2022 9th International Conference on Digital Home (ICDH) | 978-1-6654-5478-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDH57206.2022.00043

# A METHOD FOR ANALYZING TEACHER BEHAVIOR IN CLASSROOM BASED ON THE LONG- AND SHORT-TERM FEATURES OF POSE SEQUENCES

*Yuanzhong Li*[1,2]*, Zhengjie Deng*[*1,2]*, Meijun Liu*[1]*,Shuqian He*[1]*,Yizhen Wang*[3]*,Wenjuan Jiang*[1]

[1]School of Information Science and Technology, Hainan Normal University, Haikou,China.
[2]Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin, China.
[3]School of physics and Electronic Engineering, Hainan Normal University, Haikou,China.
email:737563479@qq.com, hsdengzj@163.com, liu1999312@qq.com,76005796@qq.com, wangxuesu1980@163.com,may_jwj@qq.com.

## ABSTRACT

The analysis of teachers' behaviors in classroom videos is helpful to objectively evaluate teaching, promote teaching reform, and improve teaching effectiveness. With the continuous introduction of deep learning models, there have been many analysis methods, but the recognition accuracy is not high enough. In this paper, we construct a new model combining convolutional neural network and recurrent neural network to learn and fuse short-term features with long-term features, and use the spatio-temporal information of videos to get the behavior categories. In our experiments, we collected teacher classroom videos in teaching scenarios to construct the dataset. The 2D pose heatmap sequences are extracted from the videos as model inputs, which are used to exclude the interference of environmental noise in the teaching scene. The experimental results show that this method can achieve higher recognition accuracy compared with the PoseR(2+1)D method on the teacher behavior dataset for teacher behavior analysis in teaching scenes.

***Index Terms***— behavior analysis, R(2+1)D, convLSTM, long- and short-term features, spatio-temporal information

## 1. INTRODUCTION

In recent years, the analysis of teachers' behaviors in teaching videos has gradually become a hot research topic for realizing smart education and teaching. Accurate identification of teachers' teaching behaviors in videos helps analyze and evaluate teachers' teaching styles and helps teachers improve teaching quality. At the same time, the behavior recognition analysis can get the corresponding data and suggestions sent to the teacher himself or the school, which helps to understand the teaching situation in multiple directions. Through comparison and learning, teachers are able to develop targeted improvement plans for weaknesses in classroom teaching styles. The purpose of analyzing teacher behavior in classroom scenarios is to improve the accuracy of teacher behavior recognition, to make full use of the teaching video resources to analyze overall teacher behavior, and to automatically generate teaching style assessments.

At present, in the research field of human behavior recognition, it is a very difficult task to recognize human beings in the natural environment due to complex space-time characteristics, chaotic background and various perspectives. In the research of behavior recognition methods of machine learning, most of them combine feature extraction and classifier for recognition and analysis. For example, using Histograms to manually calculate complex features, and then these handcrafted features to train classifiers, such as logistic regression and SVM.

The deep learning has also made great breakthroughs in visual tasks such as pose estimation and behavior recognition. Among the methods for behavior recognition, the main ones are 3D convolutional neural networks (3D-CNN), such as R(2+1)D [1], CSN [2], X3D [3], etc.; graph convolutional neural networks (GCN), ST-GCN [4], CTR-GCN [5], etc.; two-stream 2D convolutional neural networks (two stream 2D-CNN), two-stream [6], Slow-Fast [7], etc. where the R( 2+1)D [1] method decomposes the 3D convolution kernel of R3D [8] to derive separate spatial and temporal components, and designs the R(2+1)D spatio-temporal convolution block to replace the original 3D convolution block, and significantly improves the recognition accuracy on the public action recognition dataset. Meanwhile, some scholars have combined the three mainstream methods with each other to form a new behavior recognition network architecture, such as in PoseC3D proposed by Hao-Dong Duan et al [9], in which the skeleton sequence features previously used in graph convo-

*:Corresponding author.

Acknowledgement. This work was financially supported by Hainan Natural Science Foundation (620RC604), Science and Technology Project of Haikou (No. 2020-053, 2020-014, 2020-044), the Open Funds from Guilin University of Electronic Technology, Guangxi Key Laboratory of Image and Graphic Intelligent Processing (GIIP2012), the National Natural Science Foundation (61502127), Key R&D Projects in Hainan Province (ZDYF2019010), and Research project on education and teaching reform of Hainan Normal University (hsjg2020-20).

