Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **10**, Number **2**, May **2025**

Fusion of Computer Vision and Inertial Capture for Basketball Player Motion Recognition

Bo–Xuan Xia*

St. Paul University Philippines, Tuguegarao 3500, Philippines
18647239797@163.com

Lin–Zhen Li

School of Educational Management and Development
Mahasarakham University, Mahasarakham 44150, Thailand
chenzhi201808@163.com

*Corresponding author: Bo–Xuan Xia
Received May 1, 2024, revised August 25, 2024, accepted December 9, 2024.

Abstract. *Due to the error of visual analysis, the movement recognition technology of basketball players generally suffers from the issues of dense movement and insufficient feature extraction, which leads to the low accuracy and poor stability of traditional recognition methods. Intending to the above issues, this article suggests a basketball player action recognition method that integrates computer vision and inertial capture. Firstly, the inertial sensor is adopted to obtain the basketball player's movement data, and the collected data are processed by data denoising, error correction, gravity acceleration removal, etc. to obtain the actual movement data. Relied on the acquired human motion data, extracting the temporal characteristics of human motion by taking the extreme difference as the eigenvector value of the human body's overall motion amplitude, and then realize motion capture. Secondly, on the ground of the basketball video frames, extracting the skeletal information relied on the coordinates of the athletes' joints, construct a directed graph of the skeleton structure, and build a graph convolutional neural network to extract the spatial features of the movements. Finally, the spatial and temporal features of the action are fused by LSTM, and input into softmax for classification, so as to realize the accurate recognition of the action. The experimental outcome on the SpaceJam basketball action dataset indicates that the suggested method has higher accuracy and better stability than other recognition methods.*
Keywords: Computer vision; Inertial capture; Motion recognition; Graph convolutional neural networks; Feature fusion

1. **Introduction.** As the computer vision technology rapidly grows, video analysis technology has been more and more widely used in various sports [1] to solve the problem of relying on expensive auxiliary sensing equipment and manual guidance by coaches in traditional training methods. In sports video analysis, recognizing athletes' technical movements from videos is a crucial part [2], which can promote the development of smart sports such as automatic sports video narration [3, 4] and team analysis [5]. Basketball is a sport that requires very strict technical movements, and accurate recognition of each player's movements on the court can provide coaches and analysts with guidance on technique and is the basis for analyzing team tactics and strategies [6, 7]. Therefore, the application of deep learning and computer vision techniques to recognize the movements of players in basketball videos is of extensive and important practical significance.

1.1. **Related work.** In the field of basketball player action recognition, researchers have adopted different sensor technologies and data processing methods to improve the accuracy and efficiency of recognition. Ren and Wang [8] implemented an inertial sensor–based action recognition system for basketball players, which preprocesses the collected acceleration data, then uses wavelet transform to obtain classification features, and finally realizes offline recognition of actions. Li et al. [9] analyzed the characteristics of acceleration data of different wrist gestures and then used a discrete Hidden Markov Model (DMM) to classify different shooting actions. Sun and Ma [10] used inertial sensors to capture velocity characteristics of basketball players and employed SVM for movement classification.

Data collected from inertial sensors have complex structures, and manual feature extraction modules may limit generalization. In recent years, with advancing research in computer vision and deep learning, it has been found that deep models can automatically learn abstract features from images and videos. Wang et al. [11] proposed a Time Segmentation Network (TSN), which segments input competition video isochronously and feeds all segments into a two-stream network to predict action categories. Ahmadi et al. [12] proposed a Temporal Relation Network (TRN) based on TSN at various scales, improving overall efficiency. Li and Gu [13] introduced a Linear Dynamical System (LDS)-based framework capable of capturing temporal and spatial information of athletes. Sousa Lima et al. [14] first captured inertial data, then generated a 2D spectral image via Fourier transform of the 1D time-series signals, and used a deep neural network for automatic feature identification. Kumar et al. [15] similarly used inertial sensor signals and a deep network with Softmax for movement feature recognition. Kong et al. [16] integrated 2D convolution with SVM for player movement recognition. Zuo and Su [17] applied 3D convolution to basketball video, recognizing multiple movements of a single player, though gradient explosion and overfitting limited performance. Xiao et al. [18] combined residual structures in a 3D convolutional network. Khobdeh et al. [19] constructed an LSTM network based on skeletal information to capture co-occurrence between LSTM structure and joint data. Liu and Che [20] used a Graph Convolutional Network (GCN), representing joints as graph nodes and limbs as edges. Feng et al. [21] combined GCN with SVM to capture spatial key-point information and classify temporal features.

