Design of University Instrumental Music Performance Movement Recognition and Accurate Feedback System Integrating OpenPose and LSTM

Xuan Zhou[1], Wei He[2]
[1]Tourism and Performing Arts Academy, Guilin Tourism University, Guilin 541006, Guangxi, China
[2]School of Education and Music, Guilin University, Guilin 541006, Guangxi, China
Corresponding author: Wei He (e-mail: 15078368039@163.com)

ABSTRACT Accurate recognition of fine-grained human movements with low-latency feedback is fundamental to intelligent perception systems and real-time human–machine interaction in modern engineering applications. This study presents an end-to-end motion recognition and feedback framework for university instrumental music performance by integrating OpenPose-based pose estimation with a three-layer bidirectional long short-term memory (BiLSTM) network. The proposed approach extracts normalized skeletal keypoint sequences from performance videos and exploits bidirectional temporal modeling to capture both local motion evolution and long-range spatiotemporal dependencies. An adaptive feedback module further quantifies posture deviations through coordinate distance analysis and automatically generates structured correction instructions based on predefined performance standards. Experimental results demonstrate that the three-layer BiLSTM architecture achieves a motion recognition accuracy of 96.3%, while the optimized system reduces end-to-end latency to 48.3 ms and maintains 98.2% overall accuracy under standard teaching conditions. The framework also exhibits robust performance across challenging environments, including low illumination and complex backgrounds. By combining vision-based sensing with temporal sequence learning, the proposed method provides an effective solution for objective motion evaluation and intelligent feedback generation, while offering valuable methodological references for real-time signal interpretation, human-centered electromagnetic sensing, and next-generation intelligent monitoring systems requiring precise dynamic feature extraction.

INDEX TERMS OpenPose, bidirectional long short-term memory, motion recognition, intelligent feedback, electromagnetic sensing

I. INTRODUCTION

INSTRUCTION in music performance at the collegiate and university levels provides the basis for a music-related profession. Systematic training of instrumental skills is directly related to the formation and development of students' musical expression [1,2]. Music pedagogy theory states that learning performance skills involves a combination of muscular memory and neural coordination processes. Typically, teachers observe and listen to assess progress, which is subjective and limits the ability to accurately notice subtle differences in performance movements, including the assessment of high-frequency, small-amplitude motions [3-5]. Research on the use of human pose estimation technology in music performance analysis is growing, but current methods cannot meet the needs for higher-precision identification of instrumental motions. OpenPose and other state-of-the-art pose estimation architectures provide good performance in human motion recognition tasks [6-8], but are limited in their ability to represent the continuous temporal progression of musical performance movements. With respect to time series analysis, traditional models for temporal performance evaluation often do not adequately model the continuous dynamic nature of musical performance movements; therefore, accurately assessing rhythmic coherence and general regularity is difficult [9,10]. Instrumental performance movements include complicated displacement in space and organization of time, and singular frame of reference analysis approaches cannot represent the full features of performance technique [11-13].

Scholars have investigated instrumental performance movements extensively. Traditional computer vision-based methods can extract 'static' posture features but cannot represent the temporal organization of movements. Unidirectional LSTM (Long Short-Term Memory) can represent temporal temporal dependencies in a single direction while ignoring historical contextual information on the current motion, which marginally degrades the efficiency of recognizing continuous performance movements [14-16]. Multi-sensor fusion approaches can provide bigger data dimension, but are difficult to implement in a typical university teaching context due to high hardware implementation costs and complexity of data synchronization [17,18]. In deep learning applications, the potential of BiLSTM offers a distinct advantage for motion-recognition tasks; however, there is still complexity in the theoretical and applied understanding of optimizing BiLSTM in an informational way for performance theory [19,20].

This paper proposes an end-to-end architecture that integrates spatial action recognition from OpenPose with BiLSTM temporal modeling. By driving a three-layer BiLSTM network with a normalized string of coordinates, This paper provides a collaborative analysis of the spatiotemporal characteristics of performance actions. The system integrates spatial and temporal features to objectively assess performance movements in university music education. This fusion architecture not only overcomes the limitations of existing techniques in separating space and time, but also provides a scientific and accurate movement assessment mechanism for instrumental music education.

II. SYSTEMATIC APPROACH IMPLEMENTATION

KEY POINT DETECTION AND DATA STANDARDIZATION BASED ON OPENPOSE

The system uses the OpenPose framework to perform real-time human pose estimation on video streams of university instrumental performances. A CNN (Convolutional Neural Network) architecture is used to capture the feature representations of the performer's hands and joints [21-23]. Input video frames undergo multistage convolution processing to generate keypoint confidence maps and partial affinity fields. A nonmaximum suppression algorithm is used to locate keypoints within the confidence map and extract a two-dimensional coordinate sequence of the performer's hands and upper limb joints [24,25].

