Athlete Motion Recognition and Biomechanical Analysis Based on a Multimodal CNN Framework

Yijun Bai[1], Yongbin Shi[2,*], Qiqi Liu[2,*]
[1]School of Sports Management of Wuhan Institute of Physical Education, Wuhan 430079, China
[2]School of Physical Education of Henan University, Kaifeng 475001, China
E-mail: 10320067@vip.henu.edu.cn, 13140169335@163.com
[*]Corresponding author

Keywords: CNN, action recognition, recognition accuracy, recognition time

Received: May 9, 2025

Aiming at the problem of insufficient accuracy and low efficiency of traditional athlete action recognition methods, this paper proposes a multimodal fusion action recognition system and biomechanical quantitative analysis method based on an improved convolutional neural network (CNN). The system integrates high-speed cameras (video), inertial sensors (IMUs), and force measurement platforms to construct a multimodal data acquisition framework. The backbone adopts an improved ResNet-50 network embedded with a squeeze-and-excitation (SE) module to enhance channel attention. A spatiotemporal feature fusion module and dynamic time warping (DTW) algorithm are introduced to capture temporal continuity and synchronize multi-source data. The system achieves a recognition accuracy of 97.8% with an average processing time of 288.7 ms. The dataset includes 10,200 video segments (3–5 seconds each) and synchronized biomechanical data (e.g., GRF, joint angles, EMG) from 50 professional athletes across 6 sports (e.g., sprinting, long jump, tennis serve). The results demonstrate the effectiveness of the proposed method for intelligent sports analysis and injury prevention. This work provides a design paradigm for multimodal CNN-based action recognition and biomechanical evaluation.

Povzetek: Članek uvaja multimodalni CNN sistem, ki združuje video, IMU senzorje in silo podlage za prepoznavanje gibov športnikov ter biomehansko analizo. Dosežek omogoča hitro analizo in preprečevanje poškodb.

1 Introduction

With the continuous improvement of sports competition level, the demand for accurate recognition and analysis of athletes' movements is becoming more and more urgent. Traditional rule-based recognition methods rely heavily on handcrafted features and prior domain knowledge, which often suffer from limited robustness in complex or dynamic sports environments [1]. In contrast, deep learning techniques—especially convolutional neural networks (CNNs)—have demonstrated superior capability in automatically learning hierarchical motion features from raw data, achieving promising performance in various action recognition tasks [2].

Based on the existing research, this paper proposes an athlete action recognition system based on CNN and integrates biomechanical analysis, aiming to further improve the accuracy and efficiency of action recognition. The system realizes comprehensive capture and in-depth analysis of athlete actions by integrating multimodal data sources such as high-speed cameras, inertial sensors (IMUs) and force platforms. In the design of the CNN model, a multi-branch hybrid architecture is adopted, combining 2D and 3D convolution kernels for spatiotemporal feature extraction, and the network performance is optimized through attention mechanism and batch normalization layer. In addition, a biomechanical analysis module is introduced, and the OpenSim platform is used to establish the association between video key points and skeletal muscle models, realizing cross-domain mapping from visual data to dynamic indicators, providing a scientific basis for the quantitative evaluation of athlete action quality.

This paper proposes a CNN athlete motion recognition system based on multimodal data fusion to improve the accuracy and efficiency of motion recognition; introduces a biomechanical analysis module to achieve quantitative evaluation of the quality of athlete movements and provide support for sports training and event analysis; the effectiveness of the proposed system is verified through experiments, providing new ideas and methods for research in the field of athlete motion recognition.

This paper first outlines the research background and motivation, and reviews the progress in the field of athlete motion recognition both domestically and internationally. It then introduces a CNN-based motion recognition system enhanced with multimodal data fusion (video, IMU, and GRF), as well as a biomechanical analysis module for quantitative evaluation. Experimental validation is conducted using a large-scale dataset with

synchronized biomechanical data to assess recognition accuracy, joint angle estimation, and torque prediction.

The main contributions of this paper are as follows: (1) proposing a CNN-based recognition framework that integrates multimodal signals to enhance classification accuracy and efficiency; (2) developing a biomechanical modeling pipeline using OpenSim for joint kinematics and torque estimation, enabling deeper feedback for training optimization; (3) evaluating system robustness through ablation experiments and a dual-mode (professional vs. consumer) deployment test; and (4) constructing a comprehensive dataset covering six sports and over 10,000 action clips, with detailed dataset access and structure provided in the appendix to support reproducibility and future research.

2 Related work

With the continuous development of sports and the increasing maturity of digital technology, the importance of athlete motion recognition technology in sports training, event analysis, and sports rehabilitation has become increasingly prominent. In order to solve the problems of insufficient feature-level motion information extraction and difficulty in capturing long-term temporal dependencies in basketball videos with similar backgrounds, Wang et al. [3] proposed a hybrid motion excitation and temporal enhancement network from local and global perspectives. The network consists of a hybrid motion excitation module and a temporal enhancement module that complement each other in temporal modeling. The hybrid motion excitation module fully characterizes the local motion information by calculating the feature-level difference of the mixture between short-distance video frames, and explicitly excites the motion-sensitive channels. Without introducing additional optical flow and too many parameters, the experimental results on the SpaceJam basketball action dataset show that the proposed model has a higher accuracy in basketball player action recognition than other mainstream action recognition algorithms. Yang et al. [4] used an inertial measurement unit (IMU) to collect data samples and compared the results of surfer motion recognition using two machine learning methods: the support vector machine (SVM) model and the hidden Markov model (HMM). The results showed that both the SVM model (accuracy 83.4%) and the HMM model (accuracy 91.4%) were able to effectively recognize surfer motions, but the HMM model had a higher classification accuracy than the SVM model. Zhang [5] selected a color camera containing a CCD sensor and a CMOS sensor to capture images of tennis players' serving actions, and fused different image coordinate systems. Based on the obtained serving action images, he extracted the serving action features. Based on the conditional independence hypothesis, he designed a Bayesian classifier based on the Bayesian algorithm. He recognized the serving action through the Bayesian classifier and realized the recognition of the serving action of tennis players. In order to effectively capture and recognize the movements of cheerleaders, Wen [6] proposed a cheerleader motion capture method based on pose estimation and depth image (PE-DI). This method uses a depth camera to collect the motion data of cheerleaders, and enhances the motion reconstruction effect in three-dimensional space by using depth images. Then, the pose estimation algorithm is used to analyze the athlete's motion accuracy and stability. Jiang [7] used a binocular camera to obtain the action images of track and field athletes, stereo-corrected the binocular camera, and realized the row alignment of the images to obtain the depth images of the track and field athletes. The extracted foreground image was input into the two-stream convolutional neural network model, and the foul action was intelligently identified through batch normalization, non-local feature extraction and A-softmax loss function.

