A Spatio-Temporal Feature Representation of Multimodal Surveillance Images for Behavioral Recognition

Lei Ma
School of Artificial Intelligence
Beijing Polytechnic, China
Ma_LeiMM@outlook.com

Hongxue Yang
School of Artificial Intelligence
Beijing Polytechnic, China
yanghongxue@bpi.edu.cn

Guanghao Jin
School of Artificial Intelligence
Beijing Polytechnic, China
103126@bpi.edu.cn

Abstract: *Due to the difficulty of accurately expressing complex learning behaviors based on features obtained from a single behavioral modality, research is being conducted on a multimodal monitoring image Spatio-Temporal (ST) feature representation method for behavior recognition to improve the effectiveness of learning behavior recognition. Using an improved 3D Convolutional Neural Network (CNN) with Spatio-Temporal Pyramid Pooling (STPP), an attention based Long Short-Term Memory neural network (LSTM), and a special orthogonal popular spatial network, the RGB spatial features, RGB temporal features, and 3D skeletal features of the monitoring images are extracted from each channel; by improving the dual attention mechanism and integrating three modal features to complement each other's strengths; using bounding box regression analysis to fuse the ST features of multimodal monitoring images, the learning behavior recognition results are obtained. Experimental results have shown that this method can effectively extract ST features of multimodal monitoring images, and the edge information retention of multimodal ST feature fusion is relatively high at different lighting conditions, close to 1, indicating that the feature fusion effect is excellent and the learning behavior recognition accuracy is high, above 96%.*

Keywords: *Multimodal, surveillance images, spatio-temporal features, behavioral recognition, convolutional neural network, dual attention mechanism.*

Received May 15, 2024; accepted January 15, 2025
https://doi.org/10.34028/iajit/22/4/15

1. Introduction

1.1. Research Background

With the popularization of smart devices and the development of information technology, the demand for personalized learning for individual users is increasing. Learning behavior largely reflects individual learning characteristics and preferences, and can provide valuable information for personalized learning. Therefore, learning behavior recognition has become a key technology to realize intelligent education, personalized recommendation and other applications, and as the intersection of artificial intelligence and education, it is increasingly receiving widespread attention [8, 18]. Learning behavior recognition, that is, by analyzing and interpreting the learner's movement, expression, voice and other multimodal information, in order to understand and assess their learning state, interest, emotion and other internal psychological activities [4]. By identifying and analyzing students' learning behaviors, we can gain a deeper understanding of students' behavioral characteristics and habits in the learning process, so as to better understand students' learning styles and effects. This helps educators to adjust teaching strategies in a targeted manner and improve teaching effectiveness. According to the learning behavior characteristics of each student, it can also provide personalized teaching suggestions and learning counseling to meet the different learning needs of students, and further promote the realization of personalized education [9]. At the same time, it can also help education administrators better understand the students' learning status [16], provide scientific basis for education decision-making, and enhance the intelligent level of education management. In addition, learning behavior recognition technology can provide a large amount of data about students' learning behavior, provide rich empirical information for educational scientific research, and promote the progress of educational science [1]. In distance education and online learning, learning behavior recognition technology can help teachers find students' learning anomalies in a timely manner, such as leaving the learning state for a long time, learning progress suddenly stops, etc., so as to protect students' learning safety.

1.2. Literature Review

In the context of the above research, some scholars have conducted relevant research on learning behavior recognition, for example, proposed a behavior recognition method combining human posture estimation and object detection Mo *et al*. [11]. First, the target detector extracts a single area from the key frame as the input of the network. Then, the Multi Task Heatmap Network (MTHN) module extracts intermediate heatmaps associated with multi-scale features. Attitude estimation and target detection tasks are constructed by mapping relationships to obtain key points and target position information. Finally, key point behavior vector and measurement vector are used to model behavior, and behavior recognition is completed based on fully connected network. This method has better classification performance and robustness when recognizing different learning behaviors. However, the behavior modeling is too dependent on the information of posture key points and target positions, and other important behavior characteristics may be ignored in some complex behavior scenes, resulting in inaccurate behavior recognition. Chen [3] proposed two improved Channel Attention (CA) modules, namely the Spatio-Temporal (ST) interaction module of matrix operation and the depth separable convolution module, combined with the research of human behavior recognition. Combining the superior performance of CNN in image and video processing, a multi-scale CNN method for human behavior recognition is proposed by Chen [3]. First, the behavioral video is segmented, each video clip is low order learned, and the corresponding low order behavioral information is extracted. Then this low order behavioral information is connected on the time axis to obtain the low order behavioral information of the entire video, so as to effectively capture the behavior. However, the performance of this method using multi-scale CNNs in behavior recognition still depends on the training dataset used. If the size of the dataset is limited or the category distribution of data samples is uneven, it will lead to poor generalization performance in practical applications. Zhao *et al*. [21] proposed a feature extraction method of 3D Convolutional neural network fusion Channel Attention (3DCCA) model. The RGB video frames are preprocessed by means of the mean normalization method, and then the ST features of the input clips are extracted by 3D convolution, and the more critical features for current behavior recognition are selected from all features through CA. Finally, Softmax classifiers are used to classify and recognize human behaviors in videos. However, when this method is used to process behavior recognition tasks in complex scenes, its feature extraction is single, which makes it difficult to accurately express complex learning behaviors, and it is prone to miscalculation or decline in accuracy. Wang [13] proposed online learning behavior recognition based on image emotion. The flow of image emotion recognition is introduced in detail to facilitate the analysis of online learning behavior. The improved local binary mode and wavelet transform are used to extract key frames from face images. Next, the structure of online learning behavior analysis system is constructed, a method of learning emotion recognition based on facial expression is proposed, and an online learning image emotion classification model based on attention mechanism is established to complete learning behavior recognition. However, different individuals may have different physiological responses in the learning process, so this method will be affected by individual differences, resulting in unstable recognition accuracy. Wu [14] combines Particle Swarm Optimization (PSO) algorithm and K-Nearest Neighbors (KNN) algorithm to get PSO-KNN joint algorithm, and combines it with emotional image processing algorithm to build a classroom student behavior recognition model based on artificial intelligence. In addition, based on image processing technology, key frame detection is used for feature recognition, and the recognition process based on inter frame similarity measurement algorithm and initial cluster center selection is improved in the key frame extraction method of clustering to complete student behavior recognition. Behavior recognition usually requires comprehensive analysis of data from multiple different modes. However, this method is limited to data input of single mode, and it is difficult to obtain more comprehensive and accurate results in practical applications.

Learning behavior is a complex and multidimensional process, and it is often difficult for the information of a single modality to comprehensively and accurately reflect the real state of learners. Therefore, based on the research of the above methods, by fusing multimodal features such as time and space [7], the advantages of each modality can be comprehensively utilized to improve the accuracy and robustness of the learning behavior recognition, and the multimodal spatial and temporal features of surveillance images for behavior recognition can be studied to provide a more efficient, accurate, and comprehensive means of analyzing the learning behaviors in the field of education. The proposed method uses an improved 3D CNN that includes ST pyramid pooling, an attention based Long Short-Term Memory neural network (LSTM), and a special orthogonal popular space network to extract RGB spatial features, RGB temporal features, and 3D skeletal features of monitoring images. The three extracted modal features are fused through an improved dual attention mechanism to improve the comprehensiveness and accuracy of feature representation, providing reliable support for subsequent learning behavior recognition. Finally, based on the concept of boundary regression, bounding box regression is used to analyze the ST feature fusion results of multimodal monitoring images, achieving accurate learning behavior recognition.