To address the detection noise present in high-frequency, small-amplitude performance scenes, an adaptive sliding average filtering algorithm is applied to perform time-domain smoothing on the coordinate sequence. This algorithm dynamically adjusts the window size according to the action frequency, and the window width w is calculated using the following formula:

$$w = \mathrm{round}\left(\frac{1}{2 \times f \times \Delta t}\right) \tag{1}$$

f is the fundamental frequency of the playing action, obtained through a short-time Fourier transform, Δt is the time interval between video frames. While maintaining the temporal dynamic features of the playing action, the sliding average technique successfully reduces coordinate jitter brought on by quick finger movements.

To remove the impact of individual spatial scale differences on ensuing time series analysis, the coordinate sequence is normalized using a standardization technique based on the performers' body proportions [26,27]. The original coordinate (x_i, y_i) is converted into relative coordinate (x_i', y_i'), and the calculation formula is:

$$x_i' = \frac{x_i - x_{\mathrm{ref}}}{d_{\mathrm{ref}}}, \qquad y_i' = \frac{y_i - y_{\mathrm{ref}}}{d_{\mathrm{ref}}} \tag{2}$$

$(x_{\mathrm{ref}}, y_{\mathrm{ref}})$ represents the coordinates of the reference keypoints, with the wrist joint used as the reference point; d_{ref} represents the reference distance calculated from the spatial distance between shoulder keypoints detected in the initial video frames which is continuously updated through a weighted moving average algorithm to adapt to performer posture changes while maintaining normalization stability. To ensure that the movement characteristics of performers with different body types can be compared, the normalized coordinate sequence is distributed within the interval [-1, 1]. The coordinate range threshold is set to [-1.2, 1.2] by the system. Outlier coordinate points are those that fall outside the threshold and are fixed using linear interpolation.

BILSTM NETWORK BASED ON BIDIRECTIONAL SPATIOTEMPORAL FEATURE FUSION

The forward LSTM layer captures the temporal characteristics of performance movements in a dynamic fashion through the collaborative action of the forget gate, input gate, and output gate. The forget gate determines which portion of the output from the prior state to remember. The input gate determines the amount of new information to input into the model. The output gate is responsible for modulating how much of the current state output to relate to the next moment. The backward LSTM layer processes the sequence in reverse, detecting the impact of historical context on the current movement while allowing the network to consider features at past and future time points simultaneously. The computation of the forget gate describes the state update mechanism and is formally represented by:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{3}$$

f_t represents the forget gate activation value at the moment, σ is the sigmoid activation function, W_f is the forget gate weight matrix, h_{t-1} is the hidden state at the previous moment, x_t is the input vector at the current moment, and b_f is the forget gate bias. This gating mechanism enables the network to dynamically adjust the degree of memory of historical information based on the characteristics of the performance action. The three-layer BiLSTM structure processes the normalized coordinate sequences with standard LSTM gating mechanisms, while ReLU (Rectified Linear Unit) activation is applied in the fully connected layers following the BiLSTM network to enhance nonlinear expressive power for action recognition. The network handles variable-length sequences through time-step unfolding [28,29] to ensure adaptability to performance action segments of different lengths. The network is trained using the Adam (Adaptive Moment Estimation) optimizer with an initial learning rate of 0.001 and a batch size of 64. The network parameters are updated through backpropagation to minimize the cross-entropy loss function. Through systematic parameter exploration experiments the three-layer architecture with 256 memory cells per layer was determined to achieve optimal spatiotemporal modeling performance while maintaining real-time processing capability demonstrating the best balance between recognition accuracy and computational efficiency.

As shown in Figure 1, the keypoint coordinate sequence output by OpenPose is normalized and then fed into the forward and backward layers of a three-layer bidirectional LSTM network. Each layer contains 256 memory cells and has a dropout rate of 0.3. The features of the forward and backward layers are integrated in a fusion module through horizontal concatenation of the hidden state vectors from both directions, and the action recognition results are output through a fully connected layer and classifier.

ACCURATE FEEDBACK GENERATION MODULE FOR INSTRUMENTAL MUSIC TEACHING

After the fully connected layer and the Softmax classifier output the action category, the feedback module starts the error type mapping mechanism, and determines the error type by calculating the difference between the predefined action error feature vector and the standard playing posture model. The difference calculation uses the Euclidean distance to measure the degree of deviation of the coordinate sequence [30,31].