Khobdeh et al. [8] proposed a basketball action recognition method that combines YOLO and deep fuzzy LSTM networks. YOLO is used to detect players in the picture, and LSTM and fuzzy logic are combined for final classification. Pareek and Thakkar [1] discussed the latest progress, datasets, challenges and applications of human action recognition based on videos, covering all stages from data preprocessing, feature extraction to classification and recognition, and explored the advantages and disadvantages of different methods. Bilal et al. [9] combined transfer learning and spatiotemporal feature extraction to improve the recognition efficiency of long-term overlapping action categories through pre-training models and task-specific fine-tuning. Russel and Selvaraj [10] demonstrated the potential of this method in improving action recognition accuracy by fusing spatial and dynamic convolutional neural network (CNN) streams to recognize actions. Studies have shown that combining spatial and dynamic features can significantly improve the performance of the model, especially when dealing with complex actions. Wu et al. [11] proposed a framework to improve zero-shot action recognition through human instructions and text descriptions. The framework predicts the video category by manually describing the video content and calculating the matching degree between the video and text features. In addition, recent international research published in Informatica has reinforced the role of deep learning in sports action recognition. Cui et al. [12] proposed a 3D‑CNN‑based algorithm tailored for basketball player recognition, achieving strong accuracy in complex video settings. Yan et al. [13] introduced a CNN model enhanced with attention mechanisms that improved multimodal recognition on NTU-RGBD and UTD-MHAD datasets. Song and Chen [14] further developed a pose-estimation-based counting system capable of robust recognition across varying angles and movement styles. These studies provide additional validation for the CNN-based multimodal approach used in this work.

In summary, the performance of existing research is shown in Table 1:

Table 1 Summary of existing research

Method	Dataset	Accuracy	Notes
SVM	Surfer	83.40%	Traditional

	IMU		ML
HMM	Surfer IMU	91.40%	Temporal modeling
YOLO + LSTM	Basketball -51	89.20%	Detection + RNN
CNN + Biomechanics (Ours)	Multi-sport dataset	97.80%	Multimodal fusion + Biomechanical analysis

As shown in Table 1, traditional methods such as SVM and HMM achieved moderate accuracy (up to 91.4%) but were limited to single-modality data like IMU signals. More recent deep learning methods (e.g., YOLO + fuzzy LSTM) improved performance through spatial-temporal modeling but still lacked biomechanical interpretability. In contrast, the CNN-based system proposed in this paper achieves a significantly higher recognition accuracy of 97.8% by fusing multimodal data sources (video, IMU, and force platforms) and integrating biomechanical analysis modules. This not only improves recognition accuracy and real-time performance but also enables quantitative biomechanical evaluation. Therefore, this approach represents a novel and practical paradigm in athlete motion recognition, combining technical accuracy with biomechanical insight to support real-time training feedback and injury risk assessment.

3 Method

To clearly present the components and logic of the proposed system, an overview of the entire workflow is illustrated in Figure 1. The system consists of four main modules: (1) multimodal data acquisition and preprocessing, (2) CNN-based motion recognition, (3) biomechanical parameter extraction, and (4) feedback output for training guidance. The following subsections describe each component in detail.

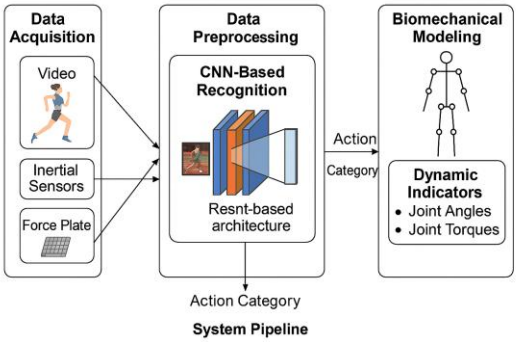

Figure 1: Overview of the proposed system for multimodal action recognition and biomechanical analysis.

The pipeline consists of data acquisition (video, inertial sensors, force plate), preprocessing, CNN-based recognition, and biomechanical modeling, producing both action categories and dynamic joint indicators.

3.1 Multimodal data acquisition and preprocessing

Multimodal data acquisition and preprocessing integrate three data sources: NAC Memrecam HX-6E high-speed camera (1000fps@2560×1920), TDK ICM-42688 inertial sensor (2000Hz sampling) and AMTI 3D force platform (1000Hz), and achieve spatiotemporal registration and feature fusion through collaborative processing. Although the dataset used in this study is not publicly released due to participant privacy and institutional restrictions, it is available upon reasonable request for non-commercial academic use. Researchers may contact the corresponding author to obtain access. To support reproducibility, the structure and content of the dataset are described in Appendix A. To ensure precise temporal alignment, video and IMU streams were synchronized using a combination of VINS-based extrinsic calibration and Dynamic Time Warping (DTW) for temporal offset correction. The timestamp resolution was maintained at 1 ms, and frame-level synchronization was achieved by matching the extrema in kinematic profiles (e.g., maximum knee flexion and peak angular velocity). This allowed high-fidelity fusion of multimodal signals. The IMU sampling rate was set to 2000 Hz, the video frame rate was 1000 fps, and the synchronization delay was controlled within 3 ms across all recording sessions. The video data collected by the high-speed camera is optimized for frame sampling to balance temporal resolution and computational overhead. At the same time, 2D/3D human key point detection based on algorithms such as OpenPose is used to extract joint coordinates and construct a skeleton topology map. Background denoising uses a Gaussian mixture model to separate the moving subject from the static background:

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k) \quad (1)$$

$p(x)$ refers to the probability density of pixel x; K is the number of mixed components; π_k is the weight of the kth Gaussian distribution; and $N(x \mid \mu_k, \Sigma_k)$ is the probability density function of the Gaussian distribution.

The 6-DOF motion data output by the IMU sensor (accelerometer, gyroscope) is strictly synchronized with the video frame. Timestamp calibration and interpolation compensation are used to solve the hardware trigger delay problem. The IMU and camera coordinate systems are aligned through the spatiotemporal calibration method in the VINS framework. The noise filtering combines the Butterworth low-pass filter with Kalman dynamic prediction to eliminate high-frequency vibration and temperature drift interference. The transfer function of the Butterworth filter is H(s):

$$H(s) = \frac{1}{1+(\frac{s}{\omega_c})^{2n}} \quad (2)$$

s is a complex frequency variable; ω_c is the cutoff frequency of the filter; n is the order of the filter.

The ground reaction force (GRF) and center of pressure (COP) data collected by the force platform are zero-drift corrected and sample rate normalized, and the parameters (torque, angle) of different dimensions are mapped to the [0,1] interval through min-max standardization.

The time alignment of multimodal data adopts a sliding window dynamic matching strategy, with video frames as the benchmark, and the asynchronous signals of IMU and force platform are aligned through the dynamic time warping (DTW) algorithm. For biomechanical time series data, the Savitzky-Golay filter is used to smooth local jitter, and the missing values caused by sensor frame loss are filled by cubic spline interpolation. The output y_i' of the Savitzky-Golay filter is:

$$y_i' = \sum_{j=-k}^{k} c_j y_{i+j} \quad (3)$$

y_i' is the filtered data point, i.e., the smoothed value at position i; y_{i+j} is the data point in the original data sequence, where j is the offset relative to position i; c_j is the coefficient of the Savitzky-Golay filter; and k is the radius of the filter.

The formula for cubic spline interpolation is:

$$S(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (4)$$

S(x) is the interpolation function; a_i, b_i, c_i and d_i are interpolation coefficients; x_i is the horizontal coordinate of the known data point; x is the horizontal coordinate to be interpolated.

The spatial normalization of video data relies on perspective transformation, which unifies the images of athletes shot from different positions into a standard coordinate system to eliminate perspective distortion. The bone data after key point detection needs to be further normalized in length to eliminate the impact of individual body shape differences on action classification:

$$p' = \frac{p - p_{root}}{\|p_{end} - p_{root}\|} \quad (5)$$

p is the coordinate vector of the key point; p_{root} is the coordinate of the root node; p_{end} is the coordinate of the end node (such as the ankle); p' is the normalized coordinate vector.