2. Learning to Recognize Behavior

The overall implementation architecture diagram of the learning behavior recognition model based on the ST features of multimodal monitoring images is shown in Figure 1.

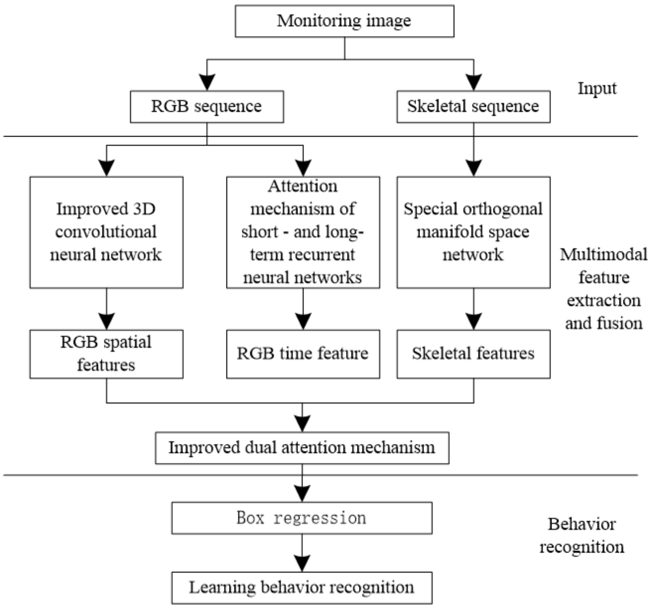

Monitoring image
RGB sequence
Skeletal sequence
Input
Improved 3D convolutional neural network
Attention mechanism of short - and long-term recurrent neural networks
Special orthogonal manifold space network
Multimodal feature extraction and fusion
RGB spatial features
RGB time feature
Skeletal features
Improved dual attention mechanism
Box regression
Behavior recognition
Learning behavior recognition

Figure 1. Learning behavior recognition model based on ST features of multi-modal monitoring images.

The extraction stage of learning behavior features in the model mainly includes three layers: improved 3D CNN, attention mechanism LSTM and special orthogonal pop space network. The specific implementation steps are described as follows: First, input a surveillance image to obtain the RGB video sequence and 3D skeleton sequence of the behavior samples. Then, in the 3D CNN network, adding ST pyramid pooling can automatically process RGB video frame sequences of any size and quickly convolve the resulting RGB spatial features. Then, the LSTM module is used to transfer the current or even earlier information to the next moment of use through its memory function, further extracting RGB temporal features. At the same time, the ST attention mechanism is used to enhance key information and obtain the final RGB temporal features. Secondly, a special orthogonal popular space network is used to represent the 3D human skeleton sequence and obtain skeletal features. Finally, the improved dual attention mechanism is used to fuse the extracted three modal features, and the final learning behavior recognition result is output by bounding box regression.

2.1. Monitoring Image Spatial Feature Extraction Based on Improved 3D CNN Network

Compared with 2D convolution network, 3D convolution network can simultaneously capture the appearance and motion information of the target in the RGB sequence of the surveillance image, and has better learning behavior recognition performance. Its structure is simpler than many current behavior recognition models, and it has a faster running speed. This is because 3D convolutional networks have a single network structure that can simultaneously process spatial and temporal information in one network, reducing model hierarchy and parameter count, thereby improving running speed. The model stacks multiple consecutive frames to form a cube, and generates multiple channel information in each frame of image. Each channel of the consecutive image frame is checked with different methods for 3D convolution [5]. The feature map obtained in this way is connected to multiple adjacent image frames, so as to obtain time information while extracting spatial information. Finally, the information on all channels is combined to calculate various types of RGB spatial features [20].

The 3D CNN structure consists of one hard wire layer, three convolution layers and two lower sampling layers. The network takes continuous frame monitoring images with a size of 60×40 as input. Among them, the hard-line layer is a special preprocessing layer that extracts feature information from each frame of the image through pre-set "hard wired kernels," converting the original monitoring image into a feature map form suitable for subsequent convolutional layer processing. The hard-line layer extracts grayscale from each frame, horizontal coordinate gradient *x*, vertical coordinate gradient *y*, light streams *x*', light streams *y*' these 5 channels generate 33 feature maps. The C2 convolutional layer uses two different 3D to check the five-channel information output from the previous layer for convolution operation. The C4 convolutional layer, on the other hand, uses three different convolutional kernels to perform convolutional operations on the feature maps respectively, thus obtaining more feature maps with both spatial and temporal dimensions [15]; for the downsampling layer, the S3 and S5 sliding window of size 2×2 and 3×3 is used to downsample each feature map obtained from the previous layer, respectively, keeping the number of feature maps constant while reducing the spatial resolution; the last convolutional layer C6, each feature map is convolved with a 7×4 2D core to obtain 128 feature maps, that is, 128 feature vectors of action information in the input image frame [19]. However, the training and testing of 3D CNN networks require the input of monitoring image frames with fixed size and scale. When inputting a surveillance image of any size, 3D CNN will crop or scale the surveillance image to produce a fixed size of input samples, which may lead to the loss and distortion of important information, thus affecting the extraction of RGB spatial features. In order to process the monitoring image frames of any size more comprehensively, the last pooling layer in 3D CNN is replaced by the Spatio-Temporal Pyramid Pooling layer (STPP) to receive inputs of different sizes and convert them into feature vectors of fixed length, while extracting more features of different time angles [2, 6].

Because the convolution layer can receive any size of input, and then produce different sizes of output. Given a RGB monitoring image sequence of any size as the input of 3D CNN, after the previous 3D convolution and ordinary down sampling, assume that the feature mapping size of the last convolution layer is $T×W×H$, of which T is the time to pool the cubes, the H and W are the height and width of the monitored image frame. Unlike the conventional sliding window pooling used in 3D CNN, STPP will dynamically adjust the size of the sliding window after giving the number of features generated by the pooling layer. Specifically, $P(p_t, p_s)$ is denoted as the ST pooling level, where p_t is the time pooling level, p_s is the spatial pooling level, hence, each pooling cube is of the size $\left\lfloor \frac{T}{p_t} \right\rfloor \times \left\lfloor \frac{W}{p_s} \right\rfloor \times \left\lfloor \frac{H}{p_s} \right\rfloor$. When p_s=4,2,1 and p_t=1, the convolution outputs with different sizes can be converted into RGB spatial feature vectors X with fixed dimensions. Each ST pooled cube maximizes the pooled response value. In this way, the improved 3D CNN configured with STPP can adapt to monitoring image frames of any size or scale, and support arbitrary scaling of frame scale.

2.2. Monitoring Image Spatial Feature Extraction Based on Attention Mechanism LSTM

Represent the RGB spatial feature sequence obtained above as U, and then use the attention mechanism LSTM of each country to extract the RGB temporal features in U. As an improvement of recurrent neural network, LSTM has a strong ability to process long time series.

