IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 3, June, 2025, pp. 1767-1778
Journal homepage: http://iieta.org/journals/ts

A Vision-Based Gesture Recognition and Student Engagement Assessment Model for Interactive Educational Environments

Xuan Zhang[1], Songlin Wang[2*]

[1] School of Economics and Management, Chuzhou University, Chuzhou 239000, China
[2] Business School, Jiangsu Open University, Nanjing 210000, China

Corresponding Author Email: wangsl@jsou.edu.cn

https://doi.org/10.18280/ts.420345

Received: 18 December 2024
Revised: 8 May 2025
Accepted: 17 May 2025
Available online: 30 June 2025

Keywords:
interactive educational environments, image processing, gesture recognition, student engagement assessment

ABSTRACT

In the context of educational digitalization, traditional classroom methods for assessing student engagement—primarily based on teachers' subjective observations—suffer from limited real-time capabilities and lack of objectivity, making it difficult to accurately capture students' true interactive behavior. Taking management classrooms as an example, student actions such as raising hands or gesturing are key indicators of learning enthusiasm. Vision-based gesture recognition technology offers a novel, objective approach to engagement assessment. However, existing methods relying on depth cameras face challenges such as high equipment costs, poor environmental adaptability, and simplistic gesture-counting models that ignore contextual semantic information, making them unsuitable for complex classroom scenarios. This study focuses on interactive educational settings and proposes a gesture recognition and engagement assessment model based on image processing. First, we optimize image preprocessing, feature extraction, and pattern recognition algorithms for standard classroom environments to achieve high-precision, real-time recognition of various gestures such as hand-raising and waving. Second, by integrating multi-dimensional data—gesture types, frequency, and duration—with instructional context, we construct a dynamic evaluation model that addresses the robustness issues of traditional methods in complex settings. The proposed approach offers a contactless solution for engagement assessment in smart classrooms, supports teachers in refining instructional strategies, and facilitates the digital transformation of educational interaction. This work holds significant implications for improving teaching quality and advancing educational technology innovation.

1. INTRODUCTION

In management classrooms, teachers often need to understand students' learning status and mastery of knowledge through interaction with students, such as asking questions, organizing discussions, and guiding students in case analysis [1-4]. Students' engagement in class not only affects their own learning outcomes but is also related to the quality and efficiency of classroom teaching [5, 6]. Traditional methods of assessing student engagement mainly rely on teachers' subjective observation and simple classroom records [7-9]. This method has strong subjectivity and limitations, making it difficult to accurately and in real-time capture students' subtle reactions and actual engagement in class. With the continuous development of educational informatization [10-13], how to use advanced technical means to achieve objective and accurate assessment of student engagement has become an important problem to be solved urgently in the field of education. In this process, gestures, as a natural and intuitive interaction method [14, 15], have important application value in educational interactive scenarios. Students' actions in class such as raising hands and gesturing often reflect their learning enthusiasm and willingness to participate. Therefore, gesture recognition technology based on image processing provides new ideas and methods for accurately assessing student engagement.

Accurately assessing student engagement in educational interactive scenarios has important practical significance for optimizing the teaching process and improving teaching quality. Through the recognition and analysis of student gestures, the learning status and needs of students can be obtained in real-time, providing a basis for teachers to adjust teaching strategies and improve teaching methods, thus achieving more personalized and precise teaching. At the same time, this research helps to better understand students' behavior patterns and learning habits in class and provides empirical support for the research and development of educational theory. In addition, gesture recognition technology based on image processing has advantages such as non-contact and strong real-time performance [16, 17], which can achieve engagement assessment without disturbing students' normal learning, and has broad application prospects. It can not only be applied to traditional classroom teaching, but also extended to online education, virtual experiments, and various other educational interactive scenarios, laying a foundation for building a smart education environment.

At present, there have been some studies on gesture recognition and student engagement assessment in educational

interactive scenarios. For example, some studies [18, 19] adopt gesture recognition methods based on depth cameras. Although they have achieved certain results in specific environments, the equipment cost is high, and they are sensitive to environmental lighting, student postures, and other factors, making it difficult to be widely applied in ordinary classroom environments. In terms of student engagement assessment, some studies [20, 21] only use simple gesture counting to judge engagement, ignoring the diversity and contextual information of gestures, resulting in insufficient accuracy and comprehensiveness of the assessment results. In addition, most existing assessment models do not fully consider the complexity of educational interactive scenarios, such as the simultaneous participation of multiple students, and classroom environment interference, which limits the robustness and applicability of the models.

The main research content of this paper includes two parts. The first part is gesture recognition in educational interactive scenarios, aiming to study image processing algorithms suitable for ordinary classroom environments, to achieve accurate recognition of various student gestures, including raising hands, waving, thumbs-up and other common classroom gestures. By improving traditional image preprocessing, feature extraction, and pattern recognition methods, the accuracy and real-time performance of gesture recognition are enhanced, and the dependence on special equipment is reduced. The second part is student engagement assessment based on gesture recognition in educational interactive scenarios, which will construct a scientific and reasonable student engagement assessment model by combining multi-dimensional information such as gesture types, frequency, and duration. This model will comprehensively consider the specific context of classroom teaching, such as teaching content and forms of teaching activities, to realize dynamic assessment and analysis of student engagement. The value of this study lies in proposing a gesture recognition and student engagement assessment model based on image processing, providing new methods and technical support for student engagement assessment in educational interactive scenarios. Through this model, teachers can more objectively and accurately understand students' participation, adjust teaching strategies in time, and improve teaching effectiveness. Meanwhile, this research enriches the research results in the field of educational technology regarding gesture recognition and engagement assessment, provides useful references for subsequent related studies, and helps to promote the development and application of smart education.

2. GESTURE RECOGNITION IN EDUCATIONAL INTERACTIVE SCENARIOS

2.1 Method framework

In educational interactive scenarios, for short-term motion features with clear instructional semantics such as raising hands, waving, and classroom demonstration gestures, the gesture recognition method proposed in this paper achieves fine-grained modeling of multi-time-scale sub-gesture actions by constructing an adaptive temporal segmentation attention module and a dynamic information fusion module. Considering the characteristics of high frequency and short duration of gesture interactions in classroom environments, the joint coordinates of the hand skeleton sequence are first transformed into high-dimensional feature vectors through a linear embedding layer to retain spatial structure information. Then, using the adaptive temporal segmentation attention module, the input feature sequence is segmented at multiple scales in the temporal dimension based on knowledge of short-term motion, capturing intra-frame joint associations and inter-frame motion trajectories of local sub-actions through different time windows. This module learns the correlation weights between different joints in adjacent frames and dynamically focuses on key semantic action segments in educational scenarios, avoiding the general treatment of entire action sequences in traditional models. Meanwhile, the dynamic information fusion module adaptively fuses the short-term features extracted by branches of different time scales through weight vector learning, which not only preserves the detailed features of individual gestures but also enhances the robustness of the model against interference factors such as complex classroom lighting and multi-student occlusion, providing high-precision basic gesture recognition results for subsequent engagement assessment. Figure 1 shows the framework of the gesture recognition model in educational interactive scenarios.

