Hindawi
Mobile Information Systems
Volume 2022, Article ID 9903342, 8 pages
https://doi.org/10.1155/2022/9903342

Research Article

Identifying and Monitoring Students' Classroom Learning Behavior Based on Multisource Information

Chuck Chung Yin Albert,[1] Yuqi Sun,[2] Guang Li,[3] Jun Peng,[4] Feng Ran,[5] Zheng Wang,[6] and Jie Zhou[4]

[1]*Faculty of Data Science, City University of Macau, Macao, China*
[2]*Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence (Ministry of Education PRC), Macao Polytechnic Institute, Macao, China*
[3]*Faculty of Data Science, City University of Macau, Macao, China*
[4]*School of Education, City University of Macau, Macao, China*
[5]*Beijing Dongcheng Academy of Educational Sciences, Beijing, China*
[6]*Shandong Youth University of Political Science, School of Information Engineering, Jinan, China*

Correspondence should be addressed to Jun Peng; d19092105030@cityu.mo

Received 17 March 2022; Accepted 9 May 2022; Published 25 August 2022

Academic Editor: Mian Ahmad Jan

Understanding human activity and behavior, particularly real-time understanding in video feeds, is one of the most active areas of research in Computer Vision (CV) and Artificial Intelligence (AI) nowadays. To advance the topic of integrating learning engagement research with university teaching practice, accurate and efficient assessment, and analysis of students' classroom learning behavior engagement is very important. The recently proposed classroom behavior recognition algorithms have some limitations, such as the inability to quickly and accurately identify students' classroom behaviors because they do not consider the motion information of students between consecutive frames. In recent years, action recognition algorithms based on Convolutional Neural Networks (CNN) have improved significantly. To address the limitations of existing algorithms, in this study, a 3D-CNN is selected as a network model for classroom student behavior recognition, which increases information multisourcing and classroom student localization with high accuracy and robustness. For better analysis of human behavior in videos, the 3D convolution extends the 2D convolution to the spatial–temporal domain. In the proposed system, first of all, a real-time picture stream of each student is obtained by combining real-time target detection and tracking. Then, a deep spatiotemporal residual CNN is used to learn the spatiotemporal features of each student's behavior, so, as to achieve real-time recognition of classroom behaviors for multistudent targets in classroom teaching scenarios. To verify the effectiveness of the proposed model, different experiments are conducted using the labeled classroom behavior dataset. The experimental results demonstrate that the proposed model exhibits better performance in classroom behavior recognition. The accurate recognition of classroom behaviors can assist the teachers and students to understand the classroom learning situation and help to promote the development of smart classroom.

1. Introduction

Artificial intelligence (AI) is an emerging field of computer science that investigates and develops ideas, techniques, technologies, and application systems for imitating and enhancing human intelligence. In recent years, the application of AI technology has flourished in the field of education. The AI development plans released by many countries mention the need to use intelligent technology to accelerate the reforms of talent training mode and teaching methods, which purpose is to build a new education system that includes intelligent learning and interactive learning. Its major goal is

to build intelligent campuses to encourage the use of AI throughout the teaching, administration, and resource development processes [1]. In recent years, Convolutional Neural Network (CNN) has become the standard technique for a wide range of computer vision applications, including image classification, object identification, semantic segmentation, and recognition of human actions [2].

Expressions and behaviors are the two parts of a student's listening status. Positive expressions and actions indicate that the students are actively engaged in classroom teaching, aware of the teacher's progress, and examining the teaching contents. However, the primary listening status must be determined by an analysis of students' behaviors. If a student engages in undesirable behaviors such as napping or turning his/her head, it means that he/she is perplexed or bored with the instructional material. Moreover, students' classroom behavior change over time, so we need to observe students' changes in listening behaviors in real time to provide timely feedback on the teaching effect. After the lesson, teachers can examine students' classroom behavior in recorded classroom videos to assist them modify their teaching approaches and obtain better teaching results. How to better identify and analyze students' classroom behaviors more effectively has become a research focus of smart education [3].

Classroom behavior analysis aims to study the mechanisms underlying teachers' teaching activities and students' academic development in the classroom, and to help teachers and students reflect on their classroom performance, so, as to promote the improvement of classroom teaching quality. Traditional classroom teaching behavior analysis approaches rely on self-evaluation, manual monitoring, and manual coding to gather and interpret the data. Traditional approaches have disadvantages like strong subjectivity of coding, small sample size, time-consuming, and leading to their interpretability which leads to low interpretability and scalability [4, 5].

