

Deep Learning-Based Quantitative Evaluation Model for Student Motor Skills and its Application Validation

Zhenpeng Yu

College of Physical Education and Sports Science, Anshan Normal University, Anshan 114007, Liaoning, China

Corresponding author: Zhenpeng Yu (e-mail: yzp7677@163.com)

ABSTRACT Quantitative assessment of student motor skills requires robust spatiotemporal feature extraction from motion video data, especially when automated evaluation is expected to reduce subjective scoring bias. This study develops a deep learning-based evaluation model integrating pose recognition, convolutional neural networks, and long short-term memory networks. Movement videos are preprocessed to extract key human-joint points, CNN layers are used to capture spatial posture features, and LSTM layers model temporal continuity and coordination during movement execution. A multi-task regression module is further introduced to quantify posture stability, movement fluency, and body coordination. The model is validated using movement data from 1200 students covering basketball, gymnastics, and long jump. Experimental results show that the mean squared error of the scoring task is 0.042, with repeated-test variation controlled within 0.002, indicating high consistency and generalization across different movement types. The proposed framework provides an engineering-oriented solution for intelligent motion assessment, video-based sensing, and automated spatiotemporal signal analysis.

INDEX TERMS deep learning, motor skill assessment, pose recognition, spatiotemporal modeling, image signal processing

I. INTRODUCTION

THE assessment of student motor skills serves as a foundational element of physical education and individual development, mirroring the rigorous precision required in testing material properties to ensure the integrity of the final material. The majority of existing evaluation tools, however, are dependent upon manual observations, and experience of teachers, which leads to high levels of subjectivity, lack of consistency in scoring scales, and a lack of quantitative aspects. The integration of the computer vision technology and the video data offers the possibility of motion analysis the new direction in the field of motion analysis with the development of the digital sports and the intelligent teaching [1-2]. The measurement of motor skills would demand not merely the reflective capture of spatial structural aspects of postures but also an emphasis on the continuity and coordination of the completion of movement thus making greater demands on the conventional manual or shallow feature-based approach. It has become an issue of research in the sphere of sports assessment how to accomplish automated, refined, and interpretable assessment without compromising the authenticity of the data.

Deep learning (DL) is developed, which offers a practical answer to the quantitative evaluation of motor skills. Its multi-layer capability in abstraction and time series model-

ing ability can be used to extract important motion features of complex motion data and gain a profound insight into the spatiotemporal coupling information. Convolutional Neural Network (CNN) has a high capability of extracting posture features, whereas Long Short-term memory (LSTM) has the ability of extracting temporal features. A mix of the two makes scoring of the motor skills of the students automatic and manual feature design is not necessary. This end-to-end training approach facilitates robust generalization and flexibility within the model, providing essential technical assistance to the intelligent assessment framework of physical education while aligning with the high-precision automated inspection standards used to detect defects in production processes [3].

II. RELATED WORK

Initial measures of motor skills were dependent upon either manual observation or analytical models of the motor skills, which were usually scored using key movement indicators and ranges of scores. Although these techniques are a reference of the standardized setting, they are poor in managing the change of posture during real movement and highly rely on the experience of the assessor. Traditional computer vision research has tried to use feature extraction algorithms like optical flow and edge detection but the weakness of

features representation capability has led to inadequate accuracy and generalization performance in action recognition.

With the popularization of deep learning (DL), researchers have introduced convolutional neural networks (CNNs) into the field of action recognition and behavior analysis. This structure performs well in spatial feature extraction and can capture the positional distribution and local texture changes between key points of the body [4]. Some studies have improved the accuracy of action discrimination by improving the convolutional

structure or adding a pose estimation module, but it still has limitations in temporal information processing and is difficult to fully describe the dynamic continuous process of action.

Long Short-Term Memory (LSTM) networks have been introducing gradually into motion assessment tasks to solve the problem of the lack of temporal modeling to capture the temporal dependencies of action evolution. Combining CNN and LSTM into a hybrid architecture has become a research trend in recent years and it allows simultaneous learning of spatial and time-related features. The current models are however largely focused in competitive sports or one action recognition, and does not provide general quantitative models of the different motor skills of students, and also, the scoring mechanism is not interpretable or open to control of objectivity, which is exactly the innovation point in this research [5-6].