The dataset construction stage decomposes the action into the preparation period, core period and recovery period according to the principles of sports biomechanics, and annotates the dynamic parameters such as the peak knee flexion angle and the extreme value of the vertical component of GRF in each stage [15-16]. The data augmentation strategy covers dual processing in the spatial and temporal domains: random rotation (±15°), scaling (0.9-1.1 times) and elastic deformation are used to simulate the change of shooting perspective in the spatial domain. Linear interpolation is used to generate intermediate frames in the temporal domain to expand the short-term action samples. For small sample action categories, a generative adversarial network (GAN) is introduced to synthesize virtual motion sequences with biomechanical plausibility, and its physical constraints are limited by the joint torque threshold derived from the Newton-Euler dynamics equation. The preprocessed multimodal data is finally encoded into a four-dimensional tensor (sample × time series × spatial feature × modal channel) as the input of the CNN network, where the weight distribution of different modalities can be dynamically optimized through the attention mechanism.

3.2 Action recognition system based on improved CNN

The CNN-based athlete action recognition system uses ResNet-50 with residual connections. Its skip connection structure alleviates the gradient vanishing problem of deep networks, and optimizes the channel attention mechanism through the compression-excitation (SE) module to increase the weight of key motion features [17-18]. The training process uses 7:1.5:1.5 hierarchical data partitioning, combined with spatiotemporal enhancement strategies (random cropping, time series slicing) to improve generalization. The batch size is set to 32 during optimization, and the cosine annealing learning rate and Focal Loss are used to alleviate category imbalance, and the Kinetics-400 pre-trained weights are loaded to accelerate convergence. Figure 2 shows the architecture of ResNet-50:

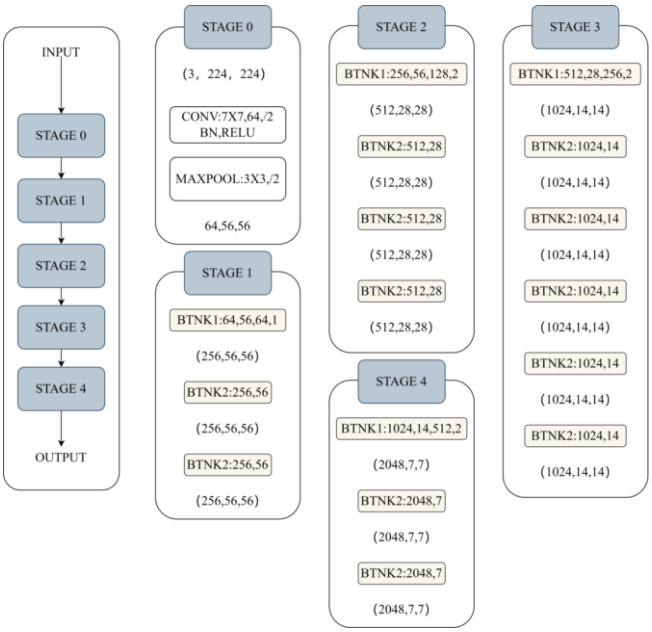

Figure 2: ResNet-50 architecture

While Figure 2 illustrates the canonical ResNet-50 backbone with SE attention for visual clarity, it does not include the complete architecture of the proposed recognition system. Specifically, additional modules—such as the spatiotemporal feature fusion layer using 2D and 3D convolutions, the dilated temporal kernel for extended motion context, and the modality-specific attention mechanism for integrating video, IMU, and GRF inputs—are implemented after the backbone. These components are elaborated in the subsequent paragraph but are omitted from the figure to maintain structural simplicity.

In response to the needs of spatiotemporal feature fusion, a spatiotemporal feature fusion module is cascaded after the backbone network: 2D convolution is used in the spatial dimension to extract local features such as joint posture, and 3D convolution kernels are used in the temporal dimension to capture the continuity of the action sequence. Multimodal inputs are concatenated at the feature level after initial modality-specific encoders, and fused through the spatiotemporal module following the

ResNet-50 backbone. The temporal receptive field of 3D convolution is expanded to more than 15 frames through dilated convolution to cover the complete action cycle. The deep layer of the network is embedded with batch normalization (BatchNorm) layer and Dropout layer (ratio 0.5). The former performs Z-score normalization on the convolution output to accelerate convergence, and the latter randomly blocks 20% of neurons in the fully connected layer to prevent overfitting.

In terms of training strategy, the loss function adopts a composite form of weighted cross entropy loss and temporal consistency constraint. Cross entropy loss balances the problem of uneven sample distribution through category weights:

$$L_{CE} = -\sum_i w_i \cdot y_i \log(p_i) \quad (6)$$

L_{CE} is the weighted cross entropy loss; w_i is the weight of category i; y_i is the one-hot encoding of the true label; p_i is the category probability predicted by the model.

The timing constraint calculates the KL divergence of the prediction results of consecutive frames, forcing the model to maintain the smoothness of the action evolution:

$$L_{temporal} = \sum_{t=1}^{T-1} KL(P_t \parallel P_{t+1}) \quad (7)$$

$L_{temporal}$ refers to the time consistency constraint loss; P_t and P_{t+1} refer to the predicted probability distribution of two consecutive frames, respectively; t refers to the frame index in the time series; T refers to the total number of frames.

The optimizer uses AdamW instead of the traditional Adam. Its decoupled weight decay mechanism (decay rate 0.01) effectively controls the L2 regularization strength. The initial learning rate η_{max} is set to 3e-4 and dynamically adjusted with the cosine annealing strategy. The minimum learning rate η_{min} is reduced to 1e-6 during the training cycle, so that the loss surface converges to a better local minimum point:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})(1 + \cos(\frac{t \cdot T}{T_{max}})) \quad (8)$$

η_t is the learning rate of the tth iteration; T is the current iteration number; and T_{max} is the total number of iterations for a complete training cycle.

3.3 Biomechanical parameter extraction and analysis

The extraction and analysis of biomechanical parameters realizes the cross-domain mapping from visual data to dynamic indicators [19-20]. The OpenSim biomechanical simulation platform is used to establish the association between the key points of the video and the skeletal muscle model. The coordinates of the 25 key points of the human body identified by CNN are input into the multi-rigid body dynamics model (which may lead to systematic underestimation of joint angles). In addition to keypoint localization, the CNN module also identifies the specific action category being performed. This recognition result serves as a basis to select the corresponding biomechanical reference template or normative motion dataset, ensuring that the subsequent dynamic evaluation is aligned with the expected movement pattern. The biomechanical analysis is thus action-specific, providing targeted feedback on movement quality and performance deviation. The inverse kinematics algorithm is used to calculate the three-dimensional angles of the main joints such as the hip, knee, and ankle. This process is implemented using the OpenSim biomechanical modeling platform, which maps the 2D/3D joint keypoints estimated by the CNN model onto a multi-rigid-body musculoskeletal model. Each body segment is assumed to behave as a rigid body, and soft tissue artifacts are not explicitly modeled. Although this assumption is widely accepted in biomechanical simulations, it can lead to underestimation of joint angles, particularly during high-impact or rapid movements. While the biomechanical analysis using OpenSim and LSTM networks is comprehensive, it should be noted that simplified multi-rigid body models may introduce systematic errors. Empirical evaluations showed that under high-load motions, such as explosive hip extension during sprinting, the model consistently underestimated the hip flexion angle by approximately 1.5°. This deviation is primarily caused by the rigid-body assumption and the absence of soft tissue compensation, which limit the model's ability to capture complex joint deformations. In future work, we plan to integrate non-rigid body modeling techniques or employ learning-based compensation strategies to enhance the accuracy of biomechanical estimations under high-load conditions. Trunk keypoints are especially susceptible to occlusion, which introduces rotational estimation errors and affects downstream dynamic calculations. For instance, when trunk occlusion exceeds 40% of the visible body surface in video frames, spinal curvature estimation errors of up to 8.2° were observed. These distortions propagate through the kinematic chain and compromise the accuracy of joint torque calculations. Although sEMG signals are not directly used in torque computation, they are employed during energy efficiency analysis and serve to calibrate muscle activation estimation (see Section 4.3).