The distance metric formula is:

$$d = \sqrt{\sum_{i=1}^{n} \left[(x_i' - x_i^s)^2 + (y_i' - y_i^s)^2 \right]} \tag{4}$$

d represents the Euclidean distance between the current playing posture and the standard model; x_i' and y_i' represent the normalized coordinates of the i-th key point; x_i^s and y_i^s represent the normalized coordinate values of the corresponding key points in the standard playing posture; and n is the total number of key points.

The calculated distance is compared with the preset threshold. Movements exceeding the threshold are judged as error types. Error types are categorized through a physics-informed framework mapping movement deviations to acoustic consequences, incorporating joint angle deviations affecting bow pressure, amplitude variations influencing dynamic range, timing discrepancies, and instrument-specific parameters including bow angle, finger curvature, and contact point precision. The system correlates movement features with sound production through a knowledge base encoding biomechanical relationships between performer actions and acoustic output. Teaching rules are retrieved based on error type and acoustic impact, ensuring feedback addresses the relationship between physical movements and sound quality. The teaching knowledge base is constructed based on the teaching standards of college instrumental music and contains standard movement parameters for playing various instruments and common error correction strategies [32-34].

The formula for generating teaching language is:

$$F = \alpha \cdot A + \beta \cdot T \tag{5}$$

F represents the generated feedback instruction, A represents the action amplitude parameter, T represents the timing feature parameter, α and β are weight coefficients, which respectively reflect the importance of action amplitude and timing features in the feedback instruction.

The quantitative parameters are fed into the teaching language generation model, which outputs structured feedback instructions containing error descriptions and correction suggestions. These feedback instructions are directly linked to the university instrumental music teaching standards [35,36]. The movement feature data output by the system is synchronously transmitted to the intelligent performance costume design module, enabling the coordinated optimization of performance movement analysis and costume functional parameters.

III. EXPERIMENT AND RESULTS ANALYSIS

DATASET CONSTRUCTION AND EXPERIMENTAL SETUP

As shown in Table 1, the dataset contains 1,200 videos of five musical instruments at three levels, totaling 420 hours in length, at 1920 × 1080 resolution and 30 frames per second. The data serves as the foundation for assessing system performance and is separated into training, validation, and test sets in a 7:2:1 ratio. The five instruments in the university instrumental performance video dataset are the cello, flute, erhu, violin, and piano. Rehearsal room environment data under various lighting conditions is included in the test set. The system is running Ubuntu 22.04 and has an Intel Core i9-11900K processor, 64GB of random access memory, and an NVIDIA RTX 4090 graphics processor.

TABLE 1. University instrumental music performance experimental dataset

Parameter catalog	Specific parameters	Numerical
	Types of musical instruments	5 types
	Performance level	3 levels
Dataset size	Total video volume	1200 paragraphs
	Total duration	420 hours
	Resolution	1920×1080 pixels
	Frame rate	30 frames per second
Dataset partitioning	Training set	70%
	Validation set	20%
	Test set	10%

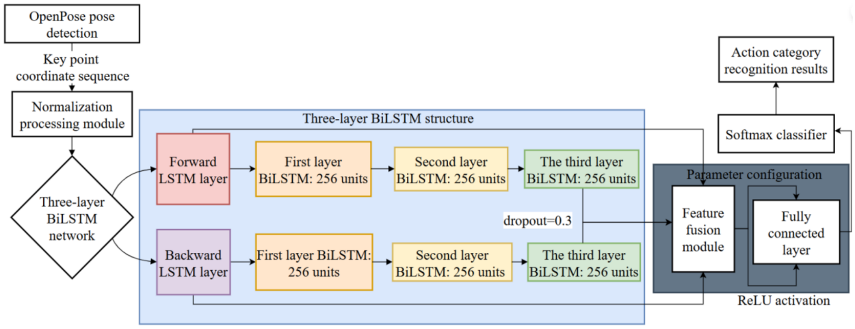

FIGURE 1. BiLSTM-based spatiotemporal feature fusion architecture for instrumental performance movements

PERFORMANCE ANALYSIS OF MULTI-INSTRUMENT ACTION RECOGNITION

The purpose of this study was to examine BiLSTM's capacity to recognize movements from university students' instrumental performances. Five representative instruments were used, which were piano, violin, erhu, flute, and cello. Each instrument exhibits a different mechanism to produce its sound: keyboard instruments rely on precise keystrokes to produce music, string instruments require suitably complex coordination of bowing and stringing, and wind instruments call for careful finger and breath coordination.

