Machine Vision and Applications (2023) 34:44
https://doi.org/10.1007/s00138-023-01396-0

ORIGINAL PAPER

Human skeleton behavior recognition model based on multi-object pose estimation with spatiotemporal semantics

Jiaji Liu[1] · Xiaofang Mu[1] · Zhenyu Liu[1] · Hao Li[1]

Received: 29 August 2022 / Revised: 18 February 2023 / Accepted: 28 March 2023 / Published online: 28 April 2023
© The Author(s) 2023

Abstract
Multi-object pose estimation in surveillance scenes is challenging and inaccurate due to object motion blur and pose occlusion in video data. Targeting at the temporal dependence and coherence among video frames, this paper reconstructs a multi-object pose estimation model that integrates spatiotemporal semantics for different scales and poses of video multi-objects. The model firstly, with an end-to-end detection framework, detects multiple targets in the video. Secondly, it enhances the positioning of key points of human body using the temporal cues among video frames and designs modular components to enrich the pose information, effectively refining the pose estimation. Finally, the improved human skeleton behavior recognition model based on pose estimation is employed to recognize the classroom behaviors of students oriented to video streams. Comparison with multiple classifiers through experiments reveals that the human skeleton behavior recognition model for multi-object pose estimation combined with spatiotemporal semantics exhibits an effectively improved accuracy.

Keywords Spatiotemporal semantics · Multi-object detection · Human key points · Behavior recognition

1 Introduction

Models for human pose estimation usually describe the location information of a set of crucial points (10–30 points) and corresponding limb orientation information to follow the human skeletal structure. The key points for skeleton are significant for describing human postures and predicting human behaviors. Thanks to the deep learning technologies in recent years, detection accuracy in key points of human skeleton has been continuously improved, especially in research and design of static images in the early stage. However, its performance is reduced in processing video stream input. In terms of multi-object pose estimation for video streams, the problems for detecting key points of human skeleton are mainly reflected in the following two aspects:

(1) Temporal information of a sequence of video frames: after a video stream is inputted, the object pose information among different frames should be considered to solve pose occlusion and motion blur, resulting in additional time clues.
(2) Multi-object multi-scale detection: the localization effects of various objects under different backgrounds should be analyzed due to flexible multiple objects in the video frame.

Introducing a recurrent neural network (RNN) to the model is the most intuitive way to solve the time series problems. Luo et al. [1] employed the long and short-term memory (LSTM) to extract the spatiotemporal information and predict the heatmap information of key points in the video stream data. Wang et al. [2] adopted a three-dimensional (3D) high-resolution network (HRNet) to extract the spatiotemporal features of video frames for pose estimation. Liu et al. [3] proposed a deep dual consecutive network for human pose estimation (DCPose) to compensate for frame degradation and built a bi-continuous network to improve the pose esti-

Xiaofang Mu, Zhenyu Liu and Hao Li have contributed equally to this work.

✉ Jiaji Liu
494882733@qq.com

✉ Xiaofang Mu
mu_xiao_fang@163.com

Zhenyu Liu
1419297125@qq.com

Hao Li
193613452@qq.com

[1] College of Computer Science and Technology, Taiyuan Normal University, Taiyuan 030000, China

mation. When it is applied in different scales, the resolution size should be considered for analysis. Relied on the featured pyramid, Newell et al. [4] proposed the Associative embedding method for key points detection grouping, and Kreiss et al. [5] put forwarded Pifpaf, for body part relocation. Besides, Papandreou et al. [6] increased the input resolution into a high-resolution heatmap. In addition, Bowen Cheng et al. [7] proposed a Higher-HRNet, which could generate high-resolution heatmap information using the high-resolution feature pyramid, solving pose estimation of multiple objects at multiple scales.

Pose estimation has been applied in robot vision and motion tracking, among which the human skeleton behavior recognition is an active research direction. With various behaviors, people can convey different information and express their emotions. Therefore, behavior recognition exhibits a wide application range such as human-computer interaction, intelligent monitoring systems, and virtual reality. Traditional methods often use red-green-blue (RGB) image sequences [8–10], video frame sequences [11], or a specific fusion of multi-modalities. For example, building a dual-stream network based on RGB images and optical flow [12, 13] has achieved some results that exceed the expectation. However, compared with skeleton data, previous methods generate more computational consumption and show an insufficient robustness to deal with complex backgrounds and changes in the human scale. In this case, the human pose estimation model is selected to output the skeleton data, and then a specific algorithm is designed to discriminate different behaviors in this paper.

Further, this paper integrates spatiotemporal semantics for multi-object pose estimation and recognizes the human behaviors based on skeleton data. Firstly, a network model that combines video frame time information and multi-object information at different scales is reconstructed. Secondly, the temporal and spatial characteristics of human key points are based to enhance the spatiotemporal semantic information of crucial point detection, estimating the attitude more accurately. Finally, the skeleton data outputted by the pose estimation model fused with the human vital points information and skeleton map features to build a specific classifier model. The model aims to strengthen the expression ability of different behaviors and effectively improve the accuracy of behavior recognition based on skeleton data.

2 Related works

This paper focuses on the multi-object human pose estimation model to locate the key points or limbs of the human body. The current mainstream methods for multi-object scene pose estimation are implemented by two ways: top–down and bottom–up. The former firstly locates the human body target frame based on the target detection algorithm and then detects the single bone joint point. The latter firstly detects the key points of multiple targets, and then connects the detected key points as the individual using the key point matching algorithm, thereby completing the multi-object human pose estimation.

2.1 Top–down methods

The top–down methods are based on the multi-object detection and single-person key point detection, focusing on spatiotemporal information and thinking about the core of crucial point detection. Aiming at the difficulty of key point detection, the convolutional pose machine (CPM) model [14] generates a large receptive field area to extract different features of key points, and then solve the weak discrimination of local information caused by background confusion. Targeting at positioning errors and repeated detection, Fang et al. [15] constructed a spatial transformation network and designed the architecture for the AlphaPose model. Chen et al. [16] designed a counter propagation network (CPN) to detect the key points of an object with different difficulties. Artacho et al. [17] proposed a unified human pose estimation framework (UniPose), with a cascade structure, which can draw a multi-scale field of view comparable to a spatial pyramid. Rafi et al. [18] put forwarded an affinity map to represent the inter-frame correspondences to recover omissions and estimate poses across video frames. These methods all exhibit high accuracy but poor real-time performance. However, they are limited to small-sized object images and lack computing resources for multi-object detection based on bounding box with high complexity.