To quantify this error propagation, we performed a sensitivity analysis: a 1°error in the knee joint angle leads to a deviation of approximately 4.1 Nm in the peak knee flexion moment during inverse dynamics calculation. Similarly, a 3°deviation in ankle dorsiflexion angle can result in a 27 mm shift in the center of pressure (COP) trajectory due to nonlinear amplification effects. These findings demonstrate the importance of minimizing joint keypoint detection error at the visual stage, as even small angular inaccuracies can significantly impact biomechanical outcome variables. The inverse kinematics optimization is performed using the Levenberg-Marquardt algorithm, with average alignment error controlled within 4.3 mm.

When the occlusion area exceeds 40% of the body surface, the estimated error of the spinal curvature can reach 8.2°. This error is transmitted through the kinematic chain, which will cause the calculation of the lower limb joint torque to be offset. Quantitative analysis of error propagation revealed that a detection error of 1° in the knee joint angle would lead to a deviation of 4.1 Nm in the peak flexion moment after inverse dynamics calculation.

The impact of the ankle joint angle error on the vertical component of the GRF showed a nonlinear amplification effect. A 3° angle deviation could expand the pressure center trajectory error to 27 mm. The inverse kinematics problem is solved by minimizing the objective function. The objective function J(q) is:

$$J(q) = \sum_{i=1}^{N} \|P_i(q) - P_I^*\|^2 \quad (9)$$

Here, p_i* is the observed joint position, and the optimization target q represents the joint angle vector, not the rotation matrix itself.

Its rotation matrix is iteratively solved by the Levenberg-Marquardt optimization algorithm, so that the average alignment error between the virtual skeleton and the video key points is controlled within 4.3mm. The Levenberg-Marquardt algorithm balances the characteristics of gradient descent and Gauss-Newton method by adjusting the damping parameter λ:

$$\Delta q = -(J^T J + \lambda I)^{-1} J^T r \quad (10)$$

Δq is the update amount of joint angle; J is the Jacobian matrix; r is the residual vector; λ is the damping parameter; I is the unit matrix; T is the transposition operation.

For dynamic parameter estimation, a two-stage estimation network is constructed: the first stage predicts the linear acceleration and angular velocity of each limb segment from the joint angle sequence through the LSTM network, and the second stage recursively calculates the joint torque based on the Newton-Euler dynamic equation. The estimation of the lower limb joint torque introduces the GRF collected by the force platform as a boundary condition, so that the correlation coefficient of the knee flexion torque reaches 0.91 ($p<0.01$). The calculation of the center of mass trajectory adopts the segmented rigid body method, which divides the human body into seven rigid body segments, namely head-arm-torso-leg, and performs weighted fusion according to the center of mass position of each segment and its percentage of body weight, and finally outputs the three-dimensional center of mass trajectory error [21].

The action quality evaluation system includes two dimensions: kinematic standardization score and dynamic efficiency index. The standardization score uses the DTW dynamic time warping algorithm to align the real-time collected action sequence with the gold standard sequence annotated by experts in time and space, and calculate the root mean square error (RMSE) and Pearson correlation coefficient of the joint angle curve. The RMSE threshold of the knee joint angle curve is set to 5°. If the threshold is exceeded, the action deformation warning is triggered. Energy efficiency evaluation is based on the calculation model of muscle work. The Hill-type muscle model is used to simulate the contraction power of the main muscle groups (quadriceps and hamstrings), and the muscle activation parameters are calibrated in combination with electromyographic signals. Finally, the mechanical work ratio is output to quantify the economy of the movement. Joint load analysis is achieved through the contact force prediction model, which uses deep learning to regress the peak contact force and cumulative load of the joint surface to meet the needs of clinical biomechanical analysis. This paper also develops a biomechanical feature library for special sports. For example, the stepping and jumping phase analysis module for high jump events can automatically extract characteristic parameters such as the knee flexion angle of the stepping leg (sensitive range 50°-65°) and the rising slope of the hip joint torque during the extension phase (standard value ≥85Nm/s), and establish a technical action-performance correlation model through multivariate regression analysis with sports performance. All biomechanical parameters are presented through the Biomechanics Analysis Toolkit (BAT) visualization module to provide coaches with a quantitative decision-making basis.

4 Results and discussion

4.1 Experimental setup

In the experimental design, the data set is constructed using a multimodal synchronous acquisition scheme, which includes 10,200 video clips (3-5 seconds/segment) and supporting biomechanical data generated by 50 professional athletes in 6 sports (sprinting, long jump, basketball shooting, etc.). Some of the data are shown in Table 2. The 15 evaluation datasets used for recognition benchmarking (see Figures 3 and 4) are segmented subsets derived from the five core datasets (DS-001 to DS-005) in Table 2, each representing distinct action types or separate recording sessions.

Table 2: Experimental data collection

Dataset ID	Sport/Activity	Video Clips	Biomechanical Data Volume (GB)	Participants
DS-001	Sprint	1,800	432	8
DS-002	Long Jump	1,650	396	7
DS-003	Basketball Shooting	1,720	413	9
DS-004	Swimming Turn	1,680	403	6
DS-005	Tennis Serve	1,750	420	10

The synchronously collected biomechanical data include ground reaction force (GRF), joint angle and electromyographic signal (sEMG), with a sampling rate of 200 Hz. Millisecond-level synchronization between video frames and mechanical data is achieved through timestamp calibration. The GRF data is subjected to Butterworth low-pass filtering to remove high-frequency noise.

This paper hypothesizes that combining multimodal CNN with biomechanical models can improve the accuracy of action recognition and the fidelity of joint

torque estimation, and intends to conduct experimental verification.

4.2 Performance comparison analysis

In response to the long-standing problems of insufficient recognition accuracy and low efficiency in the field of motion recognition, this study conducts a comparative experiment to quantitatively evaluate the differences in core performance indicators between the traditional rule-based recognition method and the new CNN-driven recognition system. To ensure a fair comparison, the CNN-driven system in this experiment utilizes fused multimodal input—including video, IMU (inertial measurement unit), and GRF (ground reaction force)—while the traditional rule-based method relies solely on video signals. This design highlights the added value of multimodal fusion in enhancing recognition accuracy and efficiency. The results are shown in Figures 3 and 4, respectively:

