Check for updates

ACM DIGITAL LIBRARY | Association for Computing Machinery | acm open

Latest updates: https://dl.acm.org/doi/10.1145/3501409.3501529

RESEARCH-ARTICLE

Student Abnormal Behavior Recognition in Classroom Video Based on Deep Learning

HUAYONG LIU, Central China Normal University, Wuhan, Hubei, China

WEIDONG AO, Central China Normal University, Wuhan, Hubei, China

JINLIN HONG, Central China Normal University, Wuhan, Hubei, China

Open Access Support provided by:

Central China Normal University

PDF Download
3501409.3501529.pdf
01 February 2026
Total Citations: 17
Total Downloads: 163

Published: 22 October 2021

Citation in BibTeX format

EITCE 2021: 2021 5th International Conference on Electronic Information Technology and Computer Engineering
October 22 - 24, 2021
Xiamen, China

https://doi.org/10.1145/3501409.3501529
ISBN: 9781450384322

Student Abnormal Behavior Recognition in Classroom Video Based on Deep Learning

Huayong Liu†
School of Computer
Central China Normal University
Wuhan Hubei PR China
lhywuhee@mail.ccnu.edu.cn

Weidong Ao
School of Computer
Central China Normal University
Wuhan Hubei PR China
718475817@qq.com

Jinlin Hong
School of Computer
Central China Normal University
Wuhan Hubei PR China
2523882977@qq.com

ABSTRACT

For the complication and low speed of traditional human behavior recognition process, a method of student abnormal behavior recognition in classroom video based on deep learning was proposed. First of all, for the poor effect of small target identification with the original network, our method introduces an cascading improved RFB module by adding a branch to the RFB to increase the reference to the peripheral visual field. This network was called Rs-YOLOv3, and it can enhance the feature extraction capability of the original network and make full use of the shallow information to improve the identification effect of small targets. Secondly, for the character occlusion due to classroom structure and student density, the resn module of Darknet-53 in YOLOv3 was replaced with SE-Res2net module. The feeling field of each layer of the network can be increased to represent feature information in more fine-grained and realize multi-layer feature multiplexing. Finally, the border regression calculation was performed by changing the border loss function to DIoU_Loss. The experimental results show that the improved network SE-Res2Net-DIoU achieves 80.1% in the accuracy of student abnormal behavior recognition, a 5.8% improvement compared with the traditional YOLOv3, and reduces the missed recognition rate.

CCS CONCEPTS

• Artificial Intelligence • Machine Learning • Education

KEYWORDS

Behavior recognition, Deep learning, YOLOv3, Classroom video

1 Introduction

With the continuous development of science and technology, the application of surveillance cameras in daily life was gradually widespread, accompanied by the explosion of video data. How to reasonably and effectively use these video data and information has become a hot research topic.

In the school, some abnormal behaviors in the classroom will affect the quality of teachers, and reduce the learning efficiency of students, which brings work pressure to the teachers. In view of such problems, the relevant departments of the school often spend lots of money to supervise classroom students and arrange professionals to manage them. In addition, the school teaching evaluation also needs to take the students' class situation as a reference index, so some auxiliary means were needed to solve these situations. But some traditional practices have problems such as large information density and difficulty in energy allocation of duty staff[1]. The continuous development of artificial intelligence technology makes the application of deep learning in the field of image processing research, and because of its breakthrough in the field of image processing, it has become a popular research direction of image segmentation and recognition. Computer vision as an important branch of its work was being studied by most people, designed to use multi-layer convolutional neural networks for the extract of target's nonlinear features by mimicking the neuronal information transfer process in the human brain [2].

In recent years, many research scholars have published a large number of AI-based target recognition and detection models. It can be roughly divided into two categories: (1) behavior identification as the first task, establishing a sample database for abnormal posture or actions, and then determining the specific behavior by human target detection, attitude estimation, action recognition, and finally determining whether it belongs to the category of abnormal behavior; (2) abnormal detection as the first task, less considering specific abnormal actions, often to determine the abnormal scenes in video by similarity comparison with normal scenes. In China, similar researches began to rise in 2012, and gradually developed into a popular research field. Due to the late development, there was still a large room for improvement [3].