2.2 Bottom–up methods

The bottom–up methods are based on multi-object key point detection and critical point clustering. The multi-object key point detection here is the same as the single-person key point detection in the top–down methods, which mainly solves the key point clustering. Xia et al. [19] introduced the image segmentation to obtain the key point information, based on which the limbs are connected through clustering explicitly. The well-known pose estimation model (OpenPose) [20, 21] adopts the custom vector Part Affinity Fields (PAFs) algorithm and the vector field to simulate the torso structure of human limbs, solving the misconnection key points. Recently, many novel methods have been developed for multi-object three-dimensional (3D) human pose estimation. The high-resolution volume heatmaps are recommended to model the joint positions effectively and to design a simple and effective compression method. The latest research of Google Research Institute has launched the gesture detection model (MoveNet) and an application programming inter-

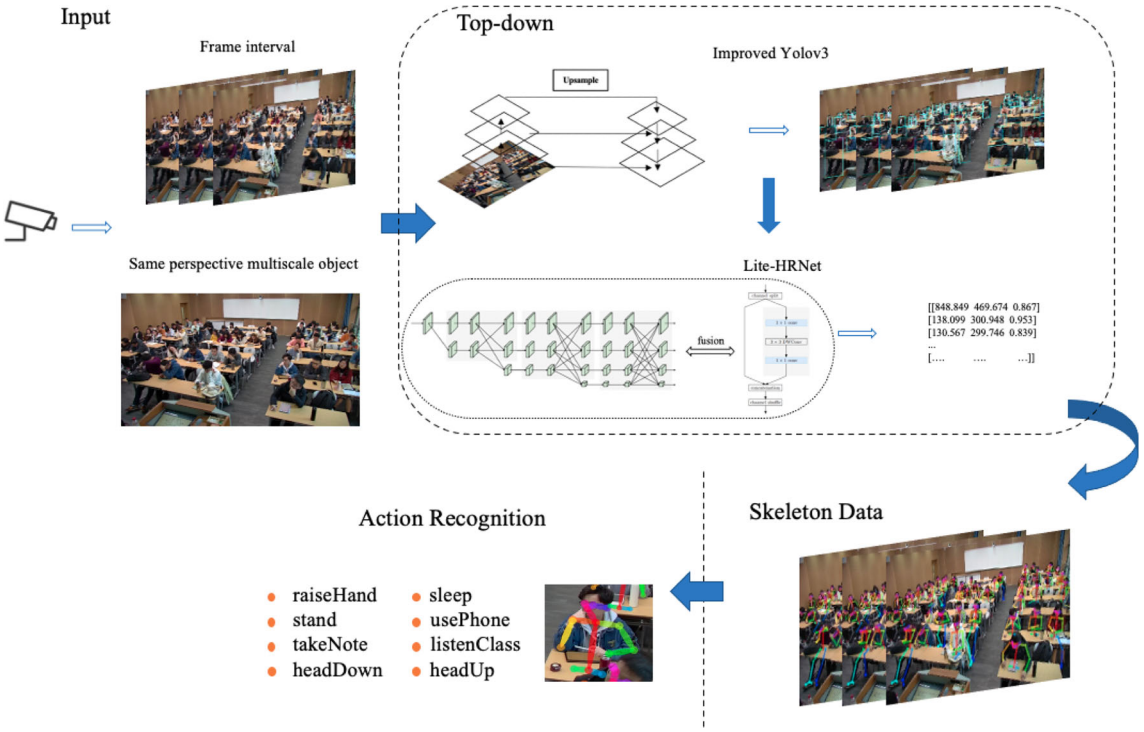

Fig. 1 Diagram of the model architecture

face (API) in TensorFlow.js to detect human key points more quickly and accurately. These methods show low accuracy but good real-time performance. Classic methods such as PersonLab and Offset can be employed to quantify the accuracy of small-size images.

3 Reconstruction the network model

Based on the design idea of top–down methods, the multi-object pose estimation network model is reconstruction in this paper. It integrates spatiotemporal semantics, focuses on the temporal continuity and dependency among frames in the video stream, and aims at solving the multi-object multi-scale problem and realizing the multi-object pose estimation based on behavior recognition of skeleton data. Overall, the network model is oriented to video stream input. Firstly, it detects multiple objects quickly based on an improved You Only Look Once (Version 3) (YOLOV3) structure. Secondly, the lightweight Lite-HRNet [22] model is selected to accurately detect the human key points and then output the multi-object skeleton data. Finally, a specific classifier is designed to classify the information about the human body key points and skeleton data to recognize the behaviors. Overall architecture of the network model is displayed in Fig. 1.

3.1 Multi-object detection

The end-to-end YOLOV3 multi-object detection structure is adopted and optimized. Through training and testing, the optimized YOLOV3 network model is composed of four parts, namely input ($416 * 416$), backbone network (DarkNet53), multi-scale detection, and prediction output. The backbone network relies on the Mish activation function to extract finer features in complex backgrounds, thereby improving the accuracy and generalization ability of multi-object detection. Before multi-scale detection, a $2 * 2$ convolution kernel is introduced for maximum pooling, and the dimension and the amount of calculation are reduced for a fast multi-object detection. Figure 2 demonstrates the improved YOLOV3 structure.

3.2 Detection of human key points

Taking multi-object bounding box among different frames in the video stream as input, the key point information of multiple targets is extracted. In consideration of scale change, a HRNet model with high resolution and sensitive to location information is selected herein. The network structure can be divided into multiple stages. In each stage, information of feature maps with different resolutions is fused through upsampling and downsampling. In this paper, the