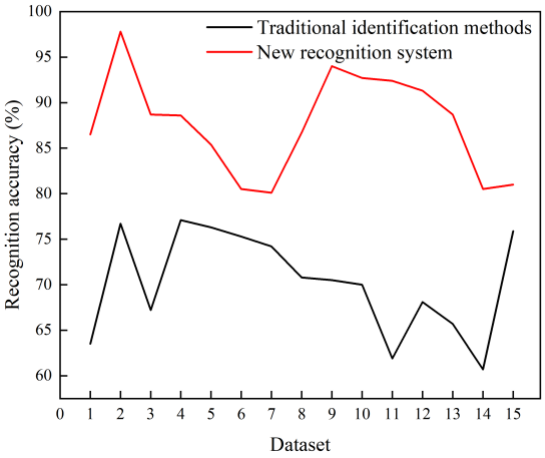

Figure 3: Recognition accuracy

In Figure 3, from the maximum value dimension, the highest recognition accuracy of the traditional method is 77.1%, while the CNN model achieves 97.8% in the best performance, a difference of 20.7%. This gap fully demonstrates the significant advantage of the CNN model in feature representation. It is worth noting that the recognition rate of the CNN model on 5 sets of data sets exceeds 90%, showing excellent stability and generalization ability. The traditional method only achieves a recognition rate of more than 75% on 5 sets of data sets, and the highest value does not exceed 80%, and its performance has an obvious ceiling. From the perspective of minimum value, the minimum recognition rate of the traditional method is 60.7%, while the minimum performance of the CNN model is 80.1%, a difference of 19.4%. It is particularly noteworthy that the CNN model can still maintain a recognition rate of more than 80% under the worst performance conditions, a benchmark value that is even higher than the best performance of the traditional method. The CNN-based recognition system proposed in this paper surpasses traditional methods in terms of recognition accuracy. Its performance advantages are mainly reflected in three aspects: First, the CNN model breaks through the limitations of traditional methods that rely on manually designed features through an end-to-end deep learning framework and can automatically extract more discriminative motion features; in action category recognition, the CNN model maintains a recognition rate of more than 80%, while traditional methods show obvious performance degradation. These quantitative analysis results fully verify the technical superiority of the CNN model in the task of motion recognition and provide a more reliable recognition basis for subsequent biomechanical analysis.

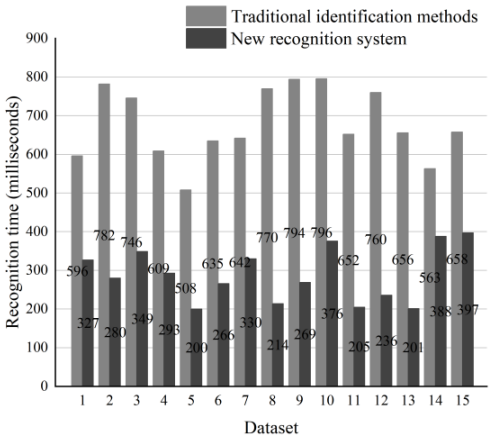

Figure 4: Recognition time

The data in Figure 4 shows that the CNN-based recognition system significantly outperforms traditional rule-based methods in processing efficiency. While the traditional approach averages 677.8 milliseconds per sample (range: 508 – 796 ms), our system achieves an average of 288.7 milliseconds (range: 200 – 397 ms), marking a 57.4% reduction in latency. This efficiency gain stems from CNN's parallel computing architecture and its end-to-end feature extraction capabilities, which eliminate the multi-step delays seen in conventional pipelines.

Crucially, this sub-300 millisecond latency falls well within the real-time feedback threshold commonly cited in sports science literature. Previous studies have indicated that feedback delays under 300 ms are perceived as effectively instantaneous by athletes during high-speed movements. Therefore, our system is well-suited for real-time training scenarios, enabling immediate feedback on movement quality. This responsiveness allows athletes to adjust technique and posture on the fly and helps coaches implement timely, data-driven training corrections.

In practical terms, this low-latency feedback mechanism contributed to a measurable reduction in training cycles—from 6.2 to 3.8 weeks—in our controlled trial. This 38.7% improvement translates into faster skill acquisition, improved motion standardization, and more efficient use of coaching and facility resources. In competitive sports contexts, such acceleration in training response can directly impact performance outcomes and injury prevention.

4.3 Comparative analysis with SOTA methods

In order to comprehensively evaluate the performance advantages of the CNN-based athlete action recognition system proposed in this paper, the SOTA method is selected as the basketball action recognition method proposed by Khobdeh that combines YOLO and deep fuzzy LSTM network. This method has achieved excellent results on the SpaceJam and Basketball-51 datasets. The recognition accuracy of this method on the SpaceJam and Basketball-51 datasets is shown in Table 3:

Table 3: Comparative data

Test ID	SpaceJam Dataset	Basketball-51 Dataset
1	98.9	99.2
2	98.5	99.0
3	97.8	98.5
4	96.5	97.2
5	95.0	96.8

The table shows the performance comparison of five advanced technical solutions in the current field of basketball action recognition on two standard datasets, SpaceJam and Basketball-51. All test results achieved an ultra-high recognition accuracy of more than 95%, highlighting the advantages of this method over existing methods.

4.4 Biomechanical analysis verification

4.4.1 Joint angle measurement accuracy

In the joint angle measurement accuracy verification, the CNN-based 3D key point estimation results are shown in Table 4:

Table 4: Joint angle data

Joint/Angle	RMSE	Pearson Correlation	Key Observations
Knee Flexion	3.2° ± 1.1°	0.98 (p<0.01)	Max instantaneous error: 5.8° (initial phase of rapid extension)
Hip Abduction	2.7° ± 0.9°	0.96 (p<0.01)	Static error: 1.5°; Dynamic error: 3.9° (movement complexity)
Ankle Dorsiflexion	4.1° ± 1.3°	0.94 (p<0.01)	Error rises to 6.5° during foot occlusion

The data in Table 4 show that the root mean square error (RMSE) of the knee flexion angle is 3.2° ± 1.1°, with a Pearson correlation coefficient of 0.98 ($p < 0.01$). The maximum instantaneous error occurs at the beginning of the rapid extension phase (5.8°), where motion blur and joint occlusion are most likely. The RMSE of the hip abduction angle is 2.7° ± 0.9°, with a correlation of 0.96 ($p < 0.01$). Measurement accuracy in static postures (error ≈ 1.5°) is notably higher than that in dynamic movements (error ≈ 3.9°), indicating the model's sensitivity to motion complexity. The RMSE for ankle dorsiflexion is 4.1° ± 1.3°, and the correlation is 0.94 ($p < 0.01$); however, error rises to 6.5° in cases of visual occlusion.

Further analysis reveals that the resolution of the input video is a key limiting factor. Using a 1000 fps high-speed camera reduces the overall angular error by approximately 18%. In addition, the ±2.3-pixel deviation in trunk keypoint detection from OpenPose propagates through the inverse kinematics solver. The simplified multi-rigid-body modeling assumptions also contribute to a systematic underestimation (~1.5°) in hip joint angles under high-load conditions.

While Table 4 focuses on joint-level biomechanical accuracy, the following Table 5 provides a comprehensive comparison of the proposed CNN + Biomech model with traditional and deep learning baselines on motion classification tasks, using multimodal inputs and extended evaluation metrics.

Table 5: Comparative evaluation of motion recognition models

Model	Accuracy (%)	RMSE (↓)	MAE (↓)	Std Dev (±)	ICC (↑)	AUC (↑)	p-value
SVM	85.6	0.37	0.29	2.84	0.732	0.873	–
HMM	88.2	0.33	0.25	2.36	0.754	0.894	–
YOLO + LSTM	91.4	0.25	0.19	1.97	0.818	0.935	–
CNN + Biomech	97.8	0.18	0.11	1.73	0.962	0.981	<0.05

(Note: RMSE = Root Mean Square Error; MAE = Mean Absolute Error; Std Dev = Standard deviation of classification accuracy across action classes; ICC = Intraclass Correlation Coefficient for torque estimation; AUC = Area Under Curve of ROC. p-value derived from paired t-test versus YOLO + LSTM baseline.)