978-1-6654-5478-0/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDH57206.2022.00043

lutional neural networks are transformed into continuous 2D pose heatmap sequences and input to a 3D-CNN model for behavior recognition, in which the R(2+1)D model is used as one of its experimental 3D-CNN models. CNN model, excellent results were achieved. By excluding the factors of environmental background interference in this way, the pure 2D pose heatmap sequences is used for feature extraction using 3D-CNN to achieve better recognition results. However, 3D-CNN is easy to lose attention among long interval tasks, which has some impact on recognition accuracy. Xi Ouyang et al [10] proposed to combine 3D-CNN with LSTM [11] to form a multi-task learning architecture based on 3D-CNN and LSTM for behavior recognition, which further improves the behavior recognition accuracy with the powerful learning ability of LSTM on long intervals. However, LSTM usually only learns in a single temporal dimension, which has an impact on behavioral recognition that requires temporal and spatial features. While convLSTM [12] is a combination of recurrent neural network and convolutional neural network, changing the traditional LSTM that can only learn in temporal dimension, convLSTM is able to extract spatio-temporal features on image sequences and capture spatio-temporal information of image sequences, which is more suitable for behavior recognition.In C3D-convLSTM [13] Model, convlstm was used to extract the temporal and spatial characteristics of cows

Inspired by the literature [9], [10], [12], a deep learning network for computing and fusing long- and short-term features based on pose sequences is proposed in this paper for teacher behavior analysis in teaching scenarios with long action durations and complex background environments. The network converts video inputs into 2D pose heatmap sequences as a way to remove the influence of the background environment on behavior recognition, and then combines the short-term feature extraction network and the long-term feature extraction network and fuses the feature information learned from both to improve the behavior characterization ability in the feature space, thus improving the accuracy of teacher behavior recognition.

## 2. TEACHER BEHAVIOR ANALYSIS NETWORK MODEL

The deep neural network designed in this paper for long- and short-term feature computation and fusion based on pose sequences is based on a modified R(2+1)D neural network model by introducing a convLSTM recurrent convolutional neural network to extract feature information for long time intervals and fuse it with short-term features using bypass connections, by which the spatio-temporal feature relationships of human behavior are captured. It is then fed into the classification network for behavior prediction. The specific model architecture is shown in Fig 1.

First, 2D pose sequences within all frames of the video will be obtained using target detection and pose estimation to construct 2D pose heatmap sequences with uniform spatio-temporal dimension as model inputs. In the main network structure, we construct a neural network combining R(2+1)D and convLSTM, which first uses a modified R(2+1)D network for local short-term feature extraction of the input to obtain a short-term feature tensor, and then uses convLSTM for feature extraction of the short-term feature tensor to obtain a long-term feature tensor of the same dimension as the short-term feature tensor. Finally, the short-term feature tensor is fused and transformed with the long-term feature tensor, and the features are pooled using the adaptive average pooling layer, and then sent to the fully connected layer for prediction and outputting the behavioral categories. By performing short-term feature learning and long-term feature learning to obtain better spatio-temporal feature information, and then combining short-term features with long-term features to get the final output, the model is enhanced to feature representation of video information, so the prediction accuracy is effectively improved.

### 2.1. Constructing 2D pose heatmap sequences

To construct a 2D pose heatmap sequence, firstly, for a video of duration T, height H and width W, using fastRCNN [14] to detect the human body in the video, and mark the area where the human body is located in the video, that is, the human body boundary box. Subsequently, HRNet [15] is used to estimate the pose of the human body in each frame and obtain the set of triples $(x_k, y_k, c_k)$ of the human skeleton coordinates of the whole video. Then the set of triads with length 48 is selected by uniform sampling. The uniform sampling will measure the number of the entire video frames against the number of video frames to be sampled and select the appropriate step size for sampling. After obtaining the set of 48 frames of the triad, using the coordinates and confidence in the set of triads, the joint heatmap $J$ of $K$ Gaussian mappings can be synthesized by this information, and the relationship is expressed as in Equation 1.

$$J_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}} * c_k \tag{1}$$

where $\delta$ controls the contrast of the Gaussian mapping, and $(x_k, y_k)$ and $c_k$ are the position and confidence of the $k$-th joint, respectively. The heatmap construction process for a single image is shown in Fig 2.

A corresponding keypoint heat point is generated in a region of 2 pixels around each keypoint coordinate, and in this way, the Gaussian heatmap of the cumulative owner can be expanded at each frame, and the final input sequence size is $K \times 48(T) \times H \times W$ by stacking the heatmaps on the time dimension.