One of the most fundamental functions of computer vision (CV) is video interpretation and identification, and recognizing actions in videos is one of the challenging and practical tasks. The content and background of videos are more complicated and varied than those of images. Further, motion recognition is also difficult due to occlusion, shaking, illumination, and perspective changes caused by shooting. Feature extraction, fusion, coding, and classification stages are used in traditional video action recognition algorithms [6]. Gradient histogram, spatiotemporal interest point detection, and optical flow histogram are widely used to extract feature representations of images and temporal sequences in traditional methods, but these features often lack flexibility and scalability [7]. To better utilize video timing features, Simonyan and Zisserman [8] proposed a dual-stream-based fusion neural network, which divides the video into "spatial and temporal" parts, and feeds RGB and optical stream images into two neural networks respectively. First of all, the video was divided into two parts: "spatial and temporal", and the RGB and optical stream images were fed into two neural networks and the final classification results were fused. After that, many researches have improved and optimized the framework of dual-stream network. For example, Wang et al. [9] proposed a network structure that can capture longer time sequences. Similarly, Xu et al. [10] proposed a framework based on dense expansion networks and explored different ways of integrating spatial and temporal branches.

Due to the success of deep learning (DL) approaches in multitarget behavior identification, a series of video behavior recognition methods based on DL have been developed. Tran et al. [11] proposed 3D Convolutional Network (3D-CNN) for behavior recognition to extract spatiotemporal features of behaviors, which can identify behaviors more accurately. Cao et al. [12] combined the inflated 3D Convolutional Network (I3D) algorithm with a dual-flow method for better recognition of students' behavior. Guan [13] used multiperson behavior recognition algorithms for the video domain based on the Real-time Multi-Person 2D Pose approximation (OpenPose) method, which can identify the position of human nodes to estimate the human pose. Some other scholars use the OpenPose algorithm to extract the location of human joint points, then use the Sort algorithm to monitor the target and compute the offset position of the joint points to obtain the human behavior category [14, 15].

There is no effective solution for online multiperson behavior detection at the present, but, the single-person behavior recognition research results are outstanding, and the residual network and convolutional network show their powerful ability on computer vision tasks. So, the main approach used in this paper is to apply the single-person behavior recognition algorithm based on a deep spatiotemporal residual CNN to the multitarget classroom teaching scenario. In a classroom teaching scenario, the behavior recognition algorithm solves the student classroom behavior recognition problem. Meanwhile, other effective video action recognition methods are being explored, such as the long short-term memory (LSTM) based recognition framework and generative adversarial neural network (GAN) based framework for video action recognition.

The rest of the paper is organized as; section 2 shows the related work of the proposed study; section 3 demonstrates the material and methods; section 4 illustrates the results and discussion. Finally, the research work is concluded in section 5.

2. Related Work

The continuous development of AI has brought great changes in people's lives and gradually entered all aspects of work and life. Smart cities, smart offices, and smart medical care all are concepts of AI that are fast emerging and developing. Advances in image recognition systems, which were frequently modified and expanded to deal with video data, have stimulated video recognition research. The computer vision community has conducted various studies on video and action recognition in the past few years. Various challenges have been offered over the years, including action identification, anomaly detection, video retrieval, event and action detection, and many others [16]. By

extending Harris corner detectors to three dimensions (3D), Dollar et al. [17] proposed spatiotemporal interest points (STIPs). Laptev and Lindeberg [18] proposed Cuboids features for behavior recognition. ActionBank was created by Sadanand and Corso to recognize actions [19]. Wang and Schmid [20] recently presented an improved Dense Trajectories (iDT), which is the state-of-the-art hand-crafted feature at the moment. The iDT descriptor is an excellent illustration of how temporal signals and spatial signals might be treated differently. Rather than extending the Harris corner detector into 3D, it starts with densely sampled feature points in video frames and tracks them using optical flows. Along the trajectory, distinct hand-crafted characteristics are retrieved for each tracker corner. Though this method gives high performance, but this method is computationally expensive and becomes unsolvable on large datasets. Convolutional neural networks (ConvNets) have made advancements in visual identification with the recent availability of powerful parallel machines (GPUs, CPU clusters) and vast volumes of training data [21]. ConvNets have also been used to solve the challenge of estimating human pose in images and videos [22].

