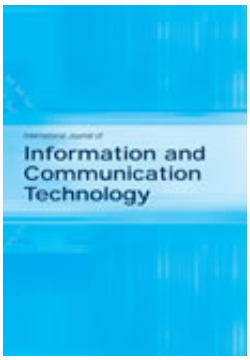
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Track and field sports skill recognition based on multimodal sensing data

Zhicong Zhou

DOI: 10.1504/IJICT.2026.10075848

Article History:

Received:	02 August 2025
Last revised:	27 November 2025
Accepted:	28 November 2025
Published online:	04 February 2026

Track and field sports skill recognition based on multimodal sensing data

Zhicong Zhou

Zhengzhou Yellow River Nursing Vocational College,
Zhengzhou 450066, China
Email: Zhoucong2025@126.com

Abstract: Track and field sports skill recognition is a key technology in intelligent sports training, but traditional methods suffer from issues such as information redundancy and poor recognition performance. To address this, this paper first proposes an adaptive selection mechanism for multimodal sensor data based on mutual information, filtering out sensor combinations that provide maximum information correlation. Then, a convolutional neural network (CNN) is combined with a long short-term memory network (LSTM) for multimodal sensor feature extraction, and a recurrent matrix-based multimodal feature fusion method is proposed. Finally, the fused feature vector is input into a fully connected layer, and the softmax function is used to calculate the score for each category of athletics skill from the output classification layer. The experimental results show that the Macro_F1 of the proposed method is improved by at least 4.01% compared to baseline methods, demonstrating good recognition performance.

Keywords: track and field sports skill recognition; mutual information; multimodal sensor; convolutional neural network; CNN; graph attention network.

Reference to this paper should be made as follows: Zhou, Z. (2026) 'Track and field sports skill recognition based on multimodal sensing data', *Int. J. Information and Communication Technology*, Vol. 27, No. 4, pp.16–31.

Biographical notes: Zhicong Zhou received her Bachelor's degree from Zhengzhou University. Currently, she works as a Lecturer in Zhengzhou Yellow River Nursing Vocational College. Her research directions include physical education, sports training, and sports teaching.

1 Introduction

As competitive sports continue to flourish, track and field, as a fundamental project in the sports domain, has consistently occupied a central position. It is not only a stage for showcasing human speed, power, endurance, and skills, but also a key area for cultivating outstanding sports talent (Pereira et al., 2015). From the Olympics to the World Championships, every breakthrough on the track and field tugs at the heartstrings of sports enthusiasts around the globe, and the level of competition directly reflects a country's comprehensive strength in its sports industry. Traditional assessments of track and field skills are highly subjective, relying on features from single-sensor data, making

it difficult to accurately capture action details and underlying patterns, and unable to comprehensively consider multi-sensor information from track and field athletes (Zheng and Man, 2022). Multi-modal sensor data integrates information from various types of sensors, such as an inertial measurement unit (IMU) that can obtain data such as body motion acceleration and angular velocity from the athlete (Gai, 2025). These multi-source heterogeneous data complement and validate each other, enabling a comprehensive and precise depiction of the athlete's motion state and skill characteristics (Mekruksavanich and Jitpattanakul, 2022). How to utilise multi-modal sensor data to achieve precise identification of track and field skills is a very important research topic.

Traditional track and field skill recognition algorithms are mainly based on machine learning algorithms to construct recognition models. Liu and Wang (2023) integrated the Canny edge detection method to extract edge contour features from original track and field technique images and used support vector machine (SVM) to output the sports skill recognition results. Cui and Wang (2025) used 2D optimal orthogonal separable directional filters for feature extraction in track and field motion images, and achieved sports skill recognition through a decision tree classifier; however, there were issues of relatively complex computation and longer time consumption. Yu and Xing (2022) applied Gabor wavelet transformation to extract the features of track and field motions and used principal component analysis to remove redundant features, thereby improving recognition accuracy. Yao and Li (2022) used Fourier transformation (Fong et al., 2021) and discrete cosine transformation (Ahmad et al., 2015) to extract frequency domain features respectively and achieved track and field motion recognition based on extreme learning machine technology, achieving a recognition accuracy of 78.35%. Liu and Chang (2022) proposed a track and field motion behaviour identification system using accelerometer data, captured statistical characteristics, used non-negative matrix factorisation for feature reduction, and finally classified using an ensemble approach based on rotation random forests. The application of traditional machine learning methods in the domain of track and field skill recognition has demonstrated the effectiveness of manual feature extraction and diverse classifiers, although they do present limitations in handling complex data and generalisation capabilities. Nevertheless, these methods have laid the foundation for subsequent deep learning approaches and provided significant reference value.

Deep learning-based recognition models can mine deep feature information from data without relying on domain knowledge and have proven effective in image recognition applications, achieving better results than manual feature extraction. The track and field skills recognition model based on deep learning quantifies or prunes the model to reduce reasoning time and meet the requirements of real-time action recognition. It also extracts the deep features of track and field skills through deep neural networks to enhance the recognition accuracy of the model. Li et al. (2023) proposed a multi-scale convolutional neural network (CNN) to capture spatial and temporal characteristics from the raw acceleration data of track and field athletes, achieving a recognition accuracy of 82.93%. Zhang (2023) proposed a general deep neural network framework applied to a single homogeneous sensor modality. Zhang (2021) used five wearable inertial sensor units to record different daily activities and sports, and then performed feature extraction and recognition using a long short-term memory network (LSTM) model, improving recognition accuracy. Hsu et al. (2019) proposed a wavelet ensemble CNN, introducing discrete wavelet transform into the convolution structure, combining the time-frequency

localisation features of wavelet transform with the self-learning ability of neural networks to realise sports skill recognition under a single sensor perception environment.