This study systematically evaluated five primary high-frequency movements: finger keystrokes, wrist oscillations, arm movements, fingering transitions, and tempo variations. These movements are relevant to technical challenges, and movement recognition accuracy matters in determining feedback system performance. Accuracy was the key metric reflecting the system's ability to capture finer gestures changes. In addition, comparing BiLSTM structures' performance differences with different layer numbers revealed effects network depth has on temporal-focusing modeling. In Figure 2, the flute had the highest recognition rate of 99.1%, making it the best-performing instrument overall. The key feature of the wind instruments is characterized by large fingering amplitudes and evident spatial features. The violin has a much lower recognition rate of 97.9%, which is largely due to the difficulty of coordinating the bow-arm to the string as well as dealing with wrist micro-movements related to playing very quickly. The three-layer BiLSTM model seemed to perform the best for all types of movement at 96.3%, an improvement of 7.1% over the single-layer model. This demonstrates that deep networks are more effective at managing the coordinate jitter in OpenPose detection during high-frequency low-amplitude movements. A wrist swing recognition rate of 98.8% further demonstrates the bidirectional nature's advantage at modeling the temporal features of continuous movement. In contrast, the fingering transitions had the lowest rate, which is likely related to brief frame drops during the rapid shifts of fingering from note to note. Overall, the experimental results show that the three-layer BiLSTM model can effectively leverage the spatial pose and temporal pose continuity to inform the researcher while assessing an instrumental musician at the undergraduate level.

VERIFICATION OF SYSTEM REAL-TIME PROCESSING PERFORMANCE

In the research undertaken to evaluate feedback latency performance, it systematically decomposed the process, from acquisition of the video to generating feedback commands, with specifics on the normalization and BiLSTM methods it developed to compress timing overhead. The video acquisition and preprocessing stages consist of real-time acquisition and basic processing of a video stream captured at 1920×1080 resolution. OpenPose's keypoint detection extracts two-dimensional coordinate sequences of the hands and joints across the video. In the normalization, it applies an adaptive sliding average filter to the coordinate sequences to remove scale differences in space. Various three-layer BiLSTM networks analyze the data to exploit both bidirectional temporal dependencies during the performance movement. During feedback command generation, it assessed error types by calculating Euclidean distances, outputting teaching commands accordingly. The latency of each stage provides a degree of accuracy for the latency measurements during the real-time evaluation of performance and comparison. The relative processing time before optimization and as optimized is presented both as a quantitative acceptance measure of the overall real-time feedback system capabilities, as well as capability for revealing analysis of the actual technical components. The absolute processing time distribution is presented following a system optimization on the left, and the latency reduction impact of the optimizations applied for each stage of processing on the right in Figure 3.

Figure 3 illustrates how the optimized system's 48.3 ms total latency is made up of 3.8 ms for video acquisition, 14.2 ms for OpenPose detection, 6.7 ms for coordinate normalization, 18.5 ms for BiLSTM analysis, and 5.1 ms for feedback generation, with the majority of the latency coming from BiLSTM analysis. During the system optimization process model compression techniques such as pruning and quantization were considered for the BiLSTM network however these approaches were not implemented due to concerns about potential accuracy degradation in capturing subtle performance movements critical for music instruc-

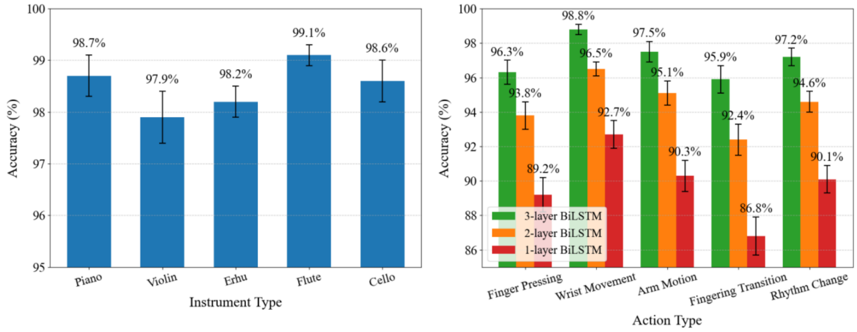

FIGURE 2. Comparison of motion recognition accuracy of different instruments and BiLSTM structures

