classified by softmax for foreground/background and sorted. The first 20000 anchors are selected, and the final candidate anchor is obtained through NMS (non-maximum suppression).

After the RPN network extracts the anchors classified as foreground, we hope to fine-tune these anchors to locate the target more accurately. In this paper, (x, y, W, H) represents an anchor, where x and Y respectively represent the abscissa and ordinate of the anchor center point, and W and H represent the width and height of the anchor. We hope to find a linear mapping to map the original anchor to a window closer to the actual window.

After the new (x, y, W, H) is generated by regression, the anchors are generated again using the proposal layer, the first 20000 anchors are extracted, and the anchors beyond the boundary are eliminated. After NMS again, extract the first 3000 with the highest score as the output of the suggestion box.

3.2.3 Classification network

The original feature map and the suggestion frame output by the RPN network are sent to the ROI pooling layer. Since the size of the candidate frame screened by the RPN is not uniform, the feature vector with the same length is required to realize the classification of the fully connected network. To enable the suggestion frames of different sizes to extract the feature map of the same size, the ROI pooling operation is added. The features in the feature map are divided into several blocks, and then each block is pooled. Finally, the pooled operation with a fixed number of blocks is used to replace the pooled operation with a fixed size to generate the pooled result with the same size.

After RoI pooling, the features are flattened into fixed-length feature vectors. In the classification network, the full connection layer and softmax function are used for classification. At the same time, the detection frame is regressed to obtain a more accurate detection frame.

3.3 Improved SEQ-NMS

3.3.1 NMS and soft-NMS

NMS eliminates redundant detection frames and finds the best position of objects. NMS algorithm sets an IOU threshold. When the IOU between detection frames is higher than this threshold, the detection frame with a lower score will be removed, and the detection frame with the higher score will be retained. The calculation consensus of IOU is shown in equation 1. The intersection and union ratio calculates the area where two frames intersect to the area where two frames merge. This index can be used to measure the proximity of two frames.

However, NMS will force the test frame score of adjacent confidence to zero, which is easy to miss detection in target intensive scenes. Soft NMS adopts a relatively acceptable method to improve this problem. The detection box with IOU greater than the set threshold is not simply deleted, but the score is reduced. There are two ways to reduce the score: linear and Gaussian.

This paper is based on the classroom scene and belongs to the target intensive task. It does not use the simple NMS algorithm but improves the SEQ NMS algorithm suitable for video target detection combined with the soft NMS algorithm. We will briefly introduce the traditional SEQ NMS algorithm and the improved SEQ NMS algorithm following.

3.3.2 Improved SEQ NMS

SEQ NMS is proposed for the video target detection task. Firstly, sequence selection is carried out, all detection frames with scores greater than the threshold are selected, the detection frames at the corresponding positions in the front and rear frames are matched, and the average value or maximum value is selected for rescoring, and then NMS is used for suppression. Hard NMS is used in Seq NMS as a way of inhibition.

To make the algorithm more suitable for classroom scenes, we do not use the average and maximum rescoring methods of traditional Seq NMS. Still, it carries out weighted rescoring based on frame spacing. At the same time, this paper combines soft NMS and seq NMS to improve traditional Seq NMS suppression steps. The improvement methods are introduced in detail below.

In this paper, the traditional Seq NMS rescoring algorithm is improved. The confidence U_n of the detection frame in the current frame is rescored as $0.8U_n + 0.1U_{n-1} + 0.1U_{n+1}$, which alleviates the false detection and missing detection caused by simply taking the average value, the erroneous detection caused by taking the maximum value and the missing detection caused by taking the minimum value. Still, it will be more dependent on the detection performance of the network.

This paper improves the traditional suppression algorithm of Seq NMS. Instead of using hard NMS as the suppression algorithm, this paper combines soft NMS and seq NMS.

3.4 Face count

In this paper, the sort algorithm is used to track the detection box to judge whether there are new students and students are leaving the classroom. With the help of the front and rear frame matching of the detection box, the counting function is realized, and the number of detection boxes judges the existing number in the classroom.

3.4.1 Sort tracking algorithm

Sort is a tracking method based on matching. The author approximates the motion of the target between frames as a linear motion independent of the motion of other objects and cameras. After obtaining the bounding box, the Kalman filter estimates the target motion state. Then the Hungarian matching algorithm is used to perform the linear motion of the position. The state of each target is expressed as x = [u, v, s, r, u ', v,' s'] t, where u and v represent the center of the detection frame, s means the size of the bounding box. R represents the length-width ratio of the bounding box, and

when associating the bounding box, the target is updated with the position of the bounding box. If there is no bounding box information, the linear model is used for prediction.