The update rules of LSTM network are as follows in Equations (1) to (6):

$$i_\tau = \psi(w_i u_\tau + b_i + w_{hi} h_{\tau-1} + b_{hi}) \quad (1)$$

$$f_\tau = \psi\left(w_f u_\tau + b_f + w_{hf} h_{\tau-1} + b_{hf}\right) \quad (2)$$

$$q_\tau = \tanh\left(w_q u_\tau + b_q + w_{hq} h_{\tau-1} + b_{hq}\right) \quad (3)$$

$$z_\tau = \psi(w_z u_\tau + b_z + w_{hz} h_{\tau-1} + b_{hz}) \quad (4)$$

$$c_\tau = f_\tau * c_{\tau-1} + i_\tau * q_\tau \quad (5)$$

$$h_\tau = z_\tau * \tanh(c_\tau) \quad (6)$$

In the formula, u_τ is input the RGB spatial feature sequence of the extracted monitoring image into the network at time τ; h_τ is hidden layer state vector of LSTM at time τ; c_τ is the memory state vector; $h_{\tau\text{-}1}$ is the hidden layer state vector at time τ-1; i_τ, f_τ, q_τ, z_τ are the output vectors of input gate, forgetting gate, memory gate and output gate of LSTM unit; $\psi(\cdot)$ is the hyperbolic tangent sigmoid activation function;* is hadamarjah. w_i, w_{hi}, w_f, w_{hf}, w_q, w_{hq}, w_z, w_{hz} all represent a linear layer; b_i, b_{hi}, b_f, b_{hf}, b_q, b_{hq}, b_z, b_{hz} are the bias term.·

LSTM mainly overcomes the problems of "gradient disappearance" and "gradient explosion" in traditional RNN training. The biggest difference between LSTM and traditional RNN is that its gating structure is more effective for extracting information from long sequences. Although LSTM has the ability to extract the temporal characteristics of long sequences, the performance of LSTM will decline rapidly when the sequence length is too long. For learning behavior recognition, only when the sequence length is long enough can it contain more learning behavior time series characteristics of the target [22]. Therefore, only using LSTM cannot extract timing features of RGB sequences of surveillance images.

In order to solve the above problems, attention mechanism and LSTM are combined to jointly grasp the temporal characteristics of long sequences. Attention mechanism used to calculate the historical learning behavior of goals $\{r_1, r_2, \ldots, r_{\tau\text{-}1}\}$ with current learning behaviors r_τ correlation of the current moment, thus constructing the context vector s_τ of the current moment. The addition of the attention module will not weaken the historical characteristics due to multi-step timing propagation, thus making up for the defect that LSTM cannot handle long sequences [17]. Context vector s_τ represents the dependence of current learning behaviors on historical learning behaviors, calculated as in Equations (7) and (8).

$$s_\tau = \sum_{k=1}^{\tau-1} \mu_k h_k \quad (7)$$

$$\mu_k = soft\max(h_\tau \varpi_s h_k) \quad (8)$$

In the formula, the h_τ, h_k are hidden layer output vector of LSTM for the current moment and, respectively, the k moment (k<n); ϖ_s is a linear layer; μ_k is the weight corresponding to the RGB timing feature at time k, that is, its correlation with the current learning behavior.

After getting the context vector s_τ, s_τ and h_τ will be spliced and input the linear layer, and use the softmax activation function to obtain the timing characteristics of the monitoring image RGB:

$$X' = soft\ max(\varpi_{out}(s_\tau \oplus) h_\tau)) \quad (9)$$

In the formula, ϖ_{out} is the output linear layer.

2.3. Skeletal Feature Extraction for Surveillance Images Based on Special Orthogonal Popular Space Network

Next, in order to further ensure the accuracy of the final learning behavior recognition results, the skeleton features of the monitoring image are extracted based on the RGB spatial features and RGB temporal features of the above extracted monitoring image. Let A=(V, E) represent a human skeleton that monitors the skeletal sequence of the image, wherein V=$\{v_1, v_2, \ldots, v_N\}$ represents the set of joints (points), E=$\{e_1, e_2, \ldots, e_M\}$ represents the set of rigid body parts (edges). Given a pair of body parts e_m and e_η, let $e_{\eta 1}$ and $e_{\eta 2}$ denote, respectively, the beginning and the end of the point for body parts e_η, l_η represents the length of e_η. In order to characterize the relative geometric relationship between them, another body part is represented in the local coordinate system of each body part e_η of the local coordinate system is obtained by rotating and translating the global coordinate system, then e_{n1} becomes the origin, and e_η side is on the axis x. Thus for 2 edges, the e_m and e_η, get 3D transform vectors $\hat{e}_m$ and $\hat{e}_\eta$ respectively.

Then, calculate the rotation matrix from e_m to $Y_{m,\eta}$ with e_η as the local coordinate system. First, calculate the axis angle (ω, θ) of the rotation matrix $Y_{m,\eta}$:

$$\omega = \frac{\hat{e}_m \otimes \hat{e}_\eta}{\|\hat{e}_m \otimes \hat{e}_\eta\|} \tag{10}$$

$$\theta = \arccos(\hat{e}_m \cdot \hat{e}_\eta) \tag{11}$$

In the formula, the ⊗ and the symbols · denote the outer and inner products respectively. This can then be easily converted to a rotation matrix $Y_{m,\eta}$ by the representation of the axial angles. Similarly, it is also possible to obtain a rotation matrix $Y_{\eta,m}$ of e_η from taking e_m as the local coordinate system. In order to be able to fully encode to $Y_{\eta,m}$, $Y_{m,\eta}$ and $Y_{\eta,m}$ are used together to describe the relative geometry of the body parts, expressing the body skeleton A at time τ as $\left(Y_{1,2}^{\eta}(\tau), Y_{2,1}^{\eta}(\tau), \ldots, Y_{M-1,M}^{\eta}(\tau), Y_{M,M-1}^{\eta}(\tau)\right)$, of which M is the number of body parts, η, $m \in M$.

The $\eta\times\eta$ matrix in Y^η form the special orthogonal group $SO(3)$, which is actually a matrix group. Thus, each action sequence of a moving skeleton can be represented as a curve in group $SO(3)\times\ldots\times SO(3)$, obtaining the 3D human skeleton representation A' within the bone sequence of the monitoring image, taking A' as the bone feature of the monitoring image.

2.4. Multimodal Surveillance Image Feature Fusion Based on Dual Attention Mechanism

Using dual attention mechanism to fuse RGB spatial features of surveillance images X, RGB timing characteristics X' with skeletal features A' of three modal features [10], followed by the use of a 1*1 convolution kernel to generate the number of target channels for the use of bounded regression, and finally the output of the coordinate frame information of the target's belonging space as well as its belonging category of learned behaviors.