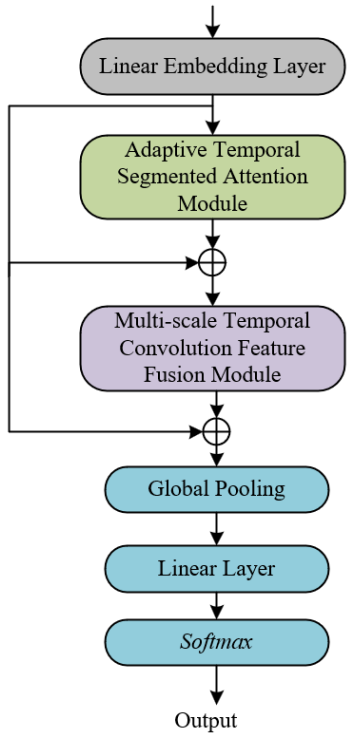

Figure 1. Framework of the gesture recognition model in educational interactive scenarios

In response to long-duration gesture interactions that may occur in educational scenarios, the method captures long-term correlations between different sub-action segments through a multi-scale temporal convolution feature fusion module. This module performs hierarchical processing on the encoded skeleton feature sequences using multi-scale temporal convolution kernels to learn both short-distance action transitions and long-distance temporal dependencies. In the L-layer stacked encoder modules, each layer achieves progressive spatiotemporal feature extraction through the cascade of the adaptive temporal segmentation attention module and the multi-scale temporal convolution feature

fusion module: the former focuses on fine modeling of local sub-actions, while the latter captures semantic correlations of cross-period action segments. Finally, after global average pooling and classification layers, the model can accurately recognize gesture categories with specific instructional intentions in educational scenarios. Through interpretable temporal segmentation attention weights and multi-scale convolution feature mapping, the gesture recognition process is transformed into a human-understandable decision logic of "sub-action - time scale - semantic label," solving the problem of insufficient interpretability of traditional models for complex gesture sequences in classroom environments.

2.2 Linear embedding layer

In educational interactive scenarios, gestures serve as an important medium of communication between students and teachers, and their core features are composed of spatial coordinate sequences of hand joints. For example, the dynamic position changes of shoulder, elbow, and wrist joints when raising a hand, or the motion trajectory of the arm when waving. The linear embedding layer proposed in this paper targets such skeleton sequences composed of a small number of joints and discards the traditional visual Transformer's approach of dividing images into fixed-size patches, directly performing linear mapping on the 3D coordinates of each joint. Specifically, through a learnable weight matrix, the coordinate vector of each joint is mapped into a d-dimensional feature vector, thereby converting the original skeleton sequence into a feature vector sequence. This joint-based direct mapping method not only retains the spatial structure information of the hand skeleton but also provides basic representations for the model to capture subtle differences in educational gestures through transformation into high-dimensional feature space. Suppose the feature vector sequence obtained after linear mapping is denoted by $\tilde{C}_0$. The frame number, number of joints, and number of feature channels are denoted by S, N, and Z, respectively. The 2D convolution with kernel size 1×1 is denoted by $CONV2D_{(1\times1)}$. The input feature vector sequence obtained by linear mapping for each joint is:

$$\tilde{C}_0 = CONV2D_{(1\times1)}\left(A_{SEQ}\right) \in \Re^{S\times N\times Z} \tag{1}$$

Although the feature vector sequence HHH obtained through linear mapping contains the spatial coordinate information of joints, it lacks two key dimensions: first, the spatial hierarchical relationship of joints in the skeleton sequence; second, the temporal order information of gesture actions. In response to the temporal dependency of gestures in educational scenarios—for example, the "raising hand to ask a question" action involves continuous stages of "lifting arm → fixing wrist → opening palm"—this paper adopts the sine-cosine positional encoding method consistent with Transformer to generate encoded vectors containing time and spatial position information, denoted as PE. Among them, spatial position encoding is generated through the hierarchical index of joints in the skeleton tree, and temporal position encoding is calculated based on the frame index of the action sequence. Suppose the position of each feature vector in the input sequence is denoted by o, and the specific index of the positional encoding vector is denoted by $u\in[0,1,\ldots,Z/2]$, the specific formula is:

$$OR(o,2u) = SIN\left(o/10000^{2u/Z}\right) \tag{2}$$

$$OR(o,2u+1) = COS\left(o/10000^{2u/Z}\right) \tag{3}$$

By adding the positional encoding vector to the feature vector after linear embedding, the model is not only provided with spatiotemporal coordinate references for gesture actions but also enhances the semantic attributes of gestures in educational scenarios. For example, through temporal encoding, the model can distinguish the engagement differences between "raising hand quickly" and "raising hand slowly"; through spatial encoding, it can recognize the action type differences between "waving with one hand" and "gesturing with both hands."

In educational interactive scenarios, gesture recognition needs to meet the requirements of real-time performance and robustness. For example, in class, teachers need to instantly capture students' hand-raising gestures to adjust the teaching pace. The design of the linear embedding layer is optimized in two aspects to adapt to this demand: First, the coordinates of joints are directly linearly mapped, avoiding the computational redundancy caused by traditional image patch segmentation, significantly reducing the input dimension and computational complexity of the model, laying the foundation for real-time processing. Second, with the supplementation of spatiotemporal information from positional encoding, the model can, when dealing with complex scenes with multiple students participating simultaneously, distinguish the gesture trajectories of different individuals through the spatial position encoding of joints, and capture the temporal correlation of continuous gestures of the same student through temporal encoding. The input feature C_{IN} after being processed by the linear embedding layer not only preserves the physical properties of educational gestures but also injects spatiotemporal coordinates consistent with the semantics of teaching scenarios, providing structured input for the subsequent encoder modules to model gesture sequences with clear educational intentions such as "raise hand - ask question" and "wave hand - speak," ultimately achieving accurate recognition and semantic analysis of diverse gestures in classroom interaction. The specific input feature formula is:

$$C_{IN} = \tilde{C}_{IN} + OR \in \Re^{S\times N\times Z} \tag{4}$$

2.3 Inter-segment self-attention mechanism

In educational interactive scenarios, the semantics of gestures are usually composed of multiple sub-actions with clear instructional intentions arranged in temporal order. For example, the "raise hand to speak" action can be decomposed into three sub-action stages: "arm lifting → palm turning outward → holding still," with each stage corresponding to a specific joint co-movement pattern. The inter-segment self-attention mechanism targets this type of temporally structured feature. By dividing the input feature sequence into fixed-length, non-overlapping temporal segments along the time dimension, it limits self-attention computation within local temporal windows, thereby explicitly modeling the interaction relationships between joints within sub-actions. Figure 2 shows the architecture of the inter-segment self-attention mechanism module. Taking the commonly seen "waving to signal" gesture in the classroom as an example, this

mechanism focuses within a single time segment on the dynamic correlation between the angle changes of the shoulder and elbow joints and the trajectory of the wrist joint during arm swinging. Through query, key, and value interaction aggregation, it captures the motion coordination between different joints within the same sub-action stage. For instance, when the arm swings forward, the model will automatically enhance the weight correlation between the shoulder joint rotation angle and the wrist joint displacement speed, suppressing interference from irrelevant joints, thereby accurately extracting the feature representation of the core sub-action "arm swinging." Specifically, suppose the length of each temporal segment is denoted by v, and the total number of joints in each temporal segment is $N_v=N\times v$, then the input feature C_{IN} is divided along the time dimension into S_v non-overlapping temporal segments of equal length:

$$C_{IN} \in \Re^{S\times N\times Z} \rightarrow \Re^{S_v\times v\times N\times Z} \rightarrow \Re^{S_v\times N_v\times Z} \tag{5}$$