III. METHODS

A. VIDEO PREPROCESSING AND KEY POINT EXTRACTION

Video data needs to be standardized before being input into the model to ensure the recognizability of motion features. The original student motion videos were sampled at 30 frames per second, and the resolution was adjusted to 640×480 to reduce data volume and maintain the integrity of motion posture. The OpenPose framework based on pose estimation algorithm was used to extract body keypoint information, covering 18 major joints of the human body, including the coordinates of limb extremities, joints, and trunk center. Temporal filtering was performed on the extracted keypoints, and the Savitzky-Golay smoothing algorithm was used to reduce trajectory noise caused by camera shake and correct the coordinate values of missing points. After normalization, the coordinates of all keypoints were mapped to the $[-1, 1]$ interval to adapt to the input format of the deep learning model [7-8]. The continuity of motion between video frames was maintained by temporal index reconstruction. To handle varying video lengths (4–8 seconds), a zero-padding and truncation strategy was implemented to standardize all input sequences to a fixed temporal dimension of $T = 240$ frames (8 seconds at 30fps), ensuring a uniform tensor shape for the LSTM input layer. After frame difference analysis, static segments were removed from the data, and only continuous intervals containing motion were retained. The processed video keypoint

sequence is stored in tensor form, where each frame contains the two-dimensional coordinates and confidence values of the keypoints, providing a high-quality data foundation for learning the spatial features and temporal information of the model.

B. CONSTRUCTION OF CNN SPATIAL FEATURE EXTRACTION MODULE

The CNN spatial feature extraction module takes frame-level images of keypoint sequences as input. The model adopts a stacked convolutional structure to enhance the recognition of local motion details. The input tensor size is set to $64 \times 64 \times 3$. Spatial texture and limb distribution features are extracted through three layers of convolution and batch normalization. The first layer has a 3×3 kernel size, 32 channels, and a stride of 1 to capture the relative position information between joints. The second layer uses $64 \ 3 \times 3$ convolutional kernels and introduces the ReLU activation function to improve nonlinear expression. The pooling window is set to 2×2 to achieve local feature compression. The third layer expands the convolution depth to 128 channels and maintains the gradient stability of deep features through the residual connection structure [9-10]. The output of each layer is mapped to a one-dimensional feature vector through a fully connected layer, so that the spatial pose distribution information can be expressed in high dimension. To improve feature robustness, a Dropout ratio of 0.3 is introduced to reduce the risk of overfitting. The feature vector output by this module has high spatial discriminativeness and can accurately capture the muscle movement relationship and local coordination between key parts of the body .

C. LSTM TEMPORAL ANALYSIS AND FEATURE FUSION MECHANISM

The LSTM time analysis module is applied to identify the dynamism in the actions of students over time. The input sequence comprises of spatial feature vectors which are extracted by a CNN, laid out in time step order to create a $T \times 128$ dimensional tensor in which T is the duration of the sequence. The model utilizes the model of a long short-term memory network structure, which has 256 hidden units. The forget gate, output gate, and input gate are equipped with independent weight matrices (W_f , W_o , W_i) to facilitate the selective flow of information, and the weights are updated with the Adam optimization algorithm in order to maintain the propagation of the temporal gradient. The hidden state vectors are unfolded in time followed by inputting them into the attention weighting layer and the intensity of contribution of each time step towards the overall performance of the motion is computed with the weight coefficients. Lastly, the output feature vector is cascaded and sufficiently mixed using the CNN spatial features using a fully connected transformation to create a fused high-dimensional motion description vector. This framework is able to describe the dynamic properties of student behaviors in spatial sense and

time sense. The structural parameters and feature fusion configuration are shown in Table 1.

TABLE 1. LSTM Structure Parameters and Feature Fusion Configuration

Layer type	Number of units	Activation function	Output Dimension	Fusion method
LSTM-1	256	tanh	$T \times 256$	Temporal feature extraction
LSTM-2	256	tanh	$T \times 256$	Deep dynamic coding
Attention layer	—	softmax	256	Keyframe weighting
Fusion layer	—	ReLU	384	CNN-LSTM cascade

This network enhances the discriminative power of temporal representation with a dynamic weighting mechanism, ensuring that the structural changes of the action flow and spatial posture features are consistently mapped in a high-dimensional representation, thereby achieving accurate modeling of the continuity of action skills.