The above recognition models based on deep learning and single sensors have issues with insufficient multi-dimensional feature extraction and unsatisfactory recognition accuracy. Recognition models based on multimodal sensors greatly enhance recognition performance by fusing features from multiple sensors. Mekruksavanich and Jitpattanakul (2022) extract movement action features from accelerometer, gyroscope, magnetometer, and barometer sensors, and input them into a CNN for action recognition, achieving an average accuracy of 84.18% for identifying four sports skills. Dahou et al. (2023) utilise combination data from accelerometers and gyroscopes, using CNNs with different kernel dimensions to capture movement features at different resolutions, achieving a recognition accuracy of 85.39%. Lee et al. (2024) propose a four-layer hybrid LSTM model that combines data from accelerometers and gyroscopes and is effectively applied in sports skills recognition tasks.

In summary, although existing sports skill recognition methods based on single modalities in athletics have achieved some results, single-modality recognition can no longer meet practical demands, making the effective fusion of heterogeneous sensor information from multiple sources through multimodal learning increasingly important. To address the problems of insufficient sensor feature extraction and unsatisfactory recognition results in current research, this article suggests a sports skills recognition approach in light of multimodal sensor data. The main work of this method is summarised into the following four aspects.

1 Aiming at the problem of redundant information in the sports behaviour data collected by multimodal sensors, this study proposes a multimodal sensor data adaptive selection method based on mutual information. By calculating the mutual information of different sensor combinations across different sports skill categories, sensor combinations that provide maximum information correlation are selected, thereby improving the efficiency and accuracy of data fusion.

2 Key behaviour features of multimodal sensor signals are extracted through CNN, and the behaviour features processed by CNN are then sequentially input into LSTM according to time order. A two-layer LSTM is used to capture temporal domain information on the contextual relationships between different signal frames, and the gate mechanism is employed to selectively retain the behavioural information obtained from the features extracted by CNN, thereby obtaining spatiotemporal features related to sports skill recognition.

3 A reasonable cross-combination of different sensor feature vectors is achieved to realise sensor fusion at the feature level. After cyclic matrix fusion, the characteristic vector is input into the dropout layer and fully linked level, and finally the softmax function is used to calculate the score for each category of sports skills from the output categorisation level, resulting in the final sports skill recognition results.

4 Experimental results show that the average recognition accuracy of the proposed method for various types of athletic sports skills is 94.17%, which is improved by 4.72%–11.38% compared with the baseline approach. The approach is able to effectively handle information interaction among heterogeneous features, achieve multimodal feature complementarity, and demonstrate good recognition performance.

2 Relevant technologies

2.1 Introduction to multimodal sensors

Common types of sensors used for human body motion skill recognition include visual sensors and wearable sensors. Visual sensors perform activity recognition by capturing human visual images. They are usually installed in fixed positions, such as surveillance cameras, which can capture human actions in a large range in real time. However, visual sensors may be limited by lighting conditions, obstacles, and privacy protection issues (Dang et al., 2020). Compared with visual sensors, wearable sensors have advantages in human activity recognition. Wearable sensors are directly attached to the human body and identify motion skills by recording motion data. Wearable sensors are not limited by line of sight and can be widely applied in fields such as sports training and health monitoring (Zhang et al., 2022). At the same time, their direct data collection method effectively protects user privacy, making users more at ease when using them.

Common types of wearable sensors include accelerometers, gyroscopes, magnetometers, etc., which are used to record human behavioural data. The human body motion skill recognition process first collects data through sensors, and the data is then transformed into multi-dimensional time series signals at a fixed frequency. Therefore, the problem of human body motion skill recognition can be regarded as a segmentation problem of multi-dimensional time series. The basic process of human body motion skill recognition includes several steps: data collection, data preprocessing, feature extraction and selection, feature fusion, and model classification.

2.2 Convolutional neural network

The key of CNN lies in the convolution operation, which extracts features from data such as images to achieve prediction and classification of different data. It is extensively utilised in the domain of picture recognition. Essentially, CNNs represent a specialised type of multi-layer perceptron that employs local connectivity and shared weights. These architectural features simultaneously reduce parameter count for more efficient optimisation while minimising overfitting risks. The convolution operation of CNN is essentially a linear transformation, while the activation function enables the network to learn complex features by introducing nonlinear factors. For instance, in track and field movement recognition, ReLU can distinguish between the force application stage of athletes, maintaining the original value of positive input and setting the negative input to zero during the relaxation stage, thereby capturing the dynamic changes of movements.

The basic framework of CNN consists of a convolutional level, a pooling level, and a fully linked level. The convolutional layer performs convolution operations on input data by sliding a convolution kernel, which can extract local features of the data (Chen et al., 2021). The pooling layer performs dimensionality reduction on the characteristics output by the convolutional layer, thereby simplifying the calculation. The most frequently employed pooling operations are max pooling and average pooling. The function of the fully linked level is to map the characteristic vector processed by the pooling layer to category labels, thus achieving classification or recognition of input data.