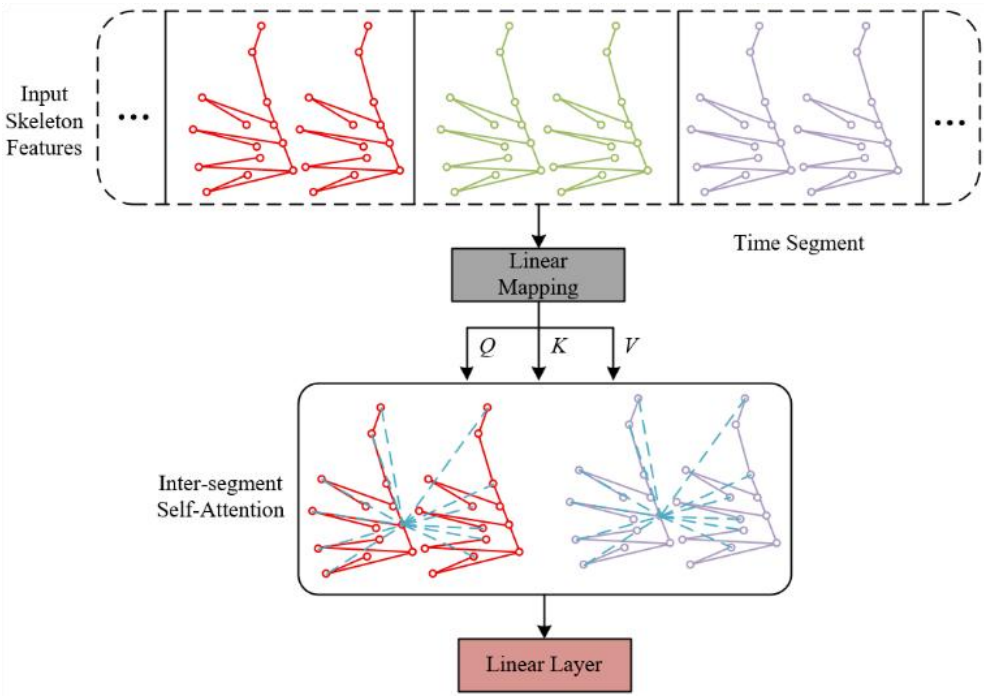

Figure 2. Architecture of the inter-segment self-attention mechanism module

Assuming that the linear mapping matrices for Q, K, and V are represented by W_Q, W_K, and W_V, respectively, Q, K, and V are obtained by applying linear mapping to C_{IN} as follows:

$$Q,K,V = W_Q C_{IN}, W_K C_{IN}, W_V C_{IN} \tag{6}$$

Then, the output features can be obtained by multiplying the attention weight matrix—calculated from the correlation between Q and K—with N. Suppose the output features of the self-attention mechanism operation are denoted by C_{AT}, the hyperbolic tangent function is denoted by *tanh*, gradient stability during training is adjusted by $(Z)^{1/2}$, and the attention weight matrix calculated from the similarity between Q and K is denoted by F, the formula is:

$$F = \tanh\left(QK^S / \sqrt{Z}\right) \tag{7}$$

In view of the inherent hinge-joint connection characteristics of the hand skeleton in educational gestures, the inter-segment self-attention mechanism introduces a learnable topological parameter matrix O to model the physical connection relationships of hand joints. For example, in the "heart gesture," the fingertip contact between the thumb and index finger relies on the spatial position constraint between them. The matrix O can automatically learn this prior structural information through training, allowing the model to prioritize the coordinate interaction between thumb and index finger joints when processing this sub-action. Specifically, during the calculation of attention weights, the feature interaction matrix F of joints is fused with the topological matrix O through weighted addition, where β is a learnable parameter, dynamically balancing the contributions of data-driven features and domain knowledge. The output feature expression of the self-attention mechanism operation is:

$$C_{AT} = V\left(\beta \times F + O\right) \tag{8}$$

Considering the diversity of gestures in educational interaction, the inter-segment self-attention mechanism adopts a multi-head attention strategy, dividing the input features into multiple subspaces for parallel processing, with each head focusing on different dimensions of sub-action features. For example, the first head may specifically capture fine movements of the fingers, while the second head focuses on the macro movement of the arms. Through the concatenation of multi-head results, the model can retain both the detailed features and the overall shape of gestures. In complex classroom scenarios with multiple students participating simultaneously, the multi-head mechanism can also distinguish the gesture trajectories of different individuals by assigning weights through different heads. For example, in the spatially overlapping area where the student on the left is "raising hand" and the student on the right is "waving," the

model can use the attention weights of specific heads to focus on their respective joint combinations. Furthermore, the fully connected layer at the end of the mechanism further performs nonlinear transformation on the multi-head output to adapt to gesture feature variations caused by lighting changes and occlusion in educational scenarios. For example, when a student's raised hand is partially occluded by a book, the fully connected layer can learn the historical trajectory of joint movement to complete the missing spatial information, thereby improving recognition robustness under complex conditions. Suppose the matrix concatenation along the feature channel dimension is denoted by *CONCAT*, and the output feature of the *g*-th attention group is denoted by C^{g}_{ATT}. Finally, the output feature C_{ATT} is obtained through concatenation:

$$C_{ATT} = CONCAT\left(C^{1}_{ATT}, C^{2}_{ATT}, ..., C^{3}_{ATT}\right) \tag{9}$$

To enhance the fitting capability of the network, the inter-segment self-attention mechanism finally adopts a fully connected layer for information processing. Suppose the 1×1 convolution operation is denoted as $CONV2D_{(1\times1)}$, its output features are obtained by the following formula:

$$C_{DJ\_ATT} = CONV2D_{(1\times1)}\left(C_{ATT}\right) \tag{10}$$

Although the inter-segment self-attention mechanism achieves explicit modeling of sub-actions through fixed time segments, the duration of sub-actions in educational gestures varies significantly. A single fixed segment length may lead to the fragmentation of long-duration sub-action features. Therefore, in educational scenario applications, the mechanism introduces a domain knowledge-guided segment length adaptation strategy: Based on prior classroom observations, the average duration of common gesture sub-actions is used as a baseline parameter for segment division, while allowing the model to dynamically adjust segment boundaries during training according to specific gesture types.

2.4 Adaptive temporal segmentation attention module

In educational interactive scenarios, the time dimension of gestures presents significant variability: instantaneous gestures such as "raise hand to ask a question" require capturing sudden joint trajectory changes during high-speed motion, while procedural gestures such as "demonstration of experimental operations" contain temporal combinations of multi-stage sub-actions like "grasping tool → adjusting gesture → demonstrating action." Traditional fixed-time-segment attention mechanisms struggle to balance feature extraction for both types of gestures. Short segments may fragment the continuity of procedural gestures, while long segments may include irrelevant frames for instantaneous gestures. To address this, the adaptive temporal segmentation attention module adopts parallel multi-branch modeling. According to different gesture motion speeds and stage features, three temporal scale branches—short, medium, and long—are designed to respectively focus on high-frequency features of instantaneous actions and temporal dependencies of long-duration actions. For example, when recognizing "group discussion gesture combinations," the short branch captures "joint velocity spikes at gesture switching moments," while the long branch models "spatial position transitions between consecutive gestures," thereby preserving the temporal semantic structure of educational gestures. The module architecture is shown in Figure 3.

