ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

# Design of a Deep Learning-based System for Recognizing and Evaluating Traditional Martial Arts Movements

*Xu Wang*[1], *Xiaoyun Fan*[2], *Biao Ma*[1] *and Jingtang He*[1,*]

[1] College of Physical Education, Huainan Normal University, Huainan, Anhui, 232038, China
[2] Art and Sports Department, North Anhui Electronic Information Engineering School, Fuyang, Anhui, 236600, China

**SUMMARY:** *Action recognition and evaluation of traditional martial arts routines has always been difficult, and an intelligent solution based on deep learning is proposed to address this problem. This solution combines the convolutional neural network as well as the long and short-term memory network algorithms, the convolutional neural network is used to extract the spatial features of each video frame, and the long and short-term memory network is used to mine the characteristics of the action sequences of the wushu routines. Meanwhile, the fusion of different scale features is carried out in the feature layer, and the public pose estimation algorithm is used to predict the position of human joints as auxiliary information, and the position of joints and the image features are weighted and fused by attention to get the final image features. Finally, the end-to-end approach is used for model training, combined with the multi-objective classification regression method to improve the classification performance and prediction effect. The test shows that the recognition accuracy of the five types of traditional martial arts movements reaches 94.2%, and the correlation between the system prediction value and the score given by the boxer is r=0.89, which indicates that this system has certain practical application value and feasibility.*

***KEYWORDS:*** *deep learning; traditional wushu; action recognition; evaluation system; multimodal fusion*

# 1 Introduction

As an important part of Chinese culture, traditional wushu carries profound historical and cultural values [1]. As a sport, traditional wushu demonstrates the wisdom of the Chinese nation in every move, which not only attracts the enthusiastic participation of the majority of Chinese people, but is also increasingly respected and admired around the world. However, most of the teaching methods of traditional wushu are based on the mode of teacher-disciple transmission, and the relative scarcity of wushu instructor resources has led to low teaching efficiency, and this mode of teaching is often difficult to meet the wide range of learning needs as well as the personalized cultivation of the instructed [2, 3]. In wushu competitions and examinations, the way of wushu movement identification and assessment is usually carried out by judges according to established standards, and this way of judging may lead to inconsistent judging results due to the influence of subjective factors [4]. Therefore, it is particularly important to explore a scientific and objective method for recognizing and assessing wushu movements. In recent years, with the rise of deep learning, human movement recognition has been developed

*hejingtanghn@163.com
https://doi.org/10.65102/is2026668
Published: 30 April, 2026

extremely fast [5]. The automatic acquisition of motion features through deep learning greatly reduces the cost of feature acquisition, and with this advantage it can play an important role in wushu movement recognition and evaluation [6]. By designing a traditional wushu movement recognition and assessment system based on deep learning, the intelligent recognition and assessment of wushu movements can be realized through the preprocessing, feature extraction and modeling of wushu movements, which can help to reduce the problems such as the lack of objectivity, insufficient teacher's power, and low motivation of students in traditional methods [7, 8].

Aiming at the application of deep learning and its related technologies in traditional martial arts movement recognition and assessment, literature [9] proposed a spatio-temporal hierarchical key point aggregation framework based on deep learning, analyzed its excellent performance in complex martial arts movement recognition and localization by considering the key points of the human skeleton as a three-dimensional point cloud, and pointed out that the model can effectively overcome the deficiencies of the traditional methods in terms of occlusion and changes in appearance, and provided a technical support for accurate assessment and training guidance. Literature [10] introduces a wushu training evaluation model based on deep learning technology, analyzes its effectiveness in dealing with massive and diversified wushu evaluation data by constructing a deep neural network and fusing multiple raw data, and points out that the model can provide decision support for offline performance prediction, personalized learning recommendation, etc., and ultimately improve the efficiency of wushu teaching and learning. Literature [11] analyzed the application of deep learning in improving human posture estimation by constructing a lightweight convolutional feature extraction network and combining it with a long and short-term memory network, pointing out that the model significantly reduces the computational complexity and improves the accuracy of action recognition, and then verifies its good performance in recognizing different martial arts actions. Literature [12] proposes a deep learning framework that integrates multi-scale and structured attention mechanisms, analyzes its advantage of dynamically focusing on key parts and spatial relationships in recognizing fine gestures in martial arts, points out that the model effectively improves generalization ability and recognition accuracy by preventing overfitting, and emphasizes its superior performance compared with traditional methods. Literature [13] describes a deep learning system that integrates multi-view posture estimation and spatio-temporal graphical convolutional networks, analyzes its application in real-time evaluation and personalized feedback of taijiquan movements, points out that the system significantly improves the training efficiency and the quality of the movements, and emphasizes the technical advantages of standardized and scalable teaching while preserving the cultural essence. Literature [14] explored the optimization method of wushu movement recognition based on deep learning algorithms, analyzed the recognition challenges of wushu movements due to the variety and nuance of differences, and examined the positive effects of the method on improving recognition accuracy and efficiency by optimizing feature extraction and model training. Literature [15] explored the application of deep learning in the recognition and evaluation of traditional wushu movements, analyzed the advantages in processing time and accuracy by reviewing the 2D posture estimation method based on RGB images and convolutional neural network (CNN), and conducted movement evaluation by using homemade datasets and measurements such as joint lengths and angles, which emphasized the importance of the method for the preservation and inheritance of traditional wushu culture. Value.