The specific steps to fuse X, X', A', using the spatiotemporal feature fusion module (CFAM) of direct attention, are as follows:

- *Step* 1: Apply the RGB spatial features X extracted in section 2.1, RGB temporal characteristics X' extracted in section 2.2, with the skeletal features A' extracted in section 2.3, stacked according to the last two dimensions to form a surveillance image feature map D.
- *Step* 2: Place D input into the two convolutional layers to generate a feature map of the surveillance image G. The gram matrix operation is performed on the G monitoring image feature map.
- *Step* 3: Mapping the surveillance image features G remodeled into a tensor F, i.e., the feature vectors of the surveillance images of each channel are reshaped into one-dimensional vectors. The remodeling tensor F of G and its transpose F^T are used by the matrix product operation, and then calculate Gram matrix ρ, the mathematical expression of which is as follows:

$$\rho = F \cdot F^T \tag{12}$$

$$\rho_{i'j'} = \sum_{\beta} F_{i'\beta} \cdot F_{\beta j'} \tag{13}$$

The matrix ρ represents RGB associated temporal features, RGB spatial features and bone features. From the mathematical expression, each element $\rho_{i'j'}$ in ρ is obtained by the inner product of mapped by vectorized features i' and j'. After calculating the gram matrix ρ, then, use Softmax to generate the CA map O for each element in the matrix. The mathematical formula is as follows:

$$O_{i'j'} = \frac{e^{\rho_{i'j'}}}{\sum_{i',j'} e^{\rho_{i'j'}}} \tag{14}$$

From the mathematical expression, each element $O_{i'j'}$ in the CA graph O is a measure of the influence of the j' th channel on the i' th channel. Therefore, the RGB spatial features X extracted according to section 2.1, RGB temporal characteristics X' extracted in section 2.2, with the skeletal features A' extracted in subsection 2.3, calculated the inter-channel dependence O of the individual channels of the multimodal fusion feature. Then use O and F matrix multiplication, as in Equation (15), will then be made so that the F' dimensional shape of the tensor of the re becomes that of F'', which makes its dimension the same as the initial input multimodal fusion feature, the effect of attention mapping on the initial multimodal fusion feature is obtained. The equations are as follows:

$$F' = O \cdot F \tag{15}$$

- *Step* 4: Combine the above calculated F'' and initial multimodal fusion feature maps F, its mathematical expression is as in Equation (16), where the parameter γ is a trainable parameter, whose value is gradually learned from 0 during the training process. The formula is as follows:

$$\xi = \gamma \cdot F'' + F \tag{16}$$

From the above mathematical expression, the CA module, each element in ξ is denoted as the final feature computed for each channel separately, i.e., the weighted sum of the original multimodal features and the features of all channels constitutes the final feature for each channel [12].

- *Step* 5: The computed feature map of the surveillance image, the ξ after two convolutional layers, the final fused multimodal ST feature map ε is generated.

Inspired by the idea of Dual Attention Network (DANet) based on the dependency of capture space and Channel global Features, an improved Dual Attention Mechanism (CFDAM) module is proposed. At the same time, two attention mechanism modules, CFAM and CFDAM, are used to fuse the spatiotemporal features of multimodal surveillance images. Finally, the spatiotemporal features of the two fused multimodal surveillance images are superimposed, and finally input into the boundary regression box, output the results of learning behavior recognition.

Based on the original spatiotemporal feature fusion module CFAM, CFDAM is extended to two attention mechanisms to fuse the spatiotemporal features of multimodal surveillance images CFAM and CFDAM are used to fuse the spatiotemporal features of multimodal surveillance images in parallel, and then the extracted spatiotemporal features of multimodal surveillance images are superimposed. The role of CFAM is to reasonably combine the RGB spatial features *X* extracted in section 2.1, RGB temporal characteristics *X'* extracted in section 2.2, with the skeletal features *A'* extracted in section 2.3, CFDAM is used to fuse RGB spatial features *X* extracted in section 2.1, RGB temporal characteristics *X'* extracted in section 2.2, with the skeletal features *A'* extracted in subsection 2.3. On the other hand, using TA attention mechanism to emphasize the interaction between multidimensional channel features will not reduce the importance of dimensions. The reason why the dual attention mechanism is used to fuse the spatiotemporal features of multimodal surveillance images is that it can better enable the channels in the spatiotemporal features of multimodal surveillance images to achieve global dependency, establish rich dependencies between the learning behavior characteristics of the front and rear image frames on local features, and more fully fuse the spatiotemporal features of multimodal surveillance images.

2.5. Learned Behavior Recognition based on Boundary Regression

Based on the boundary regression idea, the fusion results of ST features of multimodal surveillance images obtained in subsection 2.4 are processed to obtain the results of learning behavior recognition. In the learning behavior recognition model based on ST features of multimodal surveillance images, the last layer uses a 1*1 convolution kernel to generate the required number of target channels [(5*(Num+5)*H*W]. The number of channels in the final output, the[(5*(Num+5)*H*W], where the first 5 means that for each grid cell in H*W, K-means clustering algorithm is used to select 5 prior anchors on the monitoring image data set; among which, *Num* in *Num*+5 indicates that there are number of *Num* score of learning behavior classification, 5 represents 4 coordinates and 1 confidence score. Then, based on the regression of these anchors on the refined bounding boxes, non-maximum suppression control is used to remove the redundant prediction bounding boxes and retain the prediction bounding boxes with higher overlap with the true boxes. During training, the intersection ratio *IOU* between the remaining predicted bounding boxes and the true box is calculated, which is calculated as in Equation (17), the accuracy of compliance with the regression frame, and the recall of the regression frame, both of which are given in the following Equations (18) and (19).

$$IOU = \frac{\kappa(\lambda_p \cap \lambda_g)}{\kappa(\lambda_p \cup \lambda_g)} \tag{17}$$

In the formula, $\kappa(\cdot)$ denotes the area of the rectangle. λ_g denotes the true bounding box, λ_p denotes the prediction bounding box.

$$P = \frac{TP}{TP + FP} \tag{18}$$

In the formula, *TP* denotes true instances, which refers to the number of positive instances predicted by the model to be positive instances. *FP* denotes false-positive cases, which refers to the number of negative cases that the model predicts as positive.

$$R = \frac{TP}{TP + FN} \tag{19}$$

In the formula, *FN* denotes false-negative cases, which refers to the number of positive cases that the model predicts as negative.

Ultimately, based on *IOU*, accuracy and recall, determine the final learning behavior recognition bounding box to complete the learning behavior recognition.

3. Experimental Analysis

3.1. Experimental Data Preparation

In this study, the lecture monitoring video of real classroom teaching in a key university is used as the original data to analyze the characteristics of students' learning behaviors, and eight representative common college students' learning behaviors are divided, which are listening to lectures in class, looking right and left, playing with cell phones, flipping through books, sleeping, standing, and writing, and the descriptions of each kind of learning behaviors are shown in Table 1, and the corresponding students' learning behaviors are constructed to monitor the video.

In order to effectively identify students' learning behaviors in the classroom environment, it is first necessary to collect video data from multiple real classrooms. In this study, the teaching process in real classrooms of a major university is taken as the research object, and the videos of the whole teaching process recorded in six courses (two compulsory courses and four elective courses) are selected as the research data. A total of about 6000 minutes of classroom videos were collected, totaling 60 classroom data, each classroom is two sessions totaling 90 to 110 minutes. The real classroom scenario is a smart classroom with seven cameras distributed in the front and back of the classroom (four in the front of the classroom and three in the back of the classroom). Since the experiments need to identify students' learning behaviors, the experimental data are all positive information of the students in the classroom, so as to construct a video dataset of eight kinds of students' learning behaviors.

Table 1. Description of learning behavior.