1.2. **Contribution.** Existing basketball player movement recognition techniques generally suffer from background interference, dense movement and limb occlusion, resulting in low accuracy. To address these issues, this article proposes a method integrating computer vision and inertial capture. First, inertial sensor data acquisition is optimized and preprocessed for noise reduction. Then the extreme difference is used as the eigenvector value of the body's overall motion amplitude to extract temporal characteristics and realize motion capture. Second, inter-frame differencing processes basketball video to extract image shapes, constructs a directed spatial skeleton map, and builds a GCN to parameterize this map for spatial feature extraction. Finally, spatio-temporal features are fused and input into Softmax for classification, achieving accurate movement recognition.

2. Theoretical analysis.

2.1. **Inertial sensor.** The inertial sensor is a wearable device used to capture athletes' movement characteristics. It is mainly composed of an inertial data acquisition device, a data transmission device, a data receiving device, and a data processing device [22], as implied in Figure 1.

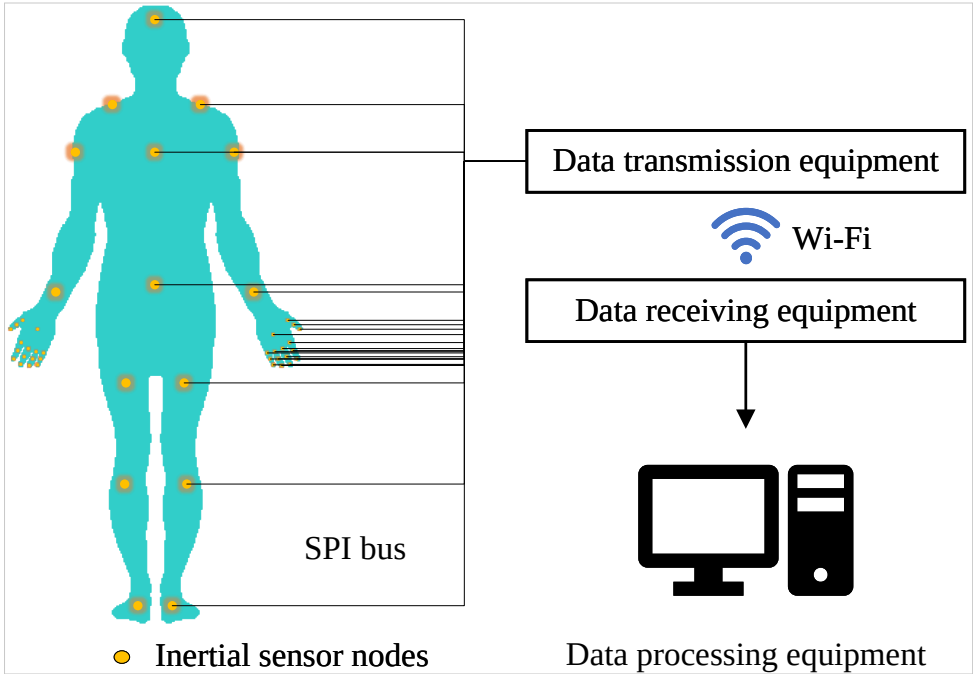

Figure 1. Components of inertial sensors.

2.2. **Graphical convolutional neural network.** Conventional convolutional neural network (CNN) can only process Euclidean data and cannot be used to process non-Euclidean structured graph data. However, the joints in skeletal data belong to a non-Euclidean structure [23]. For data consisting of graph structures, structural invariance cannot be guaranteed when performing conventional convolution operations, and therefore cannot be processed by conventional convolution. To efficiently process complex graph data, a GCN can be utilized to update the features of nodes and act as a feature extractor [24].

The graph convolution is used to update the node features by aggregating the features of neighboring nodes. In CNN, the features of neurons in the next layer are obtained by aggregating features from a region in the previous layer; applying this idea of local connection to GCN, the iteration of node features is given by

$$G^{(k+1)} = A\, G^{(k)}\, V^{(k)} \tag{1}$$

where $G^{(k)}$ is the characteristic graph of the nodes in the k-th layer, A is the adjacency matrix of the graph, and $V^{(k)}$ is the weight parameter in the k-th layer. This calculation does not account for each node itself, so A is replaced by $\tilde{A} = A + I$, where I is the identity matrix. In this way, the features of neighboring nodes are summed with the features of each node. To avoid changing the feature distribution, the rows and columns of $\tilde{A}$ are symmetrically normalized using the degree matrix C, i.e. $C^{-1/2}\,\tilde{A}\,C^{-1/2}$. Thus Equation (1) becomes

$$G^{(k+1)} = C^{-1/2}\,\tilde{A}\,C^{-1/2}\,G^{(k)}\,V^{(k)} \tag{2}$$

3. **Preprocessing of basketball players' movement data based on inertial sensors.** When using inertial sensors to collect basketball players' movement data, the inertial sensor measurements will contain errors. It is therefore necessary to remove noise, perform error correction, remove gravity acceleration, and apply other processing to obtain the actual data of the players' movements.

