Using SlowFast Network to Analyze the Relationship between Teacher-Guided Action Rhythm and Teaching Efficiency in Classroom Interactive Videos

Qing Dong
College of Cultural Creativity and Tourism, Yuncheng Vocational and Technical University, Yuncheng 044000, Shanxi, China
Corresponding author: Qing Dong (e-mail: 17835366990@163.com)

ABSTRACT To address the difficulty of quantitatively characterizing teachers' nonverbal behaviors and clarifying their relationship with teaching effectiveness in classroom interaction videos, this study employs a SlowFast dual-pathway neural network to analyze the rhythm of teacher-guided actions, providing an effective framework for intelligent video understanding and spatiotemporal signal analysis. Considering the increasing demand for real-time visual information processing and multimodal perception in advanced wireless sensing and communication systems, the proposed approach models teacher rhythm parameters by combining the Slow pathway for subtle facial expressions and gesture analysis with the Fast pathway for large-scale body motion extraction. Furthermore, a classroom behavior coding framework is integrated with student attention, interaction response time, and in-class assessment performance to evaluate teaching efficiency. Experimental results show that the SlowFast model achieves an accuracy of 89.7% in recognizing periodic teacher-guided movements, representing an 11.2% improvement over a single-pathway 3D-CNN model. Correlation analysis further demonstrates that the variance of teacher gesture rhythm is significantly negatively correlated with student attention ($r = -0.83$, $p < 0.01$), while moderate movement frequency exhibits a positive correlation with classroom interaction response speed ($r = 0.76$). The proposed framework confirms a significant relationship between teacher-guided action rhythm and teaching efficiency, providing a quantitative basis for optimizing instructional behavior and offering valuable references for intelligent visual sensing, spatiotemporal signal processing, and multimodal information analysis in advanced engineering applications.

INDEX TERMS SlowFast network, teacher-guided action rhythm, classroom video analysis, spatiotemporal signal processing, intelligent visual sensing

I. INTRODUCTION

WITH the booming development of smart education, refined and intelligent analysis of classroom teaching processes has become a key path to improving teaching quality. Classroom teaching videos serve as a valuable data source for recording the interaction between teachers and students [1,2]. The teacher's non-verbal behavior, especially the rhythm of their guiding movements, is considered an important potential factor affecting teaching efficiency [3,4]. However, traditional manual observation and evaluation methods have difficulty objectively and quantitatively describing dynamic and continuous teacher movements, and are even less able to systematically reveal the stable and intrinsic correlation mechanism between them and teaching efficiency in large-scale video data. This dual dilemma of how to quantify and how to correlate constitutes a core bottleneck that urgently needs to be overcome in the field of educational video analysis.

In the current research context of educational video analysis, scholars have mainly explored along two directions: on the one hand, focusing on the qualitative analysis of classroom interaction and the verification of teaching effectiveness, and on the other hand, striving for breakthroughs in the application of computer vision technology. Sert O. [5] revealed the influence of negative evaluation in teacher-student verbal interaction through discourse timeline analysis, and Noetel M.'s [6] systematic research confirmed the learning promotion role of video in higher education. However, these studies failed to touch upon the dynamic characteristics of teachers' non-verbal behavior; Bakla A. [7] compared the effects of different video resources in flipped classrooms, and Ramly N. [8] further verified the positive impact of educational videos in computer science courses, but their focus still remained on the teaching medium level;

Zhou J.'s [9] student abnormal behavior detection model based on multi-task deep learning, and Sumer O.'s [10] identification of student participation through multimodal technology, although showing the application potential of computer vision technology, lacked in-depth tracing of the key inducement of teacher behavior. These studies show that current educational video analysis relies too much on speech content analysis or static behavior classification, while ignoring the rich temporal information contained in teachers' body language, especially movement rhythm, which is a key factor affecting teaching effectiveness.

At the same time, the field of behavior recognition technology is undergoing profound changes, and various advanced computer vision methods have provided new possibilities for educational video analysis. Deep learning models represented by the SlowFast network have shown significant advantages in capturing spatiotemporal features due to their unique dual-path architecture, which echoes the conclusions of Le N.'s [11] survey on the application of deep reinforcement learning in computer vision. The review studies by Li Chen [12], Liang X [13] and WANG Cailing [14] have established the mainstream status of this technology in video understanding; Bayoudh K.'s [15] systematic investigation of multimodal learning has further expanded the dimension of visual analysis; Zhang Q [16] and LIU K [17] have improved the performance of the model in complex scenarios by improving the information fusion mechanism and temporal modeling capabilities; Matsuzaka Y.'s [18] exploration of Artificial Intelligence (AI)-based computer vision technology and expert systems, and Qi J.'s [19] comprehensive review of gesture recognition, have provided technical references for teacher behavior analysis. Hu Y.'s [20] research on the application of AI-based interactive visual analysis in intelligent education directly reflects the potential of combining computer vision with educational scenarios. However, these technological breakthroughs are still based on the traditional behavior classification paradigm. Whether it is Ding D.'s [21] classroom behavior recognition or ZHAO C.'s [22,23] video behavior recognition, they only focus on identifying what action is and do not delve into what are the rhythmic characteristics of the action.

The above research gaps highlight the necessity of a paradigm shift from behavior type identification to behavior rhythm quantification. Therefore, this paper proposes to innovatively apply the temporal modeling capabilities of the SlowFast network to the extraction of rhythm parameters of teacher-guided actions, aiming to solve the dual problems of the difficulty in quantifying non-verbal behavior and its unclear association mechanism with teaching efficiency. An empirical relationship is established between dynamic rhythm indicators and teaching efficiency to address the gap in quantifying non-verbal behavior rhythm in educational analysis and to advance beyond the prevalent focus on discrete action classification in behavior recognition. Ultimately, it provides a new data-driven paradigm for teaching evaluation and teacher professional development in the smart education environment.