2.1. Convolutional Neural Networks (ConvNets) for Action Recognition. Several researchers have attempted to create an efficient ConvNet structure for video action recognition [23]. On a large dataset (Sports-1M), Karpathy et al. [24] evaluated ConvNets with deep structures. By utilizing the ImageNet dataset for pretraining and calculating optical flow to explicitly capture motion information, Sun et al. [25] created two-stream ConvNets including spatial and temporal nets. Varol et al. [26] have used 3D ConvNets to learn both appearance and motion attributes using 3D convolution operations on realistic and large-scale video datasets. Pirsiavash and Ramanan [27] suggested a factorized spatiotemporal ConvNets that uses several decomposition methods for 3D convolutional kernels. Several recent studies have employed ConvNets to represent long-range temporal structure [28]. These approaches, on the other hand, worked directly on longer continuous video streams. These algorithms typically process sequences of predetermined durations ranging from 64 to 120 frames due to computational costs. Due to their limited temporal coverage, learning from a complete video is difficult for these systems. Our technique differs from these end-to-end deep ConvNets in the way that, it employs a unique sparse temporal sampling strategy, which allows efficient learning of entire films without regard to the sequence length.

2.2. Temporal Structure Modeling. The identification of human actions in the video is studied using temporal structure modeling. Modeling the temporal structure for action recognition numerous research works have been accomplished [29]. For example, Wang et al. [30] annotated each video's atomic actions and developed the Actom Sequence Model (ASM) for action identification. Sun et al. [31] suggested using latent variables to represent the temporal breakdown of complicated activities and used the Latent SVM to iteratively learn the model parameters. Similarly, He et al. [32] created a sequential skeleton model (SSM) to represent the dynamic-poselet relationship and conducted spatiotemporal action detection.

In this research work, we have designed and implemented an intelligent teaching assessment system based on student classroom behavior recognition in classroom teaching videos, which relies on video understanding technology to realize the recognition of student classroom behavior. Further, this approach can make a comprehensive assessment of students' classroom status and help classroom intelligent education. Behaviors are classified into three categories such as positive behaviors (standing and listening), neutral behaviors (head-down), and negative behaviors (head-turning). By counting the results of student classroom behavior recognition the proportion of positive, neutral, and negative behaviors is calculated. If the proportion of negative behaviors exceeds a certain value, feedback will be given to the teacher to issue an alert.

3. Material and Methods

The material and methods of the proposed research work are briefly described in the following subsections.

3.1. Dataset Collection of Student Behavior. The raw data used in this study is about 900 min of classroom videos taken in the field. To facilitate the behavior recognition work, the selected behaviors have more distinctive features and do not have ambiguous or overlapping data. In this paper, we collect 8000 pictures of students doing hand raising, 500 pictures of students bending down, 8000 pictures of students walking back and forth, 12000 pictures of students writing on the blackboard, 3000 pictures of students looking up, 3000 pictures of students looking down, 1000 pictures of standing students, and 1000 pictures of students lying on their desks.

3.2. Proposed Framework for 3D-CNN-Based Students' Behavior Recognition Model. CNN is a class of feed-forward neural networks with a deep structure including convolutional computation, which is one of the representative algorithms of DL. It has been widely used in the computer vision field because of its advantages in feature extraction and processing. In the field of image processing, CNNs are commonly dealing with a variety of complex situations like human pose, lighting situation, and complex background, etc. [33]. However, when extended from 2D image processing to video image processing, traditional CNNs using 2D convolutional kernels for feature extraction have some limitations for both preservation and processing of the 3D (temporal dimension) information. To solve such problems, 3D-CNNs with 3D convolutional kernels for feature extraction have been developed. The recognition of students' behaviors in the classroom video needs to consider the motion information between consecutive frames. So, in this paper 3D-CNN is selected for students' behavior recognition as a network model. The 3D-CNN network architecture used in this paper consists of a video stream of eight consecutive frames with an image resolution of 224×224. The input images are