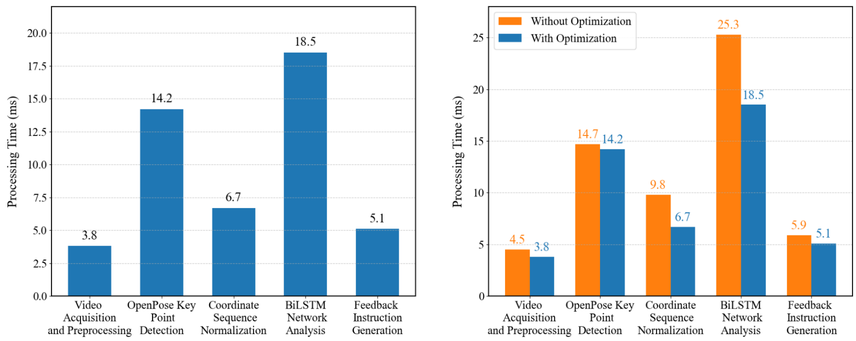

FIGURE 3. Comparison of delay time at each stage of system processing and analysis of optimization effect

tion The current three-layer BiLSTM architecture maintains the necessary spatiotemporal modeling capability while achieving acceptable latency performance Future work will investigate hardware acceleration methods and lightweight network architectures to further optimize the computational efficiency without compromising recognition accuracy for fine motor skill analysis. The normalization phase saw a notable optimization effect, reducing the time from 9.8 ms to 6.7 ms. This has to do with the adaptive sliding average filter algorithm's capacity to dynamically modify the window size in response to the motion's fundamental frequency. The latency reduction in BiLSTM analysis from 25.3 ms to 18.5 ms was achieved through optimized CUDA (Compute Unified Device Architecture) kernel implementation with memory coalescing and reduced thread divergence. The network utilized cuDNN's optimized LSTM primitives with tensor core acceleration on the RTX 4090 GPU (Graphics Processing Unit). Memory access patterns were optimized through data layout transformation from NTC to TNC format, enabling efficient utilization of GPU shared memory. Batch processing was optimized with a batch size of 64 to maximize GPU throughput while maintaining temporal coherence in sequence processing. Kernel fusion was implemented to combine multiple LSTM operations into single GPU kernel launches, reducing the number of kernel invocations required for sequence processing. When compared to the unoptimized solution, the overall system latency decreased from 60.2 ms to 48.3 ms, indicating that normalization and BiLSTM parallel computing effectively reduced timing overhead and satisfied the strict requirements for real-time feedback in university instrumental music instruction.

ADAPTABILITY TEST OF MULTI-ENVIRONMENT TEACHING SCENARIOS

In order to test the practical efficacy of the proposed system in university instrumental music instruction, the present study applied adaptability assessments in five typical environmental contexts: an ordinary rehearsal room environment, low-light illumination, high illumination with strong light interference, complex background, while the setting includes multiple people as learners. The coordinate standard deviation, a quantitative measure of keypoint detection stability, demonstrates the system has successfully captured small gradient changes in wrist movement velocity; less can indicate smoother keypoint trajectories. The feedback commands' accuracy demonstrate the ability for the system to effectively provide instructional precision in the context of two data perspectives: types of error detection and command generated for the learner. Quality of the feedback to the learner has a direct relationship to the overall evaluation criteria for movement standardization in university instrumental music instruction. Figure 4 (left) shows the fluctuations in wrist keypoint coordinates, while the right shows the accuracy of feedback commands. Together, these figures demonstrate the robustness of the system in a real-world university teaching environment.

In Figure 4, the standard deviation of the overall coordinates in a standard rehearsal room environment is 3.2 pixels, and the overall accuracy reaches 98.2%, demonstrating the

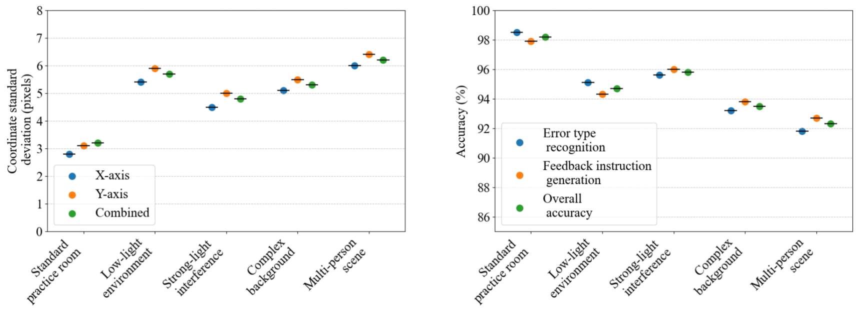

FIGURE 4. Adaptability of instrumental music teaching environment in colleges and universities: Analysis of wrist key point stability and feedback instruction accuracy