To fill the research gap from behavior recognition to rhythm quantification, the SlowFast network is introduced to quantify the rhythm parameters and establish their correlation with teaching efficiency, thus breaking through the bottleneck of difficult quantification of non-verbal behavior. The organizational structure of the subsequent sections of this paper is as follows: Section "TEACHER-GUIDED ACTION RHYTHM MODELING BASED ON SLOWFAST NETWORK" elaborates on the entire process of modeling the rhythm of teacher-guided movements based on the SlowFast network, including data preprocessing, a dual-path feature extraction framework, and the specific definition and calculation of rhythm parameters. Section "EXPERIMENTAL VERIFICATION AND ANALYSIS" focuses on experimental verification and analysis, detailing the experimental setup, model recognition results and comparison, and research findings on the correlation between movement rhythm characteristics and teaching efficiency indicators. Finally, Section "CONCLUSION" summarizes the conclusions of the entire research.

II. TEACHER-GUIDED ACTION RHYTHM MODELING BASED ON SLOWFAST NETWORK

A. CLASSROOM VIDEO DATA PREPROCESSING AND TEACHER TARGET DETECTION

This paper uses the MM-TBA dataset as an example to obtain classroom video data and performs standardized preprocessing on the original video materials. It should be noted that the teaching behaviors and their perceived rhythms captured in this dataset are situated within a specific cultural and institutional context. The definition of moderate or optimal movement frequency derived from this study may therefore reflect localized pedagogical norms and should be interpreted with caution when generalizing to different cultural settings. Future work would benefit from validating these rhythmic parameters across more demographically and culturally diverse classroom datasets. First, all video files are converted to MP4 format using Fast Forward Mpeg (FFmpeg) [24], and the frame rate is standardized to 30 fps to ensure consistency in timing analysis. The resolution of all video frames is uniformly adjusted to 1920×1080 pixels, retaining sufficient spatial details to analyze the teacher's subtle movements. To address the uneven brightness problem in some videos, histogram equalization technology [25,26] is used for image enhancement. In addition, the data is randomly divided into training, validation, and test sets in an 8:1:1 ratio to ensure a balanced distribution of teaching content and teacher groups across different sets.

The You Only Look Once version 7 (YOLOv7) target detection model [27,28] was selected to locate the teacher's position in real time in a classroom scene containing multiple people and objects. This model was chosen mainly based on its excellent balance between accuracy and speed, as well as its excellent detection performance for targets of

different scales, especially medium and large targets (such as the teacher on the podium). The YOLOv7 model was trained for transfer learning based on the MM-TBA dataset, using a stochastic gradient descent optimizer with an initial learning rate set to 0.01 and a cosine annealing strategy [29] for dynamic adjustment. The batch size was set to 16. Mosaic data augmentation and mixed color adjustment techniques were applied during training to improve the model's generalization ability to different teaching environments. It is noted that these augmentations were applied primarily during the teacher detection stage (YOLOv7 training). For the subsequent SlowFast network input (the cropped teacher region), data augmentation strategies (e.g., color jitter) were carefully calibrated to avoid excessive alterations that could degrade subtle facial features crucial for the Slow path analysis, prioritizing spatial consistency and realistic lighting variations.

In the actual application stage, the model outputs the bounding box coordinates and confidence of the teacher target for each video frame, and then uses linear interpolation [30] to perform temporal smoothing on the detection results between consecutive frames. Based on the detected bounding box coordinates, the system automatically captures the teacher region image from the original video frame and scales it to a standard size of 224 $\times$ 224 pixels as the input for the subsequent behavior analysis of the SlowFast network.

B. DUAL-PATH TIMING FEATURE EXTRACTION BASED ON SLOWFAST

A feature extraction framework based on the SlowFast dual-path neural network [31] is constructed. This framework processes two input streams with different spatiotemporal resolutions in parallel and is optimized specifically for the subtle posture changes and large and rapid movements that coexist in the teacher's movements. The entire process of dual-path input, feature extraction, and fusion in the SlowFast network architecture used in this paper is shown in Figure 1.

Figure 1 shows that in the Slow pathway, the input sequence is extracted from the preprocessed teacher region video at a low temporal sampling rate. Specifically, the frame sampling interval is set to 16 frames, meaning one frame is taken every 16 frames as input, resulting in a high-resolution sequence that is sparse in the temporal dimension but has complete spatial information. These input frames maintain a spatial size of 224 $\times$ 224 pixels. Spatial features are extracted using a two-dimensional convolutional network based on a ResNet-50 backbone. Deep convolutional layers employ dilated convolution techniques to maintain a large receptive field without sacrificing spatial detail. This technique is specifically designed to capture subtle changes in the teacher's gestures, subtle adjustments in facial expression, and gradual changes in body orientation. The Fast pathway uses a completely different sampling strategy from the Slow pathway. It is built on a lightweight ResNet-50 variant and utilizes 3D convolutional kernels that slide in time to directly learn the motion evolution between frames, thereby better representing subtle dynamic features of the action. This pathway uses a frame sampling interval of every two frames, resulting in a temporal resolution eight times that of the Slow pathway. To balance the computational load, the spatial size of the input frame is downsampled to 56 $\times$ 56 pixels, thereby effectively controlling the computational complexity of the model.

A comprehensive understanding of the action is inseparable from the dual-path feature fusion mechanism.

