Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.31260

Design and Classroom Application of Real-Time Evaluation Algorithm for Gymnastics Movements in Colleges and Universities Based on Lightweight Convolutional Neural Network

Jinghui Zhou[1*], Jingxin Zhou[2]

[1]School of Physical Education, Wuhan University of Science and Technology, Hubei, China
[2]Department of Physical Education, Panlonghu Campus, Hebei Vocational College of Resources and Environment, Hebei, China
**Corresponding Author. Email: 2433563462@qq.com*

Abstract. Against the backdrop of the advancement of educational digitalization, smart sports have become the core direction of physical education teaching reform in colleges and universities. However, the evaluation of gymnastics teaching in colleges and universities is deeply trapped in the traditional model and disconnected from the trend of precise quantification. This paper integrates video stream recognition, skeleton key point evaluation and lightweight CNN technology to design a real-time evaluation algorithm suitable for classroom scenarios. Verified by 12,000 gymnastic movement samples, this algorithm achieves a balance between lightweight and precision, providing technical support for the data-driven transformation of gymnastic teaching.

Keywords: Convolutional Neural Network Gymnastics movements in colleges and universities Algorithm design Smart Sports

1. Introduction

Against the backdrop of the in-depth advancement of the digitalization strategy in education, smart sports have become the core direction of physical education teaching reform in colleges and universities [1]. Gymnastics is a core course for physical education majors in colleges and universities and a characteristic public sports project [2]. However, the current teaching evaluation process in colleges and universities is still trapped in the shackles of traditional models. It not only has common problems such as the single method in the evaluation of gymnastics teaching in normal university physical education majors, but also contradicts the trend of precision advocated by the comprehensive evaluation system of teaching performance of college physical education teachers, and has a significant gap from the quantitative evaluation trend represented by the Referee Support System (JSS) promoted by FIG [3-5].

Against this backdrop, action recognition and skeleton key point evaluation technologies based on video streams provide a critical path for breakthroughs in this field. Models such as HRNet have

DOI: 10.54254/2755-2721/2026.31260

achieved precise positioning of key points in the human body [6], and ST-GCN has enhanced the dynamic correlation capture of joints [7]. Such technologies have been implemented in the scenarios of physical education assessment in primary and secondary schools for automated scoring, confirming the feasibility of applying visual technology in the field of sports evaluation [8].

The research progress of lightweight convolutional neural networks has laid a core foundation for the implementation in technical university classrooms. The depth-separable convolution of the MobileNet series [9], the channel rearrangement mechanism of ShuffleNetv2 [10], and the resource adaptability of MK-UNet [11] have provided technical possibilities for classroom scenarios.

However, the existing research still has obvious shortcomings. As shown in Figure 1, it is a three-level mapping framework that elucidates the corresponding relationship between macro meso micro problems. Firstly, the general action recognition model is not adapted to the professionalism of gymnastics movements and lacks customized design for the FIG scoring criteria. Secondly, the balance between lightweight and precision is insufficient. Some algorithms sacrifice evaluation accuracy for speed, making it difficult to meet the requirements of fine motion discrimination. Thirdly, there is a gap in the implementation of technology. The JSS system focuses on the Olympic appeal scenario. The verification of existing lightweight models is mostly limited to laboratory environments and has not completed scene adaptation and calibration, which forms a sharp contrast with the mature application of AI assessment in primary and secondary school physical education.

Figure 1. Schematic diagram of insufficient research(data source: hand organized)

In view of this, this paper integrates video stream recognition, skeleton key point assessment and lightweight CNN technology, draws on the CBAM attention mechanism and spatio-temporal feature modeling ideas, designs a real-time evaluation algorithm suitable for university scenarios and conducts application verification, providing technical support for the transformation of gymnastics teaching from experience-driven to data-driven, and meeting the core demands of the development of smart sports.

2. Algorithm design

2.1. Overall algorithm architecture

The algorithm adopts a six-level serial architecture consisting of data input, preprocessing, feature extraction, feature fusion, evaluation output and model optimization.