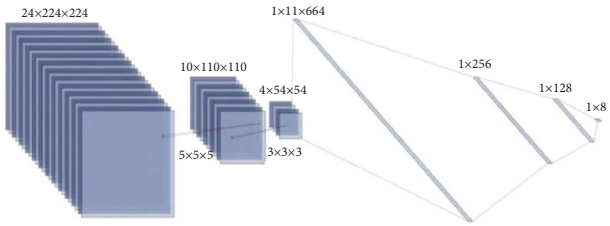

Figure 1: Network structure of 3D-CNN.

convolved by two 3D convolution kernels of $5 \times 5 \times 5$ and $3 \times 3 \times 3$ respectively, and the step size of the convolution operation is set to 2, thus reducing the dimensionality of the original input to $4 \times 54 \times 54$. After the final unfolding, the 1×8 feature vector becomes the output of three fully connected layers. Feature vector characterizes the video stream into eight categories of student behavior. The network structure of 3D-CNN is shown below in Figure 1.

3.2.1. Framework of YOLO-v5 Algorithm for Multitarget Detection and Behavior Recognition. This study also uses the YOLO-v5 algorithm as a framework for multitarget detection and behavior recognition. In the process of students' classroom behavior model construction, the teaching classroom video stream data is first scaled and regularized. After that, the features are extracted using a CNN model, followed by a fusion of high-level features using a feature pyramid network. Then the student targets are computed using a target classification network, student target locations are computed using a border regression network, and the student target frames with the highest confidence are filtered using a nonmaximum suppression algorithm. All the collected target images of students are preprocessed and normalized, and the YOLO-v5 framework is used to extract the spatiotemporal features of student behavior, and finally, students' behavior recognition is achieved through classification learning. In this paper, the YOLO-v5 model is decomposed into four parts which include the input, the main network, the feature fusion network, and the output. The input is designed with adaptive image size regardless of the image size of the dataset, the input will be automatically scaled according to the short side of the image to 608×608. If the aspect ratio of the image does not match the input, it will be automatically supplemented with black on the outer edge of the image, so, that the images input to the backbone network all matches $608 \times 608 \times 3$ network input. The flowchart of the classroom student goal tracking algorithm is shown below in Figure 2 and the YOLO-v5 specific recognition model is shown in Figure 3.

In this article, the image correlation coefficient method is used to extract key frames. The correlation coefficient is calculated using the following formula.

$$R = \sum_{N} \frac{\sum_m \sum_n AB}{\sqrt{\sum_m \sum_n A^2} \sqrt{\sum_m \sum_n B^2}}, \tag{1}$$

Where A and B represent arbitrary images, m and n are the width and height of the images respectively, and N is the

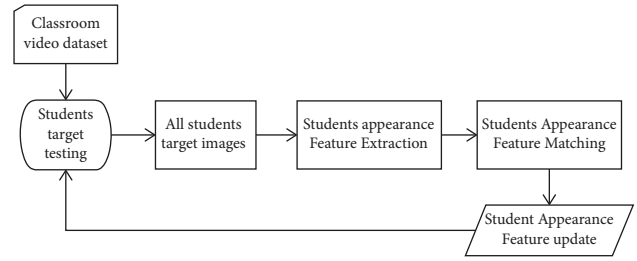

Figure 2: Flowchart of classroom student goal tracking algorithm.

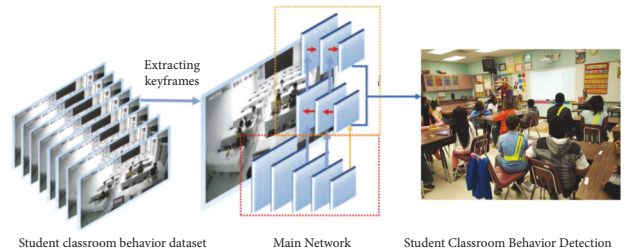

Figure 3: Student classroom behavior recognition model.

number of channels of the images. After several attempts, the keyframe obtains when the correlation coefficient (R) value becomes less than 0.8 ($R < 0.8$), which needs to be extracted. When the correlation coefficient (R) keyframe is fetched, then it can be determined that the video has a sudden change [34].