First, the high-resolution feature tensor output by the Fast path is downsampled in the temporal dimension to ensure that it is consistent with the features of the Slow path on the temporal axis. Specifically, the system uses a maximum pooling operation with a kernel size of 5 (time) $\times$ 1 (height) $\times$ 1 (width) to compress the temporal dimension of the Fast path to match the Slow path. Subsequently, the features of the two paths are spliced in the channel dimension through lateral connections to form a unified spatiotemporal feature representation. This fusion strategy not only effectively integrates the rich spatial semantic information in the Slow path with the dense temporal dynamic features in the Fast path, but also achieves deep complementarity between the two types of information through nonlinear transformation. The fused feature tensor is further fed into a temporal aggregation module composed of multiple 3D convolution blocks. This module expands the temporal receptive field through void temporal convolution [32], captures action patterns spanning longer time periods, and thus identifies the periodic rhythm features in the teacher's guidance behavior. Finally, the network outputs a comprehensive spatiotemporal feature map, which serves as a direct input for the subsequent calculation of rhythm parameters such as the teacher's standing displacement speed, gesture amplitude change frequency, and podium area movement frequency.

C. FROM FEATURES TO RHYTHM: QUANTIFICATION OF KEY ACTION RHYTHM PARAMETERS

Based on the SlowFast dual-path network's feature extraction and fusion, this paper defines three core rhythm parameters based on the high-dimensional feature tensor output by the network to accurately quantify the rhythm of the teacher's guidance movements: standing displacement speed, gesture amplitude change frequency, and podium area movement frequency. The parameter calculation is shown in Figure 2.

In Figure 2, the quantification of standing displacement speed is achieved by analyzing the coordinate changes of the center point of the teacher target detection bounding box in consecutive video frames. A first-order low-pass Butterworth filter [33] is used to smooth the velocity sequence to eliminate the noise interference caused by detection jitter. The displacement velocity v_t of the teacher in the t frame is defined as the rate of change of the coordinates of the center point of the bounding box:

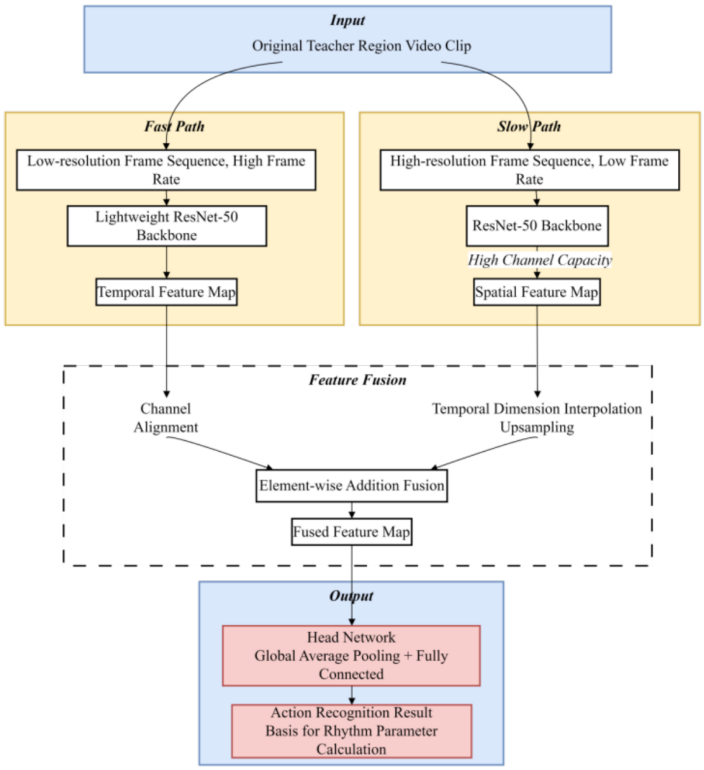

FIGURE 1. SlowFast feature extraction architecture

Note: ResNet-50 = Residual Network 50 layers.

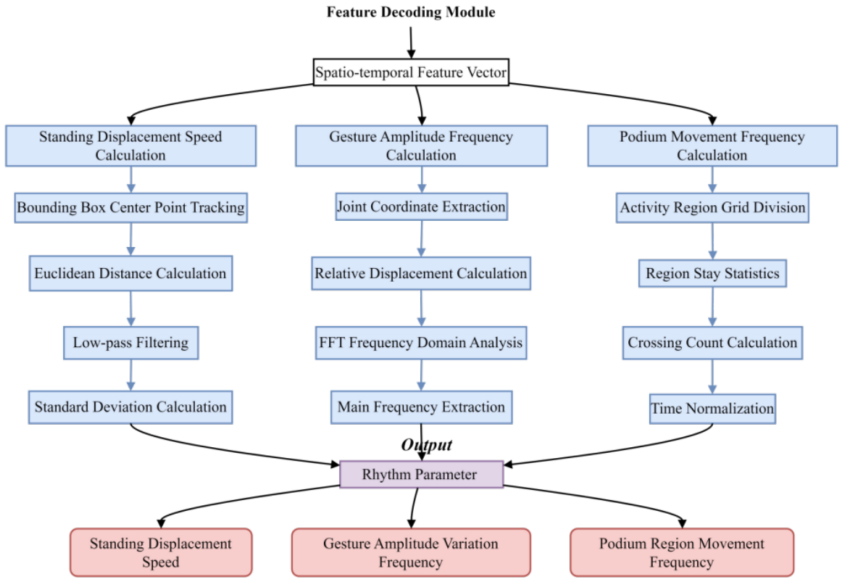

FIGURE 2. Parameter calculation

Note: FFT= fast Fourier transform.

$$v_t = \frac{\sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}}{\Delta t} \quad (1)$$

(x_t, y_t) is the center coordinate of the teacher's bounding box in frame t, $\Delta t = \frac{1}{30}$ (frame rate is 30 fps). The final rhythm parameter is the standard deviation of the speed sequence:

$$\text{Standing Displacement Variance} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(v_t - \bar{v})^2} \quad (2)$$