D. SKILL QUANTITATIVE SCORING MODEL DESIGN

The scoring model of quantification of the skills is based on the input of a fused high-dimensional feature vector with the motor skill score of the students calculated in a multi-task regression framework. The input dimension of the model is 384 and it is compressed to 128 and 64 dimensions with the help of two fully connected layers and ReLU is employed to retain the non-linear response properties. The final score is produced in linear numbers by the output layer with linear activation function. The loss function is a mix of Mean squared error (MSE) and L1 regularization, which is aimed at balancing model sparsity and fitting accuracy. The network optimization applies Adam algorithm, which is adjusted dynamically throughout the training process so that convergence becomes stable. The model employs a multi-task learning framework where auxiliary labels for posture stability, movement fluency, and body coordination were derived from expert sub-ratings during the initial manual assessment process. These branch tasks contribute to the total loss function L_{total} through a weighted summation: $L_{\text{total}} = \omega_{\text{main}}L_{\text{score}} + \sum \omega_{\text{aux}}L_{\text{aux}}$, where ω represents the task-specific attention weights. This parameter exchange mechanism ensures the shared feature backbone captures fine-grained coordination metrics essential for the final regressive score. This design attains a mapping process end to end between spatial posture and temporal movement to skill scoring to be consistent and interpretable of the scoring results regarding objectivity and fine-grained differences.

IV. RESULTS AND DISCUSSION

A. DATASET COMPOSITION AND EXPERIMENTAL SETUP

The experimental data was comprised of 8640 video clips illustrating the skills in basketball, gymnastics, and long jump of 1200 students in the middle school. The videos were 4-8 seconds and at 30fps. The clips were standardized and transformed into a standard resolution and inputted into the model. In data acquisition, simultaneous two-camera shooting at a stationary location was used so that spatial

posture data would be accurate. Three movement self-tests were done to every student and the results were assessed by three physical education teachers who were nationally qualified and graded manually. There was a scale of 0 to 100 and the average score was taken to achieve the final label which formed the supervised learning baseline; notably, to ensure numerical stability during training and provide a standardized evaluation metric, all labels were normalized to the $[0, 1]$ interval before the calculation of the Mean Squared Error (MSE). The data was randomly separated into a 70 percent training, a 15 percent validation as well as a 15 percent test. The sample balancing strategy was applied in terms of movement category and gender. All the experiments were conducted on NVIDIA RTX 4090 platform with a batch size of 32 and 100 training epochs. The optimizer was the Adam algorithm. The system itself was realized with the help of the TensorFlow framework and data augmentation policies to increase the sample diversity of the model and enhance the level of generalization.

B. MODEL PERFORMANCE EVALUATION

In order to test whether the scoring model is accurate, expert ratings served as a reference standard and a comparative experiment was applied in the rating of 300 test samples of the same batch. Three experts rated each video on its own and the average of the scores was calculated as the actual value. The model did some automatic computations with the same input of the sample. The experiment was done 5 times to ascertain the reproducibility and stability of the model. It was represented by the mean squared error (MSE) which reflected the difference between the model and the expert ratings. In this context, the ‘expert scoring reference MSE’ represents the inter-rater variability (variance) between the three individual expert scores and their collective average, serving as a baseline for human subjectivity. Table 2 shows the specific results:

TABLE 2. Comparison of Model Scoring Results and Expert Scoring

Number of repetitions	Expert scoring reference MSE	The model in this paper is MSE
1st time	0.058	0.043
2nd time	0.057	0.042
3rd	0.057	0.042
4th	0.056	0.042
5th	0.057	0.041

The results in the table show that the mean inter-rater MSE of expert human scoring is 0.057, while the mean MSE of the model in this paper is 0.042. The model’s fluctuation range was controlled within 0.002 in multiple repeated calculations, demonstrating high stability and repeatability. Compared with human scoring, the model can more accurately reflect the differences in scoring intervals, indicating that the deep learning model has strong quantitative consistency and reliability in action evaluation tasks.

C. ADAPTABILITY VERIFICATION UNDER DIFFERENT SPORTS (BASKETBALL, GYMNASTICS, LONG JUMP)

Model adaptability validation was conducted using data from three types of movements: basketball, gymnastics, and long jump. The same structure and parameter configuration were used, and no event-specific optimization was performed. 300 samples were selected for each event for independent training and testing. Output scores were compared with expert scores, and mean squared error (MSE) and standard deviation (SD) were calculated. The experiment was repeated six times to evaluate model stability. Data are shown in Table 3.