3.4.2 Counting algorithm based on classroom scene

Based on the sort algorithm, when the IOU between a target detected and the bounding box with all existing target prediction results in the selected frame is less than the specified threshold, it is considered that there is a new target to be detected. The location information of the new target is initialized with the bounding box information, and the speed is set to 0. New targets need to go through a period of undetermined time to correlate with the detection results to accumulate the confidence of new targets, effectively preventing the false alarm of target detection from creating new tracking targets.

The time interval of frame selection is enlarged to reduce the influence of occlusion in the classroom and adapt to the problem that the target in the classroom environment is relatively static. Select one frame every two seconds. When a target loses the matching frame in five adjacent frames, it is considered that the target disappears, and the face count decreases by one. On the contrary, when a target appears in five adjacent frames, it is assumed that a new target appears, and the face count increases by one.

4 Experience

4.1 Dataset

We use the dataset collected in the classroom Beijing Institute of Technology to train the model. The dataset is divided into three types by classroom size: large, medium, and trim. As shown in Figure 6. To ensure good data distribution, we sample the same number of frames in each classroom size. After extracting the optical flow data of each frame, we obtain the dataset for the video target detection task.

Fig. 6 Some pictures of classroom face dataset.

Due to the face dataset of the classroom being small, it is easy to be over-fitting when training large networks. This paper uses the widerface dataset for pre-training. Widerface dataset includes face data of various scenes, sizes, and nationalities. Also, it has a certain proportion of small target face data, as shown in Figure 7 which is consistent with the task of this paper. We use this dataset to improve the generalization ability of FFnet.

Fig. 7 Some pictures of widerface dataset.

To enable the algorithm to deal with both picture target detection and video target detection simultaneously, we use the image and video datasets to train the model. When using an image dataset, to fit the model's input shape, the image's optical flow information is set to **0**.

Because this paper uses a two-stage detection network, it is necessary to set the anchor parameters of the network structure carefully. This paper uses the K-means clustering algorithm to analyze the label box's area and aspect ratio data in the data set. Finally, it is obtained that the anchor size and aspect ratio are [25,45,70], [0.7,1.0,1.3] respectively.

4.2 Pretreatment

The dataset is expanded by image inversion, image rotation, and scaling before training to alleviate the lack of classroom face data and enhance the model's generalization. The first method is image flipping. Considering the application scene of FFnet, vertical flipping has no significance for the detection of classroom scenes, so we only flip the image horizontally. The second method is image rotation. We only

rotated the image at a small angle. The specific rotation angle value is set to {- 20, - 10, 10, 20}. After rotation, the image is inverted and filled. The third method is to zoom the image. While keeping the image size unchanged, zoom in, that is, cut the corresponding size at the center of the image, and fill the corresponding position of the image with a 0-pixel value when zooming out. The zoom ratio is {0.9, 1.1}.

4.3 parameter setting

Our experiment uses the PyTorch framework and relies on the commonly deep learning libraries such as the torch version and NumPy. The GPU of the experimental equipment is GTX 3090, and the video memory is 24GB.

This experiment involves two parts: first, the training of optical flow prediction network, and then the training of target detection network.

Firstly, this paper trains the optical flow prediction network. The results of extracting optical flow information are shown in Figure 8.

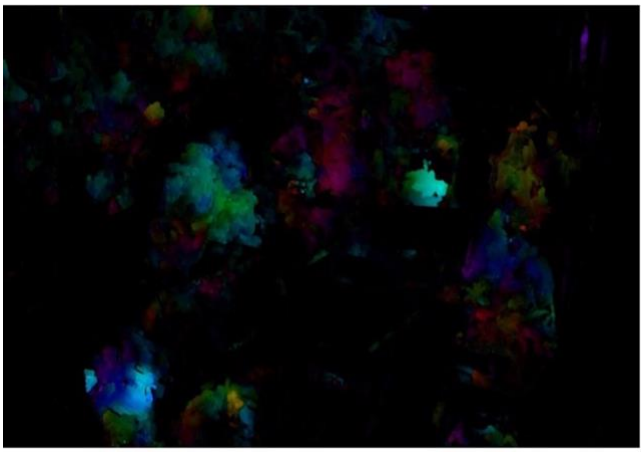

Fig. 8 The results of extracting optical flow information

In this paper, in the target detection network training process, the specific network super parameter details are as follows. When initializing the backbone network, select to import part of network model parameters pre-trained by the voc2007 dataset to accelerate the convergence speed. For the input layer, assign the pre-training values of the three channels of the pre-training model to the first three channels. The initial values of the last two channels are generated based on Gaussian distribution. The network's anchor scaling and aspect ratio parameters are shown above, and the size is set to [25,45,70]. Because this paper involves two data sets, it is necessary to train the network separately. Firstly, the widerface data set is used to supplement 0 matrices for optical flow information for 20000 iterations of training. Then the classroom face data set and video optical flow information are used for 20000 iterations of training. To balance the effect of image target detection and video target detection, 100 iterations of widerface without optical flow information and 100 iterations of classroom face data set with optical flow information are cycled, and the training is repeated and alternately trained ten times. The following figure 9 shows the loss decline curve during training.