Extracting the frequency of gesture amplitude changes relies on a detailed analysis of the teacher's upper limb movement patterns. First, based on the rich spatial semantic information contained in the Slow path feature map, a pre-trained human pose estimation model (OpenPose [34,35]) is used to extract the coordinates of key joints such as the teacher's wrist and elbow. The gesture trajectory is then constructed by calculating the relative displacement vector of the wrist joint relative to the torso center between adjacent frames. Let the displacement vector of the wrist joint between adjacent frames be:

$$\vec{d_t} = \vec{p_t} - \vec{p_{t-1}} \quad (3)$$

Its modulus length is:

$$d_t = \|\vec{d_t}\| \quad (4)$$

The FFT [36] is performed on d_t and extract its main frequency component as a quantitative indicator of the frequency of gesture amplitude change. The formula is as follows:

$$f_{\text{gesture}} = \arg\max_f \left| \sum_{t=0}^{N-1} d_t \cdot e^{-i2\pi f t/N} \right| \quad (5)$$

This parameter can effectively reflect the speed of the teacher's gesture rhythm. High-frequency changes usually correspond to urgent and scattered gestures, while low-frequency changes correspond to slow and coherent gesture patterns. Based on the teacher target detection results, the podium plane is divided into 3

$\times$ 3 grid areas. The time the center point of the teacher's bounding box stays in each grid is counted. The number of area switches is:

$$C = \sum_{k=1}^{K-1} \mathbb{I}(\text{region}(k) \neq \text{region}(k+1)) \quad (6)$$

The moving frequency is:

$$f_{\text{movement}} = \frac{C}{T_{\text{total}}} \quad (7)$$

T_{total} refers to the total duration of the class (minutes).

The 3 $\times$ 3 grid partitioning was adopted based on common podium dimensions observed in the dataset, providing a balanced spatial resolution that captures meaningful positional changes without excessive granularity. This approach accommodates typical teacher movement ranges while maintaining computational simplicity.

By analyzing the temporal patterns of the teacher's switching between different grids, the number of area crossings per unit time is calculated as a quantitative value of movement frequency. This parameter reflects the teacher's activeness within the podium. High-frequency movement generally indicates that the teacher tends to switch between different positions frequently to focus on students in different areas.

III. EXPERIMENTAL VERIFICATION AND ANALYSIS

A. EXPERIMENTAL SETUP AND MODEL IDENTIFICATION PERFORMANCE VERIFICATION

The training hyperparameter settings are shown in Table 1.

Table 1 shows the training settings in detail, where the model uses pre-trained weights and combines various data augmentations to prevent overfitting. The loss function is cross-entropy with label smoothing to alleviate class imbalance and improve model calibration.

TABLE 1. Model training hyperparameter settings

Hyperparameter	Setting value	Description
Optimizer	AdamW	$\beta_1 = 0.9$, $\beta_2 = 0.999$
Initial learning rate	0.001	Cosine annealing schedule
Weight decay	0.05	L2 regularization coefficient
Batch size	32	4 $\times$ 8 GPU parallelism
Training epochs	100	Includes 5 epochs for warming up
Data augmentation	Random flip, multi-scale cropping, color jitter	Prevents overfitting
Loss function	Label smoothing cross-entropy	$\epsilon = 0.1$
Gradient clipping	Norm threshold 1.0	Stabilizes the training process
Frame sampling strategy	Slow path $\alpha = 4$, Fast path full frame rate	Maintains temporal consistency

Note: GPU = Graphics Processing Unit.

The test set contains 360 hours of classroom video data, covering 2,880 independent instructor-guided action clips, each lasting 3-5 seconds. Three representative baseline models (M2, an interactive three-dimensional (I3D) network based on 3D convolutions; M3, a traditional 3D-CNN architecture; and M4, a single-path temporal segmentation network) were selected for comparative analysis to objectively evaluate the state-of-the-art approach M1. All comparison models were trained on the same training set and evaluated on a unified test set to ensure fair comparison. M2 was fine-tuned using weights pre-trained on the Kinetics dataset. Its two-stream architecture processes red, green, and blue (RGB) images and optical flow information, respectively. M3 extracts spatiotemporal features through successive 3D convolutional layers. M4 is based on a ResNet-50 backbone network with an additional temporal convolution module for action classification. The performance of the comparison models under the same experimental environment is shown in Table 2. It is acknowledged that more recent transformer-based architectures (e.g., Video Swin Transformer) represent the state-of-the-art in video understanding. The choice of baseline models in this study focuses on established and

representative convolutional approaches for action recognition. Future work will include comparisons with these newer architectures to further validate and potentially improve the proposed methodology.

TABLE 2. Recognition performance of different models on various teacher-guided actions

Model type	Precision (%)	Recall (%)	Inference speed (fps)
M1	88.9	87.5	24.3
M2	81.6	80.9	18.7
M3	76.8	75.3	22.1
M4	79.2	78.6	26.5

In Table 2, the SlowFast method significantly outperforms all baseline models in two core metrics (excluding inference speed). This performance advantage is primarily attributed to the SlowFast network's dual-path design, which simultaneously captures the subtle spatial features and rapid temporal changes of the teacher's movements. Single-path baseline models often struggle to account for both. In terms of inference speed, the SlowFast model achieves a processing efficiency of 24.3 fps. While slightly lower than lightweight single-path temporal segmentation networks, it fully meets the requirements of real-time analysis and strikes a good balance between accuracy and speed. Figure 3 shows the performance differences among the models.

Figure 3 clearly shows the relative performance of each model in terms of accuracy and F1 score, and the dominance of the SlowFast method is immediately apparent. Figures 3(a) and 3(b) show that compared to the next-best I3D model, the accuracy improves by 6.5 percentage points and the F1 score improves by 7.0 percentage points.