Literature [16] utilizes deep learning techniques to achieve automatic recognition of karate punches. By analyzing wrist acceleration sensor data and employing multiple convolutional neural network models, the method's high-precision performance under simulated real-world combat conditions is examined and its potential for optimizing the training process and

facilitating the application of automated learning systems is noted. Literature [17] proposed a PoseGCN model based on graph convolutional networks to address the challenge of complex spatio-temporal relationships in martial arts leg movement recognition. By integrating spatial, temporal, and contextual features and utilizing a specific attentional mechanism, we analyzed the excellent performance of the model that outperforms the existing methods on multiple datasets, and pointed out its potential in capturing subtle movements and achieving accurate recognition. Literature [18] analyzed martial arts movement decomposition node data using neural network algorithms to improve the accuracy of movement recognition and matching by capturing subtle changes and tracking joint angles, and pointed out that the method is effective in assisting training, while emphasizing its technical potential for promoting the protection and development of traditional martial arts. Literature [19] analyzed the research trend of wushu biomechanics between 2011 and 2015 through bibliometrics, examined the performance of the BP neural network model in movement classification assessment, pointed out that it was superior to methods such as Park Bayes in terms of accuracy, recall, and other metrics, and emphasized the necessity of applying deep learning techniques to innovate wushu biomechanics research. Literature [20] developed a deep learning framework combining CNNs and gated recurrent units (GRUs) for the recognition and evaluation of martial arts movements, achieving up to 94% recognition accuracy with low stance estimation error by extracting visual features and analyzing the temporal dynamics, and highlighting the potential of applying the framework in providing interpretable analysis to support fine-tuned training and skill evaluation. Literature [21] explored how to optimize and integrate the OpenPose pose estimation algorithm into a movement recognition system, analyzed its potential in improving the accuracy and robustness of traditional martial arts movement recognition, and proposed an efficient and reliable recognition framework by comparing the mainstream neural network approaches to promote the heritage and development of martial arts culture. Literature [22] proposes a 3D gesture estimation method for wushu combining hand kinematic model and CNN, analyzes its accuracy enhancement in gesture regression by extracting features and simulating joint dependencies using morphological topology, and points out that the method outperforms the existing techniques in terms of estimation error and inference speed, which demonstrates the feasibility of its application in wushu movement evaluation.

In this paper, a traditional wushu movement classification and evaluation model based on multimodal fusion is proposed. The model is jointly modeled using CNN and LSTM, with CNN capturing the spatial features of each video frame and BiLSTM capturing the temporal correlation information of the whole action. Secondly, the position information of 17 key points obtained from the open-source pose estimation algorithm is inputted into the network, and the pose information and image information are automatically assigned weights using the attention module. And a network model based on the end-to-end multi-task learning architecture is proposed to train the network with classification loss + regression loss + regularization term, as the final loss function, to optimize the evaluation sub-task while completing the recognition; a hybrid model is constructed to accurately recognize the martial arts movements and evaluate them in an all-around quantitative manner, which realizes the digital dissemination of martial arts and the application of the teaching method.

# 2 Motion recognition algorithms

## 2.1 Deep Learning Algorithms

With the successful application of deep learning in the field of computer vision, it has led to a dramatic change in the research direction of video-based action recognition methods. Feature

extraction using convolutional neural networks has become the key to this class of problems, where multiple convolutional kernels perform top-down feature abstraction of the input video sequence, obtaining low-level to high-level feature description information from shallow to deep. Due to the incorporation of a 3D convolutional neural network, the model can process information in the spatio-temporal dimension in this process, and its convolution process can be expressed as follows:

$$y_{xyz} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \sum_{d=0}^{D-1} w_{hwd} \cdot X_{(x+h)(y+w)(z+d)} \tag{1}$$

where $w_{hwd}$ denotes the 3D convolutional kernel weights, and $x$ and $y$ are the input and output feature maps, respectively.

Typical models such as ResNet3D and I3D utilize residual connectivity and dual-stream structure design to overcome the phenomenon of gradient vanishing and achieve good results in action sequence recognition. The forgetting gate of long and short-term memory networks determines the degree of retention of cell state information:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2}$$

The input gate controls the storage of new information:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

The candidate values are generated by hyperbolic tangent function:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

The cell status is updated to:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

The output gate controls the final output:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

Hidden status for:

$$h_t = o_t * \tanh(C_t) \tag{7}$$

The gated recurrent unit as a simplified version of the long and short-term memory network reduces the computational overhead while maintaining the modeling capability and shows a good balance of efficiency in action recognition tasks.

Action recognition methods based on self-attention models have also made promising progress. The Transformer model, which has been a great success in NLP, has also been migrated to video analytics due to the fact that self-attention is able to model the long-distance correlation between all the positions in a sequence very well. Vidoe Transformer treats each frame of an image or spatio-temporal voxel as each position in a sequence, and uses a multi-head self-attention mechanism to capture the complex interactions between frames, with

attention weights represented as:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where $Q$, $K$, and $V$ denote the query, key, and value matrices, respectively.

The spatio-temporal attention mechanism utilizes spatial attention to focus on the important parts in each frame and temporal attention to focus on the important parts on the moments, and its decomposition is designed to improve the understanding of complex actions. Skeleton based action recognition relies on graph convolutional networks, where the presence of a graphical structure on the human skeleton allows modeling each joint as a node and the connections between bones as edges.

Spatio-temporal graph convolutional network connects the corresponding joints of neighboring frames by constructing a spatio-temporal graph, the graph convolution operation is defined as:

$$f_{out}(v_{t_i}) = \sum_{v_{t_j} \in B(v_{t_i})} \frac{1}{Z_{t_i}(v_{t_j})} f_{in}(v_{t_j}) \cdot w(I_{t_i}(v_{t_j})) \tag{9}$$

where $B(v_{t_i})$ denotes the neighborhood of node $v_{t_i}$, $Z_{t_i}$ is the normalization term, and $w$ is the learnable weight function.