2.3 Long short-term memory network

LSTM is a special type of RNN, primarily designed to solve the problem of vanishing gradients in long sequence training (Ullah et al., 2021). Compared to RNN having only one transmission state h_t LSTM has two transmission states, a cell state C_t and a hidden state h_t. LSTM adopts gated output methods, namely the input gate, forget gate, and output gate. Taking the memory cell at time t as an example, the input of the memory cell at time t is x_t and h_{t-1} passed down from the previous state. The two input data sequentially enter the forget gate, input gate, and output gate to obtain states z^f, z^i, $\tilde{C}_t$, z^o. The two input data first enter the forget gate to obtain the information to be discarded z^f, with the calculation equation as follows.

$$z^f = \sigma\left(W_f\left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

Then it enters the input gate to obtain the information to be updated z^i and the current cell state C_t, shown in equation (2) and equation (3) respectively, where $\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$.

$$z^i = \sigma\left(W_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$C_t = z^f \cdot C_{t-1} + z^i \cdot \tilde{C}_t \tag{3}$$

Then through the output gate to determine which will be used as the current state for output, and respectively obtain C_t and h_t information. Finally perform the storage operations in equation (4) and equation (5) and input into the next neuron.

$$z^o = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{4}$$

$$h_t = z^o \cdot \tanh\left(C^t\right) \tag{5}$$

In summary, this change in network structure allows LSTM to perform better in longer sequences.

3 Design of an adaptive selection mechanism for multi-modal sensor data based on mutual information

Multi-modal sensors can capture various information of track and field athletes during their movement, including acceleration, angular velocity, physiological signals, etc., providing a rich data source for movement skill recognition. These data sources are not only numerous but also diverse in types, jointly forming a multi-dimensional information space for track and field movement skill recognition. However, due to the complexity and redundancy of multi-modal data, how to effectively extract sensor data useful for classification tasks from the data has become a hotspot and challenge in current research. The complexity of multi-modal data is reflected in aspects such as large information dimensions and complex correlations, while redundancy means that the data contains information with little contribution to classification tasks. Therefore, how to perform reasonable sensor data selection to extract the most representative sensor data subset is a problem to be solved in current research.

Current research neglects the correlations between different recognition tasks, which to some extent limits the efficiency and accuracy of sensor data selection. To address the above issues, this chapter suggests an adaptive selection approach for multi-modal sensor data based on mutual information (Hoque et al., 2014). By calculating the mutual information of different sensor combinations under different movement skill categories, sensor combinations providing maximum information correlation are selected to improve the efficiency and accuracy of data fusion. In addition, by statistically selecting the sensor combinations with the highest occurrence frequency, more reliable and effective combinations are further selected to provide a more reliable basis for track and field movement skill recognition.

Traditional research uses all sensor combinations as input for movement skill recognition. However, not all sensor modalities provide beneficial information; they may even interfere with each other, affecting the final recognition performance. Specifically, each sensor modality has its unique operating principle and range of application. For example, accelerometers are good at capturing dynamic motion information, while gyroscopes are more sensitive to body rotation and tilt. However, when all sensor data are simultaneously input into the recognition system, inconsistencies and redundancies between the data may lead to a decline in recognition performance. In addition, excessive input data can also increase the system's computational load, reduce real-time performance, and even introduce noise and interference signals. Therefore, to avoid data redundancy and excessive complexity, it is necessary to select sensor data combinations that are highly related to track and field movement skill types. At the same time, by adopting data selection and fusion strategies, the effective utilisation rate of sensor data can be improved, thus increasing recognition accuracy and efficiency.

Assume there are multiple sensor data X_1, X_2, …, X_n and track and field movement skill recognition target variables C. C has M possible values C_1, C_2, …, C_M. It is necessary to calculate the marginal probability of features belonging to the target variable, and also determine the joint probability between sensor data. This can provide key information about the association between sensor data and the target variable, as well as the interdependencies among the sensor data. Conditional mutual information is used to quantify the correlation between multiple sensor data under the track and field movement skill recognition target variable as shown below.

$$I\left(X_1, X_2, ..., X_n \mid C = C_M\right) = \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} \cdots \sum_{x_n \in X_n} p\left(x_1, x_2, ..., x_n \mid C = C_M\right) \times \log \frac{p\left(x_1, x_2, ..., x_n \mid C = C_M\right)}{p\left(x_1 \mid C = C_M\right) \times ... \times p\left(x_2 \mid C = C_M\right) \times ... \times p\left(x_n \mid C = C_M\right)} \tag{6}$$

where $I(X_1, X_2, \ldots, X_n \mid C = C_M)$ represents the joint conditional mutual information between sensor data X_1, X_2, …, X_n under a given category; $p(x_1, x_2, \ldots, x_n \mid C = \mathrm{C_M})$ represents the joint probability that sensor data x_1, x_2, …, x_n takes the value C_M under the category X_1, X_2, …, X_n; $p(x_n \mid C = C_M)$ represents the marginal probability that sensor data X takes the value x_n under the category C_M.

For a multimodal sensor data set and a movement skill recognition target variable, it is typically necessary to calculate the mutual information between sensor data to assess their correlation. The magnitude of mutual information directly reflects the degree of correlation between data. When the mutual information value is larger, it indicates a tighter correlation between the data and a stronger information dependency between

them. In short, the size of mutual information can serve as an important indicator to measure the strength of correlation among sensor data.

$$Q(n, v) = \frac{n!}{v!(n-v)!} \tag{7}$$

where n is the total amount of sensors, and v is the amount of sensor variables included in each combination.

$$D(S) = \left\{ \{X_1\}, ..., \{X_n\}, \{X_1, X_2\}, \{X_1, X_3\}, ..., \{X_{n-1}, X_n\}, ..., S \right\} \tag{8}$$

where $\{X_i\}$ is a single-sensor subset, and $\{X_i, X_j\}$ is a dual-sensor feature combination subset, S represents the set containing all sensors themselves. Pairwise conditional mutual information measures the amount of information shared between any two sensor variables under a given category; correspondingly, it measures the amount of information shared between all sensor data under a given category.

