Advances in Economics, Business and Management Research, volume 654
Proceedings of the 2022 International Conference on Urban Planning and Regional Economy (UPRE 2022)

Research on Students' Classroom Behavior Recognition Based on Pose Information Extraction and Local Feature Segmentation

Chenyi Cong *

Nanjing University of Finance & Economics, School of Applied Mathematics, Nanjing, Jiangsu, 210023
** congchenyi1016@163.com*

ABSTRACT
Based on the extraction of human posture information, students' classroom behavior will be identified after local feature segmentation. Because it is challenging to collect classroom behavior samples and the school students are numerous, the existing methods are difficult to obtain good generalization ability. This paper defines six classroom behaviors of "looking at the blackboard," "looking around," "sleeping," "playing mobile phone," "taking notes" and "reading", and uses OPENPOSE posture detection network to extract the pose information of middle school students in the image, and then identifies the head pose and the surrounding environment of hands through HPRNN and LFRCNN to obtain the student classroom behavior. Experimental verification shows that this method can identify multiple students' behaviors in the same network under the condition of ensuring recognition accuracy, which effectively alleviates the problem that neural network is difficult to train due to insufficient sample size, and avoids the decrease of network generalization ability caused by students' different clothing and posture to a certain extent.

Keywords: *OPENPOSE posture detection, HPRNN, LFRCNN, generalization ability*

1. INTRODUCTION

Behavioral Identification refers to identifying behaviors in progress in a particular environment, such as surgical action identification, driving status identification, classroom behavior identification. As we all know, if the classroom order is confusing, the acceptance of knowledge is low. With the deepening of the process of teaching informatization and intelligence, in order to make teachers more focused on teaching professional knowledge, people want to build an automated classroom teaching management system, so build a set of systems that can accurately identify students' classroom behavior is necessary.

At present, domestic and foreign scholars have been related to the recognition of student classroom behavior. Zhang Hongyu[1]extracted the feature of the human bone vector and then recognized the action vector using an SVM classifier. Dang Dongli[2]described and judged the action by extracting the optical flow feature, Zernike moment feature and global motion direction feature of the action, and then combined with the naive Bayesian classifier. Zhou Pengxiao et al.[3]obtained the data set through contour detection, face detection and subject action amplitude detection, and took Bayesian Causal Network as the reasoning model to judge the subject behavior characteristics, so as to identify the classroom teaching behavior. SHI Y[4]et al. proposed a student behavior recognition method based on Fisher Broad Learning System (FBLs), and defined seven classroom behaviors: turning head, raising hands, reading, dozing, listening, writing and standing.

Lin Canran et al.[5]used multimodal data such as human key information, RGB images, etc., and identified students in classroom behaviors. After eliminating Gaussian noise, Su Chao et al.[6]first integrated the attention machine into the target detection algorithm to detect the position of target students in the image, then extracted the human bone joint point coordinates from the detected image through the improved OPENPOSE model, and finally obtained the classification of joint point coordinates by using ST-SVM classifier, so as to quickly and accurately identify the behavior state of learners. Guo Junqi et al.[7]proposed yolo-v5 model with improved loss function to identify classroom behavior with multi-objective as the main feature. Huang Yongkang et

al.[8]combined with real-time target detection and tracked to obtain the real-time picture stream of each student, and then used the depth time and space residual convolution neural network to learn the time and space characteristics of each student behavior so as to realize the real-time recognition of classroom behavior facing multiple student targets in the classroom teaching scene.

This paper defines six classroom behaviors of "looking at the blackboard", "looking around", "sleeping", "playing mobile phone", "taking notes" and "reading", and uses OPENPOSE detection network to extract the pose information of middle school students in the image, and then identifies the head pose and the surrounding environment of hands through HPRNN and LFRCNN to obtain the student classroom behavior

2. DATA ACQUISITION AND PROCESSING

By installing cameras in the classroom and convening volunteers to conduct field simulation shooting, a series of video data containing the behaviors of "looking at the blackboard", "looking around", "sleeping", "playing with mobile phones", "taking notes" and "reading" were obtained.

This paper collects the student classroom behavior in the video in two forms: one is that each student does the same action to establish the data set (training set) for training the neural network; the other is that each student does different actions to test the final recognition effect (test set) of the trained neural network.