Given the breakthrough of deep learning, the YOLOv3 network was selected as the basic algorithm of this experiment, and optimized for the shortcomings of the original network, finally the student abnormal behavior recognition network was designed for classroom videos.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
EITCE 2021, October 22–24, 2021, Xiamen, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8432-2/21/10…$15.00
https://doi.org/10.1145/3501409.3501529

2 Self-built Classroom Video Dataset

This paper mainly studies the recognition of abnormal behavior in classroom video. Abnormal behavior refers to some behaviors that students do not meet teaching standards in class, such as sleeping, eating snacks and playing with mobile phones.

For this study, the public available online dataset was difficult to meet the needs of our experiment, so the material of this paper comes from the teaching process videos recorded by the school and other educational institutions around the campus. Dataset was made by recording the student class video and extracting the video images. With less data on abnormal behavior during class, some additional videos for abnormal behavior were recorded to supplement the dataset.

In addition, this paper uses the professional image labeling software LabelImg to label the dataset, and complete total of four behaviors: sleeping, eating, playing phone and listening.

2.1 Classroom Video Recording Process

In the experiment, the normal class environment was used to record in different classrooms and locations to expand the diversity of data, such as light source, classroom background and classroom style selection. Because students less eating in class were not conducive to model training, so the students were specially recorded for simulated video of eating in class.

2.2 Dataset Enhancement

In this experiment, 26 videos were collected. Due to the limited conditions, the length of video was different, 10 seconds or more than an hour. Considering the differences in shooting angle, light, character and distance, 22 videos were used as preliminary data, 4 of which did not participate in the training for testing, yielding a total of 53,084 images.

Students' location and behavior changes in class were too small, so it was likely to iterative learn a large number of similar images. Therefore, we selected a total of 2,679 images. However, these images, due to the single scene or small range of behavior changes, still cannot meet the needs of deep learning for the dataset. In order to make the trained network more robust and have a more effective performance, Keras platform was used to expand the dataset. It not only enriched the amount of the dataset, but also enabled the amount of individual behaviors to meet the experimental training needs.

The methods of image enhancement included flip, translation, scale, shear and contrast transformation. In addition, the pattern filling used in this experiment was nearest neighboring interpolation[4]. Some experimental results of image enhancement were shown in Figure 1.

(a)Original image **(b) Enhanced image**

Figure 1: Image Enhancement

After enhancing the image dataset, 16883 images were obtained, including training set with 7295 images, validation set with 4856 images and test set with 4732 images.

3 Student Abnormal Behavior Recognition Network

In order to recognize abnormal behavior for students in classroom videos, it requires high accuracy recognition in a specific background, but it was somewhat difficult to ensure high speed and high recognition rate in existing hardware conditions. Therefore, the network of YOLOv3, with fast speed and small memory was very suitable for the detection of abnormal behavior. However, due to the character occlusion caused by the scene and the poor small target detection effect of the traditional YOLOv3 itself, it will be improved to meet the needs of the content studied in this paper.

3.1 Traditional YOLOv3 Network

YOLOv3 combines some previous effective models and algorithms, adds multi-label classification content, draws on the resn residual structure, and builds a Darknet-53 feature extraction network, which greatly deepens the number of network layers. In addition to this, YOLOv3 combines the PFN (Feature Pyramid Networks) to constitute a target detection deep convolutional neural network in order to improve detection accuracy[5]. The backbone network structure was shown in Figure 2.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 2: Network Structure of Darknet-53

3.2 Rs-YOLOv3 Network for Enhancing Feature Extraction

The characteristics of end-to-end detection of the YOLOv3 algorithm can meet the requirements of speed, but the network has the disadvantage that the detection of small targets was poor. In our behavior detection, some tasks were more difficult, such as the image was too small due to the distance of the students, or the abnormal behavior related to the mobile phone was not detected due to its small size. In this paper, these problems were optimized based on the YOLOv3 network model by cascading the improved RFB module into the YOLOv3, and the improved RFB module structure was shown in Figure 3.

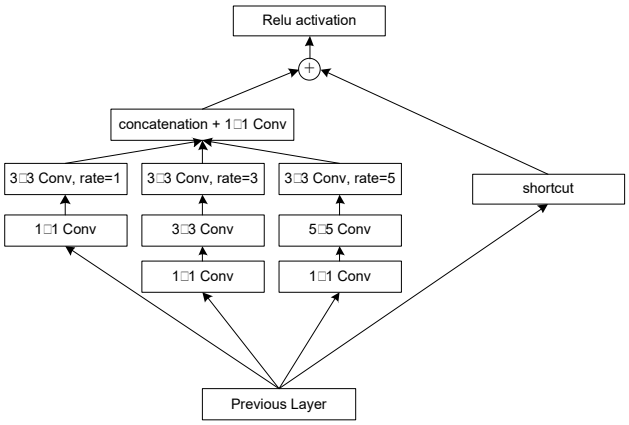


Figure 3: Improved RFB Module

The network structure designed for the recognition of abnormal behavior of students in classroom videos was shown in Figure 4 and named Rs-YOLOv3, to enhance the feature extraction capability of the original model, improve the identification of small targets and adapt them to the needs of this experiment.