The module splits the input skeleton feature sequence C_{IN} along the channel dimension into *J* parallel branches, with each branch processing *Z*/*J* dimensional features and corresponding to time segments of different lengths. Suppose the split operation along the channel dimension is denoted by *SPLIT*, then:

$$C^{1}_{IN}, C^{2}_{IN}, ..., C^{u}_{IN}, ..., C^{J}_{IN} = SPLIT\left(C_{IN}\right) \tag{11}$$

Suppose the output features of the *u*-th branch are denoted by $C^{u}_{DJ\_ATT}$, and the inter-segment self-attention mechanism using time segment length u is denoted by $DJ\_ATT_{v=u}$. C_{IN} is input to different branches for short-time motion feature modeling:

$$C^{u}_{DJ\_ATT} = DJ\_ATT_{v=u}\left(C^{u}_{IN}\right) \tag{12}$$

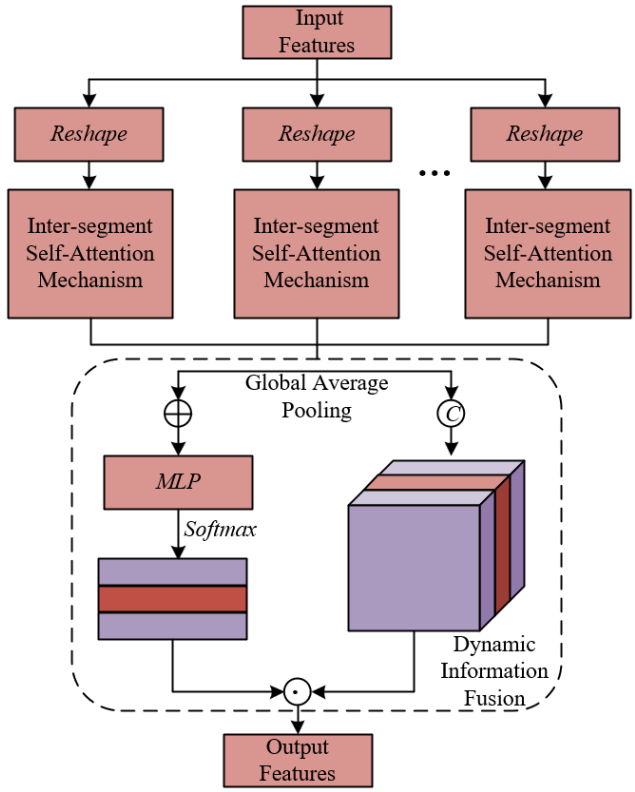

Figure 3. Architecture of the adaptive temporal segmentation attention module

This design brings two advantages. First, targeted feature extraction: Each branch independently models sub-actions at specific temporal scales through the inter-segment self-attention mechanism. For example, the short branch performs local attention computation on a 0.5-second segment of a "quick hand raise" gesture, enhancing the weight correlation of instantaneous joint coordination between the shoulder and wrist; the long branch models a 4-second segment of a "slow demonstration gesture," capturing the smooth variation pattern of palm rotation angle over time. Second, computational complexity optimization: since the channel dimension of each branch is reduced to 1/*J*, the computational cost of the attention mechanism is reduced from $O(S^2NZ)$ of a single-branch scheme to $O(S^2N(Z/J)^2)$, with a total computation of

only $1/J^2$ of the original scheme. This feature is highly suitable for real-time classroom interaction requirements. In scenarios with multiple students raising hands simultaneously, the model can process multiple skeleton data streams rapidly on ordinary computing devices, avoiding gesture recognition lag caused by computational delay.

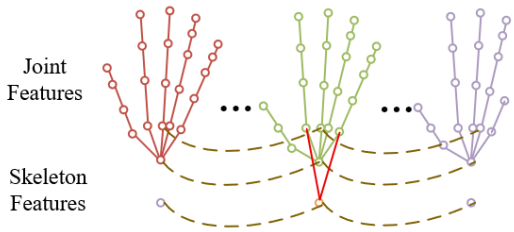

Figure 4. Feature fusion diagram of the module

To address the problem of feature fusion across different temporal scales, the module introduces a dynamic information fusion mechanism, which achieves semantic integration of multi-branch features through three stages: "feature selection - weight allocation - adaptive fusion". Figure 4 provides a diagram of the feature fusion process.

(1) Global representation generation: First, the outputs of the J branches are summed and then compressed into a compact global vector via global average pooling to extract the overall motion trend of the educational gesture.

(2) Dynamic weight computation: Using a multi-layer perceptron (MLP) and Softmax layer, weight vectors for each branch are generated based on the global representation. When the input is a "hand raise" gesture, the model automatically assigns higher weights to the short branch and lower weights to the long branch. When processing an "experimental demonstration gesture," the long branch weights increase significantly to capture the temporal association of multi-stage actions. Suppose the weight vector of the u-th branch is denoted by Q_u, then the expression for the weight vector of each set of features is:

$$Q = MLP\left(GAP\left(\sum_{u=1}^{J} Z_{DJ\_ATT}^{u} \right) \right) \in \Re^{JZ\times 1\times 1} \tag{13}$$

$$Q_1, Q_2, ..., Q_J = SPLIT\left(Q\right) \in \Re^{Z\times 1\times 1} \tag{14}$$

(3) Multi-scale feature fusion: The output features of each inter-segment self-attention mechanism module are multiplied and summed with their corresponding weight vectors to obtain the final output feature:

$$C_{DTRH} = \sum_{u=1}^{J} Q_u \otimes C_{DJ\_ATT}^{u} \tag{15}$$

Through feature fusion, the model can retain both short-term details such as "slightly spreading fingers" and long-term context such as "arm movement trajectory." This dynamic weighting mechanism is particularly suitable for the requirement of "gesture semantics changing with teaching context" in educational scenarios. For the same "waving" gesture, in the contexts of "class end signal" and "group discussion guidance," the model automatically adjusts branch weights to generate feature representations consistent with scenario semantics.

To address interferences such as lighting variation and partial occlusion that may occur in classroom environments, the module introduces residual connections at the output stage to directly add the original input features with the fused high-level features. By introducing residual connections, the basic motion information of non-occluded joints is effectively retained, avoiding recognition bias caused by local feature loss. For example, when the wrist joint is occluded, residual connections can assist in inferring the spatial position of the wrist through the historical motion trajectory of the shoulder and elbow, improving the robustness of gesture recognition in complex environments. The output feature expression of the adaptive temporal segmentation attention module with residual connection is:

$$\begin{aligned} &C_{ZS\_ATT} = DTRH \\ &\left(C_{DJ\_ATT}^{1}, C_{DJ\_ATT}^{2}, ..., C_{DJ\_ATT}^{u}, ..., C_{DJ\_ATT}^{J}\right) + C_{IN} \end{aligned} \tag{16}$$

The adaptive temporal segmentation attention module constructs a mapping channel from hand skeleton motion to educational gesture semantics through a three-layer architecture of "multi-scale branch modeling → dynamic weight fusion → residual enhancement." For the teacher side, it can parse students' instantaneous interactive behaviors such as "raise hand to ask questions" and "wave to express doubt" in real time, providing accurate references for classroom rhythm adjustment; for the student side, it can capture complex participation behaviors such as "experimental operation gesture sequence" and "group discussion gesture combination," providing fine-grained features for personalized learning analysis; for educational technology systems, its lightweight design and dynamic adaptability meet the requirements of real-time and robustness in multimodal data processing in smart classrooms, becoming a core component connecting underlying skeleton data and higher-level participation evaluation.

2.5 Multi-scale temporal convolution feature fusion module

In educational interactive scenarios, the semantic understanding of gestures not only relies on the accurate recognition of individual sub-actions but also requires capturing the temporal correlations between different sub-actions and the long-term coordinated motion features of hand joints. Although traditional short-term sub-action modeling methods can capture joint interactions within local time segments, they fail to model long-range action dependencies across segments and cannot effectively integrate overall skeleton motion with the dynamic relationship of local joints. To this end, the multi-scale temporal convolution feature fusion module combines multi-scale temporal convolution and skeleton feature fusion mechanisms to compensate for the shortcomings of short-term sub-action modeling and achieve deep analysis of long-term correlations of educational gestures. The module architecture is shown in Figure 5.

The module first performs multi-branch temporal modeling on the input features A_{IN}, using six different temporal operation branches to capture multi-scale temporal dependencies:

(1) Basic feature enhancement: 1×1 convolution adjusts channel weights to achieve cross-channel fusion of joint features, enhancing the feature expression of key joints in educational gestures. For example, in the "heart-shaped gesture," 1×1 convolution can enhance the feature weights of

the thumb and index finger joints while suppressing interference from irrelevant joints.