Adaptive GCN discards the fixed graph and learns directly from the given input to obtain an adaptive neighbor matrix to represent the relationship between the hidden joints. The introduction of multimodal information fusion greatly improves the accuracy of the model, where the color vision images represent visual features, the optical flow videos represent dynamic features, the depth videos represent 3D shape information, and the skeletal videos represent topological structure information. The dual-stream network processes the color images and optical flow and then fuses the results of the two networks, well integrating spatial appearance and temporal motion, and the multistream fusion architecture further increases the input modalities and utilizes attention to synthesize the information from the different modalities, making full use of their complementary strengths.

## 2.2 Traditional Wushu Movement Recognition

Wushu action recognition is one of the research directions combining computer vision and kinesiology, which has received extensive attention from researchers in research institutes and enterprises in the last few years and has made some progress.These research works are mainly based on traditional machine learning algorithms, i.e., manually designing some feature descriptors for complex motion description of wushu actions. The contour-based action recognition method obtains the contour of the human body from the executed action to perform the recognition process. And geometric features such as Fourier descriptors or Hu moments are used for action recognition, but these methods are sensitive to changes in lighting and background.

The optical flow technique captures the spatio-temporal features of the action by calculating the pixel motion vectors between consecutive frames with an optical flow vector of:

$$v(x,y,t) = (u(x,y,t), v(x,y,t)) \tag{10}$$

It describes the motion velocity of the pixel point $(x, y)$ at the moment $t$, and $u$ and $v$ denote the velocity components in the horizontal and vertical directions, respectively.

In recent years, the rapid development of deep learning methods provides new technical support for traditional martial arts action recognition, among which the powerful image feature learning ability of convolutional neural network makes it the main method for video action recognition. Scholars have extended the 2D convolution to the 3D time-space domain, and 3D-CNN is used to learn the original video directly to obtain the feature expression in time-space. The RGB image and optical flow are input into two different convolutional neural networks and fused at the end to obtain better spatial appearance information as well as temporal motion information, which results in higher accuracy.RNN has good sequence modeling ability and is suitable for mining the temporal information of wushu movements, and LSTM can better solve the long-range dependency, which can well capture the complex movements of wushu set Transformer process. Transformer-based visual Q&A method has high accuracy in video action classification, and Self-attention can model the correlation between any two frames, which provides a new technical idea for global spatio-temporal dependency understanding of the whole wushu action. Graph Convolutional Network is currently one of the mainstream methods in skeleton based action recognition, after building the graph structure on the human skeleton, it uses the connection between joints for feature transfer and fusion, which has a good adaptability for actions involving interactions between multiple joints such as wushu actions.

# 3 Design of a deep learning-based traditional martial arts movement recognition system

## 3.1 Overall system architecture

The overall process design of action discrimination and evaluation of traditional martial arts routines reflects the concept of combining artificial intelligence and traditional martial arts, and the overall design scheme of the system proposed by the author contains the technical route from data input to result output is shown in Fig. 1. The data input part of the IMU is a multi-source heterogeneous data fusion acquisition process, in which the RGB camera is used to capture the action video of the martial arts routines, and the Depth camera is used to obtain the spatial position of the movements of the martial arts routines. At the same time, the IMU is coupled with accelerometers and gyroscopes to capture the subtle body movement information, so as to achieve the purpose of comprehensively capturing the complex martial arts movements from different angles.

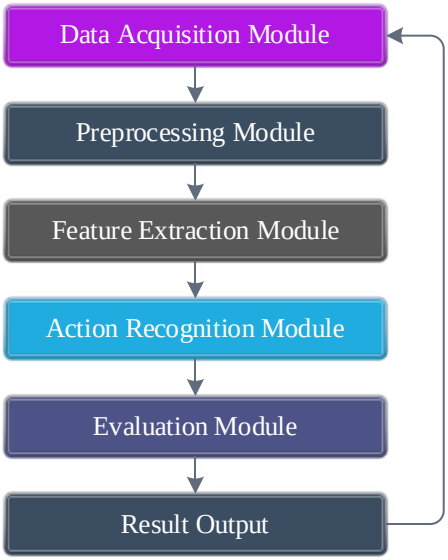


*Figure 1: Complete processing link*

The preprocessing part mainly denoises the original captured data, removes the frame jitter caused by camera instability, calculates in real time and selects an appropriate low-pass filter for filtering according to the actual scene. Frame filling can solve the broken phenomenon that occurs in the captured images due to the large shooting interval, and use image enhancement technology to enrich the image information and improve the recognition accuracy.

In the feature extraction layer, a multi-scale deep learning network is utilized to complete the function of this level, in which the bottom network captures the basic visual information such as edge texture, the middle network represents the local motion patterns, and the high level network encodes the global action semantics. Meanwhile, the design of the convolutional feature extraction network adopts a combination of residual connectivity and attention mechanism to solve the gradient vanishing problem and highlights important feature regions. Temporal feature modeling employs a bidirectional long and short-term memory network to provide adequate fusion of forward and backward information. Posture feature extraction uses a keypoint detection algorithm to obtain coordinate information of 17 major joints and computes the structured feature representation based on joint angles.

Action recognition is the most important part of the whole system, the classification network constructed in this paper uses temporal convolution of different lengths and self-attention for modeling to capture the relationship between long time intervals. Multi-tasking is introduced to train the two tasks of action recognition and temporal location together to get better results, and the schedule form is used to expand single action recognition to composite action recognition for fast learning and accuracy improvement during training. The network parameters are allowed to be updated in real time using user feedback and newly generated training data.

The evaluation unit evaluates the action effect of the recognized characters according to the multi-source scoring model, and the evaluation of the degree of action specification is to compare the similarity of each action with the template. And DTW is used to solve the problem of sequence mismatch, the evaluation of temporal consistency is to make a judgment on the smoothness and reasonableness of the action, and the evaluation of strength expression is to comprehensively judge the sense of strength of the action based on the acceleration and appearance images. The global coherence analysis includes the coherence between movements and the overall movement evaluation of body postures between movements.