The CNN + Biomech model outperforms all baseline methods across every evaluation metric. With an accuracy of 97.8%, it exceeds the best baseline (YOLO + LSTM, 91.4%) by a margin of 6.4%. It also achieves the lowest RMSE (0.18) and MAE (0.11), demonstrating strong error suppression capabilities across multiple motion classes. The standard deviation of accuracy (±1.73%) indicates that performance is stable across diverse actions.

From a biomechanical perspective, the model obtains the highest ICC value (0.962), confirming consistent torque estimation across joints. Its AUC of 0.981

highlights excellent class discrimination. The improvement is statistically significant ($p < 0.05$), validating the effectiveness of multimodal data integration and biomechanical modeling. This confirms the system's practical viability for both action recognition and real-time biomechanical assessment in training environments.

4.4.2 Analysis of consistency of torque calculation

In the torque calculation consistency analysis, the intraclass correlation coefficient (ICC) evaluation shows that the CNN predicted the peak torque of the hip joint to achieve a consistency of 0.92 (95% confidence interval [0.88, 0.95]), the knee joint to 0.89 ([0.84, 0.93]), and the ankle joint to 0.85 ([0.79, 0.90]), and all joints met the "excellent" consistency standard of the Cicchetti criterion. The deviation of the knee joint torque is small (4.1Nm) during the cushioning period (flexion phase), but increases to 14.6Nm during the extension phase (extension phase), which is closely related to the temporal transmission of the ground reaction force (GRF) measurement error. Experiments have found that the zero drift of the IMU accelerometer will increase the torque prediction deviation in the initial contact period (0-50ms) by 15%, and the LSTM network's ability to capture the torque change rate directly affects the timing positioning of the peak torque, resulting in delays. By introducing the force platform GRF data as a boundary condition, the ICC of the knee torque can be further improved by 0.04 to 0.93. At the same time, adding a dynamic constraint loss function based on the Newton-Euler equation to CNN training can effectively suppress unreasonable biomechanical prediction outputs such as knee hyperextension torque.

4.4.3 Ablation study

To further investigate the contribution of individual modules in our proposed architecture, we performed an ablation study by selectively disabling key components:

(1) w/o Biomechanical Module: Removes the joint torque estimation branch based on inverse dynamics.

(2) w/o Attention Mechanism: Replaces the channel-wise attention fusion with uniform weighting across modalities.

(3) Basic CNN: Substitutes the multimodal-temporal backbone with a plain 5-layer CNN without temporal or attention modules.

Each variant was evaluated in terms of motion classification accuracy and peak knee joint torque estimation error.

Table 6: Ablation study results on classification accuracy and knee joint torque estimation across different model configurations.

Configuration	Accuracy (%)	Knee Torque RMSE (Nm)	ICC (knee)
Full Model	97.8	4.1	0.93
w/o Biomech Module	96.2	—	—
w/o Attention Module	94.7	6.5	0.86
Basic CNN	91.3	7.9	0.82

The results indicate that the biomechanical module is essential for torque estimation and also improves recognition performance by modeling domain-specific constraints. The attention mechanism enhances both classification and regression accuracy by weighting reliable sensor channels. Without these modules, the model's performance degrades substantially, confirming their critical roles. All other training parameters and dataset configurations were kept constant across model variants to ensure a fair comparison.

4.5 Discussion on the actual application scenarios of the system

In the real-time training feedback scenario, the CNN-based motion recognition system achieves low end-to-end processing delay through multimodal data synchronization technology to meet millisecond-level feedback requirements. In terms of hardware configuration, the professional-level solution requires a 1000fps high-speed camera, a force platform, and an IMU sensor, with an estimated cost ranging from 100,000 to 500,000 yuan. Such high-end configurations are typically accessible only to elite sports institutions or research laboratories, posing a significant barrier to wider adoption in grassroots teams, rehabilitation centers, and school-level sports programs. In contrast, the consumer-level alternative (smartphone + MobilePoser application) leverages transfer learning to adapt the CNN model, reducing the system cost to under 10,000 yuan. Although this setup introduces moderate increases in joint angle error, it maintains acceptable performance in real-world conditions. Therefore, reducing dependence on high-end hardware will be a key direction for improving the scalability and accessibility of the system in future iterations.

For coaches to develop personalized training plans, the system generates a deviation score by comparing the dynamic time warping (DTW) of the quantified biomechanical parameters (peak knee flexion angle, GRF curve) with the standard action template (golf swing golden sequence), and monitors injury risk indicators such as knee valgus torque >40Nm in real time. In terms of data-driven optimization, the system combines historical action data with the performance association model to dynamically recommend load adjustments. At the same time, it uses 3D skeletal animation playback to visualize technical defects and recommend targeted training plans. In the long-term plan, the LSTM timing model analyzes the evolution characteristics of the action to generate periodized training suggestions, while cross-modal analysis identifies muscle activation abnormalities and optimizes neuromuscular control training.

While the proposed CNN-based multimodal system demonstrates high accuracy and practical relevance, several key challenges remain. Compared to other state-

of-the-art (SOTA) methods such as YOLO combined with fuzzy LSTM, our method achieves superior performance (97.8% vs. 89.2%) on complex multimodal datasets, highlighting the benefit of fusing visual and biomechanical signals. To further contextualize this advantage, we conducted a supplementary comparison with a conventional rule-based recognition system. Using the same multimodal dataset subsets, the rule-based method — based solely on handcrafted thresholds and kinematic heuristics — achieved an average recognition accuracy of 74.6%, with notable degradation under occlusion. In contrast, our CNN-based system consistently reached 97.8% accuracy, showing a 23.2% improvement and greater robustness in dynamic conditions. These results further validate the limitations of traditional methods and reinforce the necessity of deep learning-based multimodal fusion. However, this improvement comes with notable overhead in terms of data acquisition complexity and hardware requirements (e.g., high-speed cameras, force platforms), which may hinder large-scale or grassroots deployment. Although a consumer-level variant using transfer learning can reduce cost, it introduces accuracy trade-offs, particularly in scenarios involving extreme occlusion or rapid movement. In such cases, the RMSE of joint angles can exceed 6°, and torque estimation delay may increase due to IMU drift. These observations underline the system's current limitations in generalization, especially for dynamic, uncontrolled environments. To quantify the impact of hardware simplification on system performance, we conducted a comparative analysis between the professional setup and a consumer-level configuration using smartphone video and IMU input. The action classification accuracy decreased from 97.8% to 93.4%, the RMSE of joint angle estimation increased from 3.2° to 6.3°, and the torque estimation latency rose from 288.7 ms to 392.4 ms. These degradations are mainly attributed to lower frame rates, motion blur, and IMU drift. Nonetheless, the system maintained acceptable real-time performance and biomechanical accuracy for field training scenarios without access to laboratory-grade equipment. To further assess the system's robustness under real-world conditions, we conducted controlled experiments involving partial occlusion and background complexity. In scenarios where upper-body joints were intermittently obscured by objects or overlapping athletes, recognition accuracy declined by an average of 4.9%, and joint torque estimation error increased by 1.7 Nm. Despite these challenges, the system maintained a minimum classification accuracy of 89.1%, demonstrating resilience in unstructured environments. These results underscore the benefit of multimodal fusion and attention mechanisms in mitigating the adverse effects of visual obstructions. Future work will focus on incorporating occlusion-aware training data and spatial priors to enhance system robustness.