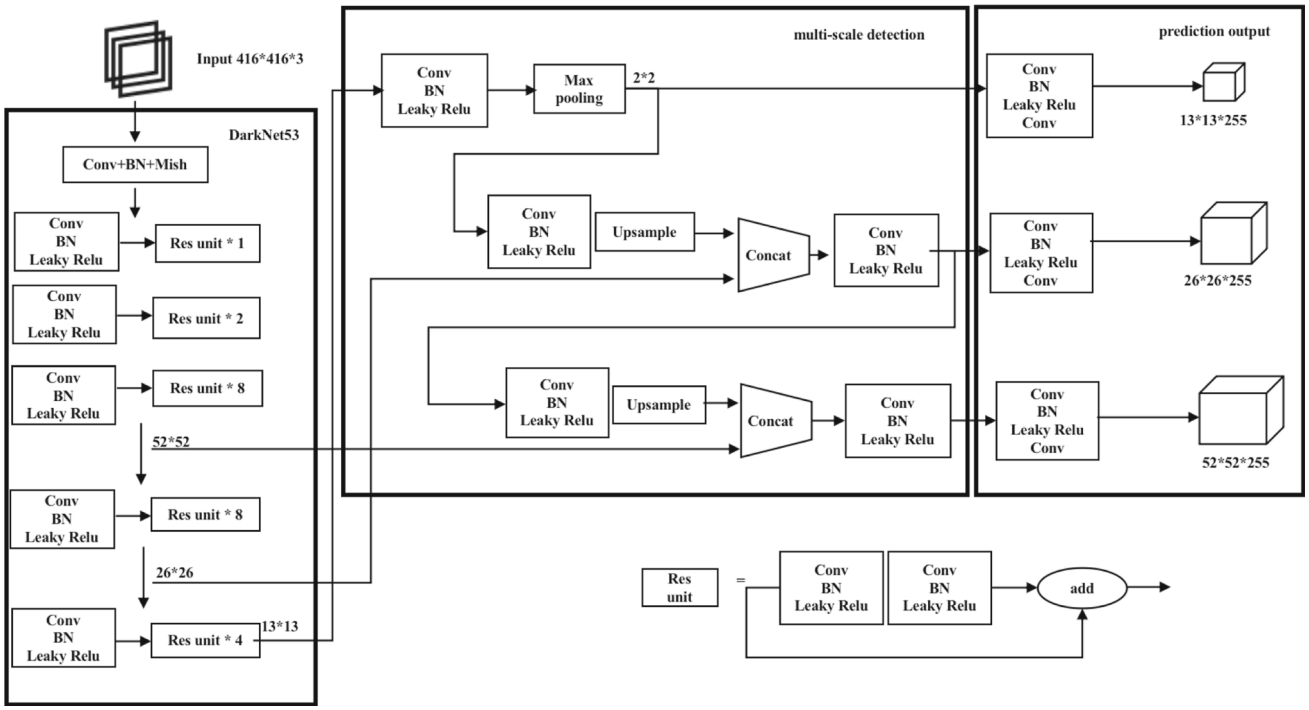

Fig. 2 Diagram of the improved YOLOV3 network structure

Lite-HRNet model that directly combines the parallel branch structure of HRNet and the shuffle module of ShuffleNet is selected after the model complexity is measured. That is, the weights are learned from the feature maps with different resolutions and their channels in the HRNet model, and different resolutions interact information with their channels through the weights according to the shuffle module structure. With multi-resolution supervised training and aggregated inference, it can more accurately solve the scale change caused by multi-object pose estimation and localization of local key point, and showing a better robustness in complex scenes.

Herein, the key point heatmap information among different interval frames is analyzed, using F_{p}, F_n, F_{f} and to represent the previous frame, the current frame, and the future frame, respectively, with an interval time of T. In this paper, based on the temporal cues, we expect to obtain richer semantic information from F_{p} and F_{f} to solve the problems of pose occlusion and motion blur. Due to the belief in time consistency, individual poses exhibit no sudden change in a short time interval. Then, a similarity-pose-temporal-merging (S-PTM) module is designed to process the pose estimation model that fuses spatiotemporal semantics. For an individual i, the result after multi-object detection is marked as $d_i\,(p, n, f)$, which serves as input of the key point detection network, and the initial key point heatmap information $h_i\,(p), h_i\,(n)$, and $h_i\,(f)$ are outputted. Difference in individual target poses is minimized in a short time interval, and all the extracted heatmap information of key points is averaged in a long time interval $f - p$, as expressed as follows:

$$H\,(p, n, f) = \frac{h_i\,(p) + h_i\,(n) + h_i\,(f)}{f - p}. \tag{1}$$

In fact, convolutions compute adjustable weights, so the convolutional neural network (CNN) are employed to implement Eq. (1). During the network calculation, all key points heatmap information will lead to feature redundancy and computational complexity when a single key point heatmap information is merged in a channel containing. In this case, the group convolution network (GCN) is performed according to the total number of key points N of the human body. For the key point j of an individual i, the heatmap information obtained in F_{p}, F_n, and F_{f} is $h_i^j\,(p)$, $h_i^j\,(n)$, and $h_i^j\,(f)$, respectively. After the connection operation and the residual module are processed effectively, output of the network model is the final individual merged key point heatmap $\varphi_i\,(p, n, f)$, which is expressed as Eq. (2) below:

$$\varphi_i\,(p, n, f) = \bigoplus_{j=1}^{N} \frac{h_i^j\,(p) \bigoplus h_i^j\,(n) \bigoplus h_i^j\,(f)}{f - p}. \tag{2}$$

Finally, $\varphi_i\,(p, n, f)$ is fed into the pose correction module, adopting the deformable convolution of Deformable Convolutional Network (DCN) network [23] to improve and fine-tune the results of multi-object pose estimation. Results of the single-person key point aggregation are visualized, as displayed in Fig. 3. F_{p}, F_n, and F_{f} in the figure refer to the images of the positioned human body target frames. The intercepted detection image of the key point of its left shoul-

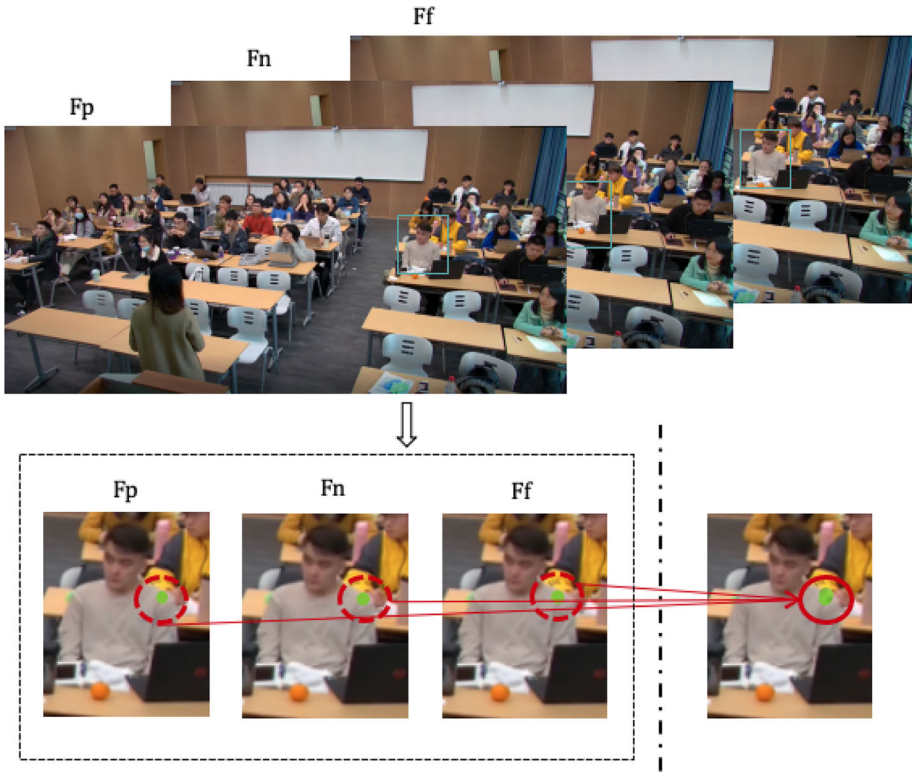

Fig. 3 Visual results of human body key point aggregation

der is given in left side of dotted line. The right side of the dotted line shows the merged key point heatmap after convolution operation using Eqs. (1) and (2).

3.3 Recognition of classroom behaviors

The key point coordinates of human 2D skeleton and the human skeleton map outputted by the reconstructed pose estimation network model are saves in this paper. Meanwhile, the two types of skeleton feature data are processed using data analysis and image processing techniques, and different behaviors of students are identified with a classifier.