(1) Data denoising. After data acquisition, the binary complement data is converted into acceleration data by a data conversion method, and data preprocessing is carried out

to lay the foundation for human motion capture. The noise signal in the human body motion data is obtained by Gaussian filtering, and the signal is decomposed according to the wavelet decomposition method [25] to calculate the noise variance in the data.

$$\begin{cases} z(s) = y(s) + \mathfrak{g}(s), \\ z_k^i = t_k^i + \mathfrak{g}_k^i, \\ \mathrm{var}[z_k^i] = \xi_j^2 \beta^{-i} + \xi_{\mathfrak{g}}^2. \end{cases} \tag{3}$$

where the noise signal of the action data is $z(s)$, the fractal noise of the data is $y(s)$, the white noise is $\mathfrak{g}(s)$, the wavelet-decomposed noise signal is z_k^i, the fractal transformation vector is t_k^i, the white noise variable vector is $\mathfrak{g}_k^i$, the wavelet scale is i, the sample point is k, the computed noise variance is $\mathrm{var}[z_k^i]$, the fractal intensity is ξ, the fractal vector is β, and the white noise intensity is $\xi_{\mathfrak{g}}^2$. The noise signal is then reconstructed according to the decomposition result to obtain the filtered data.

(2) Error correction. Set the ideal acceleration value of the sensor to be b_1 and the measured average value to be b_m. Calculate the zero-deviation error as follows:

$$b_{\mathrm{error}} = b_m - b_1 \tag{4}$$

where b_{error} is the zero-deviation error of the sensor. During data acquisition, the sensor measurement can be used directly to remove this bias error and eliminate integral accumulation.

(3) Gravity acceleration removal and displacement calculations. Since inertial sensors are affected by gravity during acquisition, an acceleration sensor is used to calculate displacement:

$$\begin{cases} \varphi_m(s) = \displaystyle\int b_m(s)\,\mathrm{d}s = \sum_{i=1}^{N} b_{m_i}\,\Delta s, \\ e_m(s) = \displaystyle\int \varphi_m(s)\,\mathrm{d}s = \sum_{i=1}^{N} \varphi_{m_i}\,\Delta s. \end{cases} \tag{5}$$

where $b_m(s)$ is the sensor acceleration, $\varphi_m(s)$ is the velocity, $e_m(s)$ is the displacement, b_{m_i} is the sampling value at time i, φ_{m_i} is the rate, and Δs is the sampling interval.

Set the 3D transformation matrix between the player's body coordinates and the horizontal coordinates as T_n^{γ}, the pitch angle as θ, and the roll angle as φ, in order to obtain the displacement in the horizontal coordinates.

$$\begin{cases} \eta_{mX} = \eta_X \cos\phi + \eta_Y \sin\phi \sin\theta - \eta_Z \sin\phi \cos\theta, \\ \eta_{mY} = \eta_Y \cos\theta + \eta_Z \sin\theta, \\ \eta_{mZ} = \eta_Z \sin\phi - \eta_Z \sin\theta \cos\phi + \eta_Z \sin\theta \cos\phi \end{cases} \tag{6}$$

where the projected components of the sensor in each direction on the horizontal coordinates are η_{mX}, η_{mY}, and η_{mZ}, respectively.

Based on the above projected components, the horizontal coordinate acceleration vector modulus η_n is obtained.

$$\eta_n = \sqrt{\eta_{mX}^2 + \eta_{mY}^2} \tag{7}$$

where the coordinate vectors are labeled in the form η_{mX}^2 and η_{mY}^2.

After the transformation of the human body coordinate state, there is an error between the data collected by the sensor and the actual data, so it is necessary to subtract the

zero bias error value calculated above to obtain accurate action data, as implied in the following equation.

$$\begin{cases} \eta'_{mX} = \eta_{mX} - \eta_{mXerror}, \\ \eta'_{mY} = \eta_{mY} - \eta_{mYerror} \end{cases} \tag{8}$$

where the data corrections after removing the gravitational acceleration are η'_{mX} and η'_{mY}, and the human body coordinate displacements are $\eta_{mXerror}$ and $\eta_{mYerror}$.

4. Fusion of computer vision and inertial capture for basketball player motion recognition.

4.1. Inertial sensor-based motion temporal feature capture for basketball players.

Based on the above acquired player motion data, this paper starts from the basketball single player scenario and integrates computer vision and inertial capture to recognize the player's movements in the basketball game video, and the network structure is implied in Figure 2. Firstly, based on the preprocessed basketball players' motion data in the previous section, collecting the temporal characteristics of the players' movements by taking the variance as the eigenvector value of the human body's overall motion amplitude. Then relied on some action pictures in the game video, the directed spatial skeleton map of the basketball player is constructed, followed by using GCN to extract the spatial features of the player's actions. Finally, the actual collected features and theoretically extracted features are fused as sample data and input into softmax to implement classification, so as to realize the accurate recognition of movements.