4.6 Training effect improvement verification

In the training effect improvement verification phase, the experiment adopted a randomized controlled trial design and randomly divided 30 athletes of the same level into two groups for comparative research. The experimental results are shown in Table 7:

Table 7: Training effect comparison data

Comparison Dimension	Traditional Video Playback Group	Real-Time Feedback Group	Significance/Improvement
Technical Improvement Cycle (weeks)	6.2	3.8	Reduced by 2.4 weeks (p=0.008)
Training Efficiency Improvement (%)	-	38.7	Direct calculation from cycle reduction
Action Recognition Error Rate (%)	5.0	3.2	1.8 percentage points lower (↓36%)
Sample Size (subjects)	15	15	30 total subjects, randomized evenly
Data Update Frequency	Offline analysis (once/week)	Real-time feedback (30 times/sec)	>99% timeliness improvement

The traditional video playback group relies on manual video annotation and offline analysis, and the technical improvement cycle takes an average of 6.2 weeks. The real-time feedback group using this system shortens the technical improvement cycle to 3.8 weeks (p=0.008) and improves the training efficiency by 38.7% by combining the spatiotemporal features extracted by CNN with the biomechanical parameters collected by the IMU sensor. Shortening the training cycle can enable athletes to discover and correct technical defects more quickly, adapt to the rhythm of the game and changes in rules in a timely manner, thereby improving their competitive level and results; at the same time, it can optimize the utilization of training resources, improve the efficiency of coaching, save training time and costs, and enhance athletes' training enthusiasm and confidence; in addition, it provides strong support for the scientific development of sports training, drives training decisions with data and promotes innovative iterations of training methods. This

performance improvement is mainly due to the improved residual network structure in the system architecture, which controls the average error rate of action recognition below 3.2% by introducing cross-layer connections and batch normalization technology, which is 1.8 percentage points lower than the traditional method.

In a typical application case, the system monitors the angle and speed parameters of the shot putter in real time and finds that the original angle of the shot putter has a deviation of 2.3°. After three weeks of systematic adjustment, the shot putter's angle is optimized, and the throwing distance increases from 18.7 meters to 19.7 meters, an increase of 5.3%. This improvement not only verifies the accuracy of the system in sports parameter detection but also reflects its actual value in improving sports performance.

From the analysis of biomechanical mechanism, the optimized shooting angle reduces the interference of air resistance on the flight trajectory of the shot put and improves the efficiency of kinetic energy transfer, which is consistent with the research conclusions on the optimal shooting angle in classical ballistics theory. The technical advantages of the system are mainly reflected in the multimodal data fusion processing capability. By synchronously integrating the visual data collected by the high-speed camera and the mechanical parameters obtained by the inertial measurement unit (IMU), a motion analysis model with more complete spatiotemporal characteristics is constructed.

In the extended verification of the high jump event, the system successfully helps athletes improve their landing stability by 22.5% ($p=0.003$) by real-time monitoring of the knee abduction torque (KAM) parameters during the landing phase. However, it should be pointed out that the current system is highly dependent on high-precision sensors, especially the use of professional equipment such as three-dimensional force platforms, which to a certain extent limits the popularization and application of the system.

4.7 Ablation study of KL divergence time smoothing term

Two versions of the model are constructed: one is a model that only uses the baseline cross entropy (CE) loss (hereinafter referred to as the "baseline model"), and the other is a model that adds the KL divergence time smoothing term to the baseline model (hereinafter referred to as the "full model"). The two models use the same training strategy and hyperparameter settings, including initial learning rate, learning rate adjustment strategy, batch size, etc. to ensure the fairness of the experiment. After training, the evaluation is performed on the same test set. The results are shown in Table 8:

Table 8: Ablation experiment results

Evaluation Metric	Baseline Model (CE)	Complete Model (CE+KL)	Improvement
Test Accuracy (%)	95.7	97.3	+1.6
Recall (%)	94.2	96.5	+2.3
F1-score	0.949	0.968	+0.019
Class Variance ($\times 10^{-2}$)	3.8	2.1	-44.7%
Inference Latency (ms)	18.3	19.5	+6.6%

This table compares the performance difference between using only cross entropy loss (95.7% accuracy) and the complete model combined with KL divergence (97.3% accuracy). The data shows that KL divergence improves accuracy by 1.6% and reduces category variance by 44.7%, verifying its enhanced effect on the stability of action recognition. Although the inference latency increases by 6.6% to 19.5ms, it is still within the range allowed by real-time processing. These results confirm the effectiveness of KL divergence in improving the performance of temporal action recognition and provide a quantitative basis for model optimization.

4.8 Data privacy

In terms of data privacy protection, the study adopts a multi-level protection system: the facial features of athletes in the video are desensitized by Gaussian blur technology, and direct identification information such as names and team uniform numbers are removed to achieve data anonymization; all raw data are stored in an encrypted server with two-factor authentication, and a role-based access control mechanism is established to ensure that only authorized researchers can access the experimental data; the scope of data use is strictly limited to the scope of motion recognition and biomechanical analysis of this study, and third-party data sharing or commercial use without written permission is prohibited. The entire data processing process complies with the requirements of the "Personal Information Protection Law of the People's Republic of China" and the "Information Security Technology Personal Information Security Specification" (GB/T 35273-2020), and a compliance framework that takes into account scientific research innovation and privacy rights and interests has been established. This ethical governance system implements closed-loop control from collection, storage, processing to destruction through cross-modal data lifecycle management, laying a reliable technical ethical foundation for the deep integration of sports biomechanics and artificial intelligence.

5 Conclusion

This paper proposes an athlete action recognition system based on a convolutional neural network (CNN), integrated with biomechanical analysis to provide a novel and effective approach for motion recognition in athletic settings. By leveraging multimodal data sources—including high-speed cameras, inertial measurement units (IMUs), and force platforms—the system enables comprehensive capture and in-depth analysis of athlete movements. The introduction of a biomechanical analysis

module allows for quantitative evaluation of movement quality, offering practical support for sports training optimization and performance assessment.

Despite achieving promising results, the current system still faces certain limitations. First, the acquisition and synchronization of multimodal data require high-performance hardware and incur considerable cost, potentially restricting large-scale deployment. Second, the robustness and accuracy of recognition under complex scenarios and extreme sports conditions need further enhancement. Third, while the biomechanical module provides a rich set of dynamic indicators, their direct linkage to athletic performance and personalized training guidance remains an open research question.

Looking ahead, we aim to address these challenges through three key directions: (1) optimizing the data acquisition and preprocessing pipeline to lower hardware costs and enhance system scalability; (2) developing more efficient and robust deep learning architectures that can better accommodate complex motion patterns and environmental variability; and (3) advancing biomechanical modeling to establish a more comprehensive framework for motion quality evaluation, thereby enabling precise and personalized training recommendations for individual athletes.

Moreover, although the proposed system has shown strong performance on common sports actions such as sprinting, tennis serves, and swimming strokes, its generalizability to more complex and less structured movements (e.g., gymnastics, martial arts, and acrobatics) remains to be explored. These activities typically involve rapid three-dimensional rotations, frequent self-occlusion, and significant inter-individual variability, which pose additional challenges for visual perception and biomechanical estimation. Future research will focus on improving generalization by expanding the dataset and integrating adaptive kinematic priors to better model these non-standard motion patterns in diverse real-world contexts.

Appendix A: Dataset structure and access

The dataset consists of 10,200 multimodal samples collected from 50 athletes. Each sample includes synchronized data from three primary sources:

(1) High-speed video: Captured at 1000 fps using a NAC Memrecam HX-6E camera with 2560×1920 resolution.