3.3.1 Data preprocessing

1. Key point coordinates of human 2D skeleton

The pose estimation algorithm tracks and outputs the key point coordinates of the human 2D skeleton, and then the critical point coordinates are extracted and normalized. The pose estimation is performed based on the 25 fundamental point model of the human body based on the COCO dataset, as shown in Fig. 4.

Since classroom behaviors of most students are based on the sitting posture, the monitoring camera mostly focuses on the upper body area of the students, and the lower body is almost blocked. Therefore, the lower body nodes of the human skeleton are removed, and the irrelevant skeleton feature points are excluded in this paper. The identified interference and the final extracted feature points are listed in Table 1. That is, the eight key points (0 - 7 in Fig. 4), including the nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, and left wrist, are selected as feature points to identify classroom behaviors of students.

Based on the coordinates of the 2D skeleton key points, the data preprocessing techniques adopted in this paper include normalizing the position of the skeleton key points and defining the connection lines of the skeleton limbs. After angles of the skeleton limbs are defined, the feature vectors of different behaviors are formed.

(1) Key point localization

Different distances between students and classroom cameras [24] cause various target sizes in each classroom video data, resulting in different coordinate dimensions and proportions of key points of the 2D skeleton. Moreover, positions of the same bone key points of different target objects will also be quite different due to the individual differences of the targets, lowering the prediction accuracy. The standardization method is adopted, and the $[x_i, y_i]$ coordinate point of the skeleton key point i is normalized by Eq. (3). The subsequent $\left[x_i^{'}, y_i^{'}\right]$ coordinates contain the features corresponding to the 8 key points. Among them, img_{width} and img_{height} refer to the width and size of the frame image, respectively. In this experiment, the resolution value is $1,920 * 1,080$, $x_{\max}$ represents the most significant value in the extracted skeleton

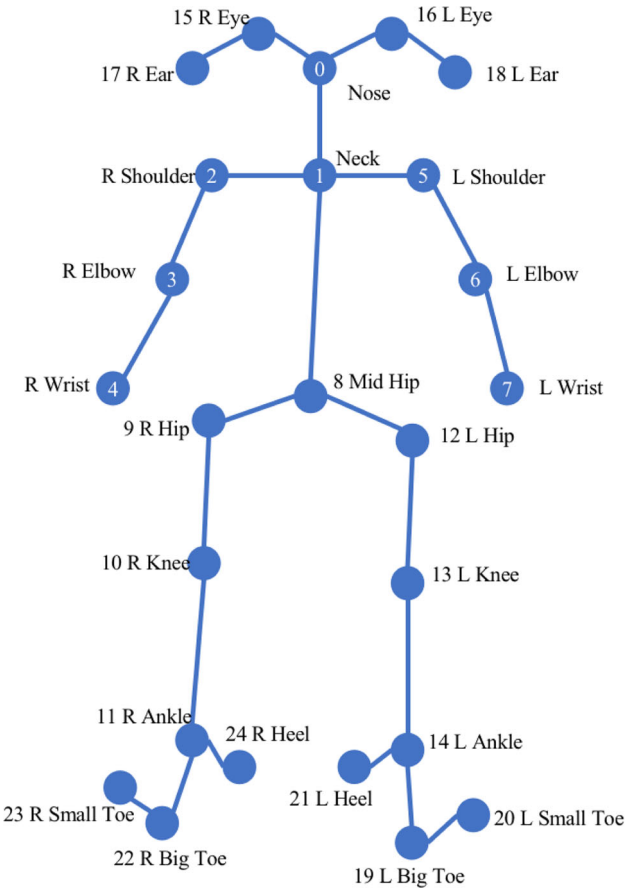

Fig. 4 25 key point model of human body

Table 1 Main human key points to identify the classroom behaviors of students

Number	Location	Number	Location
0	Nose	4	R Wrist
1	Neck	5	L Shoulder
2	R Shoulder	6	L Elbow
3	R Elbow	7	L Wrist

coordinates, and $x_{\min}$ denotes the smallest value. $y_{\max}$ and $y_{\min}$ have the same meanings.

$$\left[x_i^{'}, y_i^{'}\right] = \left[\frac{img_{\text{width}}}{2} * \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \frac{img_{\text{height}}}{2} * \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}\right] \tag{3}$$

(2) Skeleton limb connection

The Euclidean distance between any two critical points of the upper body is defined as the connection line of the skeleton limb [25]. (x_i, y_i) and (x_j, y_j) are coordinates of the critical point i and the key point j, respectively, as shown in Eq. (4). The connection of five limbs: nose and neck, right shoulder and right elbow, right elbow and right wrist, left shoulder and left elbow, and left elbow and left wrist, is shown in Fig. 5.

$$\text{dist}_{i,j} = \sqrt[2]{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{4}$$

(3) Skeleton limb angle

The angles between the limb connection line of the skeleton and the north direction are defined in Eqs. (5)–(9). In the equations, $\text{dist}_{i,j}$, $\text{horizontal}_{i,j}$ and $\text{vertical}_{i,j}$ represent the limb connection line, the horizontal distance, and the vertical distance between key points i and j, respectively. Herein, the angles between the line connecting the nose and the neck, the line connecting the right elbow and the right wrist, the line connecting the left elbow and the left wrist with the north direction, the extension of the line connecting the right shoulder and the right elbow, the extension of the line connecting the left shoulder and the left elbow with the horizontal direction are mainly described. The included angles are illustrated in Fig. 6.

$$\text{horizontal}_{i,j} = |x_i - x_j| \tag{5}$$

$$\text{vertical}_{i,j} = |y_i - y_j| \tag{6}$$

$$p = \frac{\text{vertical}_{i,j}^2 + \text{dist}_{i,j}^2 - \text{horizontal}_{i,j}^2}{2 \times \text{vertical}_{i,j} \times \text{dist}_{i,j}} \tag{7}$$

$$\text{angle}_{i,j} = \arccos(p), \text{where}[i, j] = \{[0, 1], [3, 4], [6, 7]\} \tag{8}$$

$$\text{angle}_{i,j} = 90^\circ + \arccos(p), \text{where}[i, j] = \{[2, 3], [5, 6]\} \tag{9}$$

2 Diagram of human skeleton

Based on the calibrated classroom behavior data, the object is cropped according to the object detection frame. Diagram of human skeleton of 6 classroom behaviors is shown in Fig. 7, including raiseHand, sleep(lie), stand, usePhone, takeNote, and listenClass. The feature data set of the human skeleton map is constructed according to the classroom behaviors. The numbers of samples in the training set, test set, and validation set are given in Table 2.

3.3.2 Data classifier

Aiming at the two different behavior feature data of the coordinates of 2D skeleton key points and human skeleton map, two classifiers are employed to identify the classroom behaviors of students. Three machine learning classification algorithms are adopted to classify the behavioral feature data of the coordinates of 2D skeleton key points: support vector machine (SVM) [26], decision tree [27], and random forest [28].