system's excellent performance under standard conditions. Low-light conditions increase the standard deviation to 5.7 pixels, and the overall accuracy drops to 94.7%. This is due to the principle of optical imaging that insufficient light reduces keypoint detection accuracy. The overall accuracy in strong light interference environments is 95.8%, higher than in low-light environments. This is because excessive light can actually increase image contrast in some cases. In complex background environments, the standard deviation reaches 5.3 pixels, and in scenes with multiple people in the same frame, the standard deviation increases to 6.2 pixels, with the overall accuracy dropping to 92.3%. The system implements person tracking IDs to maintain individual performer identification during multi-person scenarios, though performance degradation occurs when targets are closely spaced or occluded. Future implementations will integrate localized OpenPose runs with spatial partitioning and performer prioritization protocols to enhance multi-person scene accuracy through dedicated processing channels for each performer. These data discrepancies are caused by the particular lighting and spatial limitations of teaching instrumental music at a university. They confirm that the adaptive sliding average filter algorithm is effective at reducing coordinate jitter and that the BiLSTM can manage the temporal features of intricate learning environments. The system offers dependable real-time feedback support for teaching instrumental music while maintaining high teaching applicability in typical university settings.

IV. CONCLUSION

This research proposes an instrument performance motion recognition and feedback system, which utilizes the technologies of OpenPose and BiLSTM, a precise motion analysis framework that can be adapted to provide objective, detailed feedback on complex motor skills. This framework extracts sequences of performance postures through human keypoint detection. After coordinate normalization, the samples are then input to a three-layer BiLSTM network (Bi-directional Long Short-Term Memory network) for spatiotemporal feature fusion. The resulting motion recognition outputs engage the feedback module. Experimental results show that the finger key recognition accuracy rate is 96.3%, and the reduced system latency goes from 60.2ms to 48.3ms, demonstrating that this framework improves motion recognition accuracy and provides real-time feedback. In standard rehearsal room test conditions, the standard deviation of the integrated coordinates was 3.2 pixels, achieving 98.2%, as a total absolute accuracy. Overall, the system produced stable performance under complex conditions of university teaching environments, such as low light and strong interference light. This system tackles the issues of subjective movement guidance and delayed feedback by providing an objective, quantitative assessment of motion with timely feedback to support instrumental performance instruction in universities. This results in enhancing efficiency in music professor training and facilitating the conversion of instrumental music instruction from the experience-based practice to data-based practice. Overall, this also offers valuable technical support for instrumental music instruction to progress towards the informatization of music education in universities. The system architecture is able to integrate with intelligent performance costume design to allow for the coordinated analysis of the characteristics of performance motion and garment functional parameters.

AUTHOR CONTRIBUTIONS

Conceptualization –Xuan Zhou and Wei He; methodology – Xuan Zhou and Wei He; investigation – Xuan Zhou and Wei He; writing-original draft preparation – Xuan Zhou and Wei He. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] E. Sanchez-Escribano, F. Gertrudix, and A. Bautista, "Analyzing instrumental music education models: A fourdimension tool," Arts Education Policy Review, vol. 125, no. 3, pp. 187-197, 2024, doi: 10.1080/10632913.2022.2041139.

[2] E. Concina, "Effective music teachers and effective music teaching today: A systematic review," Education Sciences, vol. 13, no. 2, pp. 107-128, 2023, doi: 10.3390/educsci13020107.

[3] G. G. Artıktay, “Cognitive neuroscience and music education: Relationships and interactions,” International Journal of Educational Spectrum, vol. 6, no. 1, pp. 91-119, 2024, doi: 10.47806/ijesacademic.1402953.

[4] T. Ma, I. C. Sanchis, G. R. Santana, and Y. Jiang, “Neuroplasticity Mechanisms in Early Childhood Piano Education: A Literature Review from the Perspective of Educational Neuroscience,” Journal of Sociology and Education, vol. 1, no. 1, pp. 167-175, 2025, doi: 10.63887/jse.2025.1.1.19.

[5] G. Y. Duztaban, “THE PIANO EDUCATION IN EARLY CHILDHOOD: A COMPARATIVE ANALYSIS OF MUSIC EDUCATION MODELS,” LOKUM Sanat ve Tasarım Dergisi, vol. 3, no. 2, pp. 230-252, 2025, [Online]. Available: https://izlik.org/JA46DY98FX.

[6] F. Roggio, B. Trovato, M. Sortino, and G. Musumeci, “A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review,” Heliyon, vol. 10, no. 21, Art. no. e39977-e39989, 2024, doi: 10.1016/j.heliyon.2024.e39977.