This paper selected several typical scenarios (teacher's gesture explanation, podium area movement, and standing position change) from the test set for visualization, and conducted an in-depth analysis of the model's recognition effect in an actual classroom environment, as shown in Figure 4.

Figure 4 shows green boxes representing correct recognition, solid red boxes representing misidentifications, and dashed red boxes representing missed detections. In Figures 4(a) through 4(f), the horizontal and vertical axes correspond to the spatial pixel coordinates of the video frames. The grayscale background of the images represents the original video frame, and the colored rectangles mark the target teacher, student, and gesture regions detected by the model. The model achieved an overall correct recognition rate of 89.7% and an average confidence level of 88.0% for Figures 4(a) through 4(c). In Figure 4(d), the model misidentified gesture explanation as standing due to occlusion; in Figure 4(e), uneven brightness caused the model to misidentify lectern movement as gesture; and in Figure 4(f), subtle hand movements were not captured, resulting in the system failing to recognize the action. This overall performance demonstrates the SlowFast model's strengths in spatiotemporal feature fusion and its potential for improvement in complex scene perception.

B. EXPLORING THE CORRELATION BETWEEN MOVEMENT RHYTHM AND TEACHING EFFICIENCY

This paper uses the Pearson correlation coefficient as the primary statistical indicator, combined with significance tests to assess the correlation and statistical reliability between teacher behavior rhythm and teaching efficiency. Specifically, the Pearson correlation coefficient matrix is calculated for all variables, and the corresponding p-values are obtained using a two-tailed t-test to determine the significance. To control the risk of false positives caused by multiple comparisons, the q value after false discovery rate correction is further calculated to improve the robustness of the statistical conclusions. Before the formal analysis, all variables were tested for normality using the Shapiro–Wilk test and for homogeneity of variance using Levene's test to ensure that the assumptions for Pearson correlation analysis were satisfied.

Figure 5 shows the relationship between the variance of teacher gesture rhythm and student concentration.

Figure 5 represents the observation value of one class period. There is a significant negative correlation between the variance of the teacher's gesture rhythm and the students' concentration ($r = -0.83$, $p < 0.001$), indicating that the more unstable the teacher's gesture changes and the greater the rhythm variances, the lower the students' classroom concentration. The red fitting line clearly shows that as the variance of the gesture rhythm increases, the students' concentration tends to decrease. Figure 6 shows the relationship between teacher movement frequency and classroom interaction response speed.

Figure 6 shows a moderate positive correlation between the two ($r = 0.76$, $p < 0.001$), indicating that a moderate teacher movement frequency shortens response time in teacher-student interactions and improve classroom efficiency. The blue fitted line indicates that as movement frequency increases within a certain range (0.5-0.8 times/min), the response speed of interaction increases accordingly, but the improvement levels off beyond this range.

The complex relationship between multiple variables is shown in Figure 7, which also presents the joint distribution pattern of gesture rhythm variance, movement frequency and student concentration.

Figure 7 uses a three-dimensional surface model to reveal the complex coupling relationship between teacher gesture rhythm variance, movement frequency, and the overall teaching efficiency score (Z-axis; units: points).

The data shows that when gesture rhythm variance is controlled within the range of 0.15-0.25 and movement frequency is maintained between 0.5-0.8 times/min, teaching efficiency reaches a peak of 95 points. When gesture variance exceeds 0.4, teaching efficiency drops sharply to around 70 points. Movement frequencies below 0.3 or above 0.9 times/min also result in a significant decrease in effi-

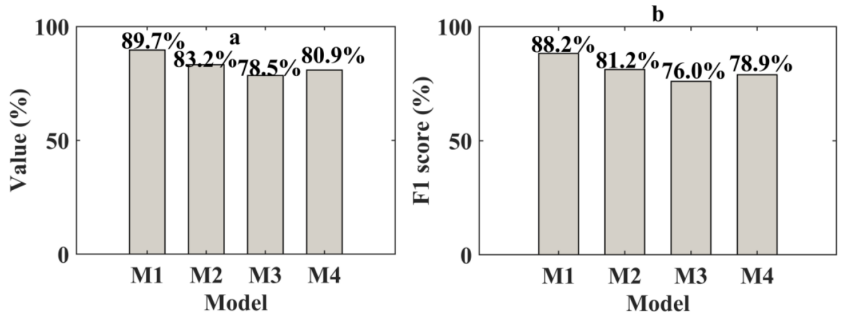

FIGURE 3. Performance comparison of teacher-guided action recognition models. Figure 3(a). Comparison of accuracy of each model. Figure 3(b). Comparison of F1 scores of various models

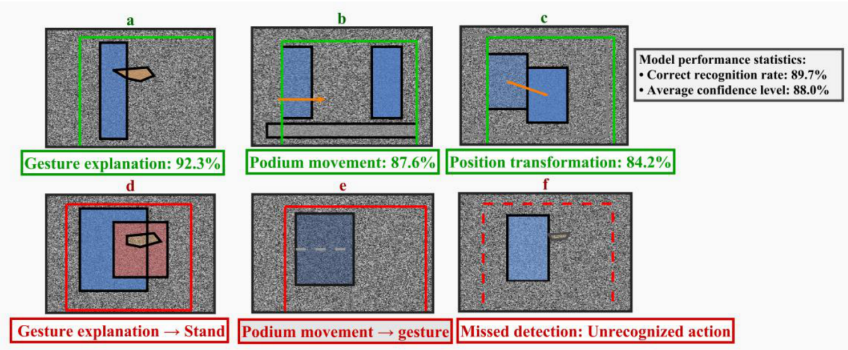

FIGURE 4. Visual analysis of typical classroom scene recognition using the SlowFast model. Figure 4(a). Correct recognition example 1: Gesture explanation; Figure 4(b). Correct recognition example 2: Podium movement; Figure 4(c). Correct recognition example 3: Position change; Figure 4(d). Recognition error example 1: Student occlusion; Figure 4(e). Recognition error example 2: insufficient light; Figure 4(f). Recognition error example 3: Action missed detection