For the behavioral feature data of the human skeleton map, the transfer learning is selected in this paper to reduce the

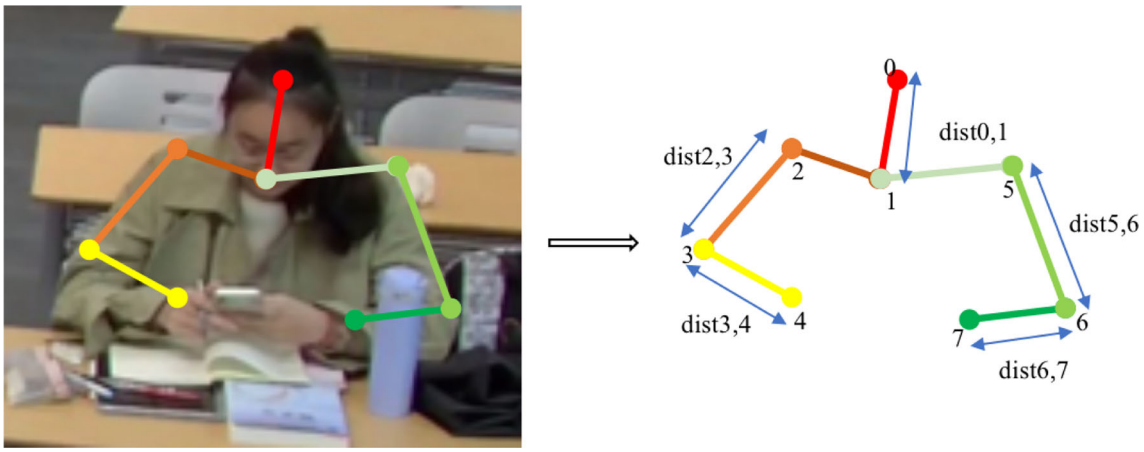

Fig. 5 Skeleton limb connection

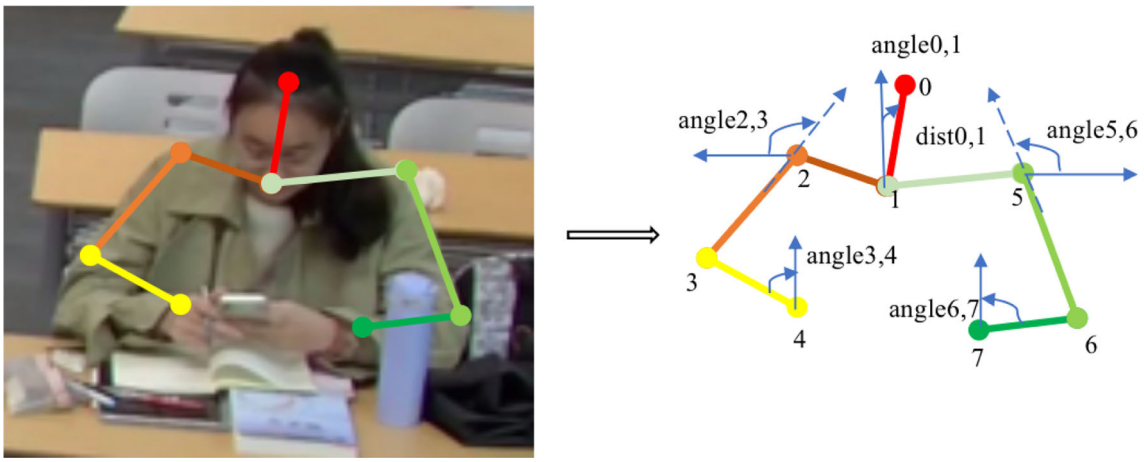

Fig. 6 Skeleton limb angles

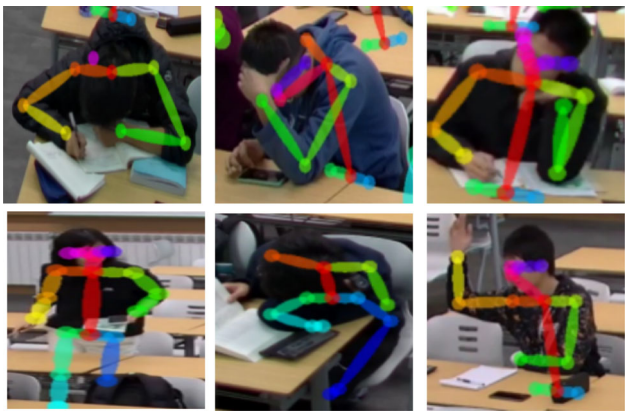

Fig. 7 Diagram of human skeleton for 6 classroom behaviors

dependence of the deep learning accuracy on the dataset and maximize the utilization of the existing dataset [29]. Transfer learning means transferring from a previous classification or prediction task to a new task. In fact, it reuses weights learned from models pretrained on large datasets and then to retrain

Table 2 Numbers of samples of human skeleton map feature dataset

Student behavior	Train sample	Test sample	Val sample
raiseHand	104	52	36
sleep(lie)	20	12	32
stand	472	256	20
takeNote	888	444	16
listenClass	176	88	48
usePhone	1164	588	100

the remaining layers or fine-tune the network. In this paper, the pre-training model of VGG16 is employed to classify the classroom behaviors of students. Figure 8 illustrates the network for recognizing the classroom behaviors of students.

4 Experiments

Experiments in this paper are conducted based on the enterprise servers, which are equipped with the operating system

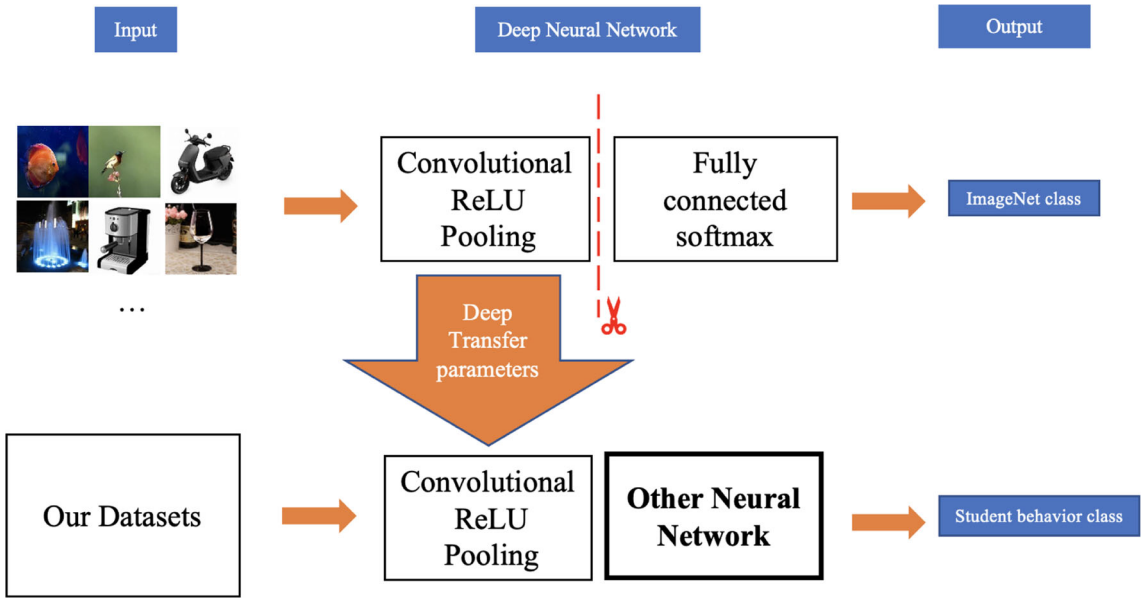

Fig. 8 Network for recognizing classroom behaviors of students

Table 3 Experimental equipment indicators

Server equipment	Equipment model
Graphics card	NVIDIA RTX 3080 S
Central processing unit	AMD R5 3700X
Memory	DDR4 3600 32 G
Disks	NVME 500 G