(2) Salient frame capture: 3×1 max pooling layer extracts key frame features in the time series, suitable for identifying peak frames of arm lift in the "raise hand" gesture or extreme points of arm swing in the "wave" gesture, avoiding redundancy dilution of core motion features.

(3) Multi-scale temporal modeling: Four 3×1 dilated convolutions with different dilation rates are used to respectively capture short-distance action transitions, medium-distance action connections, and long-distance action dependencies. For example, in the "experimental operation gesture sequence," branches with small dilation factors capture the rapid closure of finger joints during "grasping the tool," while branches with large dilation factors model the angle stability of wrist joints during "sustained gripping."

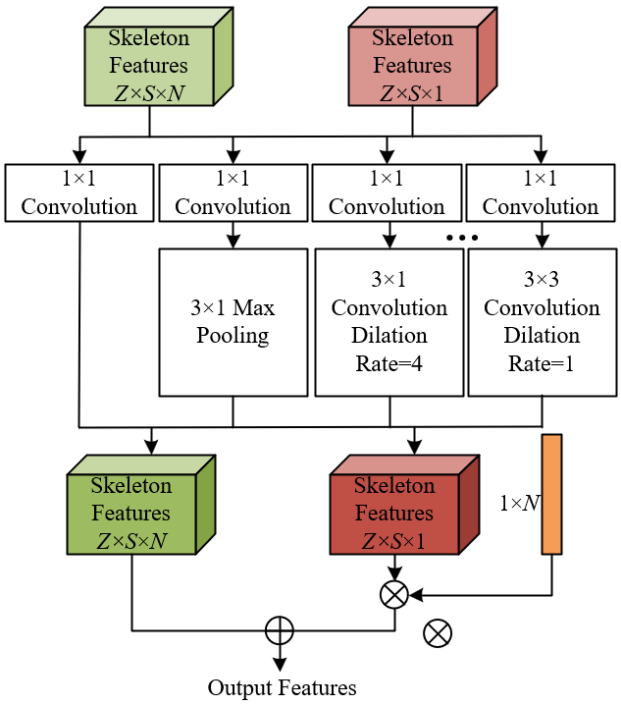

Figure 5. Architecture of the multi-scale temporal convolution feature fusion module

All branch outputs are concatenated along the channel dimension and fused through a 1×1 convolution to form an intermediate representation containing multi-scale temporal information, which not only retains short-term details such as "micro finger movements" but also extracts long-term trends such as "arm movement trajectories," providing rich temporal feature input for subsequent skeleton feature fusion. Specifically, suppose the feature of the u-th joint is denoted as A_u, and the number of joints is denoted as N. Global hand skeleton features containing global information can be obtained by applying average pooling along the joint dimension on A_{IN}:

$$A_T = \frac{1}{N}\sum_{u=1}^{N} A_u \tag{17}$$

Suppose the output features of the six branches are denoted as $A_1, A_2, \ldots, A_6$, the concatenation operation along the channel dimension is denoted as $CONCAT$, and the 1×1 convolution operation is denoted as $CONV2D$. The multi-scale temporal convolution computation process is:

$$A_{MTC} = CONV2D_{(1\times1)}\left(CONCAT\left(A_1, A_2, ..., A_6\right)\right) \tag{18}$$

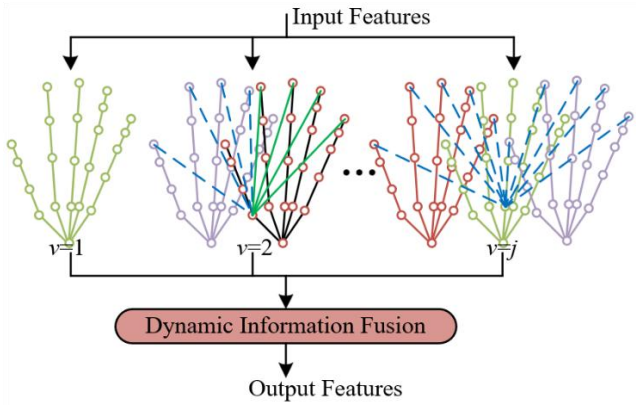

Figure 6. Schematic diagram of module dynamic information fusion

To address the correlation between overall skeleton motion and local joint points in educational gestures, the module introduces skeleton-joint feature fusion guided by dynamic weights. The schematic is shown in Figure 6. First, average pooling over the joint dimension is performed to generate skeleton features containing global information of all joints, representing the overall motion pattern of the gesture. Furthermore, the input features and skeleton features are both dimensionally reduced via 1×1 convolution, and the fusion coefficient between the two is calculated through a learnable weight vector to obtain the output feature. For example, in the "raise hand" gesture, this mechanism enhances the feature weights of dominant joints such as the shoulder and elbow, while integrating the vertical motion trend of the entire skeleton to avoid recognition bias caused by occlusion of local joints. In "group discussion gestures," guided by skeleton features, the model can capture spatial position transitions between different gestures. Suppose the learnable weight vector is denoted as Q_k, then the output feature expression of the multi-scale temporal convolution feature fusion module is:

$$A_{OUT} = MTC\left(A_{IN}\right) + MTC\left(A_T\right)Q_k \tag{19}$$

This fusion mechanism breaks the traditional method of modeling individual joints in isolation and establishes a dynamic relationship between local joint features and global skeleton motion, especially suitable for complex gestures in educational scenarios that require multi-joint coordination.

3. GESTURE RECOGNITION-BASED STUDENT ENGAGEMENT EVALUATION IN EDUCATIONAL INTERACTIVE SCENARIOS

In order to further obtain accurate student engagement evaluation results, it is first necessary to clarify the typical gestures related to student engagement and divide the evaluation dimensions according to the nature of engagement they reflect. Common educational interactive gestures include raising hands to speak, waving to ask questions, nodding to show agreement, shaking head to show confusion, making a heart gesture to show praise, spreading hands to indicate

incomprehension, etc. These gestures can be classified into three core dimensions: active participation, emotional feedback, and confusion expression. Gestures of active participation such as raising hands and actively gesturing to illustrate knowledge points directly reflect the student's willingness to actively output information and seek interaction opportunities in the classroom; emotional feedback gestures such as nodding and making a heart gesture reflect the student's emotional agreement with the teaching content and attentional investment; confusion expression gestures such as shaking head and spreading hands indicate that the student has encountered obstacles in understanding during the learning process and is in a passive or inefficient participation state. By establishing a mapping relationship between gesture types and engagement dimensions, a foundation is laid for subsequent evaluation. For example, it is stipulated that each instance of the hand-raising gesture adds +5 points to active participation, each nod adds +3 points to emotional feedback, and each spreading hands gesture adds +4 points to confusion, thus converting gestures into quantifiable engagement indicators.

After clarifying the engagement dimension corresponding to each gesture type, the frequency of each gesture needs to be counted in real time, and the weights dynamically adjusted according to different teaching scenarios. In lecture-based classes, gestures such as raising hands and asking questions should have relatively higher contribution weights to engagement, as these gestures directly reflect students' active absorption and output of knowledge; while in group discussions or interactive sessions, the weights of emotional feedback gestures and collaboration-based gestures should be increased to reflect students' depth of participation in group interactions. In addition to frequency and type, the duration of gestures is also an important indicator for evaluating engagement. For example, a longer duration of the hand-raising gesture indicates that the student is eager to participate in the interaction and has a strong willingness to engage; while a brief shake of the head may just represent a momentary confusion, a prolonged spreading hands gesture may indicate that the student has been in a state of incomprehension for a long time and has low engagement. By setting reasonable duration thresholds for each type of gesture, abnormal behaviors can be identified for cases exceeding or falling below the thresholds. For active participation gestures such as raising hands, if the duration exceeds 30 seconds, additional engagement points can be added, indicating that the student actively and persistently seeks interaction opportunities; if the duration is less than 5 seconds before putting the hand down, it may be regarded as a casual gesture, and no extra points are added or even a small deduction is applied, to distinguish between effective and ineffective participation. For confusion expression gestures such as spreading hands, if the single duration exceeds 20 seconds, a warning mechanism should be triggered to alert the teacher that the student may have a learning obstacle, and the confusion weight in the engagement score should be doubled to highlight the negative impact of prolonged confusion on engagement. At the same time, the contextual situation of the gesture should be considered. For example, the significance of the gesture duration differs between raising hands immediately after a teacher's question and raising hands suddenly during the lecture. The former more strongly reflects active response to the question and can be given a higher time-related score.