ID	Learning behavior	Status declaration
1	Write	Lower head+Pen in hand
2	Look left and right	Look left and right
3	On one's feet	On one's feet
4	Page turning	Lower head+Hand touching book
5	Play on phone	Head down+Hand touch the phone
6	Raise one's hand	Raise one's hand
7	Sleep	Lie on your stomach+Face down
8	Attend a lecture	Sitting+Looking ahead

For the collected real classroom video dataset, it is necessary to establish basic specifications and standards for data entry, and construct large-scale standard human behavior dataset in real classroom environment under the standards and specifications, so as to provide data support for the subsequent learning behavior recognition. The collected surveillance videos are screened, cut, segmented and labeled in two steps, as follows.

1. Filtering and cropping segmentation processing. As the collected classroom surveillance video time is 90 to 110 minutes, the resolution is 1920*1080, part of the surveillance video image shown in Figure 2. Due to its longer time, larger picture, it is necessary to filter the whole process of the collected long video, first select the video contains seven typical behaviors of the time period, reference to the production of the public data set, it will be cropped and segmented into a 12-second single 320*240 short video, and then 12-second video and then segmented into individual video data samples in accordance with the 6-second interval.

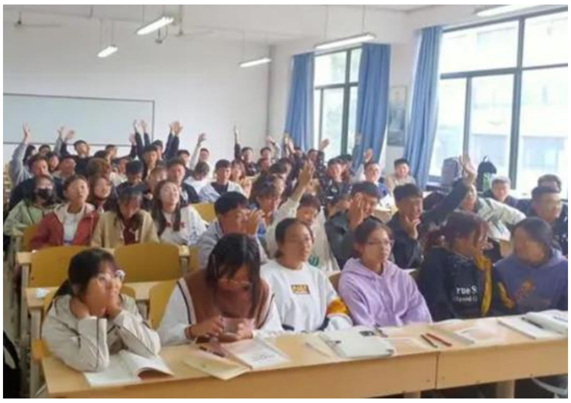


Figure 2. Part of the monitoring image.

2. Data annotation. According to the screened video data samples, the jpg format frame images are generated by segmentation at 30 frames per second, and a total of 75400 images are obtained. Then the obtained images are labeled with learning behavior according to eight categories. The labeling box is a rectangular box. At the same time, the learning behavior is partially blocked. The exported data label file is saved as a JSCM file. The data label file consists of image file information, label frame coordinate information, and the corresponding label. The labeled experimental data are randomly divided into training set, verification set and test set, with a ratio of approximately 3:1:1.

The self-constructed student classroom behavior monitoring video dataset is labeled with a total of eight categories of learning behaviors in 1% different scenarios. The monitoring image in each scene is usually 12s or 18s, and is sampled at 30 frames per second. There are 75443 frame images in total, and the size of each frame image is 60×40.

3.2. Experimental Indicators

In order to verify the recognition effect of the method in this paper, the method in the literature [3] and the method in the literature [21] are used as the comparison method, and the comparison test is carried out for the following indexes, which are described as follows.

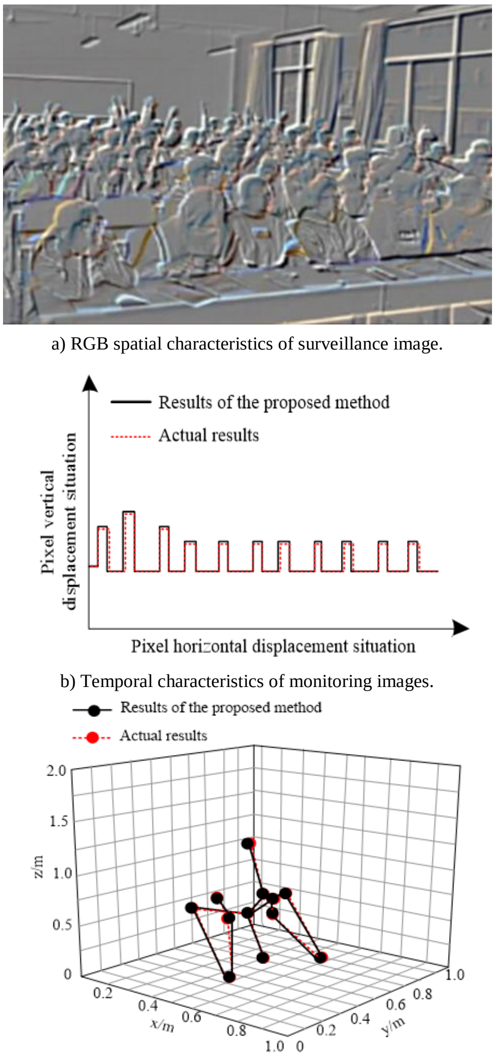
Since the result of feature fusion will directly affect the subsequent recognition effect, the edge information retention is used to measure the multimodal surveillance image feature fusion effect of each method, and the value of the edge information retention is between 0 and 1. The closer the value is to 1, the better the edge retention of the surveillance image is, and the fusion of the feature image is richer in detail information.

In order to further verify the recognition accuracy of each method, a confusion matrix was used for the analysis. The confusion matrix is a table in which the rows represent the actual learning behavior categories and the columns represent the recognized learning behavior categories. Through this table, we can see the recognition accuracy of each method for each learning behavior category, the diagonal values of the confusion matrix represent the number of learning behavior categories correctly recognized by the model, that is to say, these values represent the number of samples actually belonging to a certain category correctly recognized by the method, and the rest of the values in the rectangular box represent the number of samples actually belonging to a certain category incorrectly recognized by the method. The remaining values in the rectangular boxes represent the number of samples that the method incorrectly identifies as belonging to a category.

3.3. Analysis of Results

3.3.1. Validity Analysis

Utilize the monitoring image in Figure 2 of this article to extract multimodal spatiotemporal features and perform feature fusion. Taking the RGB spatial features of the frame image, as well as the RGB temporal and skeletal features of a student, as an example, the multimodal spatiotemporal feature extraction results and the fusion results of a student are shown in Figures 3 and 4, respectively.



a) RGB spatial characteristics of surveillance image.

b) Temporal characteristics of monitoring images.

c) Bone features of surveillance images.

Figure 3. Multi-modal ST feature extraction results.