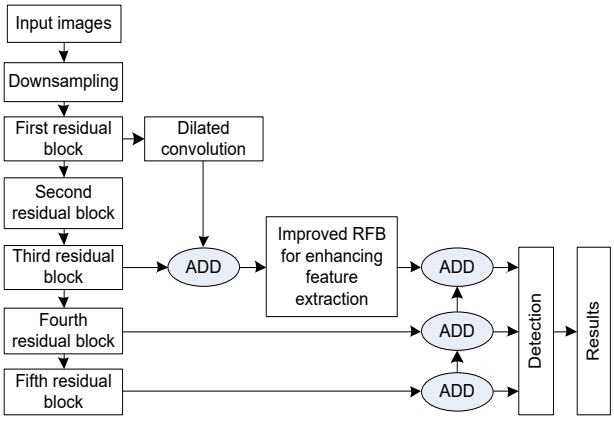


Figure 4: Rs-YOLOv3 Overall Network

In this paper, the feature map output from the first residual block of the original network was dilated convoluted with the 8 times down-sampled feature map and cascades the improved RFB modules for feature enhancement. The information of small target features in shallow layers was fully utilized to improve the detection ability of the original network for small targets. In addition, the maximum pooling operation was replaced by the 3×3 convolution layer with step size 2 to make more full use of the feature information in the pyramid.

3.3 SE-Res2Net module based on YOLOv3 Network

To solve the problem of behavior missed detection due to human and loss of feature information caused by task occlusion when the YOLOv3 network performs student abnormal behavior detection, Res2Net integrating SE (Squeeze and Excitation) will be used[6]. The structure was improved and optimized to reduce the occurrence of feature loss phenomenon.

The backbone network of YOLOv3 was Darknet-53, which referencing the ResNet structure, consisting of a 1×1 convolution and a 3×3 convolution. The structure was shown in Figure 5.

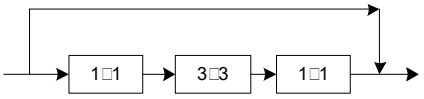


Figure 5: ResNet Basic Structure

To expand the receptive field of each network, the Darknet-53 resn module was replaced with the SE-Res2Net module to optimize the YOLOv3 network. The SE-Res2Net module structure was shown in Figure 6.

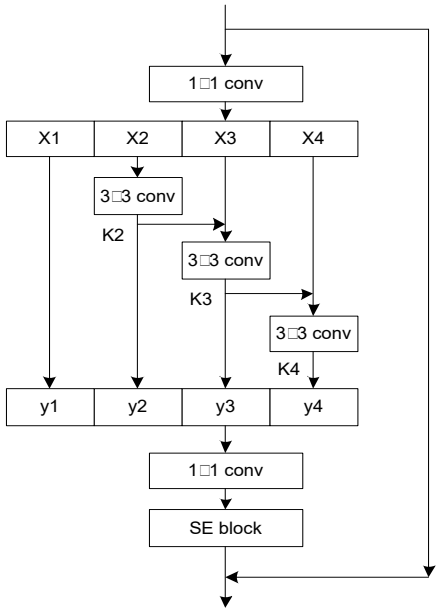


Figure 6: SE-Res2Net Structure

The SE-Res2Net first was processed by a convolutional 1×1 core, and then divided into four feature sub-graphs by channel X1, X2, X3 and X4, which representing different spatial dimensions. Then four outputs, y1, y2, y3 and y4 were obtained by convolution operations respectively. The 3×3 convolution under the four feature sub-graphs of SE-Res2Net makes use of previous features

and yields larger receptive fields. The SE-Res2Net introduces new dimension scales. The output dimensions were divided into n groups by grouping convolution. After convolution, concatenate and keep the input and output have the same dimension. Finally, the SE module was added after the 1×1 convolution, integrating the depth and local information through the local receptive field from the convolutional channel features, assigning weights to each channel adaptively to express the influence of different channels on the output results. By increasing the scale within the single layer, expanding the scope of the receptive field, making more full use of the context information, making it easier for the classifier to achieve target detection.