To achieve more accurate student engagement evaluation, gesture recognition results need to be fused with other modality data and comprehensively analyzed in conjunction with specific teaching contexts. For example, when a student raises their hand, if their facial expression is focused and their gaze is fixed on the teacher, it indicates a high level of engagement; if their eyes are wandering and their facial expression is indifferent while raising their hand, it may be a passive response and actual engagement is low. In online education scenarios, gesture recognition can be combined with students' actions in the virtual classroom, such as clicking on courseware or dragging knowledge points. For instance, if a student frequently clicks on related courseware areas while making gestures to highlight key points, it indicates that they are actively focusing on the learning content, and their engagement should be rated higher. In addition, depending on different teaching goals, the focus of contextual evaluation also varies: in knowledge-transmission courses, attention is given to the correlation between gestures such as raising hands and asking questions and the mastery of knowledge points; in practice-oriented courses, attention is paid to the coordination between gestures and the use of tools or demonstration steps. By constructing a multimodal fusion evaluation model and using machine learning algorithms to train on historical data, a complex mapping relationship between gesture features and engagement can be established, ultimately outputting student engagement evaluation results that comprehensively consider gesture types, frequency, duration, multimodal information, and teaching contexts, providing accurate evidence for teachers to adjust teaching strategies and optimize interaction design.

4. EXPERIMENTAL RESULTS AND ANALYSIS

From the normalized confusion matrix in Figure 7, it can be observed that the diagonal elements generally present high values. The prediction accuracy of original class 61 is 1, class 65 is 0.94, class 70 is 0.75, class 75 is 0.7, and class 80 is 0.98. The off-diagonal elements are all relatively low, for example, the misclassification probability of 61 as 65 is only 0.01, 65 as 66 is 0.06, and 70 as 69 or 71 is both less than 0.1. This indicates that the gesture recognition model has excellent classification performance across categories, with high recognition accuracy for common classroom gestures and low inter-class confusion. Experimental data verifies the effectiveness of the image processing algorithm proposed in the paper: by optimizing the algorithm, not only the dependence on special devices is reduced, but also the accuracy and real-time performance meet practical standards. The recognition of core classroom gestures is almost error-free, providing reliable basic data for subsequent engagement evaluation.

Table 1. Ablation experiment of inter-segment self-attention mechanism module

Time Segment Length	Training Set	Test Set
1	88.5	92.5
2	88.3	93.4
3	91.5	94.8
4	87.4	93.2
5	86.3	93.5
6	86.5	93.7

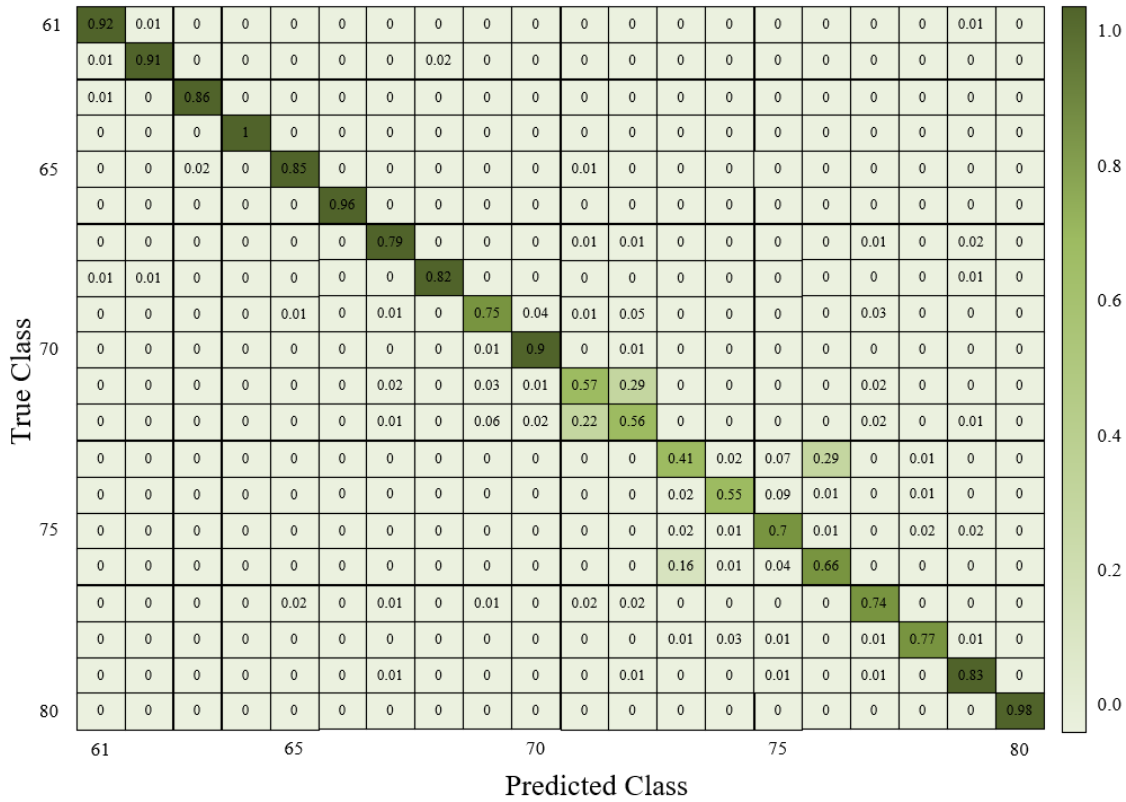

Figure 7. Normalized confusion matrix

From the data in Table 1, when the time segment length is 3, the accuracy of the training set and test set reaches 91.5% and 94.8% respectively, significantly better than other lengths. This shows that the inter-segment self-attention mechanism proposed in the paper effectively improves the ability to capture spatiotemporal features of gesture recognition by optimizing the segment length. In general classroom environments, student gestures exhibit obvious time series characteristics, and time segments of length 3 can more accurately model these dynamic features, reduce information redundancy, and avoid information loss. The experiments verify the high accuracy and real-time performance of this method in handling classroom gestures, ensuring reliable recognition of core engagement gestures such as raising hands and giving likes, and providing high-quality time series feature input for subsequent multidimensional quantitative analysis.

Table 2. Comparison of recognition performance of different fusion methods

Fusion Strategy	Training Set	Test Set
Average Fusion	91.5	94.5
Linear Fusion	91.2	94.8
Dynamic Information Fusion	92.8	94.2

From Table 2, it can be seen that the accuracy of the three fusion strategies on the test set all exceed 94%, and the training set accuracy also remains stable between 91%~93%. Among them, dynamic information fusion performs balanced on both training and test sets, demonstrating adaptability to dynamic gestures in classroom environments, consistent with the paper's design goal of "applicable to general classroom environments." The test accuracy of linear fusion is slightly higher, indicating its advantage in feature integration, while average fusion also shows reliable generalization ability. These data validate the high accuracy and real-time performance of the improved image processing algorithm proposed in the paper for gesture recognition: both static and dynamic gestures can be accurately recognized through multimodal fusion without special equipment, meeting the practical application needs of classroom scenarios. The recognition accuracy of the core classroom gesture "raising hands" is close to 95%, ensuring the quantification accuracy of the "active participation dimension" in engagement evaluation.