The openness and compatibility of the system ensures the increase of functions and

performance in the future, and the algorithms can be easily replaced. Parallel technology is used to solve the problem of high computation due to high real-time requirements, and multiple GPUs can be invoked simultaneously in the system to perform computation. In the cloud server, the tasks can be reasonably assigned to each CPU through the way of side cloud collaboration to achieve the purpose of balance. In terms of security and privacy, end-to-end encryption is used to ensure user privacy and data security, and the user interface design is based on ergonomic principles to achieve friendly interaction and visualized feedback information.

## 3.2 Data Acquisition and Preprocessing

For the study of traditional martial arts motion capture system, the article uses various types of sensors for information collection, and can collect information about the whole movement from visual, inertial and sound perspectives, and has a strong processing capability in terms of the difficulty of the movement and the influence of external conditions. For the collection of video images, the article uses a high-definition camera and a structured light three-dimensional camera to complete this part of the work. A color camera was used and the image collection rate was 60 frames/second. The information on the external morphology and details of the martial arts movements is obtained. On the other hand, the RGB-D camera can acquire millimeter-level 3D coordinates of the indoor scene for post-pose recognition computation, and can well capture the complex and fast dynamic body poses of martial arts movements.The IMU installs wearable sensors on the arms, torso and legs to detect the body movements, and acquires the data at a frequency of 200 Hz including acceleration, angular velocity as well as magnetic force. The effectiveness of high-speed dynamic feature extraction can be ensured. The voice signal acquisition uses a directional microphone to collect the sound signals such as exhalation and exertion required for the completion of martial arts movements with a resolution of 48,000 times per second (48 kHz) and 16 bit, which is a strong reference value for the analysis of the movements, and the system clock and the program set a time point for the unified timing of the various sensor signals, with an error of less than 0.01 s. This realizes the requirements of the accurate time analysis technology. The requirement of accurate time analysis technology is realized.

The design of the data preprocessing pipeline will directly affect the training of the deep learning network and the final recognition performance. The preprocessing pipeline proposed in this paper contains functional modules such as data cleaning, data transformation, data enhancement, and data labeling and verification. For the video data, firstly, the de-jittering process is carried out using the keypoint-based method to remove the effect of camera jitter, by detecting the keypoints between neighboring frames and deriving the affine transform array:

$$T = \begin{bmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

The transformation parameters $a$, $b$, $c$, $d$ together with the translation components $t_x$ and $t_y$ form the mathematical basis of motion compensation. The light normalization process improves the visual quality under different lighting conditions through histogram equalization and adaptive contrast enhancement techniques, normalized pixel values, according to the following equation:

$$I_{norm}(x,y) = \frac{I(x,y) - \mu}{\sigma} \times \sigma_{target} + \mu_{target} \tag{12}$$

The frame interpolation algorithm utilizes a bidirectional optical flow estimation method to deal with frame rate inconsistencies.

$$I_t = (1-t)I_0 + tI_1 + (1-t)tV_{t\to 0} + t(1-t)V_{t\to 1} \tag{13}$$

Depth data preprocessing consists of a joint processing mechanism of void filling and noise filtering, where void filling is based on interpolation of neighborhood information to recover missing depth values, and noise filtering uses a three-dimensional joint filter to remove measurement noise.

$$D_{Filtered}(x,y,t) = \sum_{i,j,k} w(i,j,k) \cdot D(x+i, y+j, t+k) \tag{14}$$

The inertial data preprocessing focuses on solving the sensor drift problem, and the Kalman filter performs the state estimation through the state transfer equation with the observation equation.

$$x_{k+1} = F_k x_k + B_k u_k + w_k \tag{15}$$

$$z_k = H_k x_k + v_k \tag{16}$$

The gravity compensation algorithm follows the elimination of the effect of gravitational acceleration.

$$a_{linear} = a_{measured} - R \cdot g \tag{17}$$

Semi-automatic labeling tools were utilized in the data labeling process, as well as checking and correcting by professionals in the field of martial arts to improve the labeling efficiency and accuracy. In terms of data expansion, resampling, rescaling, noise addition, time stretching and other methods were adopted to expand the original data to 5 times of the original data volume. Moreover, various automated and manual testing techniques are introduced to ensure the quality of the data, and ultimately obtain high-quality data for DL model training.

## 3.3 Deep Learning Model Selection and Training

Aiming at the characteristics of traditional martial arts movements, based on the research on deep neural networks that can effectively extract spatial information and time series correlation, and comparing the effectiveness of commonly used deep neural network models, a combined CNN-LSTM-based recognition model is designed and proposed. Since ResNet-50 residual network has good image classification ability, it is applied to the spatial information of each video frame. The residual link in the network, which solves the problem of gradient dispersion that tends to occur in deep neural networks, is expressed as:

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{18}$$

where $\mathcal{F}(x, \{W_i\})$ represents the residual mapping function and $x$ represents the input

feature vector.

Considering the complexity of martial arts movements containing multi-scale motion features, a feature pyramid network structure is integrated on the ResNet-50 infrastructure, which realizes the effective fusion of features at different levels through the top-down information propagation path and the lateral connection mechanism, and the computational formula for multi-scale feature fusion is:

$$P_i = Conv_{1\times1}(C_i) + Upsample(P_{i+1}) \tag{19}$$

where $C_i$ denotes the $i$ th layer of convolutional features and $P_i$ is the fused pyramidal feature representation.