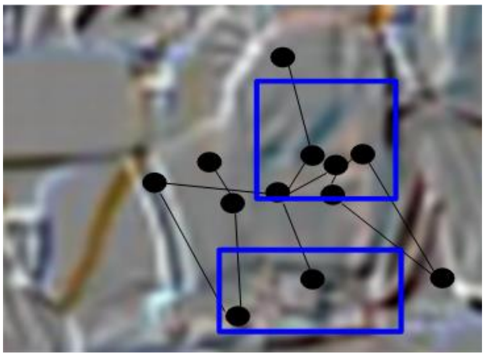
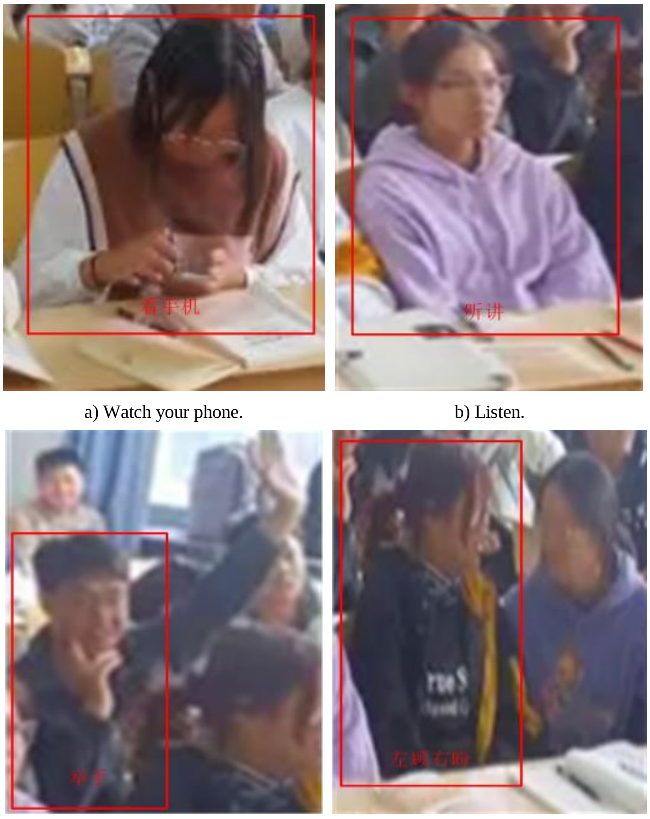


Figure 4. Fusion results of various modal features.

Figures 3-a) and (b) demonstrate the effectiveness of our method in extracting RGB features from surveillance images. From the figure, it can be seen that through the method proposed in this article, rich and effective RGB spatial features and accurate RGB temporal features can be extracted from monitoring images. These features can comprehensively describe the posture of learners, as well as the pixel movement of learners' behavior over time. Based on this, important information about learners' behavior states can be obtained. These features are of great significance for subsequent learning behavior recognition and can provide more accurate and comprehensive data support. Figure 3-c) demonstrates the effectiveness of our method in extracting skeletal features from surveillance images. From the figure, it can be seen that through the method proposed in this article, accurate skeletal features can be extracted from monitoring images, and the extracted results are basically consistent with the actual results, with small differences. These features can accurately describe the learner's actions and postures, and they are also of great significance for subsequent learning behavior recognition, providing more accurate and reliable data support. And the fusedfeatures can effectively obtain the learning state to assist in subsequent recognition. Through comprehensive analysis, it can be concluded that the method proposed in this paper has the effectiveness of extracting and fusing spatiotemporal features of multimodal monitoring images. In practical applications, this multimodal spatiotemporal feature extraction and fusion method helps to improve the accuracy and robustness of learning behavior recognition, providing strong support for personalized education and teaching evaluation.

Using the method of this paper on the surveillance image in Figure 2, learning behavior recognition, part of the learning behavior recognition results are shown in Figure 5.



a) Watch your phone.

b) Listen.

c) Raise your hand.

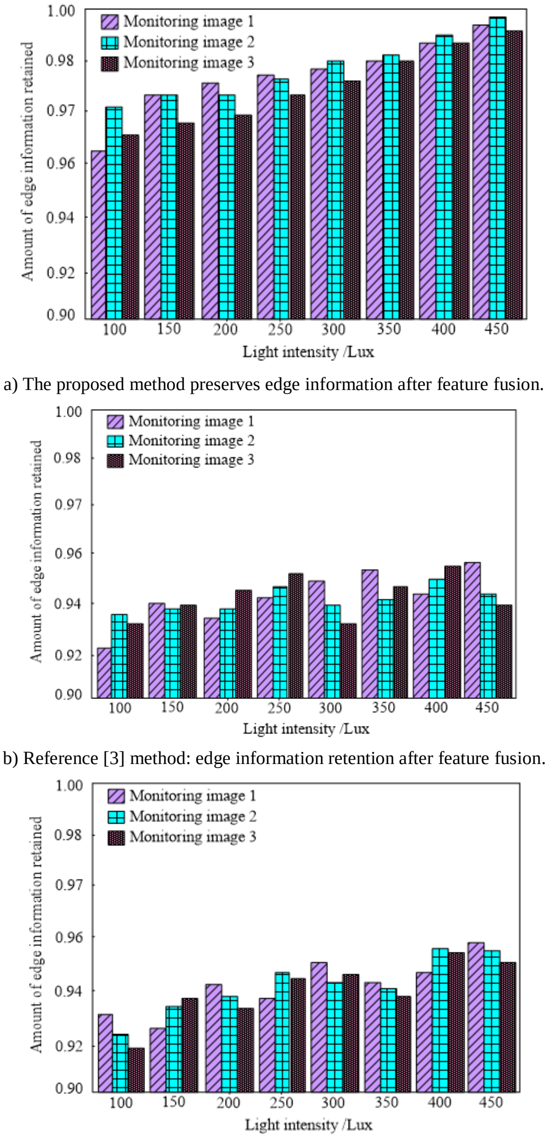
d) Look left and right.

Figure 5. Learning behavior recognition results.

Figures 5-a), (b), (c), and (d) show the application effect of this method in identifying students' learning behaviors in the classroom. It is obvious from these images that this method can accurately identify various behaviors of students in class, including looking at mobile phones, listening, raising hands, looking left and right, etc. These results of behavior recognition are very valuable for analyzing students' learning concentration in the classroom. By observing students' behavior at different times, teachers can judge students' participation and concentration in the classroom. For example, if students show active participation and concentration when listening to lectures, this may indicate that they are interested in and involved in the course content. On the contrary, if students frequently check their mobile phones or look around, it may mean that they are not interested in the course content or difficult to concentrate. By counting the learning concentration of students in the whole classroom at a specific moment, teachers can obtain the overall learning state of students at that moment. This will help teachers to understand students' learning situation in a timely manner, find out possible problems, and take corresponding measures to improve the classroom teaching effect. For example, if most students show low concentration, teachers can consider adjusting teaching methods, teaching contents or activity arrangements to attract students' interest and improve their participation. This analysis method based on the identification of students' learning behavior can not only help teachers better understand students' learning needs and problems, but also provide important reference for further improving classroom teaching strategies. By constantly optimizing teaching strategies and improving students' learning concentration and participation in the classroom, we can effectively improve the efficiency of teaching and learning and achieve better teaching results. To sum up, the method in this paper shows a significant effect in identifying students' learning behaviors in the classroom. By accurately identifying students' behaviors and analyzing their learning focus, this method can provide valuable feedback information for teachers, help them better understand students' learning status, adjust teaching strategies in time, and improve classroom teaching effects. This method is expected to become an important auxiliary tool in the field of intelligent education in the future, helping teachers and students achieve better teaching results and learning experience.

3.3.2. Comparative Analysis

Within the self-constructed video dataset of students' classroom behavior monitoring, three frames of monitoring images are randomly selected to measure the amount of edge information retained in the monitoring image feature fusion of this paper's method, the method of the literature [3], and the method of the literature [21], and the results of the analysis are shown in Figure 6.



a) The proposed method preserves edge information after feature fusion.