Considering the output-side teaching specific scenario, where students have a fixed range of activities and a small range of movements in the classroom, so, it mainly involves the improvement of the loss function. In a detection bounding box (B box), the training parameters are the position of the B box represented by (g), the classification of behavior recognition is represented by (a), and the confidence level of target prediction is represented by (c). Since a good prediction box needs to satisfy the largest overlap area, the closest distance to the center point, and the aspect ratio closest to the aspect ratio of the standard box. For the position of the geometric quantity of the B box, the Euclidean distance metric is not applied to the predicted. The difference between the predicted value and the accurate value is no longer applied to the position of the geometric quantity (B box), but the intersection of the predicted box and the accurate box is used to determine its loss. The loss is calculated using the following formula.

$$L_g = 1 - \left(R_o - \frac{d_1^2}{d_2^2} - \frac{v^2}{(1 - R_o) + v} \right). \tag{2}$$

(2) is the geometric distance between the center of the prediction box and the center of the standard box, which is used to limit the distance of the center point. The term d_2 is the diagonal length of the outer rectangle of the prediction frame and the standard frame, which is used to limit the maximum overlapping area. R_o is the ratio of the intersection of the standard frame P and the prediction frame $\hat{P}$ to the concatenated area. The following equation is used to calculate the value of R_o.

$$R_o = \frac{S_{(P \cap \widehat{P})}}{S_{(P \cup \widehat{P})}}. \quad (3)$$

V is a parameter that measures the consistency of the aspect ratio and is calculated using the below equation.

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{\widehat{p}}}{h_{\widehat{p}}} - \arctan\frac{w_p}{h_p}\right)^2, \quad (4)$$

Where w_p and $w_{\widehat{p}}$ denote the width of the standard box and the prediction box, respectively. h_p And $h_{\widehat{p}}$ denote the height of the standard box and the prediction box respectively.

4. Results and Discussion

4.1. Test Environment Results. The raw video data of teachers' teaching used in this experiment is obtained from a smart classroom. The smart classroom contains multimedia and other facilities, which are needed for the construction of an intelligent classroom system. There are two cameras above the classroom: the camera in the middle position records behavioral videos of students around the podium, and the camera in the front of the classroom is used to record behavioral videos of students walking around the classroom. In this paper, 90 classroom videos are collected and a classroom behavior library is constructed. Since it is difficult to determine the beginning and end of the action by manually labeling the videos. This paper uses a target tracking algorithm to obtain a stream of images of behavioral state changes of a single student, and then manually selects some consecutive frames from the stream as an action sample of a single person. We collected 8000 images of students raising their hands, 500 images of students bending down, 8000 images of students walking back and forth, 12000 images of students writing on the blackboard, 3000 images of students looking up, 3000 images of students looking down, 1000 images of students standing, and 1000 images of students lying on their desks for this article. The test environment and specific equipment environment are shown below in Figures 4 and 5 and the process of labeling an image is shown in Figure 6.

Performance evaluation metrics such as accuracy, recall, and average accuracy all are used in the target detection. The results of the evaluation metrics for student classroom behavior recognition are shown in Table 1.

From Table 1, we can see that both the average accuracy and recall rates of the behavioral recognition framework in identifying student classroom behaviors are quite promising. The teacher and student classroom behavior datasets were trained and evaluated individually, and on the self-developed student classroom behavior dataset, the average recognition accuracy of detecting numerous classroom actions was above 93%. Figure 7

Finally, the instructional video of the real scene is shown after the keyframe extraction. The results of detection after keyframe extraction are shown in Figure 8. From Figure 8, we can see that the student behavior recognition model has good generalization ability and can complete the recognition task more accurately in the real teaching scene.

Figure 4: The test environment.

Figure 5: The camera equipment environment.

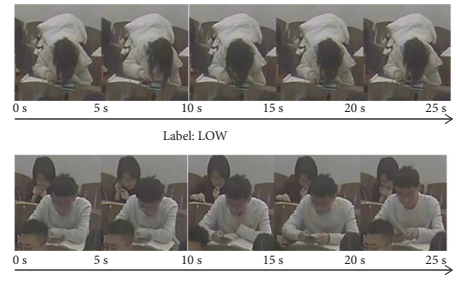

Figure 6: The process of image labeling.

Table 1: Statistical indicators of student behavior identification.