The temporal modeling adopts a bidirectional long and short-term memory network architecture, and the computation of the forward hidden state $h_t$ and the backward hidden state $h_t^{\leftarrow}$ is based on the combined analysis of the current input and historical information. The introduction of the self-attention mechanism significantly enhances the model's ability to recognize key action frames, and the attention weight is calculated by the following formula:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{K}\exp(e_{ik})} \tag{20}$$

The attentional energy function is defined as:

$$e_{ij} = v^T \tanh(W_h h_i + W_s h_j) \tag{21}$$

In this, the parameter matrices $v$, $W_h$, and $W_s$ are optimized by end-to-end training.

The model is trained in an end-to-end manner, in which about 100 to 150 videos for each action category are used for training, with a total of 25 traditional boxing styles, and a total of 3000 video samples are trained and 750 video samples are validated. All input videos were scaled to images of size 224 × 224 and were sampled into video sequences of 30 fps frame rate as well as 64 frames long to ensure consistent batch size. A composite loss function is defined to jointly optimize multiple objectives, i.e., combining the classification loss with the timing consistency constraint, then the total loss function can be expressed as:

$$L = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 + \lambda\sum_{t=1}^{T-1}\left\|h_t - h_{t+1}\right\|_2^2 \tag{22}$$

The loss function contains the mean square error term and the time-series smoothing regularization term, $\lambda$ is the balance coefficient, $N$ is the batch size, $T$ is the sequence length, and $h_t$ denotes the hidden state representation of the $t$ th frame. The optimizer chooses Adam's algorithm, which combines the technical advantages of momentum method and adaptive learning rate, and the parameter update rule is:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \tag{23}$$

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected estimates of the first-order and second-order

moments, respectively.

The training process adopts the form of course schedule, i.e., simple actions are trained first, and then more complex actions are added gradually. At the initial stage of training, only a single category of actions is used to pre-train the network, and then composite actions as well as coherent actions are added, which is a gradual approach that can improve the convergence efficiency and effectiveness. In addition, in order to further improve the robustness of the network, data expansion processing is also carried out, which is mainly realized by intercepting segments, mirror transformation, variable speed, and adding Gaussian white noise points. Time stretching was used to simulate different execution speeds with different rates, and the distortion coefficients were randomly selected from (0.8, 1.2), and the standard deviation of the Gaussian white noise was set to 0.01 so as not to affect too much of the original sound information. The label distribution after smoothing is:

$$q_i = (1-\alpha) y_i + \frac{\alpha}{K} \tag{24}$$

where $\alpha$ is the smoothing parameter and $K$ is the total number of categories. Learning rate scheduling is done using cosine annealing strategy and the learning rate is adjusted according to the following equation:

$$\eta_t = \eta_{\min} + \frac{1}{2}\left(\eta_{\max} - \eta_{\min}\right)\left(1 + \cos\left(\frac{t\pi}{T}\right)\right) \tag{25}$$

where $t$ is the current number of training steps and $T$ is the total number of training steps.

TensorBoard was utilized for training to observe the change of parameters such as loss value, accuracy, and learning rate over time, and to end the training early based on the validation set effect to avoid overfitting of the model. We set to stop training if the validation set error is not further reduced during 10 iterations. Models were saved using the checkpoint method, where a model was saved in every 5 epochs and tested with the best performing model on the validation set.

## 3.4 Assessment indicators and methodology

The evaluation method for traditional wushu movements is formulated based on the characteristics of traditional wushu, integrating the sports ergonomics in modern sports science on the basis of classical wushu, including the indicators of accuracy, fluency, strength and rhythm, and the corresponding evaluation model is established. Movement accuracy is judged mainly by the accurate comparison of postural key points, which is an algorithm for judging based on the positional relationship of each key point on the body, and a method of scoring based on the size of the similarity between the performed movement and the movement of the reference template.

The dynamic time regularization algorithm is used in the similarity calculation process to deal with the complex problem of temporal alignment, and the mathematical expression of the distance metric function is:

$$D(i, j) = \left\| P_i - Q_j \right\|_2 + \min\left\{ D(i-1, j), D(i, j-1), D(i-1, j-1) \right\} \tag{26}$$

where $P_i$ and $Q_j$ represent the corresponding joint point coordinate vectors in the execution

sequence and standard sequence, respectively.

The quantification of joint angle deviation is realized by calculating the angle difference of the corresponding joints at the same time node, which is expressed as:

$$\Delta\theta_{ijk} = \left|\theta_{ijk}^{exec} - \theta_{ijk}^{std}\right| \tag{27}$$

where $\theta_{ijk}^{exec}$ and $\theta_{ijk}^{std}$ denote the values of the angles consisting of the joints $i$, $j$, and $k$ in the executive and standard movements, respectively.

The analysis of the trajectory of the center of gravity of the body assesses the stability and coordination of the movement by calculating the change rule of the position of the center of gravity in the spatial and temporal dimensions, and the coordinates of the center of gravity are accurately calculated according to the following formula:

$$C = \frac{1}{N}\sum_{i=1}^{N} m_i p_i \tag{28}$$

where $m_i$ denotes the mass weight coefficient of the $i$ th body part and $p_i$ is the corresponding spatial position vector.

The assessment of the range of motion is achieved by in-depth analysis of how well the range of motion of the joints matches the requirements of the standard, and the range of motion metric is defined as the weighted average of the maximum angular change of each joint:

$$A = \sum_{j=1}^{J} w_j \cdot \max_t(\theta_j(t)) - \min_t(\theta_j(t)) \tag{29}$$

Among them, the weight coefficient $w_j$ needs to be set reasonably according to the important role played by different joints in a particular martial arts movement.

The evaluation index system of traditional wushu movements is shown in Table 1.