To find the maximum value among these combination conditional mutual information, it is necessary to calculate the mutual information between all possible pairs of sensor data and select the largest value to determine the most informative combination.

$$\left(i^*, ..., j^*\right) = \arg\max_{i \neq j} I\left(X_i; ...; X_j \mid C = C_M\right) \tag{9}$$

where $I(X_i; \ldots; X_j \mid C = C_M)$ denotes the combination mutual information between sensor data under the given category C_M. This is the combination index of the maximum feature under the given category, which is determined by traversing all combinations of sensor data pairs and comparing their mutual information values.

After computing the maximum mutual information between sensor data and the corresponding feature indices for each category, the most representative sensor data combination for each category is obtained. A simplified sensor data combination will more efficiently process data, improve recognition accuracy, and enhance recognition performance.

4 Track and field sports skill recognition based on multimodal sensing data

4.1 Multi-modal sensor feature extraction based on ConvLSTM

Based on the selection of key multimodal sensor data in the previous chapter, this paper proposes a athletics skill recognition method based on multimodal sensor data according to the characteristics of acceleration, angular velocity, electromyography signals, and other multimodal sensors. The overall framework consists of five parts: the input level, characteristic extraction level, feature fusion layer, and athletics skill recognition layer, as indicated in Figure 1. First, a ConvLSTM method is designed to extract spatiotemporal behaviour features from multimodal sensors, divided into the CNN and LSTM parts to extract spatial and temporal features from multimodal sensors. Then, a reasonable cross-combination of different sensor feature vectors is achieved, avoiding the enormous computational load generated by tensor fusion methods (Borsoi et al., 2024) by utilising the interaction of multimodal feature vectors, thereby achieving sensor fusion at the

feature level to generate a consistent explanation for overall recognition. The feature vector after cyclic matrix fusion is input into the dropout level and the fully linked level, and finally, the softmax function is adopted to calculate the score for each category of athletic skills in the output classification layer. The action with the highest score is recognised as the acknowledged athletic skill.

Common CNN and LSTM-based methods can automatically extract behavioural features from multimodal sensors, then perform self-learning and optimisation of model parameters via neural network gradient descent and backpropagation techniques, achieving good recognition performance. However, CNN does not further process the hidden temporal information, neglecting the continuity of human behaviour. LSTM lacks integration of sensor data, leading to relatively slow algorithm operation speed. To address the above issues, a ConvLSTM-based feature extraction method using deep learning is designed. The ConvLSTM method combines CNN and LSTM, leveraging the advantages of CNN in handling behavioural features and LSTM in managing temporal dependencies.

First, a one-dimensional convolution operation captures the temporal signal structures within the convolution kernel window, and CNN obtains the key behavioural features of multimodal sensor signals. The convolutional arithmetic is as follows.

$$C_i = f\left(\sum_{d=1}^{D}\sum_{n=1}^{N} x_d(i+n)k_d(n)\right) \tag{10}$$

where N is the length of the convolution kernel, D is the depth of multimodal sensor data and convolution kernel, $k_d(n)$ is the n^{th} weight in the d^{th} depth of one-dimensional convolution kernel, $x_d(i)$ is the i^{th} element of multimodal sensor signals at depth d, C_i is the i^{th} feature obtained through convolution of multimodal sensor data, and $f(\cdot)$ represents the activation function. The feature size after going through the pooling layer to reduce the tensor dimension in the network is as follows.

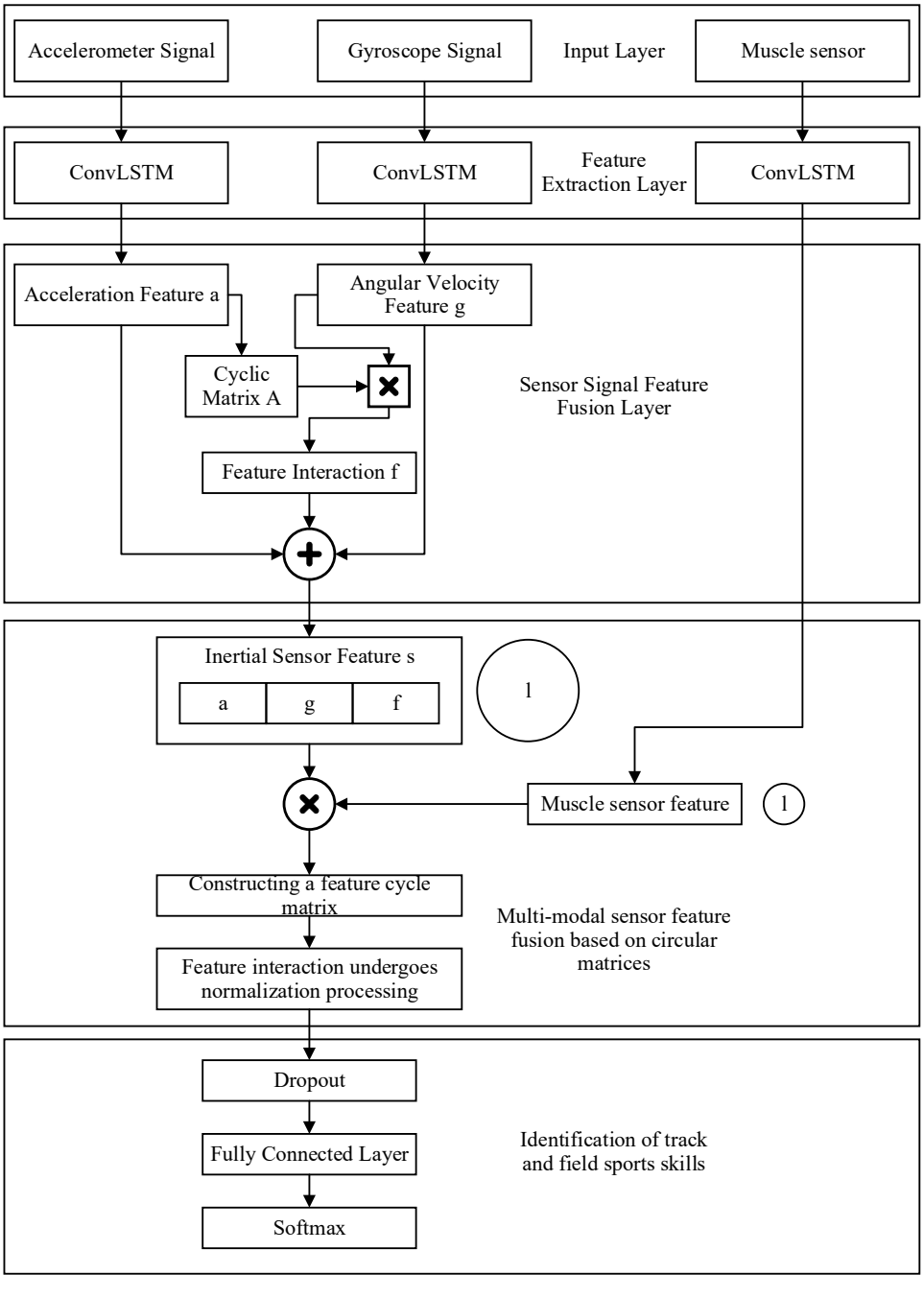
$$L_i = \frac{L_{i-1} - N + 2P}{S} + 1 \tag{11}$$