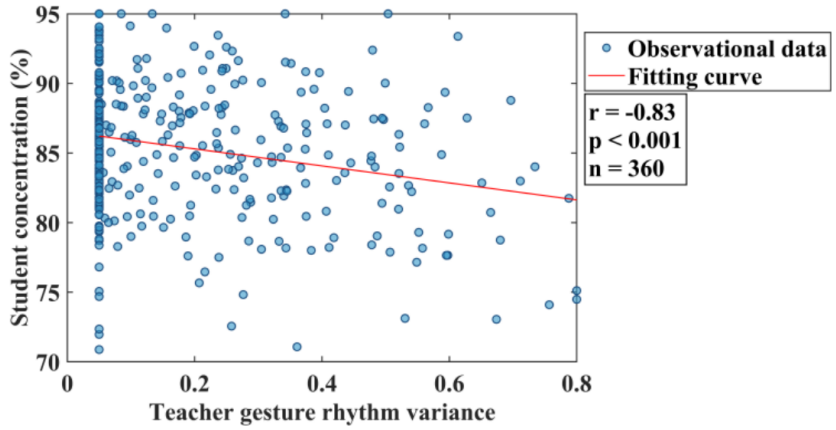

FIGURE 5. Relationship between teacher gesture rhythm variance and student concentration

ciency. The dense contour areas of the surface indicate that movement frequency within the range of 0.6-0.8 times/min has a stable effect on teaching efficiency, while increasing gesture variance is consistently associated with a loss of efficiency. These patterns of variation confirm that there is a clear optimal parameter range for teacher nonverbal behavior.

Stable and controllable gestures combined with a moderately active movement rhythm can maximize classroom teaching effectiveness, providing precise quantitative guidance for optimizing teacher behavior.

The results of the correlation analysis between teachers' movement rhythm parameters and various indicators of teaching efficiency are shown in Table 3.

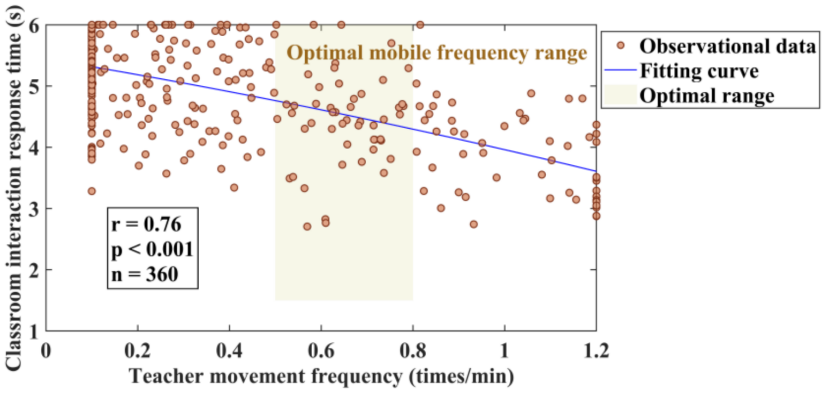

FIGURE 6. Relationship between teacher movement frequency and classroom interaction response speed

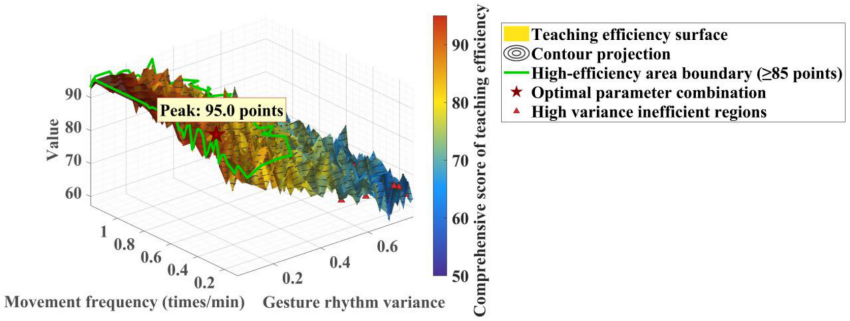

FIGURE 7. Three-dimensional relationship analysis between teacher rhythm parameter combinations and teaching efficiency

TABLE 3. Correlation analysis results between teacher movement rhythm parameters and various indicators of teaching efficiency

Rhythm parameter	Student attention	Interactive response time	In-class assessment scores
Variance of standing displacement speed	r = -0.42, p = 0.003*	r = 0.38, p = 0.008*	r = -0.31, p = 0.025*
Frequency of gesture amplitude variation	r = -0.83, p < 0.001**	r = 0.29, p = 0.037*	r = -0.67, p < 0.001**
Frequency of movement in podium area	r = 0.52, p < 0.001**	r = -0.76, p < 0.001**	r = 0.45, p < 0.001**

Note: * indicates $p < 0.05$, ** indicates $p < 0.001$.

Table 3 shows that the variance of standing movement speed showed a moderate negative correlation with student concentration ($r = -0.42$, $p = 0.003$), indicating that more unstable movement speed is associated with lower student concentration. This variance also showed a positive correlation with interactive response time ($r = 0.38$, $p = 0.008$), indicating that unstable movement prolongs student response time. It also showed a negative correlation with classroom assessment scores ($r = -0.31$, $p = 0.025$*), suggesting a negative impact on learning outcomes. The frequency of gesture amplitude variance had the strongest correlation with student concentration, showing a strong negative correlation ($r = -0.83$, $p < 0.001$), and a significant negative correlation with classroom grades ($r = -0.67$, $p < 0.001$), indicating that frequent gestures with large amplitude variance are often associated with decreased concentration and poorer grades. It also showed a small positive correlation with interaction response time ($r = 0.29$, $p = 0.037$), suggesting that chaotic gestures can also slightly increase interaction delays. Conversely, the frequency of podium movement showed a significant positive correlation with student focus ($r = 0.52$, $p < 0.001$) and classroom scores ($r = 0.45$, $p < 0.001$), and a significant negative correlation with interactive response time ($r = -0.76$, $p < 0.001$), indicating that moderate regional movement can help attract attention, shorten student response times, and improve classroom performance. In summary, stable and purposeful spatial movement is beneficial to teaching effectiveness, while frequent or fluctuating gestures and irregular movements may impair student attention and learning outcomes.