*Table 1: Evaluation index system of traditional martial arts movements*

| Evaluation dimensions | Specific indicators | Calculation method | Weight coefficient | Scoring range |
|---|---|---|---|---|
| Technical accuracy | Posture matching degree | DSS similarity | 0.25 | 0-100 |
| | Angular deviation | Angle difference statistics | 0.20 | 0-100 |
| | Gravity trajectory deviation | Trajectory distance calculation | 0.15 | 0-100 |
| | Action amplitude matching | Acceleration variance | 0.15 | 0-100 |
| Fluency of movement | Velocity smoothing | Motion range contrast | 0.20 | 0-100 |
| | Transformation nature | Bridging point analysis | 0.18 | 0-100 |
| | Rhythm consistency | Frequency domain feature extraction | 0.16 | 0-100 |
| | Physical coordination | Interlocking analysis | 0.14 | 0-100 |
| Performance of strength | Explosive indicator | Acceleration peak | 0.18 | 0-100 |
| | Control accuracy | Force curve fitting | 0.16 | 0-100 |
| | Strength change rationality | Sequential analysis | 0.12 | 0-100 |
| Overall performance | Completion evaluation | Action integrity | 0.15 | 0-100 |
| | Artistic expression | Expert subjective score | 0.10 | 0-100 |

The assessment of movement smoothness requires a comprehensive analysis of multiple sub-dimensional indicators to fully reflect the coherence characteristics and sense of natural flow embodied in the execution of wushu movements. The smoothness of velocity change is accurately measured by calculating the second-order derivatives of the velocity sequence at each joint point. The formula for the smoothness index is:

$$S = \frac{1}{T-2}\sum_{t=2}^{T-1}\left\|a_t\right\|_2 \tag{30}$$

where $a_t$ denotes the acceleration vector corresponding to the $t$ th moment.

The naturalness assessment of the action transition process focuses on the articulation quality between adjacent action segments, and is quantitatively assessed by deeply analyzing the change patterns of velocity and acceleration in the region near the transition point, and the transition smoothness is calculated using the formula:

$$T_s = \exp\left(-\lambda \cdot \max_{t\in[t_1-\delta,t_1+\delta]}\left\|v_t - v_{t_1}\right\|_2\right) \tag{31}$$

where $t_1$ represents the critical moment of action transition, $\delta$ is the size parameter of the time window, and $\lambda$ is the decay coefficient.

Rhythm consistency was assessed by analyzing the regularity of the temporal distribution pattern during the execution of the movement in depth to judge the rhythm control ability of the practitioner, and the Fourier transform technique was used to extract the characteristic information of the movement sequence in the frequency domain, and the main frequency components were precisely calculated by the integral formula:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt \tag{32}$$

The rhythmic stability index is scientifically defined based on the variance characteristics of the main frequency components. The evaluation of coordination between body parts measures the level of overall coordination and cooperation by analyzing the synchronization characteristics during the movement of different body parts, and the coordination index adopts the mutual correlation function as:

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)y(t+\tau)dt \tag{33}$$

The correlation strength between the different joint motion sequences was calculated, and higher correlation values indicated that there was good coordination between the body parts.

The evaluation of force performance requires a combination of data information collected by inertial sensors and visual characterization techniques to quantify the explosive force characteristics and precise control ability embodied in the movement, and the explosive force index is calculated by the ratio of the peak acceleration and the time required to reach the peak value:

$$P = \frac{\max\left(\left\|a\right\|_2\right)}{t_{peak}} \tag{34}$$

The assessment of control is based on the accuracy and stability characteristics of the force output process, which is quantitatively analyzed using the fit between the actual force curve and the ideal standard curve.

# 4 System testing and experimental results

## 4.1 Test Program Design

In the test case design of the traditional martial arts movement recognition and evaluation system, this paper carries out the design of test cases based on the functional realization of each module in the system requirements analysis, the coordinated operation between modules and the overall stability of the system, including single test, comprehensive test, overall test and delivery test. The specific test environment is shown in Table 2, in which the system is deployed in a hierarchical deployment mode, the GPU model is NVIDIA RTX 4090 graphics card, and the GPU contains 10,752 CUDA cores and 24G memory, which can provide sufficient arithmetic support for inference tasks even under the large-scale computing capacity of deep neural networks. Finally, the Intel Core i9-13900K CPU processor with 24 cores and 32 threads can meet the demand for multi-threaded concurrent computation. 64G of DDR5 RAM (DDR5-5600) can meet the requirements of large data volume reading. 2TB solid state drive (NVMe SSD) meets the requirements of system installation and commonly used file storage, and the 8TB hard disk (MSATA) is mainly used to place a large amount of video resources. The caching system takes into account the real-time performance and large capacity of the system. The operating system adopts Ubuntu 22.04 LTS version, which has perfect ecological support and stability in artificial intelligence algorithms, unified computing device architecture 12.1 driver, and PyTorch Deep Learning Framework 2.0 has a complete implementation of GPU acceleration. Opencv (version 4.8) is a computer vision library for preprocessing images and videos. numpy and scipy (version 1.6.0) are scientific computing libraries for numerical computation, and signal processing.