Design an attention-weighted fusion mechanism to achieve deep fusion of spatial appearance features and temporal dynamic features [12]. Let the spatial features $\left(F_S \in R^{C_S \times H \times W}\right)$ and the

DOI: 10.54254/2755-2721/2026.31260

temporal features $\left(F_T \in R^{C_T \times T}\right)$ be mapped to the same space through one-dimensional convolution:

$$F_S^{'} = \text{Conv1d}\left(F_S, C\right) \in R^{C \times T}, \quad F_T^{'} = \text{Conv1d}\left(F_T, C\right) \in R^{C \times T} \tag{1}$$

Calculate the attention weights $\left(\alpha_S, \alpha_T\right)$ through global pooling and sigmoid:

$$\alpha_S = \text{Sigmoid}\left(\text{GlobalAvgPool}\left(F_S^{'}\right)\right), \quad \alpha_T = \text{Sigmoid}\left(\text{GlobalAvgPool}\left(F_T^{'}\right)\right) \tag{2}$$

Final fusion features

$$F_{fusion} = \alpha_S \cdot F_S^{'} + \alpha_T \cdot F_T^{'} \tag{3}$$

2.2. Lightweight convolutional neural network design

A customized feature extraction network based on MobileNetV3 is divided into backbone and feature enhancement head. backbone alternately stacks depth-separable convolution and attention modules [13], and the overall parameter scale is controlled within 1M.

The traditional convolution is split into depthwise convolution and point-by-point convolution, reducing the parameters and computational load. When the input feature map $\left(\text{H} \times W \times C_{in}\right)$ and the output channel $\left(C_{out}\right)$ are used, the performance comparison of the two is shown in Table 1.

Table 1. Performance comparison

Convolution type	Number of parameters	FLOPs	Compression ratio
Traditional convolution	$\left(K^2 \times C_{in} \times C_{out}\right)$	$\left(K^2 \times C_{in} \times C_{out} \times H \times W\right)$	1:1
Depthwise separable convolution	$\left(K^2 \times C_{in} + C_{in} \times C_{out}\right)$	$\left(K^2 \times C_{in} \times H \times W + C_{in} \times C_{out} \times H \times W\right)$	$\left(\frac{1}{C_{out}} + \frac{1}{K^2} \approx \frac{1}{8}\right)$ $(K = 3, C_{out} = 8)$

Using a 3×3 deep convolution kernel, pointwise convolution is followed by BatchNorm and Hard-Swish activation [14], and the output features are:

$$F_{dw} = \text{DepthwiseConv}(F_{in}, K = 3) \setminus F_{pw} = \text{PointwiseConv}(F_{dw}, C_{out}) \setminus F_{out} = \text{Hard-Swish}\left(\text{BatchNorm}(F_{pw})\right) \tag{4}$$

Introduce a simplified CBAM channel attention module to enhance the feature response of key parts of the action. Input the feature $\left(\text{F} \in R^{C \times H \times W}\right)$, and obtain the statistical information through global pooling:

$$F_{avg} = \text{GlobalAvgPool}\left(F\right), \quad F_{max} = \text{GlobalMaxPool}\left(F\right) \tag{5}$$

DOI: 10.54254/2755-2721/2026.31260

Output the attention weights （W） by sharing the fully connected layer:

$$W = \text{Sigmoid}\left(FC\left(FC\left(F_{avg}\right) + FC\left(F_{max}\right)\right)\right) \tag{6}$$

Feature weighted screening results:

$$F_{att} = W \odot F \tag{7}$$

2.3. Real-time evaluation of model optimization

INT8 symmetric quantization Z=0 is used to achieve parameter dimensionality reduction, converting floating-point numbers （x） to integers （x_q）：

$$x_q = \text{round}\left(x/S\right), \quad x = x_q \cdot S \tag{8}$$

Taking MobileNetV3-Large as the teacher model and the model designed in this paper as the student model, the distillation loss function is:

$$\mathscr{L}_o = \alpha \cdot \mathscr{L}_{\mathscr{CE}}\left(y, y_s\right) + (1-\alpha) \cdot \mathscr{L}_{\mathscr{KL}}\left(\text{Softmax}\left(y_t/T\right), \text{Softmax}\left(y_s/T\right)\right) \tag{9}$$

Among them, （T = 4） and （α = 0.3）, the accuracy loss is reduced when the parameters are decreased.