where L_i is the feature length of the current i^{th} layer, P is the padding size, and S is the stride. Through three convolution and pooling operations, low-dimensional high-level features with temporal characteristics are generated, and then the behavioural features processed by CNN are input to LSTM in temporal order.

Two-layer LSTM captures temporal domain information on contextual associations between different signal frames. Through the gating mechanism, behavioural information is selectively preserved from the CNN-extracted features in accordance with equation (1) to equation (5), to achieve better temporal activation of the multi-modal sensor signal features and obtain spatio-temporal features relevant for motor skill recognition, thus achieving spatio-temporal feature learning.

In summary, the ConvLSTM feature extraction method avoids the missing temporal characteristics of CNN and the problems of LSTM under long-term sequences. It can shorten the spatio-temporal feature extraction time and can also be separately applied to feature extraction of multi-modal sensors.

Figure 1 Track and field skills recognition process based on multimodal sensor data



4.2 Multi-modal sensor feature fusion based on circular matrices

The spatio-temporal feature vectors extracted from multi-modal sensors by the ConvLSTM method require further feature-level fusion (Piechocki et al., 2023). In deep learning-based feature-level fusion, how to integrate multiple signal features to obtain appropriate feature maps has become a key research issue. To fully utilise the mutual influence between elements of multi-modal sensor features, a circulant matrix method is proposed to handle feature interaction. The fusion method based on circulant matrices conducts in-depth analysis of multi-modal sensor fusion information and can further improve performance compared to other methods.

To enable feature interaction among multi-modal sensor feature vectors to obtain all reasonable correlations, and at the same time address the problem of overly large fusion vectors in traditional tensor fusion methods which are difficult to train, and thereby achieve more robust and accurate behaviour recognition results. Through the feature vectors and matrix multiplication method, a circulant matrix-based fusion method is designed to extract useful feature interaction terms from motion information, achieving multi-sensor feature fusion.

Assume that the dimension of multi-modal sensor feature vectors is N. Following reference (Wu et al., 2024) a feature right circulant matrix is built. The single-modal sensor feature vector a is shifted by one element each time to generate multiple vectors, and these vectors are combined into a circulant feature matrix A (in the same way, another modal sensor feature vector g can be used to construct G). The behavioural information included in the circulant matrix is determined by a. Introducing A allows the elements in the feature vectors a and g to interact with each other in all possible ways.

$$A = Circul(a) = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_N \\ a_N & a_1 & a_2 & \cdots & a_{N-1} \\ a_{N-1} & a_N & a_1 & \cdots & a_{N-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_2 & a_3 & a_4 & \cdots & a_1 \end{bmatrix} \in R^N \tag{12}$$

Each row of the N-order circulant matrix A is the result of sequentially shifting the acceleration feature vector a to the right by one element, thereby ensuring that matrix multiplication can obtain all of the feature interactions of sensors across different-modalities to fully fuse the multi-modal feature vectors.

After reshaping the feature vector into a circulant matrix, multiplying the multi-modal feature vector g with matrix A can explore the relationship among the multi-modal sensor features. To ensure that the feature interaction f and the multi-modal sensor features play the same role in behaviour recognition and avoid the influence of different numerical ranges, the fused feature interaction f needs to be normalised, as shown below.

$$f = \frac{1}{N} A^T g = \frac{1}{N} \begin{bmatrix} a_1 g_1 + a_N g_2 + a_{N-1} g_3 + \cdots a_2 g_N \\ a_2 g_1 + a_1 g_2 + a_N g_3 + \cdots a_3 g_N \\ a_3 g_1 + a_2 g_2 + a_1 g_3 + \cdots a_4 g_N \\ \vdots \\ a_N g_1 + a_{N-1} g_2 + a_{N-2} g_3 + \cdots a_1 g_N \end{bmatrix} \tag{13}$$

Each feature crossover item of a and g is contained in f. Finally, the one-dimensional behaviour features of the multi-modal sensors are introduced into the fused features. The three feature vectors a, g, and f are concatenated to obtain the final fused feature vector fusion.

$$fusion = Concat(a, g, f) = \begin{bmatrix} a \\ g \\ f \end{bmatrix} \in R^{3N} \tag{14}$$

As shown in the above equation, the dimension of the final fused feature fusion is reduced from $(N + 1)^2$ in the two-dimensional tensor fusion method to $3N$, thus maintaining control over the feature dimensions. Since no new parameters are introduced in the circulant matrix multiplication operation, the parameters of the fusion model are effectively controlled, reducing the difficulty of training.