of Ubuntu18.04, the programming language of Python, and the deep learning framework of PyTorch [30]. According to the network model reconstructed in this paper, multi-object pose estimation is performed firstly based on the public data set. Compared with other Top–down methods, the improved network exhibits a high rationality, which has been verified by the experimental results in this paper. Secondly, after a large number of classroom behavior data sets are collected, the classroom behaviors are recognized based on a multi-object pose estimation network that integrates spatiotemporal semantics. Various classifier algorithms are employed for different output skeleton data in experiments to select the best classifier, thus improving the recognition accuracy. The experimental equipment indicators in this paper are shown in Table 3.

4.1 Experimental analysis of multi-object pose estimation

In this paper, based on the reconstructed pose estimation model, the video stream is inputted to the Improved YOLOV3 network to achieve the multi-object detection. Taking the multi-object bounding box among different frames as input, the lightweight key point detection network model is adopted to extract the critical point information of the multi-object and to output multi-object skeleton data, thus estimating the human poses accurately.

4.1.1 Experimental data

The PoseTrack2018 [31], a large-scale public dataset, which contains single-frame and multi-frame critical points in the video, is adopted for human pose estimation and key point tracking, including complex crowd movement in crowded environments. The PoseTrack2018 greatly increases the number of video clips, with a total of 1138 video clips and 15,3615 pose annotations. The training set, validation set, and test set cover 593, 173, and 384 video sets, respectively. The 30 frames in the center of the training set are densely annotated, and annotations are provided every four frames in the validation set. PoseTrack2018 identifies 15 human body key points and adds annotation labels for joint visibility. Then, the self-built video data sets are tested.

4.1.2 Parameter setting

Experiments are based on the time intervals T among F_p, F_n, and F_f, which are set to 1, 3, and 5, respectively. The backbone parameters are fixed on the pre-trained Lite-HRNet model weights. Here, all weight parameters are then initialized with a Gaussian distribution, where $\mu = 0$, $\sigma = 0.001$, and the bias parameter is initialized to 0. The stochastic gradient descent (SGD) optimizer with an initial learning rate

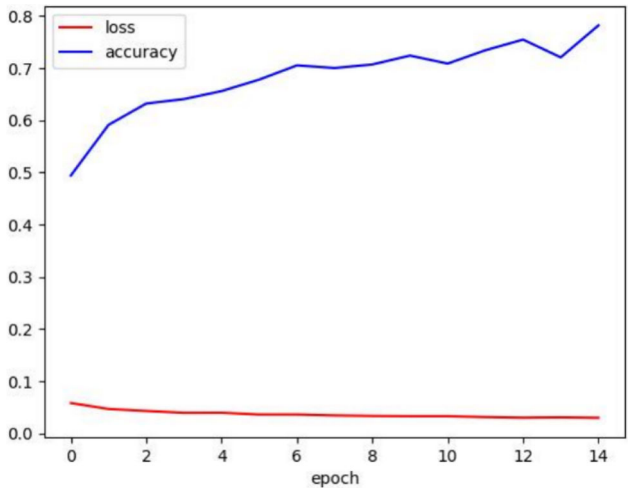

Fig. 9 Training process of the reconstructed network model

of 0.0001, 15 epochs, and a batch size of 32 sizes is selected to train the network model in this work. The standard loss function for pose estimation is used, which is defined in Eq. (10). The training goal is to minimize the Euclidean distance between the predicted heatmap and the true heatmap for all key points.

$$\text{loss} = \frac{1}{N} \|G(j) - P(j)\|^2 \tag{10}$$

where $G(j)$ and $P(j)$ denote the ground-truth heatmap and the predicted heatmap, respectively, and the total number of joints N is set to 15. The ground-truth heatmaps are generated from the 2D Gaussian centers of joint positions.

4.1.3 Analysis of experimental results

In this paper, the top–down pose estimation network model is reconstructed, and the video stream data are based on the improved YOLOV3 to realize the multi-object detection. Then, the object detection frame is inputted into the human key point detection network to compare the pose estimation results. Figure 9 illustrates the changes in loss and accuracy of the network model during the training, based on which the effectiveness of the pose estimation accuracy improvement is obtained. The accuracy of different pose estimation models is calculated and compared in Table 4. The results suggest that reconstructed model in this paper exhibits the least number of parameters and moderate number of processed images and frames per second. In addition, its accuracy reaches 80.1%, which is 12.5%, 2.1%, and 5.2% higher than that of the AlphaPose, PoseWarper [32], and HRNet [33] models, respectively.

Table 4 Accuracy of different pose estimation models

Method	Param(M)	FPS	AP(%)
AlphaPose	67.5	10.1	67.6
PoseWrapper	68.8	8.5	78.0
HRNet	63.8	9.7	74.9
The proposed model	60.7	8.9	80.1

Fig. 10 Classroom behavior states of students

4.2 Experiment analysis of behavior recognition based on multi-object pose estimation

Based on the reconstructed pose estimation model, the classroom behaviors of students in teaching videos are recognized in this paper. Meanwhile, the coordinates of 2D skeleton key points and the human skeleton map are employed to predict and classify the classroom behaviors of students. It is believed that the analysis results can effectively assist classroom teaching.

4.2.1 Dataset

The experimental research is to obtain classroom data and calibrate behaviors. Based on the constructed data sample set, machine learning and deep learning methods are adopted to train the model so that the classroom behaviors of students can be identified and classified more accurately. The preliminary task of this research is to construct a public data set of classroom behaviors. The videos for data set building are obtained from the actual teaching scene in colleges and universities. The Hikvision's high-definition camera is installed in the classroom, which can output a 1080p mp4 format file. The virtual classroom video state is shown in Fig. 10. Based on the research and calibration of video frames, 8 obvious classroom behaviors are summarized and calibrated: headUp, headDown, raiseHand, sleep(lie), stand, usePhone, takeNote, and listenClass. The final calibration data sets are listed in Table 5.