After the video data is collected, the image is extracted by uniform frame sampling. Using uniform frame sampling can retain most of the information when the video is converted into an image. Then, the obtained image containing multiple student classroom behaviors is located and cut into a single image. Because the video is continuous data, the action in the acquisition process is also continuous, which leads to the inconsistency between some images after frame sampling and the original behavior. This part of the image should be removed. For example, in the "look around" behavior, when the student's head moves to face the camera, the action will no longer be "look around" and be removed.

3. MODEL ESTABLISHMENT

3.1. Model overview

This paper identifies student classroom behavior based on pose information extraction and local feature segmentation. The schematic diagram is shown in Figure 1.

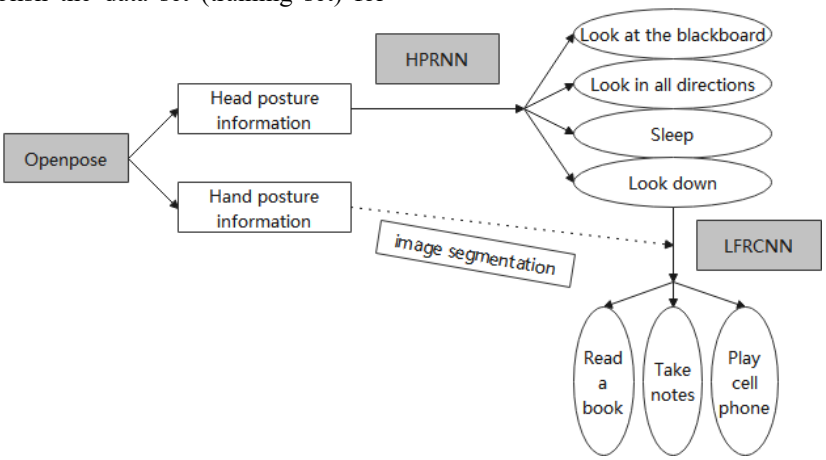

Figure 1 The schematic diagram of Openpose.

After data acquisition and processing, we have made a data set for single student classroom behavior recognition for an image containing a single student. Firstly, the OPENPOSE detection network is used to extract the pose information of students in the image, which is characterized by the coordinates of a group of body parts in the image:

$$S_0 = \left\{ (p_1, p_2, ..., p_i, ..., p_n) \middle| p_i \in R^2 (i = 1,2,..,n) \right\},$$

where, p_i represents the coordinates of the i-th part of the body.

Considering that the student attitude information is independent of the student position in the image, a group of vectors is used to replace the specific coordinates:

$$S_g = \left\{ (v_1, v_2, ..., v_i, ..., v_{n-1}) \middle| v_i \in R^2 (i = 1,2,..,n-1) \right\}$$

, where, v_i represents the attitude vector of the i-th part of the body.

Among them, "looking at the blackboard" is the front face, "looking around" is the side face, "sleeping" is unable to recognize the facial features, and "playing with the mobile phone", "taking notes" and "reading" are all bow their heads. Next, this paper constructs a

head pose recognition neural network (HPRNN) to preliminarily recognize the image. The recognition results are "looking at the blackboard", "looking around", "sleeping" or "looking down".

If the recognition result is bow, segment the image around the hand according to the student attitude information, input the segmented image into the local feature recognition convolutional neural network (LFRCNN) constructed in this paper for further recognition, and output the final recognition result of "playing mobile phone", "taking notes" or "Reading". In practical application, OPENPOSE can effectively extract the attitude information in multi-person images. Both HPRNN and LFRCNN process the extracted results, which has nothing to do with the number of students in the image, so this method is still applicable to images containing multiple students. The reason for making single student classroom behavior data set is that it supervises HPRNN and LFRCNN in this method.

3.2. OPENPOSE attitude detection network

The OPENPOSE detection network[10]is used to extract the pose information of each person in the multi-person image. The network structure is shown in Fig. 2.