Action	Accuracy	Recall	Average accuracy
Hands up	0.965	0.953	0.971
Head down	0.956	0.933	0.973
Walking back and forth	0.921	0.870	0.880
Standing	0.979	0.994	0.994
Looking up	0.937	0.882	0.916
Looking down	0.926	0.851	0.914
Lying on desks	0.891	0.950	0.934
Writing on the blackboard	0.942	0.964	0.952

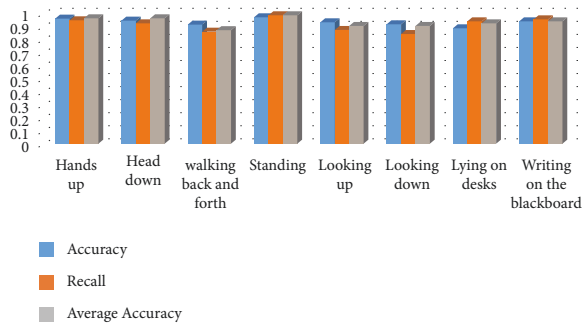

Figure 7: Statistical indicators of student behavior identification model.

Figure 8: Results of the student behavior recognition model in real scene.

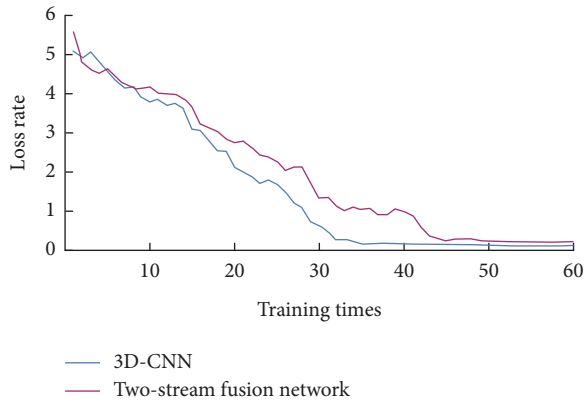

Figure 9: Comparison of loss rate curves of 3D-CNN and dual-stream-based fusion networks.

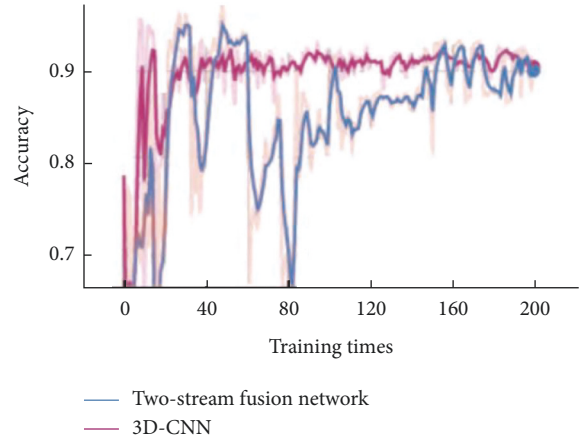

Figure 10: Accuracy curves of 3D-CNN and dual-stream-based fusion network.

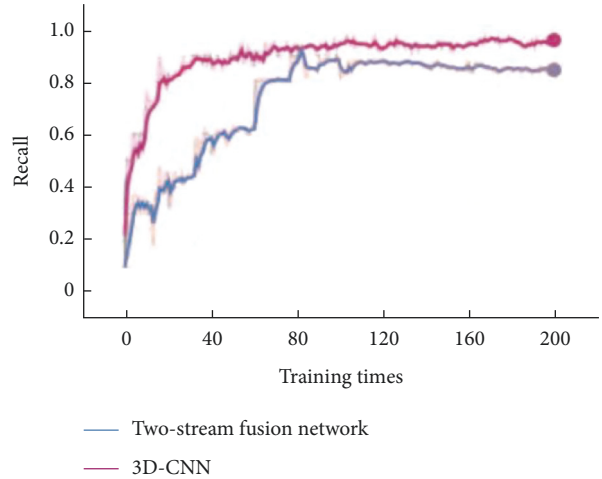

Figure 11: Recall curves of 3D-CNN and dual-stream-based fusion networks.

4.2. Comparison of Performance of Our Proposed Model (3D-CNN) and Two-Stream Fusion Network. A dual-stream fusion neural network, which is commonly used in the field of behavior recognition, compared with the 3D-CNN network that is used in this paper for the students' behavior recognition. In dual-stream-based fusion network the video input is divided into two parts: spatial (RGB images) and temporal (optical stream images), so, that the information between consecutive frames in the video is extracted and processed. The recognition network in this paper and the dual-stream fusion neural network which is used for comparison are trained and tested separately on the same dataset. In addition, the accuracy of 3D-CNN is compared with the dual-stream based fusion network for video behavior recognition. After running on this labeled dataset, the recognition accuracy of the dual-stream-based fusion network is 86.43%, and that of the 3D-CNN is 93.43%. The accuracy of 3D-CNN is better than the dual-stream-based fusion networks, which can provide better recognition efficiency for teaching video behavior recognition. The compared accuracy, Recall, and loss rate convergence of behavior recognition are shown in the following figures. Figures 9–11