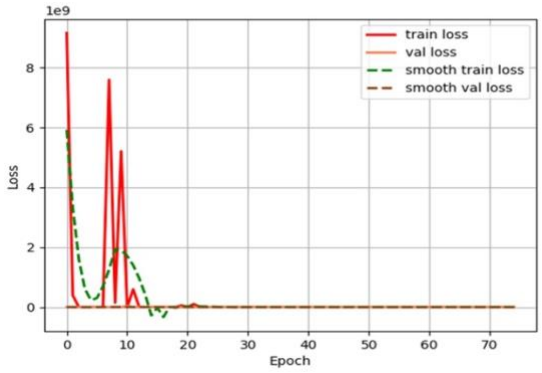

Fig. 9 The loss decline curve during training

4.4 Result

We use optical flow information to assist target detection to realize the function of face detection and counting in the classroom scene. The detection results and counting results are shown below.

4.4.1 detection result

Table 1 AP of several neural networks on widerface

	AP
Faster RCNN（resnet）	0.604
Faster RCNN（RPN）	0.731
FFnet	0.834

Table 2 Performance of several networks on single frame classroom face data set

	AP	**FPS**	**SIZE**
FFnet	0.912	10.6	198
Faster RCNN(FPN)	0.901	10.2	103
Faster RCNN(ResNet)	0.794	15.3	93

To evaluate the effect of the network model, we evaluate the network model from the aspects of accuracy, speed, and model size. In this paper, on the classroom face data set, the single-frame accuracy map reaches 0.912. Combined with the video optical flow information, the map reaches 0.932. The map on the widerface dataset reaches 0.834. The figure 4, 5 and table 1 show the test results on the widerface dataset. It can be considered that the network structure proposed in this paper has better detection performance for small targets and can extract more effective features for small targets.

Fig. 10 Widerface detection results of FFnet

Fig. 11 Widerface detection results of Faster RCNN(FPN).

Figure 10 shows the detection results of FFnet on the widerface. Figure 11 shows the detection results of fast RCNN using the FPN (feature pyramid networks). It can be seen from figures 10, 11, and Table 1 that the detection results of FFnet on a single frame image are relatively good after combining the optical flow information, and it also has better detection performance in small target detection. In contrast, faster RCNN has some missed detection. Because the model proposed in this paper needs to use the optical flow network to extract the optical flow information, which combines the target detection network and the optical flow extraction network, it makes a sacrifice in size.

It can be seen from table 2 that FFnet combines the optical flow extraction network and uses the network with more parameters for feature extraction, which sacrifices the model size. Still, the detection rate of a single frame image is slightly higher than that of fast RCNN using FPN. Because the image size of the data set in this paper is large (1920 * 1080), the speed of network processing pictures is slow. In addition, compared with the other two networks, the architecture of the FFnet network is more complex, so it does not have an advantage in model size and detection rate. In the index AP to measure the model's accuracy, FFnet is better than the traditional two-stage target detection network, and its performance is much better than the fast RCNN model using Resnet. One reason is that ffnet uses a backbone network with more parameters, and after using the optical flow information, it has been improved more obviously. The main reason is that some occlusion problems are alleviated by combining the semantic information between the front and back frames, improving the traditional NMS algorithm, and adopting a post-processing method more suitable for the classroom scene.

4.4.2 count result

This paper uses the output detection box and the counting algorithm suitable for counting classroom scenes. When the picture is input, the counting result equals the number of detection boxes. The counting result is calculated using the above counting algorithm when the continuous video frame is input.

Fig. 12 Face counting results of large classroom scenes.