*Table 2: Test environment configuration*

| Configuration category | Component name | Specific specifications | Performance indicators |
|---|---|---|---|
| Hardware environment | Processor | Intel Core i9-13900K | 24 cores /32 threads, 5.8GHz |
| | Graphics card | NVIDIA RTX 4090 | 10752 CUDA core, 24GB video memory |
| | Memory | DDR5-5600 | 64GB capacity, dual-channel |
| | System storage | Samsung 980 PRO | 2TB NVMe SSD, 7000MB/s |
| | Data storage | Western Digital Gold | 8TB HDD, 7200RPM |
| | Network equipment | Intel I225-V | Gigabit Ethernet card |
| Software environment | Operating system | Ubuntu 22.04 | Long-term support version |
| | Cuda drive | LTS CUDA | GPU computing platform |
| | Python environment | Toolkit 12.1 | Interpreter version |
| | Depth learning framework | Python 3.10.6 | Neural network training |
| | Computer vision | PyTorch 2.0.1 | Neural network training |
| | Scientific calculation | OpenCV 4.8.0 | Numerical computing library |
| | Data analysis | NumPy 1.24.3 | Data processing framework |
| | Visual tool | Pandas 2.0.2 | Chart drawing library |
| Data acquisition equipment | Color camera | Matplotlib 3.7.1 | 4096×2160@30fps |
| | Depth camera | Logitech BRIO 4K | 1280×720@90fps |
| | Inertial sensor | Intel RealSense D455 | 9-axis inertial measurement unit, 2000Hz |
| | Audio device | Xsens MTi-680G | 48kHz/24bit sampling |

In order to ensure the scientific, representativeness and diversity of the test dataset, five major categories of traditional martial arts routines, namely, Taijiquan, Bagua Palm, Xingyiquan, Shaolin Fist, and Wudang Sword, were collected, each with 15-20 typical movements, and each of which was continuously recorded 30-50 times by practitioners of different levels. This constitutes a complete dataset of 4500 videos. In the experiment, the environmental parameters were controlled, such as setting the indoor light intensity at 1000 lx, using a white screen as the test background to avoid the influence of other colors, and fixing the camera at a position parallel to the ground at 3 m in front of the subjects, so as to ensure that the environmental parameters of the tests were the same and that the data obtained had a certain degree of reliability. Moreover, the participants covered beginner, intermediate and advanced levels as well as the overall skill level of the trainers, with an age range of 18-65 years old, equal proportion of men and women, and a height and body mass within the general population, which diversified the data sources and effectively guaranteed the universality and reliability of the test results. The video action sequences were labeled by three professional teachers with many years of experience in teaching wushu, and the content of the labeling included the name of the action, the beginning and end moments of the action, typical action moments, and the evaluation of the level of the action, etc. The Fryes' kappa value was used to check the degree of consistency among the annotators, and when the kappa value was greater than 0.85, it meant that the labeling results were reliable and acceptable in this experiment. At the same time, the dataset is divided into 7:2:1 to obtain the training set, validation set and test set, the training set is used for model tuning, the validation set is used for hyper-parameter tuning and model selection, and the test set is strictly retained for the final performance evaluation, which ensures the objectivity and credibility of the test results.

## 4.2 Experimental results and analysis

The overall test results of the traditional martial arts movement recognition and evaluation system are shown in Table 3, in which good technical results were obtained, with an average recognition correctness rate of 94.2% for the five categories of traditional martial arts movements in the test dataset. The slow and posture-varying movement characteristics of Taijiquan yielded the highest recognition correctness rate of 97.8%, and the fast-paced movement variations of Shaolin Kung Fu resulted in a lower recognition correctness rate of 89.6%. However, this figure is still within a manageable range. The correct rates of discrimination for athletes of all skill levels in order were 96.5% for the high level group, 94.8% for the medium level group, 92.1% for the average level group, and 88.7% for the low level group, indicating that the high level group had relatively strong technical standardization and stability, while the other three groups had technical variability to varying degrees. The system shows good robustness under complex environmental conditions, the impact of lighting changes on the recognition accuracy is controlled within 3%, and the performance degradation caused by background interference is no more than 2.5%, and the impact of different camera angles is effectively mitigated by data enhancement technology, with a recognition accuracy of 94.2% at the frontal shooting angle, and 91.8% at the side shooting at a 45-degree angle.

*Table 3: Experimental Results of the Traditional Martial Arts Movement Recognition System*

| Martial arts schools | N | Recognition accuracy rate (%) | Average processing time (ms) | Recall rate (%) | F1 score (%) |
|---|---|---|---|---|---|
| Tai Chi | 6 | 97.8 | 85.2 | 96.9 | 97.3 |
| Ba gua | 5 | 93.5 | 92.7 | 92.8 | 93.1 |
| Xingyi Quan | 5 | 91.2 | 88.9 | 90.7 | 90.9 |
| Shaolin Quan | 4 | 89.6 | 95.3 | 88.9 | 89.2 |
| Wudang Sword | 5 | 95.4 | 91.8 | 94.7 | 95.0 |
| Overall average | | 94.2 | 90.8 | 93.6 | 93.9 |
| Test results of different skill levels | | | | | |
| High level | 25 | 96.5 | 87.4 | 95.8 | 96.1 |
| Intermediate level | 25 | 94.8 | 89.2 | 94.1 | 94.4 |
| Average level | 25 | 92.1 | 91.6 | 91.5 | 91.8 |
| Low level | 25 | 88.7 | 94.3 | 87.9 | 88.3 |
| Environmental robustness test results | | | | | |
| Standard environment | 25 | 94.2 | 90.8 | 93.6 | 93.9 |
| Low-light environment | 25 | 91.5 | 96.2 | 90.8 | 91.1 |
| Strong light environment. | 25 | 92.8 | 93.7 | 92.1 | 92.4 |
| Complex background | 25 | 91.7 | 98.5 | 90.9 | 91.3 |
| Side Angle | 25 | 91.8 | 102.3 | 91.2 | 91.5 |