The proposed intelligent teaching assessment system based on student classroom behavior recognition mainly contains four functional modules: classroom student goal detection module, classroom student goal tracking module, classroom student target real-time detection module, and classroom student target tracking module. The classroom student target real-time detection module uses a deep convolutional neural network (DCNN) model to extract

useful features of the classroom. The classroom student target detection module uses a DCNN model to extract features from classroom videos and then uses the feature pyramid network to fuse the learned high-level features. The classroom student target tracking module uses a target detection algorithm to obtain all the student target images. Further, this module uses all of the student target images obtained by the target detection algorithm to calculate the features of each student and then match them with the existing features. The classroom behavior recognition module uses a deep spatiotemporal residual CNN to learn the spatiotemporal features of students', and then a logistic regression algorithm is used to calculate the students' behavior probability. In the classroom assessment module, the proportion of positive, neutral, and negative behaviors is calculated. If the proportion of negative behaviors exceeds a certain value, then the feedback will be given to the teacher in real-time.

5. Conclusion

In this paper, we proposed a teacher–student classroom behavior recognition model based on a DL optimization model, which relies on video understanding technology to realize the recognition of student classroom behavior. The recognition of students' behaviors in the classroom video needs to consider the motion information between consecutive frames. So, a 3D-CNN model is proposed in this study for the student's behavior recognition as a network model. It combines tracking and objects detection technology to make a connection between the spatiotemporal features of the same students and achieve the multistudent action recognition goal in the classroom. The YOLO-v5 framework is used to extract the spatiotemporal features of student behavior, and finally, student behavior recognition is achieved through the classification model. For the experimental purpose, the teacher and student classroom behavior datasets were used. The teacher and student classroom behavior datasets were trained and tested separately, and the average recognition accuracy of identifying multiple classroom behaviors on the self-developed student classroom behavior dataset was over 93%. Hence, the experimental results prove the effectiveness of the proposed method. In addition, in this study, we also constructed an intelligent teaching evaluation model for smart education based on the proposed student classroom behavior recognition model, which improves the teaching quality. In the future, we will improve the generalization of the classroom behavior recognition method and will try to expand the teacher and student behavior quantity to provide more diverse data for smart classroom analysis.

Data Availability

The data used to support the findings of this study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper was supported by Beijing Municipal Educational Science "Thirteenth Five Year Plan" General Project in 2019 "Research on Evaluation Index System in the Context of General High School History Curriculum Reform" (Project no. CDDB19182), the 2021 General Project of Beijing Society of Education "Development and Application of WeChat Official Account for Departmental History Textbook Aided Teaching" (Project no. DCYB2021-076), and 2021 Higher Education Fund of the Macao SAR Government (no. HSS-CITYU-2021-07).

References

[1] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, IEEE, Oxford, England, May 2017.

[2] W. Chen, Z. Jiang, H. Guo, and X. Ni, "Fall detection based on key points of human-skeleton using OpenPose," *Symmetry*, vol. 12, no. 5, p. 744, 2020.

[3] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U R. Acharya, "ABCDM: an attention-based bidirectional CNN-rnn deep model for sentiment analysis," *Future Generation Computer Systems*, vol. 115, pp. 279–294, 2021.

[4] Y. Zhang and G. Wang, "Research on application of intelligent analysis in monitoring of classroom teaching," in *Proceedings of the 2021 3rd international conference on advances in computer technology, information science and communication (CTISC)*, pp. 253–257, IEEE, Shanghai, China, April 2021.

[5] S. Smys, J. M. R. S. Tavares, and R. Bestak, *Computational Vision and Bio-Inspired Computing: ICCVBIC 2020*, Springer Nature, New York, NY, USA, 2021.

[6] T. Poggio, V. Torre, and C. Koch, "Computational vision and regularization theory," *Readings in computer vision*, pp. 638–643, 1987.

[7] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia*, pp. 357–360, ACM, New York, NY, USA, May 2007.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," 2014, https://arxiv.org/abs/1406.2199.