3.4 Improvement of the YOLOv3 Loss Function

3.4.1 Loss Function of the YOLOv3. The expression for the loss function of YOLOv3 was shown in formula (1), where λ_{coord} and λ_{noobj} are the balance coefficients, *S* represents the size of the feature diagram, and *B* represents the number of anchor boxes, I_{ij}^{obj} determining whether the *j* anchor box in the *i* grid was responsible for the target or not (0 or 1). x_i, y_i, w_i, h_i and $\hat{x}_i$, $\hat{y}_i$, $\hat{w}_i$, $\hat{h}_i$ represent the position coordinates and width/height of the anchor box and the real box, respectively. C_i^j and $\hat{C}_i^j$ represent the predicted value and real value of the target in the anchor box respectively. p_i^j and $\hat{p}_i^j$ represent the predicted value and real value of the classes respectively[7].

$$
\begin{aligned}
Loss = \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(x_i^j - \hat{x}_i^j\right)^2 + \left(y_i^j - \hat{y}_i^j\right)^2 \right] \\
+ \lambda_{coord} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\left(\sqrt{w_i^j} - \sqrt{\hat{w}_i^j} \right)^2 \right. \\
\left. + \left(\sqrt{h_i^j} - \sqrt{\hat{h}_i^j} \right)^2 \right] \\
- \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\hat{C}_i \log\left(C_i^j\right) \right. \\
\left. + \left(1 - \hat{C}_i\right) \log\left(1 - C_i^j\right) \right]^2 \\
- \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{noobj} \left[\hat{C}_i \log\left(C_i^j\right) \right. \\
\left. + \left(1 - \hat{C}_i\right) \log\left(1 - C_i^j\right) \right] \\
- \sum_{i=0}^{s^2} I_{ij}^{obj} \sum_{c \in classes} \left[\hat{p}_i^j \log p_i^j \right. \\
\left. + \left(1 - \hat{p}_i^j\right) \log\left(1 - p_i^j\right) \right]
\end{aligned}
\tag{1}
$$

The sum of square of the difference was used for the error loss function of central and width-height coordinates. The binary cross-entropy loss function was used for behavioral confidence error, while 0 indicates absence and 1 indicates existing. The classification error was also calculated by using the cross-entropy loss, which representing the true class *j*th target in the predicted target boundary box i. Because the same target can be classified as multiple classes, this method can cope with more complex scenarios. By analyzing the deficiency of YOLOv3 loss function, it was replaced with the more accurate DIoU loss function, which increasing the accuracy of abnormal behavior recognition in the classroom scene and the detection results were closer to the real situation.

3.4.2 The DIoU_Loss with the Penalty Term. YOLOv3, calculating positional coordinate error, simply assumes the bounding box as four independent variables, using the sum of square of the difference as loss function. But regression of the bounding box from four points alone has disadvantages, it was unable to accurately describe the IoU relationship between borders and had no scale invariance[8]. To better represent the overlap relationship with the predicted and bounding boxes, border regression was performed with IoU with the formula as follows.

$$\mathrm{IoU} = \frac{\mathrm{M} \cap \mathrm{N}}{\mathrm{M} \cup \mathrm{N}} \tag{2}$$

$$\mathrm{IoU\_Loss} = 1 - \mathrm{IoU} \tag{3}$$

M refers to the coordinates of the predicted box, and N was the real box coordinate. The loss function can directly reflect the distance between the real box and the predicted box. However, it can only be reflected where there was an overlap. If there was no common area in the two boxes, the value of IoU was 0, and reverse gradient update cannot be performed, then the network cannot be trained. There was also a case that when the true box, the predicted box, and the intersecting part were determined, the true intersection of the two boxes cannot be reflected[9].

In practical scenarios, when two different objects were very close, only one detection box remains after NMS processing due to the relatively large IoU value, missing detection occurs. DIoU does not only consider the distance between the target and the bounding box, but also the scale and overlap rate. In cases where the IoU was larger between two boxes and the distance between the two boxes was larger, it may be considered the box of two objects without being filtered out. So this paper uses DIoU for border regression to make the calculation more reasonable, as shown in equation (4).

$$\mathrm{DIoU\_Loss} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} \tag{4}$$

In the above formula, a penalty term was added to the original IoU. b and b^{gt} represents the center point of the predicted box and the real box, respectively. ρ represents the euclidean distance of the two midpoints, and *c* represents the diagonal distance of the minimum closure region capable of containing both the predicted box and the real box, as shown by the green line. The model diagram was shown in Figure 7.