[7] Y. Liu, T. Zhang, Z. Li, and L. Deng, “Deep learning-based standardized evaluation and human pose estimation: A novel approach to motion perception,” Traitement du Signal, vol. 40, no. 5, pp. 2313-2320, 2023, doi: 10.18280/ts.400549.

[8] S. Kulkarni, S. Deshmukh, F. Fernandes, A. Patil, and V. Jabade, “Poseanalyser: A survey on human pose estimation,” SN Computer Science, vol. 4, no. 2, pp. 136, 2023, doi: 10.1007/s42979-022-01567-2.

[9] K. Esaki and K. Nagao, “An efficient immersive self-training system for hip-hop dance performance with automatic evaluation features,” Applied Sciences, vol. 14, no. 14, pp. 5981, 2024, doi: 10.3390/app14145981.

[10] X. Ma, “Deep learning models combining stereo vision for dance movement evaluation,” International Journal of Information and Communication Technology, vol. 26, no. 11, pp. 69-85, 2025, doi: 10.1504/ijict.2025.146102.

[11] Y. Zhang, “Virtual Concerts in Learning Oboe-Played Chinese Folk Music: Impact on Performance Proficiency, Perceived Aesthetic Qualities, and Students ’ Motivation,” International Review of Research in Open and Distributed Learning, vol. 26, no. 3, pp. 61-82, 2025, doi: 10.19173/irrodl.v26i3.8358.

[12] N. Moura and S. Serra, “Saxophone players ’ self-perceptions about body movement in music performing and learning: An interview study,” Music Perception: An Interdisciplinary Journal, vol. 41, no. 3, pp. 199-216, 2024, doi: 10.1525/mp.2024.41.3.199.

[13] L. Zhou, “Construction and Practice of Erhu Performance Course for Musicology Majors in Local Universities—A Case Study of Sichuan Minzu College,” Frontiers in Educational Research, vol. 7, no. 2, pp. 71-75, 2024, doi: 10.25236/FER.2024.070211.

[14] N. Hassan, A. S. M. Miah, and J. Shin, “A deep bidirectional LSTM model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition,” Applied Sciences, vol. 14, no. 2, pp. 603, 2024, doi: 10.3390/app14020603.

[15] N. Zerrouki, F. Harrou, A. Houacine, R. Bouarroudj, M. Y. Cherifi, A. D. A. Zouina, et al., “Deep learning for hand gesture recognition in virtual museum using wearable vision sensors,” IEEE Sensors Journal, vol. 24, no. 6, pp. 8857-8869, 2024, doi: 10.1109/JSEN.2024.3354784.

[16] J. Wu, P. Ren, B. Song, R. Zhang, C. Zhao, and X. Zhang, “Data glove-based gesture recognition using CNN-BiLSTM model with attention mechanism,” PLoS One, vol. 18, no. 11, Art. no. e0294174-e0294195, 2023, doi: 10.1371/journal.pone.0294174.

[17] S. Sahoo, “Sensor Fusion and Virtual Sensor Design for Enhanced Multi-Sensor Data Accuracy in Autonomous Systems,” International Journal on Smart & Sustainable Intelligent Computing, vol. 1, no. 2, pp. 21-39, 2024, doi: 10.63503/j.ijssic.2024.31.

[18] Y. Zhang, B. Zhang, C. Shen, H. Liu, J. Huang, K. Tian, et al., “Review of the field environmental sensing methods based on multi-sensor information fusion technology,” International Journal of Agricultural and Biological Engineering, vol. 17, no. 2, pp. 1-13, 2024, doi: 10.25165/j.ijabe.20241702.8596.

[19] P. Lalwani and R. Ganeshan, “A novel CNN-BiLSTM-GRU hybrid deep learning model for human activity recognition,” International Journal of Computational Intelligence Systems, vol. 17, no. 1, pp. 278-297, 2024, doi: 10.1007/s44196-024-00689-0.

[20] Z. Chen, Q. Xie, and W. Jiang, “Hybrid Deep Learning Models for Tennis Action Recognition: Enhancing Professional Training Through CNN-BiLSTM Integration,” Concurrency and Computation: Practice and Experience, vol. 37, no. 6-8, Art. no. e70029, 2025, doi: 10.1002/cpe.70029.

[21] A. Tharatipyakul, T. Srikaewsiew, and S. Pongnumkul, “Deep learning-based human body pose estimation in providing feedback for physical movement: A review,” Heliyon, vol. 10, no. 17, Art. no. e36589-e36604, 2024, doi: 10.1016/j.heliyon.2024.e36589.