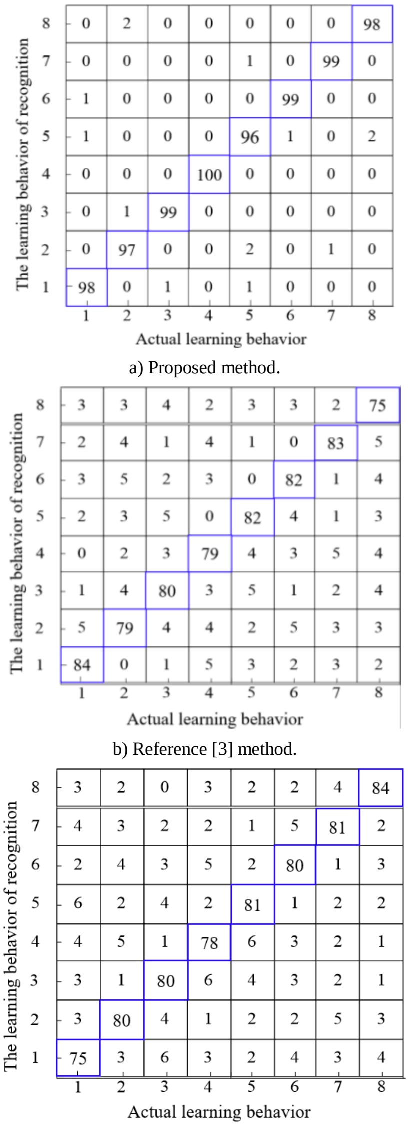
b) Reference [3] method: edge information retention after feature fusion.

c) Reference [21] method: edge information retention after feature fusion.

Figure 6. The fusion effect of monitoring image features.

From Figure 6, it can be seen that for three frames of surveillance images, after the fusion of the ST features of surveillance images by the method of this paper, the retention of the edge information of the feature maps grows with the improvement of the illumination level. This trend shows that the method can effectively process the surveillance images under different lighting conditions and obtain richer and more accurate edge information. The edge information retention after feature fusion of the other two methods does not follow the above trend, and the edge information retention of both methods is maintained at about 0.96 at most. The lowest edge information retention of surveillance images 1, 2 and 3 of this method is close to 1, which indicates that the fused ST feature maps of the multimodal surveillance images in this method have better edge retention. This means that the method can effectively retain the edge details in the original surveillance images, which is crucial for the subsequent image processing and recognition tasks. In summary, the ST feature maps of multimodal surveillance images fused by the method in this paper show good edge retention and rich detail information. This indicates that the method has good fusion effect and can improve the comprehensiveness of surveillance image features.

Within the self-constructed video dataset of students' classroom behavior monitoring, 800 frames of monitoring images are randomly selected, in which the number of monitoring images corresponding to each learning behavior is 100 frames, and this 800 frames of monitoring images are used for learning behavior recognition using the method of this paper, and the confusion matrix is used to analyze the learning behavior recognition accuracy of this paper's method, the method of the literature [3] and the method of the literature [21], and the results of the analysis are shown in Figure 7.



The learning behavior of recognition

8	0	2	0	0	0	0	0	98
7	0	0	0	0	1	0	99	0
6	1	0	0	0	0	99	0	0
5	1	0	0	0	96	1	0	2
4	0	0	0	100	0	0	0	0
3	0	1	99	0	0	0	0	0
2	0	97	0	0	2	0	1	0
1	98	0	1	0	1	0	0	0
	1	2	3	4	5	6	7	8

Actual learning behavior

a) Proposed method.

The learning behavior of recognition

8	3	3	4	2	3	3	2	75
7	2	4	1	4	1	0	83	5
6	3	5	2	3	0	82	1	4
5	2	3	5	0	82	4	1	3
4	0	2	3	79	4	3	5	4
3	1	4	80	3	5	1	2	4
2	5	79	4	4	2	5	3	3
1	84	0	1	5	3	2	3	2
	1	2	3	4	5	6	7	8

Actual learning behavior

b) Reference [3] method.

The learning behavior of recognition

8	3	2	0	3	2	2	4	84
7	4	3	2	2	1	5	81	2
6	2	4	3	5	2	80	1	3
5	6	2	4	2	81	1	2	2
4	4	5	1	78	6	3	2	1
3	3	1	80	6	4	3	2	1
2	3	80	4	1	2	2	5	3
1	75	3	6	3	2	4	3	4
	1	2	3	4	5	6	7	8

Actual learning behavior

c) Reference [21] method.

Figure 7. Confusion matrix for learning behavior recognition.

Figure 7 shows the results of using this paper's method, the literature [3] method and the literature [21] method to identify the students' learning behaviors in this classroom. As can be seen from the figure, literature [3] method and literature [21] method have some deviation, and both of them have the highest recognition accuracy of 84%; while most of the recognition results of learning behaviors in this paper's method are consistent with the actual learning behaviors, which are more than 96%, indicating that this paper's method has a high recognition accuracy. However, there are some recognition errors, among which the recognition error of cell phone playing behavior is the largest. Specifically, four samples of cell phone playing behaviors were incorrectly identified as other behaviors, including writing, looking left and right, and sleeping. These errors may be due to the fact that some minor movements or expressions of the students while playing cell phones are similar to these behaviors, which led to the misclassification of the method in this paper. Although there are some recognition errors, in general, the method in this paper has a high accuracy in recognizing learning behaviors. This is due to the effectiveness of the method in the design of ST feature extraction and fusion of multimodal surveillance images, which fully considers the comprehensiveness of the characteristics of students' learning behaviors.

4. Conclusions

Traditionally, analyzing classroom teaching information relies on direct observation and manual recording of students' performance, which is not only time-consuming and laborious, but also inefficient. In order to do so, a multimodal ST feature representation method for behavior recognition is developed to enhance the comprehensiveness of surveillance image features by fusing the ST features and skeletal features of multimodal surveillance images, and then improve the accuracy of learning behavior recognition. Experimentally, it is proved that the fusion of multimodal ST features has high edge information retention, and can effectively identify learning behaviors with high recognition accuracy. With the development of technology and in-depth research, the learning behavior recognition method based on multimodal ST features of surveillance images is expected to achieve greater breakthroughs and applications in the future. For example, by designing a finer model structure and optimization algorithm, the learning behavior can be identified more accurately, and the accuracy and robustness of the classifier can be improved. The integration of other modal data, such as sound and physiological signals, is also considered to provide more comprehensive learning behavior features, which can help identify and analyze learning behaviors more accurately and provide an effective learning behavior analysis tool for the education field.