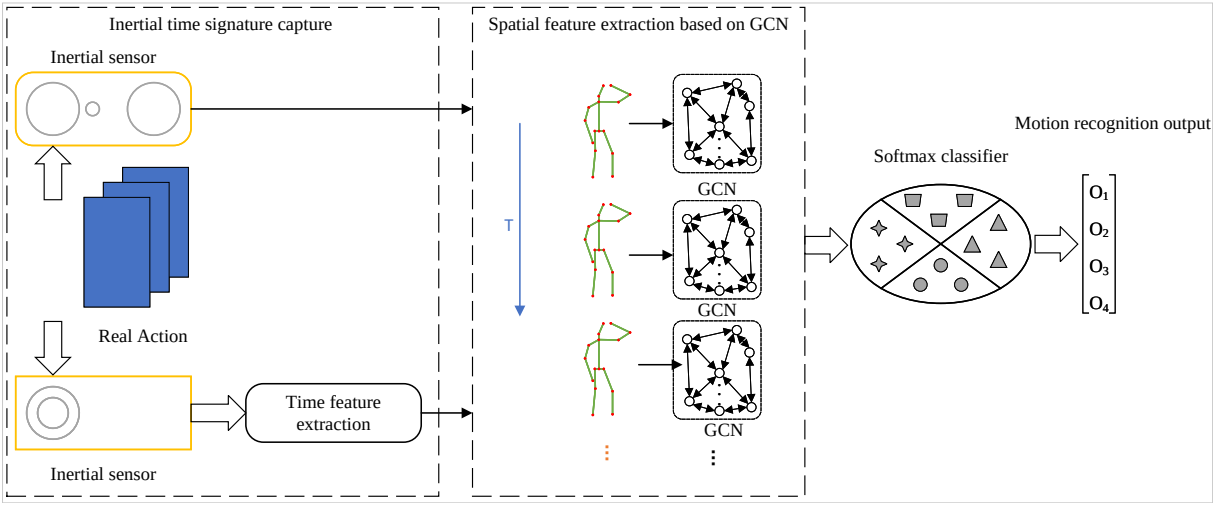

Figure 2. The network structure of the suggested method.

Based on the exercise data processed in the previous section, the variance was selected as the eigenvector value of the overall human exercise amplitude.

$$\varepsilon_h = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(L_i - L)^2} \tag{9}$$

where the acceleration modulus of the data is L_i, the number of time windows is M, the mean acceleration value is L, and the eigenvector value of the overall amplitude of the human body motion is ε_h.

The results of the above calculations are used to distinguish the walking and non-walking behaviors of the players. The interrelationship D is used to separate the dribbling, passing, and shooting actions of basketball players as follows.

$$D_{yz} = \operatorname{cov}(y, z) \,/\, (\zeta_y, \zeta_z) \tag{10}$$

where the number of interrelationships in the y, z directions in the coordinates of the player's motion is D_{yz}, the standard deviation of the directions is ζ_y, ζ_z, and the covariance of the acceleration is $\operatorname{cov}(y, z)$.

The acceleration characteristics of the axes of the player's motion are then extracted using temporal polar capture as implied in the following equation.

$$f_g = g_{\max} - g_{\min} \tag{11}$$

where the z-axis polarization of the player coordinates within the time window is C_g, the maximum y-axis acceleration value is $g_{\max}$ and the minimum acceleration value is $g_{\min}$.

Finally, the time characteristics of the player's movements can be effectively captured by the polar difference between the difference and the gyroscope pitch angle.

$$f'_g = t(\theta_{\max} - \theta_{\min}) \tag{12}$$

where $\theta_{\min}$ denotes the gravitational acceleration, $\theta_{\max}$ denotes the pitch angle maximum, and t denotes the time to capture the player's movement.

4.2. **Construction of spatial skeleton map of basketball players and spatial feature extraction.** In the basketball video scene, the raw data is a series of frames, each frame contains the coordinates of human joints, and the skeletal information is extracted based on the coordinates of joints [26]. The skeletal information is extracted based on the coordinates of joints [26]. The skeletal joints are divided into different parts, including two arms, two legs and one torso, to express the information structure of the skeleton in the basketball player's playing action. Taking the 3D skeleton data as an example, the joint coordinates are denoted as (x, y, z). For a bone, the source joint $w_s = (x_s, y_s, z_s)$ can be represented in the following way, the target joint of the basketball player is denoted as $w_{s'} = (x_{s'}, y_{s'}, z_{s'})$, and the parameters of the player's skeleton are expressed as bellow.

$$E_{w_s, w_{s'}} = (x_s - x_{s'},\, y_s - y_{s'},\, z_s - z_{s'}) \tag{13}$$

Based on the above process, the skeleton structure is constructed as a directed graph, and the information of each joint and bone is extracted by using GCN in the neighboring joints and bones. The properties of connected edges and vertices are updated in real time in the skeleton directed graph by the update function [27], which is expressed as bellow.

$$e_i^{in} = \frac{\sum_{k=1}^{K} \alpha_k^n \, h^{in}(R_{ik}^{in}) \, \alpha_k}{\sum_{k=1}^{K} \alpha_k} \tag{14}$$

where α_k represents the aggregation function of the k-th joint, h^{in} represents the attribute parameters contained in the edges, and R_{ik}^{in} represents the output edges of the network nodes.