[22] W. Zhang, “Dynamic pose recognition based on deep learning: Developing a CNN model for choral conductor pose recognition,” Journal of Computational Methods in Sciences and Engineering, vol. 25, no. 4, pp. 3856-3874, 2025, doi: 10.1177/14727978251323068.

[23] M. Sui, L. Jiang, T. Lyu, H. Wang, L. Zhou, and Chen P.et al, “Application of deep learning models based on efficientdet and openpose in user-oriented motion rehabilitation robot control,” Journal of Intelligence Technology and Innovation, vol. 2, no. 3, pp. 47-77, 2024, doi: 10.30212/JITI.202402.012.

[24] L. Xiao, Y. Cao, Y. Gai, E. Khezri, J. Liu, and M. Yang, “Recognizing sports activities from video frames using deformable convolution and adaptive multiscale features,” Journal of Cloud Computing, vol. 12, no. 1, pp. 167-186, 2023, doi: 10.1186/s13677-023-00552-1.

[25] S. Lin and W. Hou, “Efficient Sampling of Two-Stage Multi-Person Pose Estimation and Tracking from Spatiotemporal,” Applied Sciences, vol. 14, no. 6, Art. no. 2238, 2024, doi: 10.3390/app14062238.

[26] L. Bellier, A. Llorens, D. Marciano, A. Gunduz, G. Schalk, P. Brunner, et al., “Music can be reconstructed from human auditory cortex activity using nonlinear decoding models,” PLoS biology, vol. 21, no. 8, Art. no. e3002176-e3002186, 2023, doi: 10.1371/journal.pbio.3002176.

[27] L. Wang, “Neural Networks and Ensemble Model to Automatic Music Coordination: A Performance Comparison,” Information Technology and Control, vol. 54, no. 2, pp. 520-535, 2025, doi: 10.5755/j01.itc.54.2.36737.

[28] N. Mohammadi Foumani, L. Miller, C. W. Tan, G. I. Webb, G. Forestier, and M. Salehi, “Deep learning for time series classification and extrinsic regression: A current survey,” ACM Computing Surveys, vol. 56, no. 9, pp. 1-45, 2024, doi: 10.1145/3649448.

[29] X. Li, Z. Luo, J. Lv, C. Yang, S. Yan, J. Xu, et al., “m5U-HybridNet: Integrating an RNA Foundation Model with CNN Features for Accurate Prediction of 5-Methyluridine Modification Sites,” Journal of Chemical Information and Modeling, vol. 65, no. 15, pp. 8079-8096, 2025, doi: 10.1021/acs.jcim.5c01237.

[30] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, “A review of convolutional neural networks in computer vision,” Artificial Intelligence Review, vol. 57, no. 4, pp. 99-141, 2024, doi: 10.1007/s10462-024-10721-6.

[31] Q. Lei, H. Li, H. Zhang, J. Du, and S. Gao, “Multi-skeleton structures graph convolutional network for action quality assessment in long videos,” Applied Intelligence, vol. 53, no. 19, pp. 21692-21705, 2023, doi: 10.1007/s10489-023-04613-5.

[32] Y. Han, L. Han, C. Zeng, and W. Zhao, “The innovation path of VR technology integration into music classroom teaching in colleges and universities,” Scientific Reports, vol. 15, no. 1, pp. 12200-12216, 2025, doi: 10.1038/s41598-025-97003-5.

[33] Y. Jiang and D. Dumlavwalla, “Pre-college piano teaching practice in China and the United States: An international comparative study from the piano teachers ’ perspectives,” International Journal of Music Education, vol. 43, no. 3, pp. 519-533, 2025, doi: 10.1177/02557614231208236.

[34] Y. Li, “Interactive methods for improving musical literacy among students with preschool education Majors at Teacher Training Universities: The effectiveness of the Kodály method,” Education and Information Technologies, vol. 28, no. 10, pp. 12807-12821, 2023, doi: 10.1007/s10639-023-11636-5.

[35] L. R. de Bruin, “Instrumental music teachers ’ development of feedback across the lifespan: A qualitative study,” International Journal of Music Education, vol. 42, no. 1, pp. 32-46, 2024, doi: 10.1177/02557614231151445.

[36] L. R. de Bruin, “Feedback in the instrumental music lesson: A qualitative study,” Psychology of Music, vol. 51, no. 4, pp. 1259-1274, 2023, doi: 10.1177/03057356221135668.