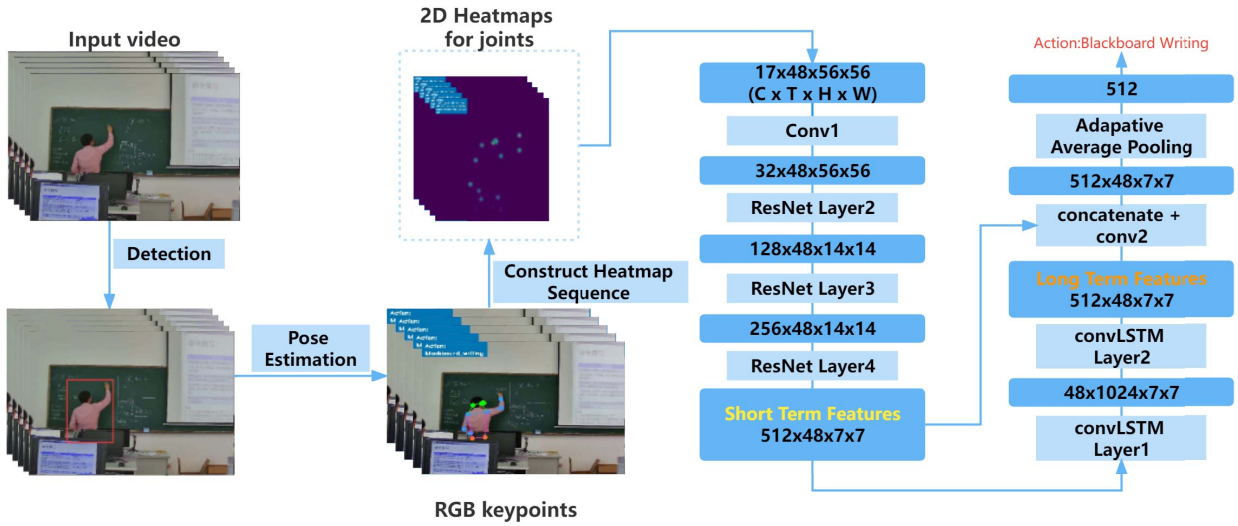


**Fig. 1**. The overall architectural flow of the model

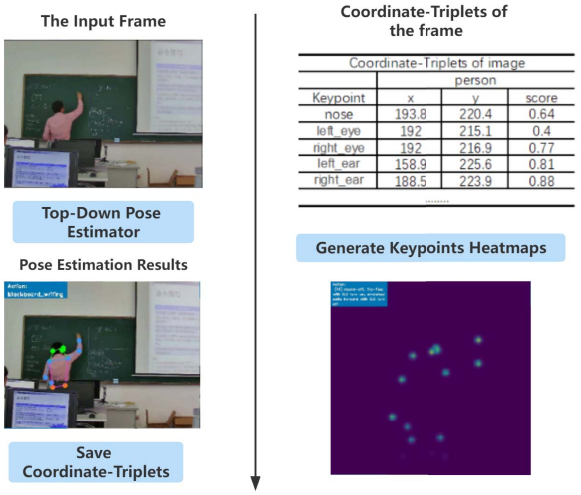


**Fig. 2**. Heatmap construction process for images

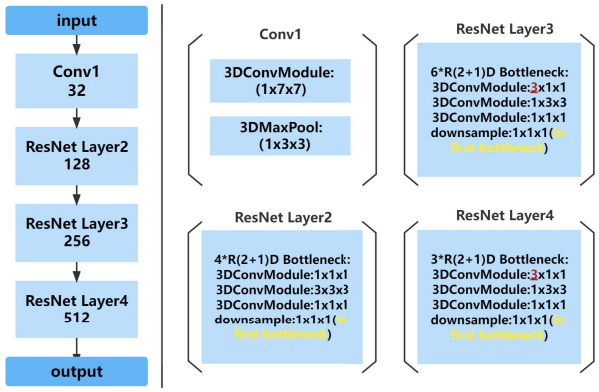


**Fig. 3**. R(2+1)D module structure

### 2.2. Short-term feature extraction module

Based on the previously obtained sequence of pose heatmap, in order to obtain the behaviorally relevant intermediate feature tensor, we refer to the poseC3D model and reduce the channel width to half (64→32) for the characteristics of this experimental data, while removing the first stage of R(2+1)D. In this paper, the structure of the R(2+1)D-50 module used is shown in Fig 3.