On the basis of the above, a spatial graph convolutional neural network is established to perform feature extraction of the skeleton topology in the following steps.

(1) Denote the defined graph as G. In the spatial dimension, define the graph convolution operation as bellow.

$$f_{aut}(w_{r_i}) = \sum_{w_j \in S_g} \frac{1}{T_g} f_{in}(w_{r_j})\, \nu\big(l_i(w_{r_j})\big) \tag{15}$$

where w represents the vertices on the spatial graph, f_{in} represents the feature mapping parameters of the target, w_j represents the set of neighboring nodes of node j, and ν represents the weight function.

(2) The spatial dimension transformation is performed on the skeleton graph G with the following equation.

$$f_{out} = \sum_{k=1}^{K} \nu_k\big(f_{in}(A_k \mid V_k)\big) \tag{16}$$

where k represents the convolution kernel size, A represents the normalization parameter of the adjoint matrix, V represents the weight matrix, and $|$ represents the dot product.

(3) Multi-attention mechanism is introduced to optimize the connectivity graph to get a more suitable graph structure for description, in order to better complete the player action feature extraction and get the preliminary extracted spatial features.

$$f' = \sum_{k=1}^{i} \nu_k\big(f_{in}(A'_k + B_k)\big) \tag{17}$$

where B_k represents the attention matrix.

(4) To better extract the action features of a player playing basketball, an attention mechanism is introduced to enhance the features [15]. Since the key information of each channel is different, it is necessary to set different weights, each weight mainly represents the degree of contribution of a channel parameter to key features [28]. The enhanced features are represented as follows.

$$f_c = \frac{1}{H \times V} \sum_{i=1}^{H} \sum_{e=1}^{V} m_c(i_e f') \tag{18}$$

where H and V represent the two weight matrices, m_c represents the squeezing operation parameters, and i_e represents the weight calculation parameters for the e-th channel.

4.3. **Feature fusion and action recognition.** After obtaining the temporal and spatial features of the basketball player's movements, the spatio-temporal features of the movements are fused by using LSTM [29], and the dependency between different sequences is established by performing full join operation on them.

$$f^* = f'_g + V_k \cdot f_c \tag{19}$$

where f^* is the fusion feature of the player's action and $\cdot$ denotes the dot product operation.

The fused features f^* are then input into the LSTM model for training and learning to obtain the final predicted output y^*. Using the softmax function [30] it is possible to derive the predicted distribution of this recognized target type. The softmax function of the model classifier in this article is defined as below.

$$P(y^* = M) = \text{softmax}(y^*) = \frac{\exp(y^* T, m)}{\sum_{m'} \exp(y^* T, m')} \; (m' \in M) \tag{20}$$

Finally, the average of the classification probabilities corresponding to each feature extraction point is obtained, and then the distribution category with the highest average probability is selected to define the class label to which the current video belongs. The training of the model is completed by using the multi-class cross-entropy loss function method, and under the condition of minimizing the objective function. The deviation between the probability distribution and the actual distribution obtained from the current training is evaluated by the cross-entropy loss function, which reflects the distance between the predicted value and the actual value, and is defined as follows.

$$E(t^*, y) = -\sum_j t_j^* \log(y_j) \tag{21}$$

where t^* represents the data label of the actual value and y represents the predicted value. In the recognition and classification stage, the trained LSTM network model can calculate the output of the test samples after forward propagation and establish their corresponding type labels for the recognition task.

5. **Performance testing and analysis.**

5.1. **Comparison and analysis of recognition performance.** To estimate the effectiveness of the suggested recognition method, this article conducts training and evaluation on the SpaceJam basketball action dataset [31]. The dataset is a collection of video clips from NBA basketball games, each clip is 176×128 in length and width, and the number of frames is 16. There are 32,560 sample sequences of basketball actions in the dataset, and six types of basketball professional actions, namely, passing, running, dribbling, shooting, defending, and no action. The dataset is randomly divided into training set, validation set and test set in the classical 7:2:1 manner. The experiments are written based on the Pytorch framework, and the operating system of the experimental platform is Ubuntu 16.04. In order to prevent the overfitting phenomenon, the Dropout is set to 0.8, the total iteration period is 100 epochs, the batch size is 8, the initial learning rate is 0.001, and the decay rate is 0.1 for every 10 epochs.