4.3 Motor skill recognition and loss function

Following characteristic extraction, a classification module projects the high-dimensional action features into a low-dimensional space for athletic movement recognition. The multidimensional features are vectorised and subsequently transformed through a stacked neural network structure. The output of the j^{th} neuron in the neural network is as follows, where $fusion_i$ represents the i^{th} input, ω_{ij} is the weight of the i^{th} output of the j^{th} neuron, and θ is the bias.

$$y_j = f\left(\theta + \sum_{i=1}^{n} \omega_{ij} fusion_i\right) \tag{15}$$

The final motion skill recognition output unit requires a softmax function for numerical processing. The softmax output represents the relative probability between different categories, and motion skill recognition is obtained in light of the output scores produced by the softmax level. The softmax score for the i^{th} action is as follows.

$$softmax\left(y_j\right) = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}} \tag{16}$$

Identifying motion skills from different track and field actions is a binary classification problem. The cross-entropy loss function is selected as the classification loss function, as shown below.

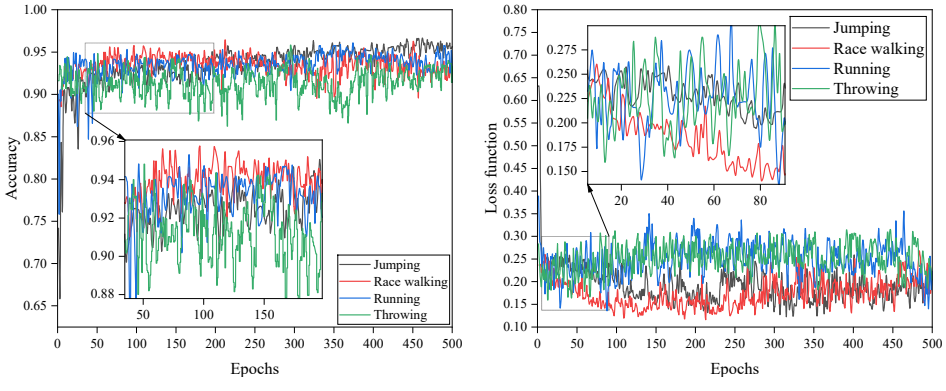
$$Loss = -\frac{1}{v}\sum_{j=1}^{v}\sum_{i=1}^{z}\left(y\log(\tilde{y}) + (1-\tilde{y}) + \chi l_2(\phi)\right) \tag{17}$$

where y is the real label of the input sample, $\tilde{y}$ is the predicted label of the model, v is the size of the sample in the batch, z is the number of categories, $l_2(\phi)$ is the regularisation term, and χ is its coefficient. The cross-entropy loss operation can directly assess the difference among model predictions and real labels, effectively solve the gradient vanishing problem, and performs well in classification problems.

5 Experimental results and analyses

This article uses the 5,397 track and field motion dataset collected in Zhang (2023). This dataset integrates multiple sensors such as IMU, pressure sensor, heart rate belt, and so on. The dataset is classified into the training set, validation set, and test set at a ratio of 8:1:1. The track and field motion skill categories include running, jumping, throwing, and race walking. The method was deployed on an Ubuntu 18.04 operating system with an NVIDIA GTX 1080Ti for experimentation, with a GPU memory size of 12 GB. The CPU used is an Intel Core i7-9700K, with a CPU memory size of 32 GB. The neural network model is built using the PyTorch deep learning framework, the compilation environment is PyCharm, the programming language is Python, and the dependencies include numpy, opencv, torchvision, etc. The batch sample size for the experiment is 100, and the number of iterations is set to 200, and it is reduced to half of the current value every 50 iterations. During the training process, the learning rate of the optimiser is set to 0.0016, the momentum value is set to 0.9, and the weight decay is set to 10^{-4}.

Figure 2 Comparison curve of recognition accuracy and loss function (see online version for colours)



The identification accuracy and loss function comparison curve of the proposed method MConvLSTM for four track and field motion skill types are shown in Figure 2. As can be seen from Figure 2, jumping motion skill types have better recognition capability than throwing motion skill types. This is because the characteristics of jumping motion are easier to identify. The recognition accuracy of various methods for throwing motion skill types is relatively low, as the action amplitude of throwing is smaller compared to jumping. Additionally, MConvLSTM can better distinguish between various track and field motion skill types, with a more stable model training process and faster convergence of the loss function.

For ease of analysis, this article selects WSA-DCNN (Hsu et al., 2019), MLCNNwav (Dahou et al., 2023), MFLSTM (Lee et al., 2024) as comparative methods, and the evaluation metrics selected are recognition accuracy (A), Macro_Precision (Macro_P), Macro_Recall (Macro_R), Macro_F1 (Zhang, 2022). The comparison of recognition accuracy for various track and field motion skill types among different methods is shown in Table 1. The average recognition accuracy of MConvLSTM for various track and field motion skill types is 94.17%, while the average recognition accuracy of WSA-DCNN, MLCNNwav, and MFLSTM are 82.79%, 84.13%, and 89.45%, respectively. The average

recognition accuracy of MConvLSTM is 4.72%–11.38% higher than that of the baseline methods, indicating higher recognition accuracy.