The input is first passed through a 3-dimensional convolutional layer with a convolutional kernel of $1 \times 7 \times 7$ to obtain a feature map with channel number $C$ (set to 32 in this paper), and then fed into 3 convolutional residual blocks, each residual layer containing 4, 6, and 3 residual bottleneck units, respectively. The first residual block has three residual bottleneck units composed of 3D convolution, and the three convolution kernels of the residual bottleneck unit are $1 \times 1 \times 1$, $1 \times 3 \times 3$, and $1 \times 1 \times 1$, respectively, and the final output expands the number of channels of the input of the previous layer by a factor of 4 and reduces the size by a factor of 1; the 2nd and 3rd residual blocks have 4 and 6 residual units composed of (2+1)D convolution respectively, and the convolution kernels are all set to $3 \times 1 \times 1$,$1 \times 3 \times 3$, and $1 \times 1 \times 1$ and the final output expands the number of channels of the previous layer input by a factor of 2 and reduces the size by a

**Table 1**. Comparison of the structure of the R3D model and the model in this paper

| stage | R3D-50 | our model |
|---|---|---|
| input layer | data(3×T×H×W) | |
| Stem layer | Conv1: 1×7×7, 32 | |
| Stage1 | $\begin{bmatrix} 1\times1^2, 32 \\ 3\times3^2, 32 \\ 1\times1^2, 128 \end{bmatrix} * 4$ | None |
| Stage2 | $\begin{bmatrix} 1\times1^2, 64 \\ 3\times3^2, 64 \\ 1\times1^2, 256 \end{bmatrix} * 4$ | $\begin{bmatrix} 1\times1^2, 32 \\ 3\times3^2, 32 \\ 1\times1^2, 128 \end{bmatrix} * 4$ |
| Stage3 | $\begin{bmatrix} 1\times1^2, 128 \\ 3\times3^2, 128 \\ 1\times1^2, 512 \end{bmatrix} * 6$ | $\begin{bmatrix} 3\times1^2, 64 \\ 1\times3^2, 64 \\ 1\times1^2, 256 \end{bmatrix} * 6$ |
| Stage4 | $\begin{bmatrix} 1\times1^2, 256 \\ 3\times3^2, 256 \\ 1\times1^2, 1024 \end{bmatrix} * 3$ | $\begin{bmatrix} 3\times1^2, 128 \\ 1\times3^2, 128 \\ 1\times1^2, 512 \end{bmatrix} * 3$ |
| output layer | $1024 \times T \times \frac{H}{16} \times \frac{W}{16}$ | $512 \times T \times \frac{H}{8} \times \frac{W}{8}$ |

factor of 1. The final output feature size of the R(2+1)D module is $512 \times T \times H_8 \times W_8$ obtained by this series of feature extraction.

A comparison of the model used in this paper with the R3D-50 model architecture is shown in Table 1.

### 2.3. Long-term feature extraction module

To better learn the long interval features in the input features, we use convLSTM as a main global long-term feature extraction process with a bypass connection on top of it and a convolutional operation with a convolutional kernel of $1\times1\times1$ as a way to achieve a fusion transformation of long-term features with short-term features. convLSTM is an evolution of LSTM for image sequences. By extending the LSTM to perform convolution in the input-to-state and state-to-state transitions, the LSTM is able to use 2D images as observations of the model, which allows the LSTM to solve the prediction problem of spatial sequences, which was originally done for spatial sequences. convLSTM reconstructs all inputs $X1 \ldots Xt$, cell outputs $C_1 \ldots C_t$, hidden states $H_1 \ldots H_t$, and threshold switches $i_t, f_t, o_t$, transforming them into a 3-dimensional tensor whose last two dimensions are the spatial dimensions $(H, W)$. In the convLSTM implemented in this paper, the equation for each key variable is shown in Equation 2 as follows.

$$\begin{aligned} i_t &= Sigmoid(Conv(x_t; w_{xi}) + Conv(h_{t-1}; w_{hi}) + b_i) \\ f_t &= Sigmoid(Conv(x_t; w_{xf}) + Conv(h_{t-1}; w_{hf}) + b_f) \\ o_t &= Sigmoid(Conv(x_t; w_{x0}) + Conv(h_{t-1}; w_{ho}) + b_o) \\ g_t &= Tanh(Conv(x_t; w_{xg}) + Conv(h_{t-1}; w_{hg}) + b_g) \\ c_t &= f_t \bigodot c_{t-1} + i_t \bigodot g_t \\ h_t &= o_t \bigodot Tanh(c_t) \end{aligned} \tag{2}$$