The loss function has no relation to the scale of regression, which can still provide the movement direction to the boundary box when there were no overlaps between the boundary box with the target box. The distance between the two target boxes can be directly minimized, which making the network convergence faster. Not only that, the regression was fast in the horizontal and vertical cases. Our method of replacing the YOLOv3 border regression loss function with DIoU_Loss, not only refers to the distance between the real box and the predicted box, but also makes the regression of the target box more stable, considering the overlap rate and

aspect ratio. Our method is called SE-Res2Net-DIoU, and it solves the problem of network training divergence, makes the prediction box positioning more accurate, increases the convergence speed of the network, and thus improves the positioning accuracy of the model.

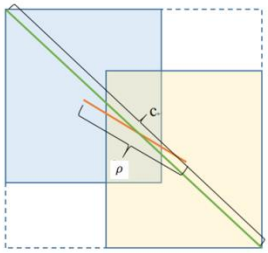


Figure 7: Target Bounding Box Predicted Map

4 Experimental Results and Analysis

4.1 Selection of dataset and the performance indicator settings

In addition to the self-built dataset, we also test the public dataset PASCAL VOC 2007+2012, which contains 20 categories and sufficient annotation information for many experimental tasks of image processing. There were 9963 images in VOC2017, while 5011 images in training set and validation set, and 4952 images in test set. There were 11540 images in the validation set while no test set in VOC2012.

The performance of the network was measured by the mAP of behavioral recognition[10], it was calculated as shown in formula (5).

$$\text{mAP} = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{5}$$

Where C represents the number of categories, and AP represents the average accuracy of each category.

4.2 Network Training

The original YOLOv3 network generated the anchor box by clustering on the COCO dataset, but this paper requires PASCAL VOC, it was necessary to re-cluster to obtain the optimal value.

Given the small size of the self-built dataset, it takes a great effort to train the dataset directly using random initial network weights with gradient dispersion. Therefore, during actual training, the original YOLOv3 network and the designed network were pre-trained on the PASCAL VOC, so that initial weights of the model were as close to the optimal value of the training as possible to prevent the gradient dispersion. After 30,000 iterations of training, the accuracy was 78.2% on the test dataset, and the 48.3 FPS speed was reached. Table 1 constructs the performance of our network Rs-YOLOv3, the original network and several other deep convolutional neural networks on the VOC dataset.

Table 1: Performance of Different Algorithms on PASCAL VOC Set

Algorithm	Rs-YOLOv3	YOLOv3	Faster-RCNN	Fast-YOLO
mAP (%)	78.2	75.4	71.2	53.7
Speed (FPS)	48.3	47.6	4.2	45.8

The key point of our paper was the target behavior detection, thus the Rs-YOLOv3 network was tested on public dataset. Ten behaviors in the PASCAL VOC dataset were selected to experiment on our improved model, containing both jumping, reading, phoning, playing instrument, riding horse, using computer, walking, running, riding bike and taking photo. The results were shown in Table 2.

Table 2: Average Accuracy (%) of Behavior Recognition under Different Algorithms on Pascal VOC Behavior Test Set

Behavior	Oquab[11]	Hoai[6]	ResNet-50[8]	Rs-YOLOv3	SE-Res2Net-DIoU
jumping	74.8	82.3	88.9	76.3	76.5
reading	45.3	53.6	75.9	72.8	73.7
phoning	46	52.9	75.2	69.8	67.9
playing instrument	75.6	84.3	91	83.2	83.1
riding horse	95	96.1	97.5	94	96.5
using computer	66.7	76	84.9	89.8	89.3
walking	69.5	72.9	53.9	73.1	74.5
running	86.5	89.7	81.1	92.3	91.7
riding bike	93.5	95.6	95.8	93.8	92.6
taking photo	49.3	60.4	69.5	58.9	60.3
mAP	70.2	76.3	81.4	80.4	81.2

It was seen from the table that the mAP values of PASCAL VOC in Rs-YOLOv3 networks were not very different from the mAP tested in other algorithms and were thus suitable for behavioral detection. It can also be learned that the Rs-YOLOv3 and SE-Res2Net-DIoU network proposed in this paper was flat or even better than the other algorithms in terms of each behavior recognition rate. While the overall recognition rate may be lower than some algorithms, its speed of computing images was close to 50 FPS, while the others were mostly below 10 FPS.