A lightweight LFPN was designed to extract action features of different scales, including an input layer, three feature enhancement blocks and a fusion layer. The parameter configuration is shown in Table 2.

Table 2. Parameter configuration

Feature Level	Input Size	Core Module	Output Channel Number	Function
P1	320×240	1×1 Convolution + Depthwise Separable Convolution	64	Extract action detail features
P2	160×120	Depthwise Separable Convolution + Attention Module	128	Extract action local features
P3	80×60	Depthwise Separable Convolution + Global Pooling	256	Extract action global features
Fusion Layer	-	Upsampling + Element-wise Addition	128	Fuse multi-scale features

Fusion calculation of features at each layer （P_1, P_2, P_3）：

$$P_3' = \text{Upsample}(P_3, scale = 2),$$

$$P_2' = P_2 + P_3' \setminus P_2'' = \text{Upsample}\left(P_2', scale = 2\right), \tag{10}$$

$$F_{LFPN} = P_1 + P_2''$$

DOI: 10.54254/2755-2721/2026.31260

3. Experimental design

Taking the core basic movements of the compulsory gymnastics courses in colleges and universities as the core collection objects, covering 8 typical movements such as forward rolls, 200 gymnastics learners from 3 colleges and universities of different levels were selected as the data collection subjects, including three groups: beginners, moderately proficient, and proficient. The sample proportions of each group were 40%, 35%, and 25% respectively.

Two 1080P high-definition cameras were used to synchronously collect RGB video streams from both the front and side perspectives. Eventually, a dedicated dataset containing 12,000 valid action samples was constructed, which was divided into a training set, a validation set and a test set in a 7:2:1 ratio. All samples were labeled by two gymnastics professional teachers according to the FIG scoring criteria.

The core performance indicators selected include the accuracy rate of action recognition, F1 value, single-frame reasoning delay, model parameter scale, teaching efficiency improvement rate, and the rate of students' action skills reaching the standard.

The traditional evaluation methods are based on the manual scoring results of college gymnastics teachers, and the deep learning methods select the MobileNetV3-Large, ShuffleNetv2, SqueezeNet and ST-GCN models.

4. Experimental results

As shown in Figure 2, the performance comparison results show that the scheme proposed in this paper outperforms traditional manual scoring and most baseline models in terms of performance, only slightly higher than MobileNetV3-Large, but the parameter scale is only 17.6% of it, and the inference delay is reduced by 31.2%. Compared with ShuffleNetv2, the proposed algorithm has a 2.8% increase in accuracy and a 21.4% reduction in inference latency, achieving a balance and fully verifying the effectiveness of customized design for lightweight networks.

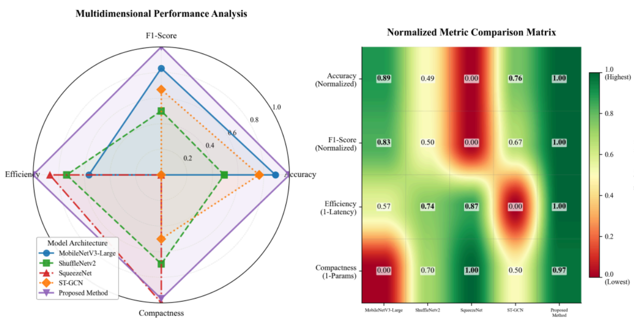

Figure 2. Performance comparison(data source: calculated)

The ablation experiment clearly quantified the contributions of each core module, as shown in Figure 3. This figure uses multidimensional ablation to analyze the performance indicators of various model configurations for gymnastics evaluation algorithms. After the depth-separable convolution was removed, the parameters and delay increased significantly, but the accuracy remained basically unchanged. The attention module can increase the accuracy rate by 3.4%, verifying its role in strengthening key action features. Although the LFPN module adds a few parameters and delays, it can improve the accuracy rate by 2.9%, proving the necessity of multi-scale feature fusion for the recognition of gymnastic movement details.