IV. CONCLUSION

A SlowFast dual-path neural network is used to construct a quantification model for the rhythm of teacher-guided movements, effectively addressing the difficulty in accurately measuring the relationship between nonverbal behavior and teaching efficiency in classroom interaction videos. Experimental results demonstrate that the model achieves 89.7% accuracy in identifying periodic teacher-guided movements, validating its superior performance in complex classroom scenarios. Further correlation analysis reveals that the variance of teacher gesture rhythm is significantly negatively

correlated with student focus (r = -0.83, p < 0.01), indicating that more erratic gesture rhythm is associated with greater student distraction. On the other hand, moderate movement frequency is positively correlated with classroom interaction response speed (r = 0.76), indicating an association between appropriate physical activity and enhanced classroom participation and communication. It should be noted that this correlation does not imply causation; further controlled experiments are needed to establish a causal link. These results reveal the intrinsic coupling between teacher movement rhythm characteristics and teaching efficiency, providing data-driven support for the quantitative assessment of teacher behavior and the optimization of teaching strategies in smart education environments.

AUTHOR CONTRIBUTIONS

Qing Dong designed, collected and analyzed the data, and drafted the manuscript. Qing Dong conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Qing Dong participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] K. C. Li and B. T. M. Wong, "Research landscape of smart education: a bibliometric analysis," Interactive Technology and Smart Education, vol. 19, no. 1, pp. 3-19, 2022, doi: 10.1108/ITSE-05-2021-0083.

[2] J. C. K. Tham and G. Verhulsdonck, "Smart education in smart cities: Layered implications for networked and ubiquitous learning," IEEE Transactions on Technology and Society, vol. 4, no. 1, pp. 87-95, 2023, doi: 10.1109/TTS.2023.3239586.

[3] Z. Dai, C. Sun, L. Zhao, and X. Zhu, "The effect of smart classrooms on project-based learning: A study based on video interaction analysis," Journal of Science Education and Technology, vol. 32, no. 6, pp. 858-871, 2023, doi: 10.1007/s10956-023-10056-x.

[4] Z. Zhan, Q. Wu, Z. Lin, and J. Cai, "Smart classroom environments affect teacher-student interaction: Evidence from a behavioural sequence analysis," Australasian Journal of Educational Technology, vol. 37, no. 2, pp. 96-109, 2021, doi: 10.14742/ajet.6523.

[5] O. Sert, A. Gynne, and M. Larsson, "Developing student-teachers' interactional competence through video-enhanced reflection: A discursive timeline analysis of negative evaluation in classroom interaction," Classroom Discourse, vol. 16, no. 2, pp. 142-171, 2025, doi: 10.1080/19463014.2024.2337184.

[6] M. Noetel, S. Griffith, O. Delaney, T. Sanders, P. Parker, B. D. P. Cruz, et al., "Video improves learning in higher education: A systematic review," Review of Educational Research, vol. 91, no. 2, pp. 204-236, 2021, doi: 10.3102/0034654321990713.

[7] A. Bakla and E. A. Mehdiyev, "A qualitative study of teacher-created interactive videos versus YouTube videos in flipped learning," E-Learning and Digital Media, vol. 19, no. 5, pp. 495-514, 2022, doi: 10.1177/20427530221107789.

[8] N. Ramly, A. N. Rosli, S. Suhaimi, M. H. A. Wahab, and A. H. Ariffin, "The effects of using educational videos in online learning: a case study for basic computer science subject," International Journal of Recent Technology and Applied Science (IJORTAS), vol. 5, no. 1, pp. 12-23, 2023, doi: 10.36079/lamintang.ijortas-0501.479.

[9] J. Zhou and N. Herencsar, "Abnormal behavior determination model of multimedia classroom students based on multi-task deep learning," Mobile Networks and Applications, vol. 28, no. 3, pp. 900-913, 2023, doi: 10.1007/s11036-023-02187-7.

[10] O. Sumer, P. Goldberg, S. D'Mello, P. Gerjets, U. Trautwein, and E. Kasneci, "Multimodal engagement analysis from facial videos in the classroom," IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1012-1027, 2021, doi: 10.1109/TAFFC.2021.3127692.

[11] N. Le, V. S. Rathour, K. Yamazaki, K. Luu, and M. Savvides, "Deep reinforcement learning in computer vision: a comprehensive survey," Artificial Intelligence Review, vol. 55, no. 4, pp. 2733-2819, 2022, doi: 10.1007/s10462-021-10061-9.

[12] C. Li, M. He, Y. Wang, L. Luo, and W. Han, "Review of video action recognition technology based on deep learning," Application Research of Computers, vol. 39, no. 9, pp. 2561-2569, 2022, doi: 10.19734/j.issn.1001-3695.2022.03.0077.

[13] X. Liang, W. Li, and H. Zhang, "Review of research on human action recognition methods," Application Research of Computers, vol. 39, no. 3, pp. 651-660, 2022.

[14] C. Wang, J. Yan, and Z. Zhang, "Review on Human Action Recognition Methods Based on Multimodal Data," Computer Engineering and Applications, vol. 60, no. 9, pp. 1-18, 2024.