Table 1. Motion Recognition Accuracy

Serial number	**Category**	**Accuracy/%**		
		FCVIC	MGRIS	IPBVU
1	Pass	91.67	79.38	84.93
2	Run around	96.57	83.67	83.59
3	Dribble	90.33	78.41	81.72
4	Shoot	91.41	76.49	85.44
5	Defense	93.84	81.74	83.25
6	No action	83.95	75.22	80.67
	mAP	91.30	79.15	83.27

For the goal of evaluating the recognition performance of the suggested method, this article is compared with the existing methods for recognizing basketball players' movements. For the convenience of analysis, this paper's method is denoted as FCVIC, the method in literature [15] is denoted as MGRIS, and the method in literature [21] is denoted as IPBVU. The evaluation metrics are used as the mean recognition accuracy (mAP), accuracy, recall, precision, F1-score, Mean Absolute Percentage Error (MAPE) and correlation coefficient R. The recognition accuracies of different recognition methods for various types of basketball actions are shown in Table 1. mAP of FCVIC for the

six types of actions is 91.3%, which is 12.15% and 8.03% higher than that of MGRIS and IPBVU, respectively. mAP of FCVIC for the category of no-action recognition is not high, which is due to the fact that the change of player's actions in the video is not particularly obvious, and the player is mostly in the same posture in each video frame, and the correlation between each frame is not strong, so the recognition accuracy is not significantly improved. However, when recognizing passing, running, dribbling, shooting and defending, the player's motion changes greatly when doing these types of actions, which contains rich dynamic information, and the temporal correlation between frames is obviously increased, so FCVIC can fully extract the action features in the video frames to improve the performance of the network.

Table 2. Performance comparison of different action recognition methods

Method	Precision	Recall	F1-score	MAPE	R
FCVIC	92.19	93.47	92.83	2.81	96.27
MGRIS	77.48	79.63	78.54	11.59	88.32
IPBVU	84.36	85.21	84.78	6.92	79.84

A comparison of the recognition accuracy performance of the different methods is shown in Table 2. FCVIC outperforms MGRIS and IPBVU in all the indexes, where the F1-score and R of FCVIC are 92.83% and 96.27%, respectively, which are improved by 14.29% and 7.95% compared to MGRIS, and by 8.05% and 16.43%. This is due to the fact that MGRIS does not process the inertial sensors with denoising, error correction, and gravity acceleration removal, which makes the extracted motion features differ from the actual ones, resulting in a low recognition accuracy. IPBVU only extracts the spatial features of the skeletal nodes of the basketball players by using the GCN, and it does not take into account the temporal features of the inertial sensors, which results in a recognition accuracy lower than that of the FCVIC. The MAPE is a reflection of the degree of discrepancy between the estimation quantity MAPE is a measure that reflects the difference between the estimated quantity and the estimated quantity, and the smaller the MAPE is, the higher the accuracy of the experimental data described by the recognition method. The MAPE of FCVIC is 2.81%, which is 8.78% and 4.11% lower than that of MGRIS and IPBVU, respectively, and shows the best recognition performance.

5.2. **Stability analysis.** The stability of the recognition methods directly affects the accuracy of feature extraction, so the box-and-whisker plots are used to compare the stability of FCVIC with the other two methods. The upper and lower endpoints of the box-and-whisker plot represent the maximum and minimum values of the stability parameters; the upper and lower sides represent the maximum and minimum values of the stability parameters without considering the error; the red line represents the average value of the stability parameters. The comparison results are shown in Figure 3. The maximum and minimum values of the stability parameters of the FCVIC method are relatively stable, while the maximum and minimum values of the stability parameters of the other two methods fluctuate. This is because FCVIC reduces the data error through pre-processing such as denoising, which is conducive to improving the stability of inertial time feature extraction. In addition, the stability of spatial feature extraction is improved by GCN and the attention mechanism, which verifies that the stability of FCVIC method for feature recognition is better.

Figure 4 implies that the recognition time of FCVIC method is less than the other two methods, the most recognition time is only 1.3 min, the average recognition time is not more than 1 min, and the recognition speed is relatively average, which indicates that