References

[1] Abdallah T., Elleuch I., and Guermazi R., "Student Behavior Recognition in Classroom Using Deep Transfer Learning with VGG-16-ScienceDirect,"

Procedia Computer Science, vol. 192, no. 4, pp. 951-960, 2021. https://doi.org/10.1016/j.procs.2021.08.098

[2] Agyeman R., Rafiq M., Shin H., Rinner B., and Choi G., "Optimizing Spatiotemporal Feature Learning in 3D Convolutional Neural Networks with Pooling Blocks," *IEEE Access*, vol. 9, pp. 70797-70805, 2021. DOI: 10.1109/ACCESS.2021.3078295

[3] Chen Y., "Human Behavior Recognition Based on Multiscale Convolutional Neural Network," *IEEE Access*, vol. 11, no. 2, pp. 13533-13544, 2023. DOI:10.1109/ACCESS.2022.3209816

[4] Chonggao P., "Simulation of Student Classroom Behavior Recognition based on Cluster Analysis and Random Forest Algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 2421-2431, 2021. DOI:10.3233/JIFS-189237

[5] Damaneh M., Mohanna F., and Jafari P., "Static Hand Gesture Recognition in Sign Language Based on Convolutional Neural Network with Feature Extraction Method Using ORB Descriptor and Gabor Filter," *Expert Systems with Applications*, vol. 211, pp. 118559, 2022. https://doi.org/10.1016/j.eswa.2022.118559

[6] Gendy G., Sabor N., Hou J., and He G., "Balanced Spatial Feature Distillation and Pyramid Attention Network for Lightweight Image Super-Resolution," *Neurocomputing*, vol. 509, pp. 157-166, 2022. https://doi.org/10.1016/j.neucom.2022.08.053

[7] Gomez L., Biten A., Tito R., Mafla A., Rusiol M., Valveny E., and Karatzas D., "Multimodal Grid Features and Cell Pointers for Scene Text Visual Question Answering," *Pattern Recognition Letters*, vol. 150, pp. 242-249, 2021. https://doi.org/10.1016/j.patrec.2021.06.026

[8] Han X., Huang D., Eun-Lee S., and Hoon-Yang J., "Artificial Intelligence-Oriented User Interface Design and Human Behavior Recognition Based on Human-Computer Nature Interaction," *International Journal of Humanoid Robotics*, vol. 20, no. 6, pp. 2250020, 2023. https://doi.org/10.1142/S0219843622500207

[9] Li G., Liu F., Wang Y., Gou Y., Xiao L., and Zhu L., "A Convolutional Neural Network (CNN) Based Approach for the Recognition and Evaluation of Classroom Teaching Behavior," *Scientific Programming*, vol. 2021, pp. 1-8, 2021. https://doi.org/10.1155/2021/6336773

[10] Li X., Ding M., and Pizurica A., "Spectral Feature Fusion Networks with Dual Attention for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2022. DOI: 10.1109/TGRS.2021.3084922

[11] Mo J., Zhu R., Shou Z., Yuan H., and Chen L., "Student Behavior Recognition Based on Multitask Learning," *Multimedia Tools and Applications*, vol. 82, no. 12, pp. 19091-19108, 2023. https://doi.org/10.1007/s11042-022-14100-7

[12] Sheng W., Sun Y., and Zhang H., "Multi-Focus Image Fusion Algorithm Based on Sparse Theory and FFST-GIF," *Journal of Jiangsu University: Natural Science Edition*, vol. 43, no. 2, pp. 195-200, 2022. DOI: 10.3969/j.issn.1671-7775.2022.02.011

[13] Wang S., "Online Learning Behavior Analysis Based on Image Emotion Recognition," *Traitement du Signal*, vol. 38, no. 3, pp. 865-873, 2021. https://doi.org/10.18280/ts.380333

[14] Wu S., "Simulation of Classroom Student Behavior Recognition based on PSO-KNN Algorithm and Emotional Image Processing," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 4, pp. 7273-7283, 2021. https://doi.org/10.3233/JIFS-189553

[15] Wu S., Jin S., Liu W., and Bai L., et al., "Graph-based 3D Multi-Person Pose Estimation Using Multi-View Images," *arXiv Preprint*, vol. arXiv:2109.05885v1, pp. 1-13, 2021. https://doi.org/10.48550/arXiv.2109.05885

[16] Xie Y., Zhang S., and Liu Y., "Abnormal Behavior Recognition in Classroom Pose Estimation of College Students Based on Spatiotemporal Representation Learning," *Traitement du Signal: Signal Image Parole*, vol. 38, no. 1, pp. 89-95, 2021. https://doi.org/10.18280/ts.380109

[17] Xu C., Gao Z., Zhang H., Li S., and De Albuquerque V., "Video Salient Object Detection Using Dual-Stream Spatiotemporal Attention," *Applied Soft Computing*, vol. 108, pp. 107433, 2021. https://doi.org/10.1016/j.asoc.2021.107433

[18] Zhang L., "Enterprise Employee Work Behavior Recognition Method Based on Faster Region-Convolutional Neural Network," *The International Arab Journal of Information Technology*, vol. 22, no. 2, pp. 291-302, 2025. https://doi.org/10.34028/iajit/22/2/7

[19] Zhang L., Song H., Aletras N., and Lu H., "Node-Feature Convolution for Graph Convolutional Networks," *Pattern Recognition*, vol. 128, no. 8, pp. 108661, 2022. https://doi.org/10.1016/j.patcog.2022.108661

[20] Zhang Y., Guan S., Xu C., and Liu H., "RETRACTED: Based on Spatio-Temporal Graph Convolution Networks with Residual Connection for Intelligence Behavior Recognition," *International Journal of Electrical Engineering Education*, vol. 60, no. 1S, pp. 52-59, 2021. https://doi.org/10.1177/0020720921996600

[21] Zhao H., Liu J., and Wang W., "Research on Human Behavior Recognition in Video Based on 3DCCA," *Multimedia Tools and Applications*, vol. 82, no. 13, pp. 20251-20268, 2023.

https://doi.org/10.1007/s11042-023-14355-8

[22] Zhao L., Mo C., Ma J., Chen Z., and Yao C., "LSTM-MFCN: A Time Series Classifier Based on Multi-Scale Spatial-Temporal Features," *Computer Communications*, vol. 182, pp. 52-59, 2022. https://doi.org/10.1016/j.comcom.2021.10.036

Lei Ma graduated with a bachelor's degree in Computer Application Technology from China Agricultural University in 2002, graduated with a Master's degree in Computer Technology from Beihang University in 2008. Currently serving as a professor at the School of Artificial Intelligence, Beijing Polytechnic University. Won the Silver Award in the World Vocational College Skills Competition in 2024. In 2023, the course "Web Front end Technology Development" won the Special Award for Teaching Design Results of the Beijing Vocational College Curriculum Ideological and Political Benchmark Achievement Selection Tree; The classroom teaching and assessment results won the first prize. Has been awarded titles such as school level teaching and education pioneer, teacher ethics pioneer, excellent homeroom teacher, outstanding individual in news and publicity, excellent teacher, and teacher ethics pioneer multiple times.

Hongxue Yang Current Deputy Secretary of the Party Branch of Beijing Polytechnic University. Dean of the School of Integrated Circuits. Lead the overall administrative work of the college. Mainly engaged in research in the field of ence, with over ten academic papers published.

Guanghao Jin graduated from Peking University in June 2002 with a bachelor's degree in Computational Mathematics and Applied Softwares; Graduated from China Academy of Engineering Physics in June 2005 with a Master's degree in Computer Software and Theory; Graduated from Tokyo Institute of Technology in March 2014 with a PhD in Mathematical and Computing Sciences. From April 2014 to November 2015, worked as a postdoctoral researcher at Tokyo Institute of Technology. From March 2016 to December 2020, worked as a lecturer at Tiangong University, research in the fields of big data and artificial intelligence. From January 2021 to present, Associate Professor at Beijing Polytechnic University, research in the fields of big data and artificial intelligence. Hosted and participated in more than 6 projects of the National Natural Science Foundation and provincial and ministerial plans, and published over 20 academic papers including SCI in well-known domestic and foreign academic journals in the fields of Artificial Intelligence, Big Data Analysis, and Computer Vision.