Table 1 Accuracy rate of identifying various types of track and field skills

Method	*WSA-DCNN*	*MLCNNwav*	*MFLSTM*	*MConvLSTM*
Running	81.93	87.31	89.55	93.72
Jumping	84.61	82.54	91.68	96.59
Throwing	80.94	82.27	88.34	91.36
Race walking	83.66	84.39	88.23	95.02

The Macro_P, Macro_R, and Macro_F1 indicators for different methods are compared in Table 2. The Macro_P and Macro_R of MConvLSTM are 94.18% and 92.63%, respectively, which are improved by 12.23% and 12.32% compared to WSA-DCNN, and by 7.69% and 9.07% compared to MLCNNwav, and by 5.76% and 2.25% compared to MFLSTM. Comparing Macro_F1, MConvLSTM is improved by 12.28%, 8.41%, and 4.01% compared to WSA-DCNN, MLCNNwav, and MFLSTM, respectively. WSA-DCNN requires multi-scale decomposition of the number of motion sensors through wavelet transform, generating subbands of different frequencies. This process significantly increases the data dimension, causing CNN to process more feature maps and exponential growth in computational workload, hence the recognition efficiency is lower than that of the other three methods. MLCNNwav is a recognition model based on multi-scale CNN. Although it considers multi-modal sensor features, it does not consider their temporal features, so the recognition accuracy is lower than that of MConvLSTM. MFLSTM proposes a four-layer hybrid LSTM recognition model. Although it considers the temporal features of multi-modal sensors, its ability to process original sensor signals is poor, so the recognition performance is lower than that of MConvLSTM. Compared to the CNN and LSTM methods, MConvLSTM can better handle the spatiotemporal behaviour information in sensor data. It uses convolution to extract features from raw signals and then further processes temporal features using an LSTM model, thereby enhancing the recognition effects.

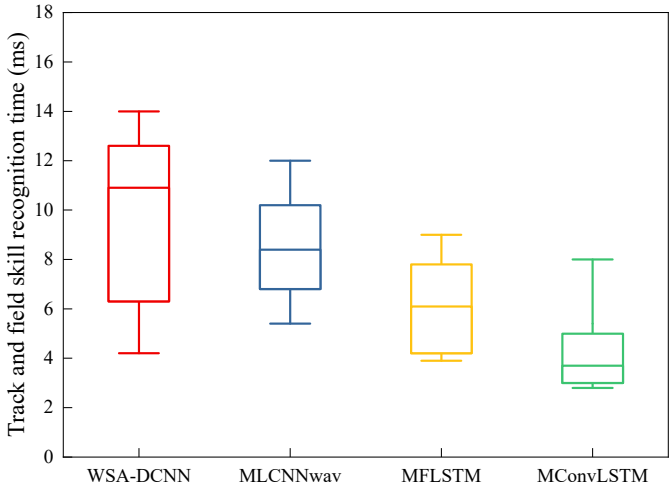
Table 2 Recognition performance indicators of different methods

Method	*WSA-DCNN*	*MLCNNwav*	*MFLSTM*	*MConvLSTM*
Macro_P (%)	81.95	86.49	88.42	94.18
Macro_R (%)	80.31	83.56	90.38	92.63
Macro_F1 (%)	81.12	84.99	89.39	93.40

The comparison of identification times for different methods is shown in Figure 3. MConvLSTM demonstrates significant advantages in the recognition of athletics movement skills. The shortest recognition time for MConvLSTM is 4.37 ms, while the shortest recognition times for WSA-DCNN, MLCNNwav, and MFLSTM are 6.05 s and 7.42 s, respectively. Compared to two literature methods, the shortest recognition times for athletics foul action recognition are 10.19 ms, 7.47 ms, and 6.05 ms, respectively. This indicates that MConvLSTM has a higher computational efficiency and recognition speed when handling athletics movement skill recognition tasks. MConvLSTM can significantly shorten the time required for the recognition process while maintaining high

recognition accuracy, thus improving the real-time and application efficiency of the proposed method.

Figure 3 Comparison of recognition times for different methods (see online version for colours)



6 Conclusions

Accurate recognition of athletics movement skills is a core part of scientific training and performance improvement. To address issues such as information redundancy, inadequate feature extraction from sensors, and low recognition accuracy in current research, this paper proposes an athletics movement skill recognition method based on multimodal sensor data. First, a multimodal sensor data adaptive selection mechanism based on mutual information is proposed. By calculating the mutual information of different sensor combinations under different movement skill categories, the sensor combinations that provide the maximum information correlation are selected, thereby improving the efficiency and accuracy of data fusion. Then, CNN is combined with LSTM for multimodal sensor feature extraction, and a multi-sensor feature fusion method based on a recurrent matrix is proposed, achieving effective fusion of behaviour features from multimodal sensors. Finally, the integrated characteristic vector is input into a fully connected layer, and the softmax function is used to compute the scores for each category of athletics movement skills from the output classification layer. Experimental outcome implied that the suggested approach achieves an average recognition accuracy of 94.17% for various athletics movement skill types, with the shortest recognition time of 4.37 ms, enabling accurate and real-time recognition of athletics movement skills.

In summary, the proposed method in this paper demonstrates good recognition performance. However, due to the limited knowledge and time available, the analysis of many situations is not thorough enough, and further research is needed.

1 For redundant information in multimodal heterogeneous sensor data, adaptive selection based solely on mutual information is insufficient. An attention mechanism could be introduced to focus on the main behavioural features of multimodal data, thereby ignoring unnecessary information redundancy.

2 This paper only utilises multimodal sensor data. In the future, more heterogeneous information, such as depth images and skeletal data, could be incorporated to further improve the accuracy of athletics movement skill recognition.

Declarations

All authors declare that they have no conflicts of interest.