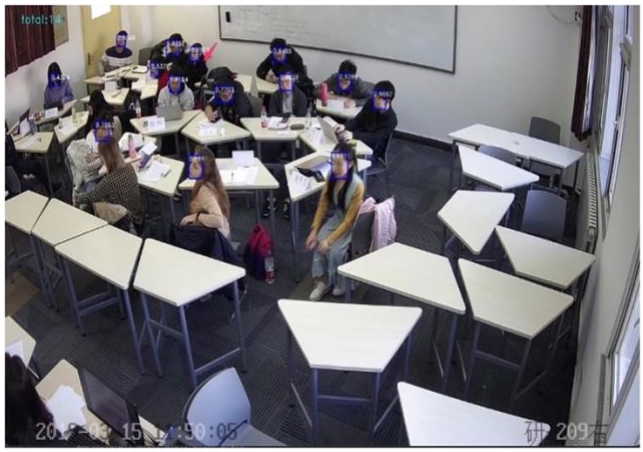

Fig. 13 Face counting results of tiny classroom scenes.

Figures 12 and 13 show the face counting effect of FFnet in the classroom scene and reflect the face detection effect. It can be seen from the figure that FFnet is also suitable for continuous video frames combined with optical flow information and face detection and counting in the classroom scene. It can be seen from figure 14 that the fast RCNN model using the FPN network still has a certain degree of missed detection on the class face dataset, while the FFnet network performs better in the classroom scenario. However, in large classrooms with smaller targets, some targets with severe occlusion cannot be detected, as shown in Figure 15.

Fig. 15 Small targets with severe occlusion.

Fig. 14 LEFT: The performance of fast RCNN (FPN) on class face dataset. RIGHT: The performance of FFnet on the class face dataset.

For small targets with severe occlusion, the following improvements are considered in subsequent experiments:

1. The original image is super divided to increase the feature information of small targets.

2. Face critical points are detected to alleviate the problem of local face occlusion.

3. The optical flow extraction network is improved to make the optical flow extraction more suitable for micromotion scenes in the classroom environment.

5 Conclusion

Based on processing RGB three-channel information of the image, we add two-dimensional optical flow information to enrich semantic knowledge between frames and relieve the detection problem caused by occlusion. By extracting optical flow information, the existing classroom face datasets are enriched. Referring to the soft NMS algorithm, we improve the Seq NMS algorithm to make it suitable for complex scenes with high target overlap in the classroom. We propose FFnet, which uses optical flow information to assist target detection. It achieves an accuracy of 0.912 on the classroom face dataset and can effectively detect faces in the classroom scene. Using a sort algorithm for tracking, a student counting algorithm suitable for the classroom scene is proposed to judge whether new students and students are leaving the classroom.

In the follow-up work, we will focus on further improving the detection accuracy, optimizing the algorithm model structure to make the algorithm more suitable for the classroom scene, and putting forward a new optimization scheme for the complex problem of dense optical flow calculation. In the next step, we plan to put forward a more efficient method for the target with severe occlusion to improve the detection performance of FFnet in practical application as much as possible.

References

1. A. Neubeck, L. Van Gool: Efficient Non-Maximum Suppression. 18th International Conference on Pattern Recognition (ICPR'06), 850-855(2006)
2. Goodfellow I J, Pouget-Abadie J, Mirza M, et al: Generative Adversarial Networks[J]. Advances in Neural Information Processing Systems, 2672-2680(2014).
3. Ren S, He K, Girshick R, et al: Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 28: 91-99(2015).
4. Cao J, Cholakkal H, Anwer R M, et al: D2det: Towards high quality object detection and instance segmentation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11485-11494(2020).
5. Bochkovskiy A, Wang C Y, Liao H Y M: Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:10934(2020).
6. Chen Q, Wang Y, Yang T, et al: You only look one-level feature[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13039-13048(2021).
7. Zhu X, Xiong Y, Dai J, et al: Deep feature flow for video recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2349-2358(2017).

8. Griffin B A, Corso J J: Depth from Camera Motion and Object Detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1397-1406(2021).
9. Nam H, Baek M, Han B: Modeling and propagating cnns in a tree structure for visual tracking[J]. arXiv preprint arXiv:1608.07242(2016).
10. Ashraf M W, Sultani W, Shah M. Dogfight: Detecting Drones from Drones Videos[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7067-7076(2021).
11. Yang T, Xu P, Hu R, et al: ROAM: Recurrently optimizing tracking model[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6718-6727(2020).
12. Jiang S, Lu Y, Li H, et al: Learning Optical Flow from a Few Matches[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16592-16600(2021).
13. Fischer P, Dosovitskiy A, Ilg E, et al: Flownet: Learning optical flow with convolutional networks[J].arXiv preprint arXiv:1504.06852(2015).
14. Zhao S, Sheng Y, Dong Y, et al: Maskflownet: Asymmetric feature matching with learnable occlusion mask[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6278-6287(2020).
15. Yang G, Ramanan D: Upgrading optical flow to 3d scene flow through optical expansion[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1334-1343(2020).