Table 3. Ablation experiment of multi-scale temporal convolution feature fusion module

Branch Configuration	Training Set	Test Set
4 Branches	91.5	93.5
6 Branches	91.6	93.4
4 Branches + Fixed Weights	91.7	94.8
4 Branches + Dynamic Weights	91.5	94.2
6 Branches + Dynamic Weights	92.3	94.5

From Table 3, it can be seen that different branch structures and weight strategies significantly affect recognition performance. Among them, the 6-branch + dynamic weight strategy achieves the best performance on both the training and test sets, and the difference between training and test accuracy is smaller than other strategies, reflecting stronger generalization ability. Specifically, increasing the number of branches improves training accuracy, indicating that multi-scale feature extraction is more comprehensive and enhances the representation of complex classroom gestures. The dynamic weight strategy optimizes the adaptability of feature fusion by incorporating teaching contexts. For example, during the lecturing stage, the feature weights of "raising hands" are enhanced, increasing the recognition priority of such gestures, which is highly consistent with the paper's design goal of "combining teaching content and activity

forms." These experimental data validate the effectiveness of the improved image processing algorithm proposed in the paper. Through multi-scale temporal convolution feature fusion and dynamic weight adjustment, high-precision recognition of classroom gestures is achieved without the need for special equipment, meeting the application needs of real-time classroom scenarios.

Table 4. Accuracy comparison of different methods on SCB-dataset and HaGRID datasets

Category	Method	Parameters/*M*	Computation /*GFLOPs*	*SCB-dataset*		*HaGRID*	
				Training Set	Test Set	Training Set	Test Set
GCN	*GAT*	3.2	4.1	82.3	87.5	71.5	72.5
	GIN	3.4	3.8	87.5	94.5	78.5	82.6
	Gated GCN	0.7	2.4	91.5	95.2	84.5	88.6
	APPNP	1.3	1.7	91.5	95.2	87.5	91.5
	Shift-GCN	2.5	2.2	85.6	93.5	77.5	78.5
	Gr-GCN	2.7	5.1	92.3	95.5	85.6	87.5
	R-GCN	1.4	1.6	91.5	95.4	88.6	91.5
	LightGCN	2.8	1.1	92.5	95.6	86.5	88.5
	SGCN	2.4	2.9	91.5	95.5	87.5	91.5
	ADGCN	2.1	2.8	91.6	95.2	88.1	91.5
	LSGCN	1.6	1.5	92.4	96.5	88.5	92.4
	CKGCN	5.5	4.1	92.8	96.6	92.5	91.5
Transformer	*STAR-Transformer*	11.9	248.5	88.5	95.4	81.5	83.5
	TimeSformer	4.2	62.3	92.5	95.4	85.6	88.4
	Video Swin Transformer	1.8	22.6	91.5	95.8	87.5	91.5
MLP	*STACNN*	2.1	9.1	914	96.5	88.6	91.6
	CG-MLP	0.6	0.7	91.5	95.8	88.4	91.8
Proposed Method		2.4	5.1	92.6	96.8	88.9	92.5

From the data in Table 4, the test accuracy of the proposed method on SCB-dataset and HaGRID datasets is 96.8% and 92.5%, respectively, significantly better than similar methods. On the SCB-dataset, the test accuracy of the proposed method is higher than GCN-based methods such as CKGCN, and the computation and parameter size are balanced, reflecting high-precision recognition capability for classroom gestures. On the HaGRID dataset, the 92.5% test accuracy also outperforms most comparison methods, verifying the algorithm's generalization to gestures in different scenarios. Through improvements in image preprocessing, feature extraction, and pattern recognition, the proposed method achieves low device dependency, high accuracy, and real-time performance, providing reliable gesture feature input for engagement evaluation. From Table 5, it is seen that the proposed method achieves a Top-1 accuracy of 94.5% and a MeanTop1 accuracy of 92.5% on the HandPose-v3 dataset, significantly outperforming GCN, CNN, and Transformer-based methods. This indicates that the image processing algorithm proposed in this paper achieves high-precision recognition of classroom gestures by improving image preprocessing, feature extraction, and pattern recognition, without the need for special equipment and meeting real-time requirements.

From the data in Table 6, it can be seen that the proposed method achieves a Top1 accuracy of 87.9% on the ChaLearnLAP-IsoGD dataset using skeleton modality, significantly higher than similar CNN and Transformer methods. This result indicates that by utilizing the skeleton modality combined with improved image preprocessing, feature extraction, and pattern recognition strategies, the proposed algorithm has better recognition ability for common gestures such as raising hands and waving in general classroom environments. For example, the skeleton modality can more stably capture the spatiotemporal dynamics of gestures, and can still maintain high accuracy even in classroom scenes with changing light or complex backgrounds, and does not require special equipment, meeting real-time requirements. This validates the effectiveness of the proposed recognition method in complex educational interaction scenarios and provides a reliable gesture feature basis for subsequent engagement evaluation, ensuring the accuracy of quantitative analysis.

Table 5. Accuracy comparison of different methods on HandPose-v3 dataset

Category	Method	*Top*-1	*Mean Top*1
GCN	*CKGCN*	35.6	24.5
	KE-GCN	93.4	65.4
CNN	*RepViT*	68.5	51.2
	OverLoCK	73.5	62.3
	WeConv	85.2	75.8
	SENet	86.4	78.5
	CBAM	86.4	81.5
	SKNet	87.5	82.6
	Swin-T	86.4	85.4
Transformer	*Switch Transformer*	87.1	85.9
	Gradformer	88.2	91.2
Proposed Method		94.5	92.5

Table 6. Accuracy comparison of different methods on ChaLearn LAP IsoGD dataset

Category	Method	Data Modality	*Top*1
CNN	*WeConv*	*RGB*	83.5
	SENet	*RGB*	87.5
	CBAM	*RGB*	81.6
	SKNet	*RGB*	85.2
Transformer	*Switch Transformer*	*RGB*	84.5
	Gradformer	*RGB*	86.5
Proposed Method		Skeleton	87.9

Based on the high accuracy of gesture recognition, the engagement evaluation model can perform deep fusion analysis of multidimensional gesture features with teaching contexts. The dynamic weight strategy makes the evaluation

results more aligned with actual classroom scenarios. For example, in experimental operation classes, the weight of "operation demonstration gesture" can be raised to 0.5 to highlight practical engagement. Gesture duration can be captured through skeleton modality to achieve fine-grained quantification of students' engagement status. For example, if a student frequently raises their hand in class, consistently gives likes, and seldom shrugs, a calculated engagement score can be obtained by combining the weights of different teaching segments. This result reflects that the student has a high level of active participation, positive emotional investment, and low confusion, providing teachers with personalized teaching suggestions. For students with high confusion levels, targeted tutoring plans can be designed; for students with insufficient active participation, more interactive tasks can be added. The teaching rhythm can be adjusted according to the emotional feedback score to improve overall engagement.

5. CONCLUSION

This paper focused on the educational interactive scenario and constructed a complete system for gesture recognition and student participation evaluation. At the gesture recognition level, by improving image processing algorithms, a Top-1 accuracy of 87.9% to 94.5% was achieved across multiple datasets, adapted to ordinary classroom environments, providing high-precision gesture feature support for participation evaluation. At the participation evaluation level, a multidimensional contextualized model was proposed, combining gesture types, frequency, duration, and dynamic weights of teaching sessions to realize fine-grained quantification of active participation, emotional feedback, and expression of confusion, which provided teachers with real-time and actionable participation analysis. The research value was reflected in: (1) technological innovation: breaking through the application limitations of traditional gesture recognition in educational scenarios, improving algorithm robustness and generalization; (2) educational empowerment: through data-driven participation evaluation, assisting classroom interaction optimization and enhancing student learning experience; (3) methodological contribution: constructing a closed-loop framework of "gesture recognition – participation evaluation – teaching feedback," which provided a model for the implementation of intelligent educational technology.