DOI: 10.54254/2755-2721/2026.31260

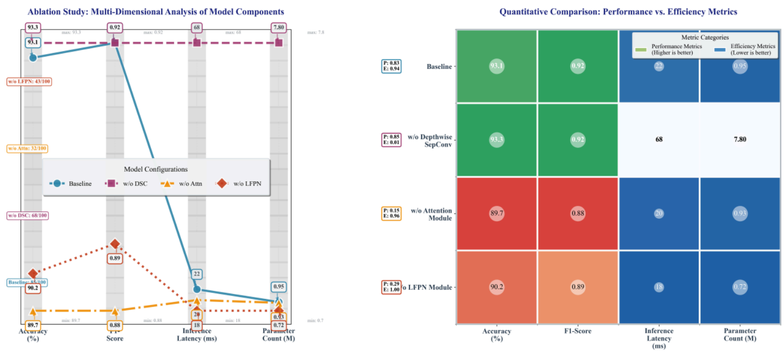

Figure 3. Ablation experiment(data source: calculated)

5. Conclusion

This paper designs a real-time evaluation algorithm that integrates lightweight convolutional neural networks, attention mechanisms and multi-scale feature fusion in view of the limitations of the traditional model of gymnastics teaching evaluation in colleges and universities. Experimental verification shows that this algorithm significantly reduces the parameter scale and inference delay while maintaining high recognition accuracy, effectively balancing lightweight and performance requirements. It is adapted to the classroom scenarios of colleges and universities, providing reliable technical support for the transformation of gymnastics teaching from experience-driven to data-driven, and meeting the development demands of smart sports.

References

[1] Jianpeng, H. (2024). The significance and practical path of constructing a smart physical education teaching system in universities. Curriculum and Teaching Methodology, 7(3), 170-174.
[2] Zhou, M. (2025). Research on the Integration Pathways of Gymnastics Curriculum and Ideological and Political Education: Based on the Integrated Construction of Primary, Secondary, and Tertiary Education Systems. Frontiers in Educational Research, 8(4).
[3] Williams, S. M., & Lacy, A. (2018). Measurement and evaluation in physical education and exercise science. Routledge.
[4] Dai, G. S. (2025). Integrated Teaching Evaluation of Physical Education in Colleges and Universities Based on Multi-class Support Vector Machine. International Journal of High Speed Electronics and Systems, 2540311.
[5] Siedentop, D., & Van der Mars, H. (2022). Introduction to physical education, fitness, and sport. Human kinetics.
[6] Seo, S. D., Madusanka, N., Malekroodi, H. S., Na, C. S., Yi, M., & Lee, B. I. (2024). Accurate Acupoint Localization in 2D Hand Images: Evaluating HRNet and ResNet Architectures for Enhanced Detection Performance. Current Medical Imaging, 20.
[7] Wang, Q., Zhang, K., & Asghar, M. A. (2022). Skeleton-based ST-GCN for human action recognition with extended skeleton graph and partitioning strategy. IEEE Access, 10, 41403-41410.
[8] You, Y. (2025). Lightweight graph convolutional network with multi-attention mechanisms for intelligent action recognition in online physical education. PeerJ Computer Science, 11, e3050.
[9] Yang, K. (2021). Research on the Basketball Goal Recognition Method Based on Improved MobileNet. Scientific Programming, 2021(1), 5862037.
[10] Zhao, G. (2024, April). Research on real-time athlete tracking method based on computer deep learning. In 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology (ICCECT) (pp. 1063-1068). IEEE.
[11] Cai, G., & Zhang, G. (2025). Application of Swin-UNet and pose estimation-based athlete motion quality detection device in IoT systems. Alexandria Engineering Journal, 128, 100-116.
[12] Xiao, D., Zhu, F., Jiang, J., & Niu, X. (2023). Leveraging natural cognitive systems in conjunction with ResNet50-BiGRU model and attention mechanism for enhanced medical image analysis and sports injury prediction. Frontiers in neuroscience, 17, 1273931.

[13] Xie, X. (2024, December). Yoga Pose Detection using MobileNetV3+ pose comparison and instruction. In 2024 4th International Conference on Information Technology and Contemporary Sports (TCS) (pp. 7-12). IEEE.
[14] Zalluhoglu, C., & Ikizler-Cinbis, N. (2020). Collective sports: A multi-task dataset for collective activity recognition. Image and Vision Computing, 94, 103870.