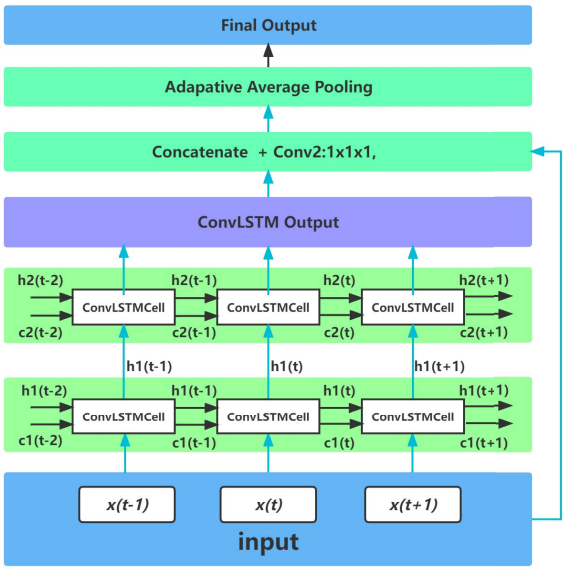


**Fig. 4**. convLSTM module structure

In this paper, the structure of the convLSTM module is designed as shown in Fig 4.

convLSTM is designed as two layers, including 1024 node hidden layer and 512 node output layer. The short-term features are dimensionally transformed ($C \times T \times H \times W \rightarrow T \times C \times H \times W$) and fed into the convLSTM for global long-term feature learning, and the output layer outputs a long-term feature tensor of dimension $T \times C \times H \times W$. After that, perform a pooling operation with a convolution kernel of $(1 \times 3 \times 3)$, and then stack it with the short-term feature tensor output from the previous module in the channel dimension (to obtain a tensor of $2C \times T \times H \times W$), and finally fuse and refine the features through the convolution operation with a convolution kernel of $1 \times 1 \times 1$, send it to the classification network of self-adaptive average pooling layer and full connection layer for behavior prediction, and output the final predicted behavior classification.

## 3. EXPERIMENT AND ANALYSIS

The proposed teacher behavior analysis network is trained and tested using a dataset of teacher behavior in teaching scenarios. It will be presented in the following 3 sections.

### 3.1. Building the teacher behavior dataset

The teacher behavior dataset in teaching scenes is designed with six behavior categories that often appear in modern multimedia teaching scenes, containing a total of 873 actions, which are blackboard board, operating multimedia, walking

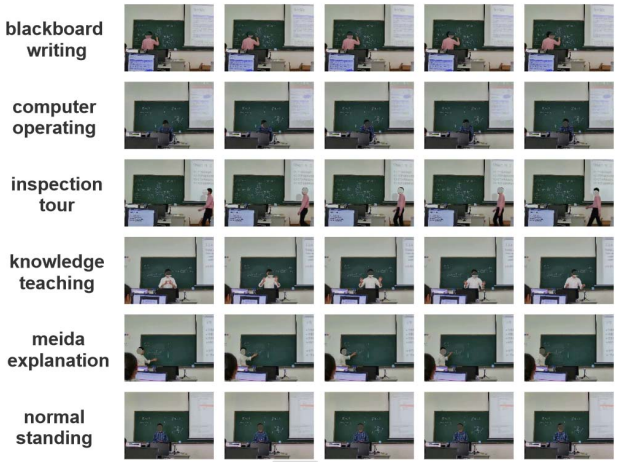


**Fig. 5**. Examples of teacher behavior

around, knowledge explanation, media explanation, and regular standing, and the number of samples in each category are 140, 200, 101, 198, 186, 48, and the average duration of samples is about 5-6 seconds. In order to enhance the richness of the data set, the data are processed with random cropping, flipping and other operations to improve the generalization and robustness of the network model. Some of the sample teacher behavioral actions are shown in Fig 5.

### 3.2. Training settings

The experimental environment of this paper is: Ubantu 18.04 OS, 32G RAM, Intel i7-10700K CPU, 1 GeForce RTX 3090 graphics card (24G video memory), and Pytorch-1.9.1 deep learning framework.