The two pre-trained weights were applied to training in the self-built dataset of this experiment. It was noteworthy that the last layer should be modified according to the number of behavior categories. A total of 40 epochs were used in the training process. Batch sizes were set to 1. The initial learning rate was set to 1e-4, and gradually decreasing, while a minimum of not less than 1e-6. The optimized network model was obtained by training of up to 13 hours, 50,000 times.

The loss function image of this training experiment shows the large changes before 10000 times, gradually stabilized after 20000 times, and basically unchanged after 50000 times, so the number of experimental iterations in this paper was 50000. Since there were some unavoidable problems in the dataset production process,

such as some differences in the behavior of the same class, the loss value was impossible to eventually drop to 0[12].

4.3 Results and Analysis on Self-built Dataset

To demonstrate that the network proposed in our paper was suitable for student abnormal behavior detection in classroom videos. It was determined to test the YOLOv3, Rs-YOLOv3 and SE-Res2Net-DIoU networks on the test set of the self-built dataset respectively, and to compare and analyze the experimental results.

Applying the YOLOv3 network to the student behavior dataset produced in this paper, the model average accuracy can reach 74.3%, and the average accuracy of various categories was shown in Table 3, and the accuracy of each behavior was listed as follows: playing phone 71.4%, eating 80.2%, sleeping 89%, and listening 56.3% respectively.

Applying the Rs-YOLOv3 network to the student behavior dataset produced in this paper, the final model average accuracy can reach 78.9%, and the average accuracy of various categories was shown in Table 3. Among these abnormal behaviors, the recognition rate of sleeping was highest, reaching 93.7%, while the recognition rate of playing phones and listening was low. The possible reasons included insufficient data annotation or complex behavioral characteristics.

From Table 3, it can be seen that SE-Res2Net-DIoU with the improvement of the loss function is effective, and the average model accuracy can reach 80.1%.

Table 3: Average Accuracy (%) of Abnormal Behavior Recognition under Different Algorithms in Self-built Dataset

Behavior	YOLOv3	Rs-YOLOv3	SE-Res2Net-DIoU
playing phone	71.4	75.5	73.2
eating	80.2	84.4	85.6
sleeping	89.0	93.7	92.3
listening	56.3	61.9	69.2
mAP	74.3	78.9	80.1

It can be seen from the two results (a) and (b) in Figure 8 that most of the students in the front row are detected by YOLOv3, and the students with distant lens and small imaging in the back row are not detected. Therefore, it can be seen that the detection effect of the original network on small targets is poor and the missed detection is serious.

In view of the poor detection of small targets by YOLOv3 , the test results on Rs-YOLOv3 after enhanced feature extraction are shown in Figure 9.

As shown in Figure 9 (a), the number of students detected has increased significantly, and the students in the back row have been detected. In such a large classroom as Figure 9 (b), the behavior detection of students has also been realized.

(a) **(b)**

Figure 8: Results of YOLOv3

(a) **(b)**

Figure 9: Results of Rs-YOLOv3

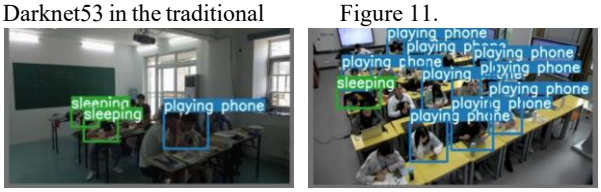
The network can achieve high recognition rates in scenarios with small numbers and poor occlusion. However, in real scenes, with the increase of the number of people and the occlusion, the recognition rate decreased, and the number of miss-detected students also increased.

As shown in Figure 10, in the scene (a), many students are blocked, so many behaviors are not recognized; The same is true in (b). There should have been another student sleeping behind the sleeping student, but it could not be detected because of the occlusion.

(a) **(b)**

Figure 10: Results of Rs-YOLOV3 under Occlusion

For the above problems, the SE-Res2Net-DIoU was proposed to optimize the residual structure of the Darknet53 in the traditional YOLOv3 network, and the experimental results were shown in Figure 11.