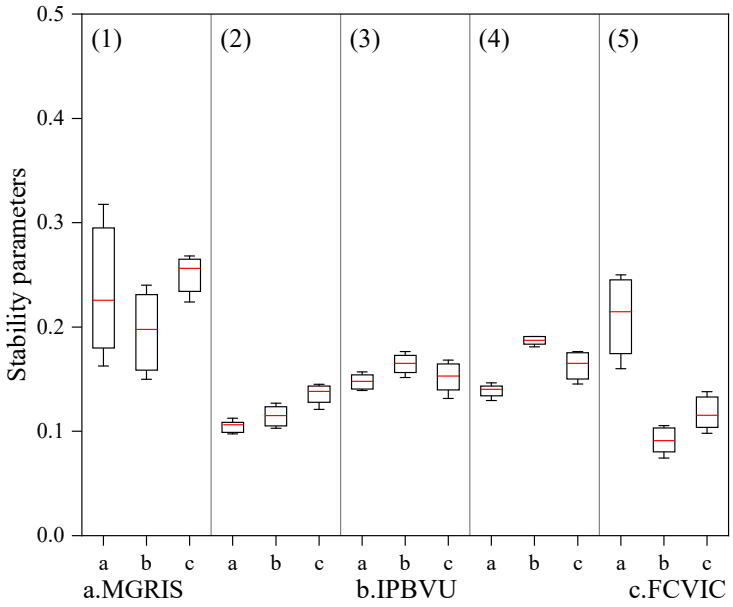

Figure 3. Method Stability Comparison Box-and-Whisker Plots

the method has a good stability. The recognition time of MGRIS reaches a maximum of 6.8 min, and the average recognition time is 4.2 min, and through the observation of the image curve, it can be found that the method recognition time fluctuation is large, and the stability is poor in practical applications. IPBVU is more stable, but the time consumed is much higher than the method in this paper. Therefore, compared with the three methods, it can be seen that the method in this paper is superior and has a certain degree of stability.

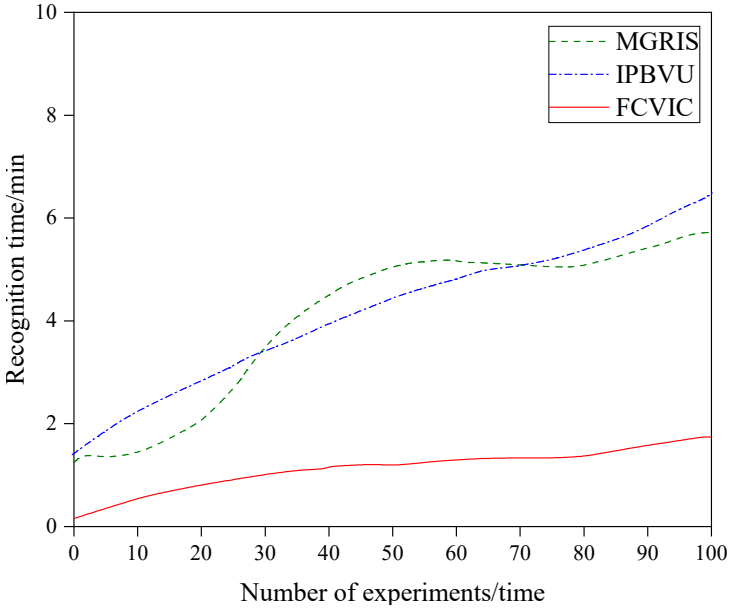

Figure 4. Motion recognition time comparison

6. **Conclusion.** Aiming at the issues of insufficient feature extraction and low accuracy of the existing basketball player movement recognition methods, this paper suggests a basketball player movement recognition method integrating computer vision and inertial

capture. Firstly, the collected data are processed with error correction and gravity acceleration removal to obtain the actual movement data. On this basis, the polar deviation is used as the eigenvector value of the human body's overall motion amplitude, and the temporal features of the movements are extracted. Secondly, the inter-frame difference method is used to process the basketball video, extract the corresponding shapes of the images to characterize the changing patterns of the movements, construct a directed spatial skeleton map, establish a GCN network, parameterize the spatial skeleton map and extract the spatial features. Finally, the temporal and spatial features are fused and inputted into softmax for classification, so as to realize the accurate recognition of human movement. Simulation outcome implies that the suggested method has high recognition accuracy and low recognition time, and can be better applied to the recognition of basketball players' movements.

REFERENCES

[1] Z. Li, "Feature extraction and data analysis of basketball motion postures: acquisition with an inertial sensor," *Journal of Engineering and Science in Medical Diagnostics and Therapy*, vol. 4, no. 4, 041006, 2021.

[2] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, vol. 26, pp. 1–25, 2022.

[3] M. Chuang and P. Narasimhan, "Automated viewer-centric personalized sports broadcast," *Procedia Engineering*, vol. 2, no. 2, pp. 3397–3403, 2010.

[4] G. Lee, V. Bulitko, and E. A. Ludvig, "Automated story selection for color commentary in sports," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 6, no. 2, pp. 144–155, 2013.

[5] J. Gao, H. Zou, F. Zhang, and T. Y. Wu, "An intelligent stage light-based actor identification and positioning system," *International Journal of Information and Computer Security*, vol. 18, no. 1/2, pp. 204–218, 2022.

[6] T.-Y. Wu, H. Li, S. Kumari, and C.-M. Chen, "A Spectral Convolutional Neural Network Model Based on Adaptive Fick's Law for Hyperspectral Image Classification," *Computers, Materials & Continua*, vol. 79, no. 1, pp. 19–46, 2024.