In the hyperparameters of the model, we use SGD optimizer for the model with $lr$ set to 0.01, $momentum$ set to 0.9, and $weight\ decay$ set to 0.0003. The learning rate is decayed by cosine annealing. The total number of iterations is 24 epochs. For the input sequence, $clip\_len$ frames are collected by uniform sampling operation for each video, and then resized into a fixed size heatmap image. In this paper, the $clip\_len$ is set to 48, and for the training set, the size after resize is set to 56; for the validation set, the size after reszie is set to 64.

### 3.3. Experimental results and analysis

Table 2 shows the comparison of the recognition accuracy of different models on the teacher behavior dataset.

In this paper, in addition to using the R(2+1)D module directly, we also experimented the way of using the residual blocks of the 3 stages after R3D as the short-term feature learning module. As can be seen from Table 2, the model proposed in this paper converts video into 2D pose heatmap sequence as input, and learns short-term features and long-term features through R(2+1)D+convLSTM, and then fuses the two features. The recognition accuracy of 84.0% is achieved on the teacher behavior dataset, which is 4.57% higher than that of 79.43% using poseR (2+1) D alone; Similarly, in the comparison of using R3D as a short-term feature learning module, the accuracy of this method can reach 83.43%, which is 5.72% higher than that of 77.71% using R3D alone, which proves that adding long-term feature learning and integrating it with short-term features can make full use of spatio-temporal information and achieve a better recognition effect. In contrast, the R(2+1)D-34 model has only 63.31% accuracy and the R3D-34 model has only 72.66% accuracy, which indicates that the conversion to 2D pose heatmap sequences can also help to improve the recognition accuracy in the process of recognizing teacher behavior. Thus, it can be seen that the teacher behavior analysis network model constructed in this paper has a good recognition effect on the teacher behavior dataset.

**Table 2**. Accuracy, performance comparison of different models on the teacher behavior dataset (where pose extract is the pose heatmap sequence extraction operation)

| Method | Input size | Accuracy | params | Pose extract |
|---|---|---|---|---|
| R3D-34 | $3 \times 48 \times 112 \times 112$ | 72.66% | 33.18M | $\times$ |
| R(2+1)D-34 | $3 \times 48 \times 112 \times 112$ | 63.31% | 33.18M | $\times$ |
| PoseR3D-50 | $17 \times 48 \times 56 \times 56$ | 77.71% | 88.28M | $\surd$ |
| PoseR(2+1)D-50 | $17 \times 48 \times 56 \times 56$ | 79.43% | 87.46M | $\surd$ |
| PoseR3D-50 +convLSTM(ours) | $17 \times 48 \times 56 \times 56$ | 83.43% | 88.28M | $\surd$ |
| PoseR(2+1)D-50 +convLSTM(ours) | $17 \times 48 \times 56 \times 56$ | 84.00% | 87.46M | $\surd$ |

## 4. CONCLUSION

In this paper, we propose a deep neural network model for computing and fusing long- and short-term features based on pose sequences, which improves the representation of behavioral features by combining R(2+1)D and convLSTM to allow the model to perform short-term feature and long-term feature learning and fuse the two features to make full use of the spatio-temporal information of the video. In the applied teaching scenario with the teacher behavior dataset, the experimental results show that the short-term feature learning and long-term feature learning and fusion with each other are effective in improving the accuracy of teacher behavior recognition. The follow-up work lies in processing the samples of the dataset to enhance the completeness of the dataset, as well as adjusting the network parameters of the architecture to improve the accuracy of teacher behavior analysis and optimize the operation speed.

## 5. REFERENCES

[1] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and et al., “A closer look at spatiotemporal

convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[2] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli, "Video classification with channel-separated convolutional networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5552–5561.

[3] Christoph Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[4] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[5] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13359–13368.

[6] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3154–3160.

[9] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin, "Pyskl: Towards good practices for skeleton action recognition," *arXiv preprint arXiv:2205.09443*, 2022.

[10] Xi Ouyang, Shuangjie Xu, Chaoyun Zhang, Pan Zhou, Yang Yang, and et al., "A 3d-cnn and lstm based multi-task learning architecture for action recognition," *IEEE Access*, vol. 7, pp. 40757–40770, 2019.

[11] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and et al., "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[13] Yongliang Qiao, Yangyang Guo, Keping Yu, and Dongjian He, "C3d-convlstm based cow behaviour classification using video data for precision livestock farming," *Computers and Electronics in Agriculture*, vol. 193, pp. 106650, 2022.

[14] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[15] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.