(a) **(b)**

Figure 11: Results of SE-Res2Net-DIoU under Occlusion

It can be seen from Figure 11 (a) that another sleeping student behind first sleeping students is detected under the new model. At the same time, when there are many people in the classrooms in Figure 11 (b), more students' abnormal behaviors can still be recognized.

5 Conclusions

The continuous development of surveillance video has made the research of intelligent video receive more attention, such as the abnormal behavior detection of characters in some public places. Abnormal behavior detection through the computer ability to deal with massive information, determine whether there was a target of interest in the specified picture or video, if any, its behavior was detected and its location and behavior were judged. This paper designed a deep learning-based convolutional neural network for the detection of abnormal behavior in the school classroom.

Because the scale of the data set was not large enough, the condition constraints, too tedious video shooting task and data annotation task, failed to be multi-angle recording of each scene, the statistics of abnormal behavior was not comprehensive enough, and the training network can not reach the ideal experimental effect. Therefore, it needs to be more collected and label more abnormal behavior to meet the needs of realistic needs. Moreover, this paper sacrifices the speed of a network. In reality, image processing, target detection can be faster. In the later research, we should focus on ensuring the accuracy improvement, while the speed should also meet the needs. Finally, although this paper puts forward a solution to the character occlusion problem, in the number of students and serious occlusion, such situation was further studied to improve the recognition rate of the network and reduce the missing rate.

The current research on human behavior testing has not been perfect, and more researchers were still needed to explore and innovate. It was believed that in the subsequent development, more and more people invest in research in this field, and new breakthroughs in the field of behavior detection to provide more practical technology for building intelligent classroom.

ACKNOWLEDGMENTS

Foundation support: Staged research achievement for “Research on Mobile Visual Search Method of Digital Library Based on Deep Learning”, funds from the Humanities and Social Sciences of China MOE (21YJA870005). And this work is also supported by the Fundamental Research Funds for the Central Universities (CCNU20TS027).

REFERENCES

[1] Dongyin Liu. 2018. *Research on Detection of Abnormal Behavior in Classroom Monitoring Video*. University of Electronic Science and Technology of China, Chendu.

[2] Samitha Herath, Mehrtash Harandi, Fatih Porikli. 2017. Going Deeper into Action Recognition: A Survey. Image and Vision Computing, 60, 4-21.

[3] Xuewei Su. 2019. *Research on Human Abnormal Behavior in Video Surveillance Based on Deep Learning*. Xi’an University of Science and Technology, Xi’an.

[4] Adeshina Sirajdin Olagoke, Ibrahim Haidi, Teoh Soo Siang, Hoo Seng Chun. 2021. Custom Face Classification Model for Classroom Using Haar-Like and LBP Features with Their Performance Comparisons. Electronics, 10(2).

[5] Wang Z, Mirbozorgi S A, Ghovanloo M. 2015. Towards a Kinect-based Behavior Recognition and Analysis System for Small Animals. IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE.

[6] Cheng Xingliang, Xu Mingxing, Zheng Thomas Fang. 2020. A Multi-branch ResNet with Discriminative Features for Detection of Replay Speech Signals. APSIPA Transactions on Signal and Information Processing, 9(28).

[7] Liu Tao, Pang Bo, Ai Shangmao, Sun Xiaoqiang. 2020. Study on Visual Detection Algorithm of Sea Surface Targets Based on Improved YOLOv3. Sensors, 20(24).

[8] Zhao Lei, Yang Fei, Bu Lingguo, Han Su, Zhang Guoxin, Luo Ying. 2021. Driver Behavior Detection via Adaptive Spatial Attention Mechanism. Advanced Engineering Informatics, 48.

[9] Liu X Q, Sun S L. 2018. Research on Abnormal Behavior Detection based YOLO Network. Electronic Design Engineering.

[10] Moechammad S, Cahya R, Berkah A N A. 2021. Detecting Body Parts from Natural Disaster Victims using You Only Look Once (YOLO). IOP Conference Series: Materials Science and Engineering, 1073(1).

[11] Jiapei FENG, Xinggang WANG, Wenyu LIU. 2021. Deep Graph Cut Network for Weakly-supervised Semantic Segmentation. Science China (Information Sciences), 64(03), 57-68.
[12] Le Wang, Hongxia Le, Wenjing Li, Mehan Zhang. 2021. Mask Detection Algorithm Based on Improved YOLO Lightweight Network. Computer Engineering and Applications. 57(8).