References

Ahmad, T., Rafique, J., Muazzam, H. and Rizvi, T. (2015) 'Using discrete cosine transform based features for human action recognition', *Journal of Image and Graphics*, Vol. 3, No. 2, pp.96–101.

Borsoi, R.A., Usevich, K., Brie, D. and Adali, T. (2024) 'Personalized coupled tensor decomposition for multimodal data fusion: uniqueness and algorithms', *IEEE Transactions on Signal Processing*, Vol. 73, pp.113–129.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S. and Miao, Y. (2021) 'Review of image classification algorithms based on convolutional neural networks', *Remote Sensing*, Vol. 13, No. 22, pp.47–62.

Cui, G. and Wang, C. (2025) 'The machine learning algorithm based on decision tree optimization for pattern recognition in track and field sports', *Plos One*, Vol. 20, No. 2, pp.31–44.

Dahou, A., Al-Qaness, M.A., Abd Elaziz, M. and Helmi, A.M. (2023) 'MLCNNwav: multilevel convolutional neural network with wavelet transformations for sensor-based human activity recognition', *IEEE Internet of Things Journal*, Vol. 11, No. 1, pp.820–828.

Dang, L.M., Min, K., Wang, H., Piran, M.J., Lee, C.H. and Moon, H. (2020) 'Sensor-based and vision-based human activity recognition: a comprehensive survey', *Pattern Recognition*, Vol. 108, pp.10–21.

Fong, D.T., Ko, J.K. and Yung, P.S. (2021) 'Using fast Fourier transform and polynomial fitting on dorsal foot kinematics data to identify simulated ankle sprain motions from common sporting motions', *Journal of Mechanics in Medicine and Biology*, Vol. 21, No. 4, pp.35–47.

Gai, X. (2025) 'Application of flexible sensor multimodal data fusion system based on artificial synapse and machine learning in athletic injury prevention and health monitoring', *Discover Artificial Intelligence*, Vol. 5, No. 1, pp.31–46.

Hoque, N., Bhattacharyya, D.K. and Kalita, J.K. (2014) 'MIFS-ND: a mutual information-based feature selection method', *Expert Systems with Applications*, Vol. 41, No. 14, pp.6371–6385.

Hsu, Y-L., Chang, H-C. and Chiu, Y-J. (2019) 'Wearable sport activity classification based on deep convolutional neural network', *IEEE Access*, Vol. 7, pp.170199–170212.

Lee, S., Lim, Y. and Lim, K. (2024) 'Multimodal sensor fusion models for real-time exercise repetition counting with IMU sensors and respiration data', *Information Fusion*, Vol. 104, pp.17–23.

Li, J., Tian, S. and Charoenwattana, S. (2023) 'Smart IoT-based visual target enabled track and field training using image recognition', *Soft Computing*, Vol. 27, No. 17, pp.12571–12585.

Liu, C. and Chang, Z. (2022) 'Sensor and attitude analysis of track and field training action recognition based on artificial intelligence', *Journal of Sensors*, Vol. 20, No. 1, pp.62–70.

Liu, Z. and Wang, X. (2023) 'Action recognition for sports combined training based on wearable sensor technology and SVM prediction', *Preventive Medicine*, Vol. 173, pp.75–82.

Mekruksavanich, S. and Jitpattanakul, A. (2022) 'Multimodal wearable sensing for sport-related activity recognition using deep learning networks', *Journal of Advances in Information Technology*, Vol. 13, No. 2, pp.132–138.

Pereira, J., Hastie, P., Araújo, R., Farias, C., Rolim, R. and Mesquita, I. (2015) 'A comparative study of students' track and field technical performance in sport education and in a direct instruction approach', *Journal of Sports Science & Medicine*, Vol. 14, No. 1, pp.11–28.

Piechocki, R.J., Wang, X. and Bocus, M.J. (2023) 'Multimodal sensor fusion in the latent representation space', *Scientific Reports*, Vol. 13, No. 1, p.2005.

Ullah, M., Yamin, M.M., Mohammed, A., Khan, S.D., Ullah, H. and Cheikh, F.A. (2021) 'Attention-based LSTM network for action recognition in sports', *Electronic Imaging*, Vol. 33, pp.1–6.

Wu, Q., Xia, J., Dai, P., Zhou, Y., Wu, Y. and Ji, R. (2024) 'Cycletrans: Learning neutral yet discriminative features via cycle construction for visible-infrared person re-identification', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 36, No. 3, pp.5469–5479.

Yao, J. and Li, Y. (2022) 'Youth sports special skills' training and evaluation system based on machine learning', *Mobile Information Systems*, Vol. 10, No. 2, pp.60–72.

Yu, W. and Xing, J. (2022) 'Sports event model evaluation and prediction method using principal component analysis', *Wireless Communications and Mobile Computing*, Vol. 4, No. 2, pp.93–102.

Zhang, L. (2022) 'Applying deep learning-based human motion recognition system in sports competition', *Frontiers in Neurorobotics*, Vol. 16, pp.86–94

Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y. and Alshurafa, N. (2022) 'Deep learning in human activity recognition with wearable sensors: a review on advances', *Sensors*, Vol. 22, No. 4, pp.14–26.

Zhang, X. (2021) 'Application of human motion recognition utilizing deep learning and smart wearable device in sports', *International Journal of System Assurance Engineering and Management*, Vol. 12, No. 4, pp.835–843.

Zhang, Y. (2023) 'Track and field training state analysis based on acceleration sensor and deep learning', *Evolutionary Intelligence*, Vol. 16, No. 5, pp.1627–1636.

Zheng, S. and Man, X. (2022) 'An improved logistic regression method for assessing the performance of track and field sports', *Computational Intelligence and Neuroscience*, Vol. 2, No. 1, pp.63–75.