[15] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," The Visual Computer, vol. 38, no. 8, pp. 2939-2970, 2022, doi: 10.1007/S00371-021-02166-7.

[16] Q. Zhang and H. Sang, "SlowFast information fusion action recognition network based on deeply nested attention mechanism," Journal of Electronic Measurement and Instrumentation, vol. 38, no. 3, pp. 159-166, 2024.

[17] K. Liu, W. Wang, H. Shen, H. Hou, M. Guo, and Z. Luo, "Behavior recognition based on time-dependent attention," Chinese Journal of Liquid Crystals and Displays, vol. 38, no. 8, pp. 1095-1106, 2023.

[18] Y. Matsuzaka and R. Yashiro, "AI-based computer vision techniques and expert systems," Ai, vol. 4, no. 1, pp. 289-302, 2023, doi: 10.3390/ai4010013.

[19] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: a review," Complex & Intelligent Systems, vol. 10, no. 1, pp. 1581-1606, 2024, doi: 10.1007/s40747-023-01173-6.

[20] Y. Hu, Q. Q. Li, and S. Hsu, "Interactive visual computer vision analysis based on artificial intelligence technology in intelligent education," Neural Computing and Applications, vol. 34, no. 12, pp. 9315-9333, 2022, doi: 10.1007/S00521-021-06285-Z.

[21] D. Ding, Y. Zhao, J. Zhang, J. Liu, J. Liu, H. Yang, et al., "A New Method of Classroom Behavior Recognition Based on WS-FC SLOWFAST," International Journal of Gaming and Computer-Mediated Simulations (IJGCMS), vol. 17, no. 1, pp. 1-28, 2025, doi: 10.4018/IJGCMS.371423.

[22] C. Zhao, X. Feng, Y. Dong, X. Wen, and R. Cao, "Video Action Recognition Based on Two-stream Feature Enhancement Network," Journal of Taiyuan University of Technology, vol. 56, no. 3, pp. 495-505, 2025, doi: 10.16355/j.tyut.1007-9432.20230692.

[23] C. Zhao, X. Feng, and R. Cao, "Video action recognition based on spatiotemporal two-stream feature enhancement network," Computer Engineering and Design, vol. 46, no. 3, pp. 871-878, 2025, doi: 10.16208/j.issn1000-7024.2025.03.031.

[24] J. Ozer, "How to Script for Ffmpeg," Streaming Media, vol. 20, no. 2, pp. 92-99, 2023.

[25] D. Vijayalakshmi and M. K. Nath, "A novel contrast enhancement technique using gradient-based joint histogram equalization," Circuits, Systems, and Signal Processing, vol. 40, no. 8, pp. 3929-3967, 2021, doi: 10.1007/s00034-021-01655-3.

[26] S. Agrawal, R. Panda, P. K. Mishro, and A. Abraham, "A novel joint histogram equalization based image contrast enhancement," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 4, pp. 1172-1182, 2022, doi: 10.1016/j.jksuci.2019.05.010.

[27] O. E. Olorunshola, M. E. Irhebhude, and A. E. Evwiekpaefe, "A comparative study of YOLOv5 and YOLOv7 object detection algorithms," Journal of Computing and Social Informatics, vol. 2, no. 1, pp. 1-12, 2023.
[28] J. Chen, R. Wen, and L. Ma, "Small object detection model for UAV aerial image based on YOLOv7," Signal, Image and Video Processing, vol. 18, no. 3, pp. 2695-2707, 2024, doi: 10.1007/s11760-023-02941-0.
[29] C. Zhang, Y. Shao, H. Sun, X. Lei, Z. Qian, and L. Zhang, "The WuC-Adam algorithm based on joint improvement of Warmup and cosine annealing algorithms," Mathematical Biosciences and Engineering: MBE, vol. 21, no. 1, pp. 1270-1285, 2022, doi: 10.3934/mbe.2024054.
[30] M. V. Nevskii, "On the Minimal Norm of a Projection Operator for Linear Interpolation on an-Dimensional Ball," Mathematical Notes, vol. 114, no. 3, pp. 415-418, 2023, doi: 10.1134/S0001434623090146.
[31] C. Shen and J. Xu, "Convolutional neural network with deep and shallow paths," Computer Applications and Software, vol. 41, no. 5, pp. 298-303, 2024, doi: 10.3969/j.issn.1000-386x.2024.05.043.
[32] Z. Lai, R. Chen, J. Jia, and Y. Qian, "Real-time micro-expression recognition based on ResNet and atrous convolutions," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 11, pp. 15215-15226, 2023, doi: 10.1007/s12652-020-01779-5.
[33] R. Daryani, B. Aggarwal, and M. Gupta, "Design of fractional-order Chebyshev low-pass filter for optimized magnitude response using meta-heuristic evolutionary algorithms," Circuits, Systems, and Signal Processing, vol. 42, no. 5, pp. 2507-2537, 2023, doi: 10.1007/S00034-022-02227-9.
[34] E. D'Antonio, J. Taborri, I. Mileti, S. Rossi, and F. Patané, "Validation of a 3D markerless system for gait analysis based on OpenPose and two RGB webcams," IEEE Sensors Journal, vol. 21, no. 15, pp. 17064-17075, 2021, doi: 10.1109/JSEN.2021.3081188.
[35] M. Xu, L. Guo, and H. C. Wu, "Robust abnormal human-posture recognition using OpenPose and multiview cross-information," IEEE Sensors Journal, vol. 23, no. 11, pp. 12370-12379, 2023, doi: 10.1109/JSEN.2023.3267300.
[36] W. A. Mahmoud, "Computation of wavelet and multiwavelet transforms using fast fourier transform," Journal Port Science Research, vol. 4, no. 2, pp. 111-117, 2021, doi: 10.36371/port.2020.2.7.