TABLE 3. Model Adaptability Validation Results for Different Sports

Number of repetitions	Basketball MSE	Gymnastics MSE	Long jump MSE	Average MSE	Standard deviation SD
1st time	0.043	0.040	0.044	0.042	0.0017
2nd time	0.042	0.041	0.043	0.042	0.0012
3rd	0.044	0.039	0.042	0.042	0.0018
4th	0.041	0.040	0.043	0.041	0.0013
5th	0.043	0.041	0.042	0.042	0.0014
6th	0.042	0.040	0.044	0.042	0.0015

The model's average MSE for basketball, gymnastics, and long jump remained between 0.041 and 0.042, with a maximum standard deviation of only 0.0018, indicating minimal fluctuation in the model's output across different motion types. The slightly lower MSE in gymnastics reflects the model's better feature capture ability for motion control and posture precision-based sports. The relatively similar MSEs in basketball and

long jump indicate stable performance in extracting features related to rapid movements and long-distance displacements. Overall, the network structure demonstrates good generalization and robustness to dynamic features across different sports types, without exhibiting any event-dependent bias.

V. CONCLUSION

The field of quantitative evaluation of student motor skills exhibits significant application potential through deep learning, mirroring the advanced pattern recognition used to analyze material morphology and material spatial distribution in modern engineering. It can combine the capabilities in both CNNs and LSTMs, as it is able to capture both the spatial and temporal characteristics of movements, which allows it to create an automated judgment process on video data to infer quantitative scores. The model is able to record the consistency of movement performance, posture control, and body coordination during the learning process, hence, minimizing subjective variations in human scoring, and refining and making the evaluation of motor performance more objective. This process proves the consistent generalization capacity in various forms of sports, which confirms that the network structure is highly adaptable to action situations and cross-category discrimination skills. The training data is mostly based on common action samples, which is why the model is not well-adapted to the complicated group movements and special postures. The integration of multimodal

sensor measurements and self-guided learning processes—comparable to the real-time monitoring of material tension and spindle vibration in high-speed spinning—offers a path for future studies to expand dynamic scene recognition and investigate the application of instantaneous movement guidance.

AUTHOR CONTRIBUTIONS

Zhenpeng Yu designed, collected and analyzed the data, and drafted the manuscript. Zhenpeng Yu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Zhenpeng Yu 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] W. Zhang, "An exploration of the influence of different physical education teaching models in colleges and universities on the mastery of sports skills by college students," *Heilongjiang Science*, vol. 16, no. 13, pp. 93-95, 2025, doi: 10.3969/j.issn.1674-8646.2025.13.028.
- [2] X. Xie, "Application of fun teaching in primary school physical education in cultivating students motor skills," *Physical Education Teaching*, vol. 45, no. S1, pp. 67-68+71, 2025.
- [3] L. Xin and H. Jiang, "Compilation and verification of evaluation indicators for the learning input of college students in sports skills," *Sichuan Sports Science*, vol. 44, no. 3, pp. 28-35, 2025, doi: 10.13932/j.cnki.sctyx.2025.03.06.
- [4] X. Zhang, "How to improve students motor skills through task-driven learning in primary school physical education classes," *Physical Education Teaching*, vol. 45, no. S1, pp. 49-50, 2025.
- [5] Y. Lu, L. Yao, and B. Wang, "A study on the influence of college students athletic skills on physical fitness," *Bulletin of Sports Science and Technology*, vol. 29, no. 1, pp. 49-50, 2021, doi: 10.19379/j.cnki.issn.1005-0256.2021.01.022.
- [6] P. Purwono, A. Maarif, W. Rahmانيar, et al., "Understanding of convolutional neural network (cnn): A review," *International Journal of Robotics and Control Systems*, vol. 2, no. 4, pp. 739-748, 2022, doi: 10.31763/ijrcs.v2i4.888.
- [7] J. Ayeni, "Convolutional neural network (CNN): the architecture and applications," *Applied Journal of Physical Science*, vol. 4, no. 4, pp. 42-50, 2022, doi: 10.31248/AJPS2022.085.
- [8] X. Xie, G. Cheng, J. Wang, et al., "Oriented R-CNN and beyond," *International Journal of Computer Vision*, vol. 132, no. 7, pp. 2420-2442, 2024, doi: 10.1007/s11263-024-01989-w.
- [9] A. A. Elngar, M. Arafa, A. Fathy, et al., "Image classification based on CNN: a survey," *Journal of Cybersecurity and Information Management*, vol. 6, no. 1, pp. 18-50, 2021, doi: 10.5281/zenodo.4897990.
- [10] Z. J. Wang, R. Turko, O. Shaikh, et al., "CNN explainer: learning convolutional neural networks with interactive visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1396-1406, 2020, doi: 10.1109/TVCG.2020.3030418.