Table 5 Datasets for classroom behaviors

Student behavior	Sample number	Student behavior	Sample number
headUp	2096	raiseHand	156
headDown	3742	sleep	188
Description: the sample is marked as single-object multi-action		stand	235
		takeNote	1506
		usePhone	1038
		listenClass	4911

Table 6 Accuracy of machine learning classification models

Machine learning classification algorithms	Accuracy (%)
Poly kernel SVM	73.66
Linear kernel SVM	67.95
RBF kernel SVM	37.63
Sigmoid kernel SVM	37.55
Decision tree	58.64
Random forest	62.86

4.2.2 Machine Learning algorithms

The classifier with higher accuracy is selected in this paper. Decision tree and random forest are looped through parameters to filter the best parameter settings. SVM selects the best parameters using the cross-validation and grid search methods. Four mainstream kernel functions are set, which are Poly, Linear, Gaussian, and Sigmoid. The penalty coefficient C is initialized to 0.001. The size is increased by ten times, the highest poly degree is initialized to 1, and the training is increased by 1 step. The accuracy of the final classification test results is calculated, as listed in Table 6.

4.2.3 Deep learning network

The human skeleton image only contains the skeleton information of the upper bodies of students. In consideration of unbalanced samples in the dataset, the sample size for training data is expanded by using some data enhancement techniques such as image flipping and scaling, adding noise, and blurring color transformation. The input image size is $224 * 224$, the batch_size is 20, the training epochs is set to 200, the optimizer is Adam, the learning rate is initialized to 0.001, and the cross-entropy loss function is adopted. The training process is illustrated in Fig. 11. The line chart shows that the test accuracy is as high as 70%, and the loss is as low as 0.05, suggesting that the model can effectively identify the classroom behaviors of students.

4.2.4 Experimental results

The results of 2D skeleton coordinates and human skeleton images are obtained based on the reconstructed pose estimation model. After several experiments, accuracy of the two types of classifiers under different feature data is calculated and analyzed, as displayed in Table 7. The behavior vector is formed by superimposing a variety of feature data. It observes that the accuracy is up to 73.66%, which is 9.1% higher than that of the original skeleton coordinates. A network model based on deep learning is constructed and the number of trainings of human skeleton images is increased after random rotation ($-45° - 45°$), random zoom, and random horizontal flip. In this way, the results are more generalized, and the recognition accuracy is improved to 70.15%.

The classification reports of six behaviors of students in the classroom in the real teaching scene are displayed in Table

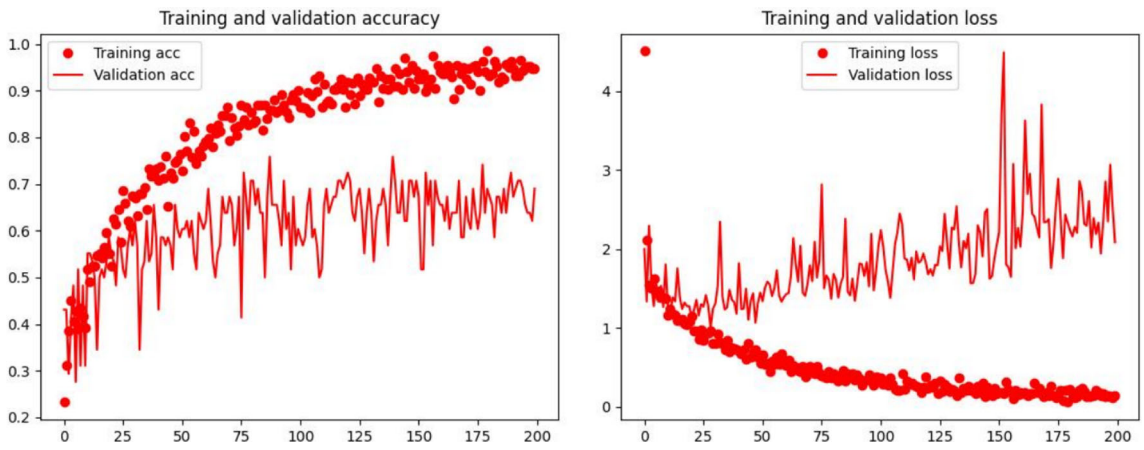

Fig. 11 Training process of deep learning network model

Table 7 Comparison of the results of the two types of classifiers

Classification algorithm	Feature data	Accuracy (%)
Machine learning algorithms	Original skeleton coordinates	64.56
	Key point location	60.75
	Key point location and limb connection	68.59
	Key point position and limb connection and limb angle	**73.66**
Deep learning network	Human skeleton picture	64.06
	Human skeleton picture after data enhancement	**70.15**

Table 8 Classification reports of the six classroom behaviors

Index	Precision	Recall	$F1$-score	Support
raiseHand	0.67	0.75	0.71	8
sleep	0.33	0.33	0.33	3
stand	0.90	0.69	0.78	13
takeNote	0.60	0.60	0.60	40
listenClass	0.68	0.94	0.79	200
usePhone	0.65	0.51	0.57	77
accuracy	0.74			341

Table 9 Classification reports of the other two classroom actions

Index	Precision	Recall	$F1$-score	Support
HeadDown	0.88	0.75	0.81	964
HeadUp	0.72	0.87	0.79	730
Accuracy	0.80			1694

8, namely raiseHand, sleep(lie), stand, usePhone, takeNote, and listenClass. The report shows that the prostrate sleep will cause missed and misjudgment due to occlusion of the target object. The precision rate and recall rate of the other five behaviors are relatively high, and the $F1$ score is basically maintained above 0.55, which effectively proves the strong applicability of proposed method.

The headUp and headDown are directly classified and recognized by calculating and judging the angles of the key points of the 2D skeleton, and the specific classification report is shown in Table 9. The precision rate, recall rate, and $F1$ score for classifications of headUp and headDown are all above 0.70, and the overall recognition accuracy is as high as 80%. Compared with the above six behaviors, these two behaviors that are only judged based on the angle of the head are easier to distinguish.

By visualizing the recognition effect of classroom behaviors with a frame interval of 1 s, Fig. 12 shows the recognition results of headUp and headDown under 4 consecutive frames (a, b, c, and d). Figure 13 compares the three successive frames (a, b, and c). Herein, the machine learning algorithms are employed to classify the 2D skeleton coordinates, and the deep learning network-based classification methods of human skeleton images to identify 6 classroom behaviors