The current research has the following limitations: (1) scenario adaptability of gesture recognition: there is still room to improve recognition accuracy under extreme lighting or complex backgrounds, requiring optimization of anti-interference algorithms; (2) depth of multimodal fusion: participation evaluation mainly relies on gesture features, and in the future, combining facial expressions, voice tone, and other multimodal data to build a more comprehensive evaluation system was needed; (3) model interpretability: visualization techniques are needed to enhance teachers' understanding of the participation calculation logic and improve acceptance in practical applications. Future research directions include: (1) cross-scenario algorithm optimization: designing adaptive gesture recognition models for online-offline hybrid classrooms, laboratory courses, and other scenarios to enhance environmental robustness; (2) multimodal participation modeling: integrating visual, auditory, and interactive behavior data to construct a "multidimensional participation index system," achieving more comprehensive student behavior analysis; (3) intelligent teaching closed loop: developing automatic teaching strategy generation systems based on participation evaluation results, forming an intelligent teaching closed loop of "evaluation-feedback-optimization." Exploration in these directions could further improve the intelligence level of educational interaction, promote the deep integration of precision teaching and personalized learning, and provide new breakthroughs for the development of future educational technology.

ACKNOWLEDGMENT

This research was funded by the Key Project of Scientific Research in Higher Education Institutions of Anhui Province (Grant No.: 2024AH052932); and the Smart Curriculum Project of Chuzhou University (Grant No.: 2024szkc001); and the Research Project of Chuzhou University (Grant No.: 2023qd62); and the Education Science Planning Project of Jiangsu Province (Grant No.: C/2023/01/126); and the Project of Social Science Foundation of Jiangsu Province (Grant No.: 23GLD002).

REFERENCES

[1] Nührenbörger, M., Steinbring, H. (2009). Forms of mathematical interaction in different social settings: Examples from students', teachers' and teacher–students' communication about mathematics. Journal of Mathematics Teacher Education, 12: 111-132. https://doi.org/10.1007/s10857-009-9100-9

[2] Bai, L., Wang, Y.X. (2022). In-class and out-of-class interactions between international students and their host university teachers. Research in Comparative and International Education, 17(1): 71-88. https://doi.org/10.1177/17454999211038774

[3] Denessen, E., Keller, A., van den Bergh, L., van den Broek, P. (2020). Do teachers treat their students differently? An observational study on teacher-student interactions as a function of teacher expectations and student achievement. Education Research International, 2020(1): 2471956. https://doi.org/10.1155/2020/2471956

[4] Gebresilase, B.M., Zhao, W., Taddese, E.T., Elka, Z.Z., Feng, Y. (2025). The mediating role of academic self-efficacy in the relationship between student teacher interaction and students' university academic achievement. Cogent Psychology, 12(1): 2500181. https://doi.org/10.1080/23311908.2025.2500181

[5] Choi, L.L.S., Brochu, N. (2022). English-as-an-additional-language nursing student support group: Student leadership and engagement. Nursing Education Perspectives, 43(1): 41-43. https://doi.org/10.1097/01.NEP.0000000000000746

[6] Handelsman, M.M., Briggs, W.L., Sullivan, N., Towler, A. (2005). A measure of college student course engagement. The Journal of Educational Research, 98(3): 184-192. https://doi.org/10.3200/JOER.98.3.184-192

[7] Rimm-Kaufman, S.E., Baroody, A.E., Larsen, R.A., Curby, T.W., Abry, T. (2015). To what extent do teacher–student interaction quality and student gender

contribute to fifth graders' engagement in mathematics learning? Journal of Educational Psychology, 107(1): 170-185. https://doi.org/10.1037/a0037252

[8] Baroody, A.E., Rimm-Kaufman, S.E., Larsen, R.A., Curby, T.W. (2016). A multi-method approach for describing the contributions of student engagement on fifth grade students' social competence and achievement in mathematics. Learning and Individual Differences, 48: 54-60. https://doi.org/10.1016/j.lindif.2016.02.012

[9] Karabchuk, T., Roshchina, Y. (2023). Predictors of student engagement: The role of universities' or importance of students' background? European Journal of Higher Education, 13(3): 327-346. https://doi.org/10.1080/21568235.2022.2035240

[10] Bhattacharya, S., Nath, S. (2016). Intelligent e-learning systems: An educational paradigm shift. International Journal of Interactive Multimedia and Artificial Intelligence, 4(2): 83-88. http://doi.org/10.9781/ijimai.2016.4212

[11] Huang, Z.H., Shi, X.D., Chen, Y.D. (2015). Intelligent cloud learning model for online overseas Chinese education. International Journal of Emerging Technologies in Learning, 10(1): 55-59. http://doi.org/10.3991/ijet.v10i1.4284

[12] Kerimbayev, N., Adamova, K., Shadiev, R., Altinay, Z. (2025). Intelligent educational technologies in individual learning: A systematic literature review. Smart Learning Environments, 12(1): 1. https://doi.org/10.1186/s40561-024-00360-3

[13] Singh, N., Gunjan, V.K., Mishra, A.K., Mishra, R.K., Nawaz, N. (2022). SeisTutor: A custom-tailored intelligent tutoring system and sustainable education. Sustainability, 14(7): 4167. https://doi.org/10.3390/su14074167

[14] Wu, H., Wang, Y., Liu, J., Qiu, J., Zhang, X. (2020). User-defined gesture interaction for in-vehicle information systems. Multimedia Tools and Applications, 79: 263-288. https://doi.org/10.1007/s11042-019-08075-1

[15] Schwenderling, L., Kleinau, A., Herbrich, W., Kasireddy, H., Heinrich, F., Hansen, C. (2023). Activation modes for gesture-based interaction with a magic lens in AR anatomy visualisation. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11(4): 1243-1250. https://doi.org/10.1080/21681163.2022.2157749

[16] Al Farid, F., Hashim, N., Abdullah, J., Bhuiyan, M.R., et al. (2022). A structured and methodological review on vision-based hand gesture recognition system. Journal of Imaging, 8(6): 153. https://doi.org/10.3390/jimaging8060153

[17] Kadhim, M.K., Der, C.S., Phing, C.C. (2025). Enhanced dynamic hand gesture recognition for finger disabilities using deep learning and an optimized Otsu threshold method. Engineering Research Express, 7(1): 015228.

[18] López, A.C. (2015). Hand recognition using depth cameras. Tecciencia, 10(19): 73-80. https://doi.org/10.18180/tecciencia.2015.19.10

[19] Nahapetyan, V.E., Khachumov, V.M. (2015). Gesture recognition in the problem of contactless control of an unmanned aerial vehicle. Optoelectronics, Instrumentation and Data Processing, 51: 192-197. https://doi.org/10.3103/S8756699015020132

[20] Heraz, A., Bhyravabhottla, K.K.A., Sajith, N. (2024). Predicting user engagement levels through emotion-based gesture analysis of initial impressions. Electronic Commerce Research. https://doi.org/10.1007/s10660-024-09915-5

[21] Trautman, C.H., Rollins, P.R. (2006). Child-centered behaviors of caregivers with 12-month-old infants: Associations with passive joint engagement and later language. Applied Psycholinguistics, 27(3): 447-463. https://doi.org/10.1017/S0142716406060358