[7] F. Zhang, T.-Y. Wu, J.-S. Pan, G. Ding, and Z. Li, "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, 40, 2019.

[8] H. Ren and X. Wang, "Application of wearable inertial sensor in optimization of basketball player's human motion tracking method," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2021.

[9] H. Li, S. Derrode, and W. Pieczynski, "Lower limb locomotion activity recognition of healthy individuals using semi-Markov model and single wearable inertial sensor," *Sensors*, vol. 19, no. 19, 4242, 2019.

[10] C. Sun and D. Ma, "SVM-based global vision system of sports competition and action recognition," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 2265–2276, 2021.

[11] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[12] A. Ahmadi, E. Mitchell, C. Richter, F. Destelle, M. Gowing, N. E. O'Connor, and K. Moran, "Toward automatic activity classification and movement assessment during a sports training session," *IEEE Internet of Things Journal*, vol. 2, no. 1, pp. 23–32, 2014.

[13] J. Li and D. Gu, "Research on basketball players' action recognition based on interactive system and machine learning," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 2029–2039, 2021.

[14] W. Sousa Lima, H. L. de Souza Bragança, K. G. Montero Quispe, and E. J. Pereira Souto, "Human activity recognition based on symbolic representation algorithms for inertial sensors," *Sensors*, vol. 18, no. 11, 4045, 2018.

[15] P. Kumar, S. Mukherjee, R. Saini, P. Kaushik, P. P. Roy, and D. P. Dogra, "Multimodal gait recognition with inertial sensor data and video using evolutionary algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 5, pp. 956–965, 2018.

[16] L. Kong, D. Huang, J. Qin, and Y. Wang, "A joint framework for athlete tracking and action recognition in sports videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 532–548, 2019.
[17] K. Zuo and X. Su, "Three-dimensional action recognition for basketball teaching coupled with deep neural network," *Electronics*, vol. 11, no. 22, 3797, 2022.
[18] J. Xiao, W. Tian, and L. Ding, "Basketball action recognition method of deep neural network based on dynamic residual attention mechanism," *Information*, vol. 14, no. 1, 13, 2022.
[19] S. B. Khobdeh, M. R. Yamaghani, and S. K. Sareshkeh, "Basketball action recognition based on the combination of YOLO and a deep fuzzy LSTM network," *The Journal of Supercomputing*, vol. 80, no. 3, pp. 3528–3553, 2024.
[20] J. Liu and Y. Che, "Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network," *Journal of Electronic Imaging*, vol. 30, no. 3, pp. 033017–033017, 2021.
[21] T. Feng, K. Ji, A. Bian, C. Liu, and J. Zhang, "Identifying players in broadcast videos using graph convolutional network," *Pattern Recognition*, vol. 124, 108503, 2022.
[22] D. K. Shaeffer, "MEMS inertial sensors: A tutorial overview," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 100–109, 2013.
[23] M. O'Reilly, B. Caulfield, T. Ward, W. Johnston, and C. Doherty, "Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review," *Sports Medicine*, vol. 48, pp. 1221–1246, 2018.
[24] M. Ishimaru, Y. Okada, R. Uchiyama, R. Horiguchi, and I. Toyoshima, "Classification of Depression and Its Severity Based on Multiple Audio Features Using a Graphical Convolutional Neural Network," *International Journal of Environmental Research and Public Health*, vol. 20, no. 2, 1588, 2023.
[25] S. Soltani, "On the use of the wavelet decomposition for time series prediction," *Neurocomputing*, vol. 48, no. 1–4, pp. 267–277, 2002.
[26] Z. Hao, X. Wang, and S. Zheng, "Recognition of basketball players' action detection based on visual image and Harris corner extraction algorithm," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 4, pp. 7589–7599, 2021.
[27] X. Wang, S. Lu, R. Zhou, and H. Wang, "Skeleton estimation of directed acyclic graphs using partial least squares from correlated data," *Pattern Recognition*, vol. 139, 109460, 2023.
[28] G. Huang, J. Zhu, J. Li, Z. Wang, L. Cheng, L. Liu, H. Li, and J. Zhou, "Channel-attention U-Net: Channel attention mechanism for semantic segmentation of esophagus and esophageal cancer," *IEEE Access*, vol. 8, pp. 122798–122810, 2020.
[29] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.
[30] D. Zhu, S. Lu, M. Wang, J. Lin, and Z. Wang, "Efficient precision-adjustable architecture for softmax function in deep learning," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 3382–3386, 2020.
[31] S. Babaee Khobdeh, M. Yamaghani, and S. Khodaparast, "A novel method for clustering basketball players with data mining and hierarchical algorithm," *International Journal of Applied Operational Research—An Open Access Journal*, vol. 11, no. 4, pp. 25–38, 2023.