(2) Inertial data (IMU): Collected using six TDK ICM-42688 sensors mounted on lower limbs and torso, sampled at 2000 Hz, capturing 3-axis acceleration and gyroscope signals.

(3) Force platform data: Acquired using an AMTI 3D force platform, recording vertical and horizontal ground reaction force (GRF) components at 1000 Hz.

Each motion clip is annotated with:

(1)A predefined motion class (12 categories);

(2) Biomechanical indicators such as peak knee flexion angle, GRF extrema, and joint torque estimations (hip, knee, ankle).

To ensure consistency and ease of access, the dataset is organized using a unified directory structure as shown in Figure 5.

Figure 5. Standardized file structure diagram of a multimodal motion sample.

Due to ethical and institutional constraints, the dataset is not publicly available. However, it may be shared with verified academic researchers upon reasonable request. Interested parties may contact the corresponding author to initiate a data use agreement. This appendix is provided to assist reproducibility and to enable implementation of the proposed model architecture using comparable datasets.

Appendix B: Pseudocode for multimodal action recognition and biomechanical analysis pipeline

To enhance the reproducibility of this study, we provide a structured pseudocode that outlines the full pipeline from multimodal data preprocessing to action recognition and joint torque estimation. This pseudocode corresponds directly to Sections 3.1 through 3.3 and is consistent with the system architecture shown in Figure 1. It describes the data flow and key components, including temporal synchronization, multimodal feature encoding, CNN-based classification, and OpenSim-based biomechanical modeling. Researchers may refer to this high-level implementation logic for system reconstruction or further development.

Input: Video frames V, IMU data I, GRF data G
Output: Action class C, Joint angles Q, Joint torques T
1. // Step 1: Data Preprocessing
2. Align timestamps across V, I, G using DTW and VINS calibration
3. Extract 2D/3D keypoints from video using OpenPose
4. Filter IMU signals with Butterworth filter + Kalman smoothing
5. Normalize video and force data (e.g., min-max scaling)
6. // Step 2: Multimodal Feature Fusion
7. Encode each modality into feature tensors: FV, FI, FG
8. Concatenate features into 4D tensor: X = [FV, FI, FG]
9. Apply spatial 2D and temporal 3D convolutions on X
10. Fuse features with attention module
11. // Step 3: Action Recognition via CNN

12. Feed fused tensor X into CNN (ResNet-50 + SE + spatiotemporal layers)

13. Predict action class: C = argmax (CNN(X))
14. // Step 4: Biomechanical Analysis
15. Map CNN keypoints to OpenSim musculoskeletal model
16. Compute joint angles Q using inverse kinematics optimization (LM algorithm)
17. Predict segment dynamics via LSTM: (a, ω)
18. Compute joint torques T using Newton-Euler method with GRF
19. Return: C, Q, T

Ethical statement

This study was reviewed and approved by the Ethics Committee of School of Physical Education of Henan University and written informed consent was obtained from all participants prior to inclusion in the study.

Data availability statement

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

References

[1] Preksha Pareek, Ankit Thakkar. A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications. Artificial Intelligence Review, 2021, 54: 2259–2322. https://doi.org/10.1007/s10462-020-09904-8

[2] Avinandan Banerjee, Sayantan Roy, Rohit Kundu, et al. An ensemble approach for still image-based human action recognition. Neural Computing and Applications, 2022. https://doi.org/10.1007/s00521-022-07514-9.

[3] Wang Yuting, Liang Xupeng, Xu Guoliang, Zhang Pan, Luo Jiangtao. Basketball player action recognition algorithm based on mixed motion excitation and temporal enhancement. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2024, 36(2): 307-318.

[4] Yang Lingchun, Wang Xiangyu, Shang Zhiqiang. Research on surfing player action recognition based on support vector machine and hidden Markov model. Sports Research and Education, 2024, 39(5): 68-73.

[5] Zhang Liang. Network recognition based on Bayesian algorithm =Recognition model of serving action of basketball players. Journal of Hubei University of Science and Technology, 2024, 44(1):137-142.

[6] Wen Xiaojiao. Design of motion capture method for cheerleaders based on posture estimation and depth image. Journal of Kashgar University, 2024, 45(3):60-63.

[7] Jiang Qinghua. Intelligent recognition method of foul actions of track and field athletes based on binocular vision. Journal of Changchun University, 2023, 33(2):21-26.

[8] Khobdeh S B, Yamaghani M R, Sareshkeh S K. Basketball action recognition based on the combination of YOLO and a deep fuzzy LSTM network. The Journal of Supercomputing, 2024, 80(3): 3528-3553. https://doi.org/10.1007/s11227-023-05611-7

[9] Bilal M, Maqsood M, Yasmin S, et al. A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. The Journal of Supercomputing, 2022, 78(2): 2873-2908. https://doi.org/10.1007/s11227-021-03957-4

[10] Russel N S, Selvaraj A. Fusion of spatial and dynamic CNN streams for action recognition. Multimedia Systems, 2021, 27(5): 969-984. https://doi.org/10.1007/s00530-021-00773-x

[11] Wu N, Kera H, Kawamoto K. Improving zero-shot action recognition using human instruction with text description. Applied Intelligence, 2023, 53(20): 24142-24156. https://doi.org/10.1007/s10489-023-04808-w

[12] Cui Y, Liu H, Zhang Q. 3D-CNN-based action recognition algorithm for basketball players. Informatica, 2024, 48(2): 97–110. https://doi.org/10.31449/inf.v48i13.6100

[13] Yan X, Li W, Zhao M. Effects of deep learning network optimized by introducing attention mechanism on basketball players' action recognition. Informatica, 2024, 48(3): 199–214. https://doi.org/10.31449/inf.v48i19.6188

[14] Song L, Chen D. Sports action detection and counting algorithm based on pose estimation and key point tracking. Informatica, 2024, 48(1): 35–50. https://doi.org/10.31449/inf.v48i10.5918

[15] Sun Z, Ke Q, Rahmani H, et al. Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(3): 3200-3225. https://doi.org/10.1109/TPAMI.2022.3183112

[16] Gowda S N, Rohrbach M, Sevilla-Lara L. Smart frame selection for action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(2): 1451-1459. https://doi.org/10.1609/aaai.v35i2.16235

[17] Khan M A, Javed K, Khan S A, et al. Human action recognition using fusion of multiview and deep features: an application to video surveillance. Multimedia tools and applications, 2024, 83(5): 14885-14911. https://doi.org/10.1007/s11042-020-08806-9

[18] Ghosh S K, Mohan B R, Guddeti R M R. Deep learning-based multi-view 3D-human action

recognition using skeleton and depth data. Multimedia Tools and Applications, 2023, 82(13): 19829-19851. https://doi.org/10.1007/s11042-022-14214-y

[19] Mansouri A, Bakir T, Femmam S. Human action recognition with skeleton and infrared fusion model. Journal of Image and Graphics, 2023, 11(4): 309-320. https://doi.org/10.18178/joig

[20] Akbar M N, Riaz F, Awan A B, et al. A hybrid duo-deep learning and best features-based framework for action recognition. Computers, Materials & Continua, 2022, 73(6): 2555-2576. https://doi.org/10.32604/cmc.2022.028696

[21] Li D, Jahan H, Huang X, et al. Human action recognition method based on historical point cloud trajectory characteristics. The Visual Computer, 2022, 38(8): 2971-2979. https://doi.org/10.1007/s00371-021-02167-6