Fig. 12 Visualized recognition results of headUp and headDown

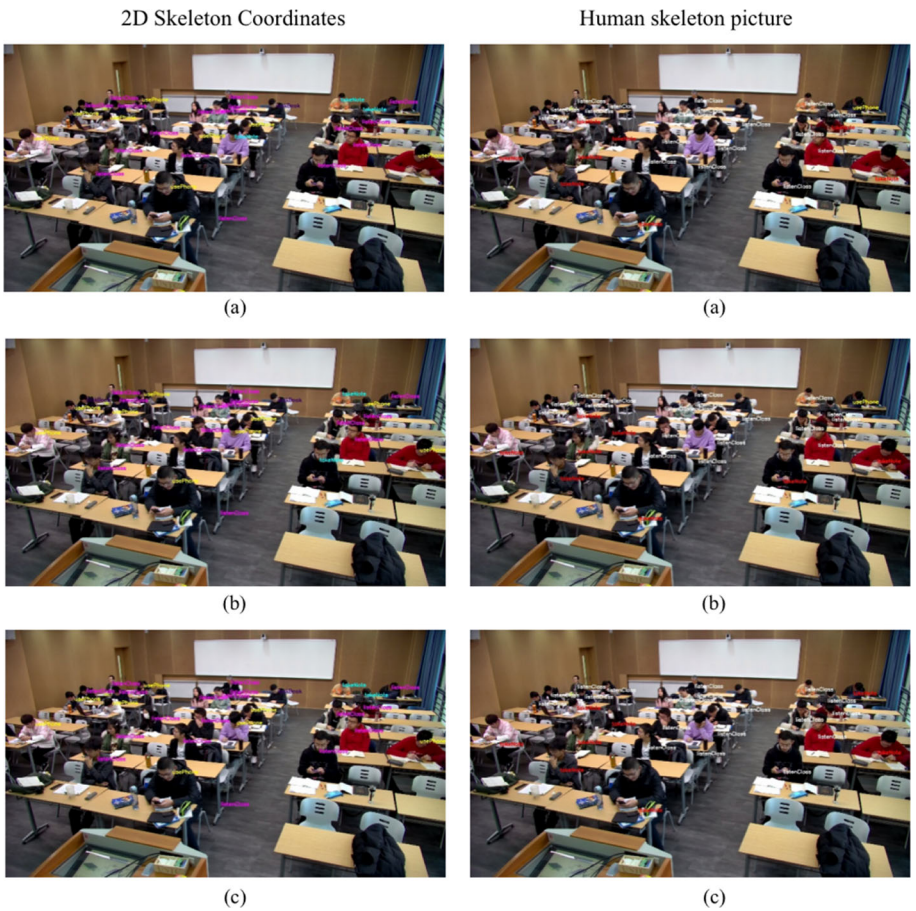

Fig. 13 Visualized recognition results of the two-class classifiers

of students. The visualization results verify the accuracy of the selected model in identifying the classroom behaviors of students based on pose estimation.

5 Conclusion

A human skeleton behavior recognition model for multi-object pose estimation that integrates spatiotemporal semantics is built in this paper. Firstly, the multi-object detection model of the YOLO series is improved based on the Top–down design. Secondly, the lightweight HRNet is selected to estimate the multi-object poses more accurately. Finally, the multimodal data saved are analyzed and outputted by the pose estimation model; that is, the 2D skeleton coordinates and skeleton picture information are combined to predict the recognition effect of the classroom behaviors of students.

Different from traditional classroom behavior recognition algorithms, the rich skeleton data adopted in this paper overcome the ambiguity of multi-class classroom behavior performance. The results of comparative experiments prove that the feature vectors and skeleton pictures introduced into the design can describe the students' behaviors better, making their behaviors more specific and visualized. Besides, the model is trained and optimized for different classroom scenarios to continuously strengthen the generalization ability of the model, improving its recognition effect under the clear monitoring perspective of a single camera. The conclusion in this paper can support an in-depth analysis of classroom teaching and help the high-quality development of education and teaching.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Luo, Y., Ren, J., Wang, Z., et al.: LSTM pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5207–5215 (2018)
2. Wang, M., Tighe, J., Modolo, D.: Combining detection and tracking for human pose estimation in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11088–11096 (2020)
3. Liu, Z., Chen, H., Feng, R., et al.: Deep dual consecutive network for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 525–534 (2021)
4. Newell, A., Huang, Z., Deng, J.: Associative embedding: end-to-end learning for joint detection and grouping. Adv. Neural Inf. Process. Syst. **30** (2017)
5. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: composite fields for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11977–11986 (2019)
6. Papandreou, G., Zhu, T., Chen, L.C., et al.: Personlab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–286 (2018)
7. Cheng, B., Xiao, B., Wang, J., et al.: Higherhrnet: scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5386–5395 (2020)
8. Lin, J., Gan, C., Han, S., TSM: temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7083–7093 (2019)
9. Feichtenhofer, C., Fan, H., Malik, J., et al.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
10. Xu, C., Govindarajan, L.N., Zhang, Y., et al.: Lie-x: depth image based articulated object pose estimation, tracking, and action recognition on lie groups. Int. J. Comput. Vis. **123**(3), 454–478 (2017)
11. Baek, S., Shi, Z., Kawade, M., et al.: Kinematic-layout-aware random forests for depth-based action recognition. arXiv preprint arXiv:1607.06972 (2016)
12. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)
13. Wang, L., Xiong, Y., Wang, Z., et al.: Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision, pp. 20–36. Springer, Cham (2016)
14. Wei, S.E., Ramakrishna, V., Kanade, T., et al.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
15. Fang, H.S., Xie, S., Tai, Y.W., et al.: RMPE: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2334–2343 (2017)
16. Chen, Y., Wang, Z., Peng, Y., et al.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112 (2018)
17. Artacho, B., Savakis, A.: Unipose: Unified human pose estimation in single images and videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7035–7044 (2020)
18. Rafi, U., Doering, A., Leibe, B., et al.: Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In: Computer Vision-ECCV: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16, vol. 2020, pp. 36–52 . Springer (2020)
19. Xia, F., Wang, P., Chen, X., et al.: Joint multi-person pose estimation and semantic part segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6769–6778 (2017)
20. Cao, Z., Simon, T., Wei, S.E., et al.: 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 (2017)
21. Cao, Z., Hidalgo, G., Simon, T., et al.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
22. Yu, C., Xiao, B., Gao, C., et al.: Lite-hrnet: a lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10440–10450 (2021)
23. Zhu, X., Hu, H., Lin, S., et al.: Deformable convnets v2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9308–9316 (2019)
24. Yuting, Bai: Research on student action recognition method based on video. Instrum. User **27**(1), 10–12 (2020)
25. Lin, F.C., Ngo, H.H., Dow, C.R., et al.: Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. Sensors **21**(16), 5314 (2021)
26. Xue, H., Yang, Q., Chen, S.: SVM: support vector machines. In: The Top Ten Algorithms in Data Mining, pp. 51–74. CRC, Chapman and Hall (2009)
27. Rokach, L., Maimon, O.: Decision trees. In: Data Mining and Knowledge Discovery Handbook, pp. 165–192 (2005)
28. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
29. Abdallah, T.B., Elleuch, I., Guermazi, R.: Student behavior recognition in classroom using deep transfer learning with VGG-16. Procedia Comput. Sci. **192**, 951–960 (2021)
30. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. **32** (2019)
31. Andriluka, M., Iqbal, U., Insafutdinov, E., et al.: Posetrack: a benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5167–5176 (2018)
32. Bertasius, G., Feichtenhofer, C., Tran, D., et al.: Learning temporal pose estimation from sparsely-labeled videos. Adv. Neural Inf. Process. Syst. **32** (2019)
33. Sun, K., Xiao, B., Liu, D., et al.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.