[9] L. Wang, Y. Xiong, Z. Wang et al., "Temporal segment networks: towards good practices for deep action recognition, Computer Vision - ECCV 2016," in *Proceedings of the European Conference on Computer Vision*, Springer, Cham, Switzerland, pp. 20–36, July 2016.

[10] B. Xu, H. Ye, Y. Zheng, H. Wang, T. Luwang, and Y Jiang, "Dense dilated network for video action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4941–4953, 2019.

[11] D. Tran, L. Bourdev, and R. Fergus, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, IEEE, Santiago, Chile, December 2015.

[12] Z. Cao, T. Simon, and S. E. Wei, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, IEEE, Honolulu, HI, USA, July 2017.

[13] C. Guan, "Realtime multi-person 2d pose estimation using shufflenet," in *Proceedings of the 2019 14th International Conference on Computer Science & Education (ICCSE)*, pp. 17–21, IEEE, Canada, CA, USA, August 2019.

[14] H. Wang, W. P. An, and X. Wang, "Magnify-net for multi-person 2d pose estimation," in *Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, San Diego, CA, USA, July 2018.

[15] M. Nakai, Y. Tsunoda, and H. Hayashi, "Prediction of basketball free throw shooting by openpose," in *Proceedings of the JSAI International Symposium on Artificial Intelligence*, pp. 435–446, Springer, Cham, Switzerland, June 2018.

[16] M. Bendersky, L. Garcia-Pueyo, J. Harmsen, V. Josifovski, and D. Lepikhin, *Up Next: Retrieval Methods for Large Scale Related Video Suggestion*, pp. 1769–1778, ACM SIGKDD, New York, NY, USA, 2014.

[17] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the ICCV VS-PETS*, IEEE, Beijing, China, October 2005.

[18] I. Laptev and T. Lindeberg, "Space-time Interest Points," in *Proceedings of the ICCV*, IEEE, Nice, France, October 2003.

[19] S. Sadanand and J. Corso, "Action bank: A High-Level Representation of Activity in Video," in *Proceedings of the CVPR*, IEEE, Providence, RI, USA, June 2012.

[20] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *Proceedings of the ICCV*, IEEE, Sydney, NSW, Australia, June 2013.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, https://arxiv.org/abs/1311.2524v5?source=post.

[22] A. Jain, J. Tompson, Y. LeCun, and C. B. Modeep, "A Deep Learning Framework Using Motion Features for Human Pose Estimation," in *Proceedings of the ACCV*, Springer, Cham, Switzerland, April 2014.

[23] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[24] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Largescale video classification with convolutional neural networks," in *Proceedings of the CVPR*, pp. 1725–1732, Columbus, OH, USA, June 2014.

[25] L. Sun, K. Jia, D. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the ICCV*, pp. 4597–4605, Santiago, Chile, December 2015.

[26] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *CoRR abs*, vol. 1604, Article ID 04494, 2016.

[27] H. Pirsiavash and D. Ramanan, "Parsing videos of actions with segmental grammars," in *Proceedings of the CVPR*, pp. 612–619, Columbus, OH, USA, June 2014.

[28] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2782–2795, 2013.

[29] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Computer Vision - ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6312, pp. 392–405, Springer, Heidelberg, Germany, 2010.

[30] L. Wang, Y. Qiao, and X. Tang, "Video action detection with relational dynamic-poselets," in *Computer Vision - ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8693, pp. 565–580, Springer, Heidelberg, Germany, 2014.

[31] P. Sun, R. Zhang, and Y. Jiang, "Sparse r-cnn: end-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14454–14463, IEEE, New York, NY, USA, June 2021.

[32] K. He, L. Ji, and C W D. Wu, "Using SARIMA-CNN-LSTM approach to forecast daily tourism demand," *Journal of Hospitality and Tourism Management*, vol. 49, pp. 25–33, 2021.

[33] Y. Huang and M. Liang, "Spatio-temporal attention network for student action recognition in classroom teaching videos," *EURASIP Journal on Image and Video Processing*, vol. 2022, 2022.

[34] S. Kapania, D. Saini, and S. Goyal, "Multi object tracking with UAVs using deep SORT and YOLOv3 RetinaNet detection framework," in *Proceedings of the 1st ACM Workshop on Autonomous and Intelligent Mobile Systems*, pp. 1–6, ACM, New York, NY, USA, June 2020.