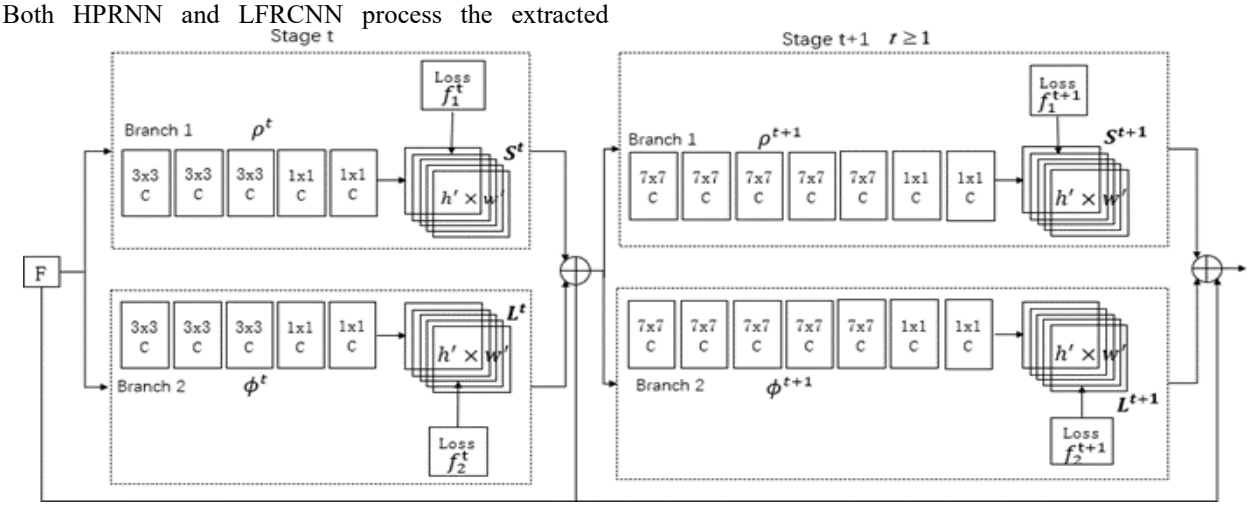

Figure 2 The network structure of Openpose.

The image input OPENPOSE first enters the convolutional neural network composed of VGG-19 top 10 layers of fine-tuning, and obtains the features F of the image, and F enters the two branches of the network for multiple iterative predictions. The first branch is used to predict the position of the body location in the image (Part Confidence Maps, i.e., the joint node), and the second branch is used to predict the part affinity of the body (Part Affinity Fields, the limb torso). The input of the first phase of the network is characteristic F . After processing, S_1 *and* L_1 are obtained, and the subsequent stage, the original image feature is connected to the previous phase, and the affinity domain is connected to generate more accurate prediction results:

$$S^t = \rho^t\left(F, S^{t-1}, L^{t-1}\right), \forall t \geq 2$$
$$L^t = \phi^t\left(F, S^{t-1}, L^{t-1}\right), \forall t \geq 2$$

Gaussian distribution will be generated. The generation formula for the first joint point of the first person:

$$S_{j,k}^*(p) = \exp\left(-\frac{\left\|p - x_{j,k}\right\|_2^2}{\sigma^2}\right)$$

When adding the confidence map generated by each part of each individual, the maximum operator should be used:

$$S_j^*(p) = \max_k S_{j,k}^*(p)$$

The ground truth $L_c^*(p)$ of the connection vector (affinity) is calculated as follows. If the second limb trunk c of the person k is determined in the image, for any position p in the image：

$$L_{c,k}^*(p) = \begin{cases} v, \text{if } p \text{ on limb } c,k \\ 0, \text{otherwise} \end{cases}$$

The true distribution of affinity domains of body parts is the average distribution of affinity domains of all people in the image:

$$L_c^*(p) = \frac{1}{n_c(p)}\sum_k L_{c,k}^*(p)$$, where $n_c(p)$ is the

number of non-zero vectors of all people at the point p.

The key function of part affinity fields is to judge whether the two parts are connected,

$$E=\int_{u=0}^{u=1} L_c\left(p(u)\right)\cdot\frac{d_{j2}-d_{j1}}{\left\|d_{j2}-d_{j1}\right\|_2}du$$

where $p(u)=(1-u)d_{j1}+ud_{j2}$.

If the direction of $\overrightarrow{d_{j1}d_{j2}}$ is consistent with $L_c^*(p)$, the value of E will be too large. It is very likely that the position is a torso.

In order to judge whether the network converges, the loss function is defined: $f=\sum_{t=1}^{T}\left(f_S^t+f_L^t\right)$, in every stage:

$$f_s^t=\sum_{j=1}^{J}\sum_{p}W(p)\cdot\left\|S_j^t(p)-S_j^*(p)\right\|_2^2;$$

$$f_L^t=\sum_{c=1}^{C}\sum_{p}W(p)\cdot\left\|L_c^t(p)-L_c^*(p)\right\|_2^2$$

Because the data annotation may be incomplete, some joint points are not labeled correctly, when $S_j^*(p)=L_c^*(p)=0,\ W(p)=0,$ so as to avoid the loss function value.

3.3. HPRNN recognition head posture

In 3.2, OPENPOSE depicts the pose information of the human head with the coordinates of six parts: left eye, right eye, left ear, right ear, nose and neck. Since the pose is independent of the position of the human body in the image, this paper converts the coordinate information into vector information to represent the head pose. The specific vectors include: left ear - left eye, left eye - nose, right ear - right eye, right eye - nose, nose-neck. After visualization, it is shown in Figure 3, and its network structure is shown in Figure 4:

Figure 3 Head posture diagram.

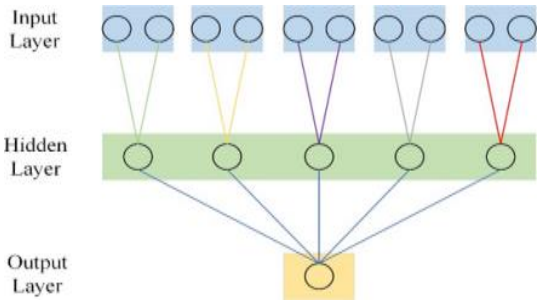

Figure 4 HPRNN structure.

3.4. LFRCNN hand surrounding image feature recognition

LFRCNN is a fine-tuning version of ResNet18 model. In this paper, the categories of its output layer are adjusted to 3, namely "playing mobile phone", "taking notes" and "reading".

The image information around the hand is segmented from the original image. Assuming that $(wl_x, wl_y), (wr_x, wr_y)$ they are the coordinates of the left and right-hand wrists in the image, the image segmentation formula is as follows:

$$w_{x\max}=\max\{wl_x, wr_x\}, w_{x\min}=\min\{wl_x, wr_x\};$$
$$w_{y\max}=\max\{wl_y, wr_y\}, w_{y\min}=\min\{wl_y, wr_y\}$$
$$M'=M_{(w_{x\min}-a:w_{x\max}+b)\times(w_{y\min}-c:w_{y\max}+d)}$$
$$M'_{128\times128}=\mathrm{Resize}(M')$$
M' is the partial image extracted,
$a=0, b=0, c=10, d=10.$

When the hand is blocked or not obvious in the image, the extracted self-love information will default. At this time, the image segmentation formula needs to be modified to estimate the hand position.

When the right wrist coordinate defaults:

$$M'=M_{(wr_x-a:wr_x+b)\times(wr_y-c:wr_y+d)},$$
$$a=60, b=0, c=10, d=10.$$

When the left wrist coordinate defaults:

$$M'=M_{(wl_x-a:wl_x+b)\times(wl_y-c:wl_y+d)},$$
$$a=0, b=60, c=10, d=10.$$

When both wrist coordinates are default:

$$M'=M_{(a:b)\times(c:d)}, a=1, b=128, c=91, d=128.$$

4. MODEL SOLVING

In this paper, a series of experiments were carried out using PyTorch on the collected data sets.

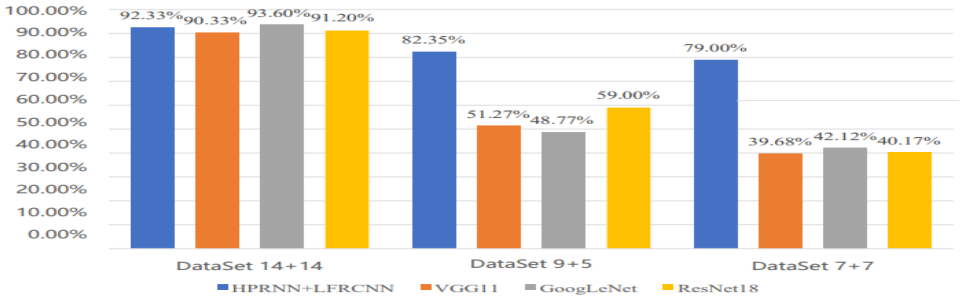

Figure 5 Comparison diagram of generalization performance of different models.

4.1. Comparison of generalization ability

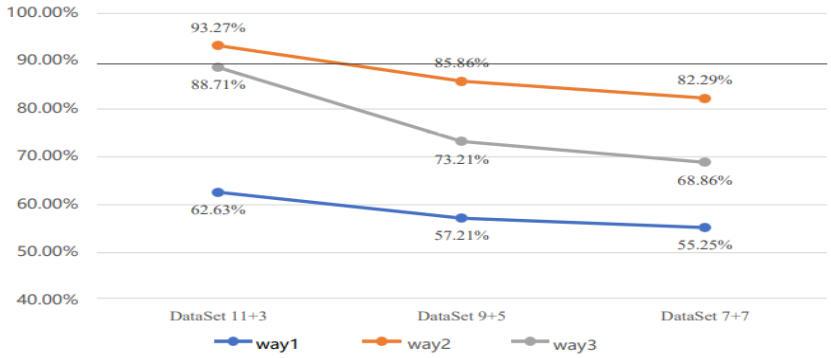

Figure 6 Variation diagram of LFRCNN generalization recognition accuracy under different interception ways.

In order to compare the performance of this paper and other methods in students' classroom behavior, this paper constructs three different partitioning methods of training set and test set, and tests the generalization performance of different methods. The first data set randomly disrupts the collected data sets, and both the training set and the test set contain images of 14 students; the training set of the second data set contains images of 9 students and the test set. It contains images of 5 students different from the training set; the training set of the third data set includes images of 7 students, and the test set contains images of 7 students different from the training set. This paper uses the three most commonly used convolutional neural network models to compare with the methods proposed in this paper, which are VGG11, ResNet18 and GoogLeNet.

4.2. Local feature selection and its effect comparison

The first method is to intercept the information below the left and right elbows in the original image, the second method is to intercept the information around the left and right wrists, and the third method is to fuse the local image information, that is, the image information below the left and right elbows, the image information around the left and right wrists and the image information around the right-hand row fusion. Figure 6 shows the change of LFRCNN generalization recognition accuracy of three local feature selection methods under three data sets. It can be seen from the figure that the second method can obtain better generalization ability than the other two methods.

5. CONCLUSION

For a small number of samples, the existing methods have low generalization accuracy in student classroom behavior recognition. This paper proposes a student classroom behavior recognition method based on pose information extraction and local feature segmentation. The method is based on the OPENPOSE detection network, which detects the posture information of all students in the image, and uses HPRNN to process the posture information. First, head posture recognition is performed, and then LFRCNN is used to further perform local feature recognition to identify the classrooms of all students in the image behavior.

REFERENCES

[1] Zhang Hongyu. Design and Implementation of Classroom Learning Behavior Measurement System[D]. Huazhong University of Science and Technology, 2016.

[2] Dang Dongli. Human behavior recognition and its application in education recording and broadcasting system[D]. Shangxi: Xi'an University of Science and Technology, 2017.

[3] Zhou Pengxiao, Deng Wei, Guo Pei Puert. Research

on S-T Behavioral Identification in Classroom Teaching Video [J]. Modern Educational Technology, 2018,28 (06): 54-59.

[4] Shi Y, Wei Y, Pan D, et al. Student body gesture recognition based on Fisher broad learning system [J]. International Journal of Wavelets, Multiresolution and Information Processing, 2019, 17 (1): 1950001.

[5] Lin Canran, Xu Weiliang, Li Yi. Exploration of Classroom Student Behavior Recognition Technology Based on Multimodal Data[J]. Modern Computer, 2020 (6): 69-75.

[6] Su Chao, Wang Guozhong. Research on Student Behavior Recognition Based on Improvement Open POSE[J]. Research on Computer Application: 1-8 [2021-09-26].

[7] Guo Junqi, Lu Jiazhen, Wang Yinhan, Xiong Qingyun, Zhang Shifeng, Hu Kang Ying.Dexue Study Model Driven Teachers and Students' Classroom Behavior [J]. Journal of Beijing Normal University (Natural Science Edition): 1-13 [2021-09-26].

[8] Huang Yongkang, Liang Meiyu, Wang Xiao Xiao, Chen Zheng, Cao Xiaowen. Classroom teaching behavior in classroom teaching video based on depth time and space residual convolution [J]. Computer Application: 1-9 [2021-09- 26].