The performance of the movement evaluation system was obtained after comparative analysis with the scoring of the martial arts experts. The Pearson correlation coefficient between the automatically generated scores of the system and the scoring of the experts reached 0.89, and the root-mean-square error of the difference in scores between the two was 4.2 points, which was within the acceptable error range. The highest correlation coefficient for the correctness judgment of posture was only 0.92, which was due to the fact that posture was more objective and easy to quantify. And the correlation coefficient of artistic expression evaluation is relatively low at 0.78, reflecting the difficulty of the subjective evaluation index itself. For simpler, simpler, and complex actions, the judgment accuracy is 95.3%, 91.7%, and 87.9% respectively, which basically decreases with the increase of the complexity of the action; in the case of 30 frames per second, it takes about 90.8ms to classify and score a single action, and in the case of batch processing, it takes about 65.4ms to classify and score a single action.The model during the running process The maximum memory requirement for normal-sized videos is 2.8GB, and the GPU memory requirement on the graphics card is 4.2GB, which does not pose too much of a burden on current computer configurations; the model prediction effect varies for different boxing styles, with Taijiquan achieving the best recognition accuracy due to its slower speed and more prominent pose characteristics. Since Shaolin boxing has a faster transition speed, which increases the difficulty of the algorithm to a certain extent, different types of martial arts disciplines may require different improvement programs. The effect of people with different skill levels doing the same action on the recognition results is that the standardized and unified action of professional athletes can provide clearer and more accurate feature information for the model to learn, while amateur athletes have difficulties in model recognition due to the low degree of completion of the action, which can be added to the self-learning function to improve the accuracy of the model during the subsequent development process; in terms of the intensity of the light, the comparison found that it has no significant effect on the recognition rate of the model. In terms of light intensity, the comparison found

that it has no significant effect on the recognition rate of the model. This is due to the fact that we have carried out light normalization before the experiment, to a certain extent, to reduce the impact of the background and the shooting angle, but in the subsequent version of the need to increase its anti-interference ability. The evaluation system has a high correlation with the experts' scores, which proves that the proposed algorithm is reasonable; at the same time, there are still deficiencies in some aspects with more subjective factors, such as the degree of artistic expression, which need to be further combined with the experts' experience and optimized to improve the evaluation system. After comparing and analyzing the effects of different types of deep learning algorithms, it can be seen that the hybrid network of convolutional neural network combined with LSTM achieves a better balance between accuracy and speed; a single convolutional network is faster, but the expression of the time series is not comprehensive enough, and it is difficult to extract the void information by using the recurrent network alone. The above results are important guidance for the model modification in the following.

# 5 Conclusion

In this paper, a CNN-LSTM-based traditional martial arts movement recognition and evaluation method is proposed, based on which the corresponding system framework is designed. The experimental results show that the average recognition rate is 94.2% in 5 traditional martial arts movement categories, among which the recognition rate of Taijiquan reaches 97.8%. Using ResNet-50 to extract spatial information and combining bidirectional LSTM and self-attention for feature extraction improves the recognition accuracy of multi-step movements to a certain extent. The action evaluation indexes constructed in this study include postural accuracy, movement coherence, and strength and speed, and the correlation between machine scoring and manual scoring is as high as 0.89, which proves the effectiveness of the evaluation. Experimental results show that the recognition accuracy of the system for trained professionals can reach 96.5%, while the recognition accuracy for beginners is 88.7%, and can achieve good results in both different lighting environments and different backgrounds, and the single picture recognition time is 90.8ms, which can realize real-time processing.

# About the Author

Wang Xu was born in Taihe county, Anhui, Female, People's Republic of China, 1989. She obtained a bachelor's degree from chengdu sports university and a master's degree from Qinghai Normal University. She received his doctorate in 2021 and works in Huainan Normal University, title of lecture. Her research interests are sports training and physical education.

# References

[1] Moenig, U., Kim, M., & Choi, H. M. (2023). Traditional martial arts versus martial sports: the philosophical and historical academic discourse. Revista de Artes Marciales Asiáticas, 18(1), 41-58.

[2] Vertonghen, J., Theeboom, M., & Cloes, M. (2012). Teaching in martial arts: the analysis and identification of teaching approaches in youth martial arts practice. Archives of Budo, 8(4).

[3] Jia, Y., Theeboom, M., & Dong, Z. (2020). Teaching traditional Chinese martial arts to

contemporary Chinese youth–a qualitative study with youth wushu coaches in China. Archives of budo, 16, 1-10.

[4] Romanenko, V., Cynarski, W. J., Tropin, Y., Kovalenko, Y., Korobeynikov, G., Piatysotska, S., ... & Gaziyev, S. (2025). Methodology for Assessing Spatial Perception in Martial Arts. Applied Sciences, 15(6), 3413.

[5] Husheng, Z. (2022). Martial arts moves recognition method based on visual image. Journal of Information Processing Systems, 18(6), 813-821.

[6] Gu, F., Chung, M. H., Chignell, M., Valaee, S., Zhou, B., & Liu, X. (2021). A survey on deep learning for human activity recognition. ACM Computing Surveys (CSUR), 54(8), 1-34.

[7] Pang, Y., Wang, Y., Wang, Q., Li, F., Zhang, C., & Ding, C. (2025). Applications of AI in martial arts: A survey. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 239(2), 301-331.

[8] Patne, G. V., & Govindasamy, P. (2023, August). Martial arts pose identification using deep learning. In AIP Conference Proceedings (Vol. 2790, No. 1, p. 020016). AIP Publishing LLC.

[9] Wu, B., & Zhou, J. (2024). Video-Based Martial Arts Combat Action Recognition and Position Detection Using Deep Learning. IEEE Access.

[10] Jin, M. (2022). Application and Implementation of Deep Learning for Evaluation of Martial Arts Trainings. Mobile Information Systems, 2022(1), 3979817.

[11] Chen, D., & Zhang, S. (2025). Deep Learning-Based Involution Feature Extraction for Human Posture Recognition in Martial Arts. Informatica, 49(12).

[12] Wang, Y., An, N., & Liu, C. (2025). Attention mechanisms in deep neural networks for fine-grained martial arts gesture recognition. Discover Computing, 28(1), 238.

[13] Zhao, X. (2025). AI-Driven Tai Chi mastery using deep learning framework for movement assessment and personalized training. Scientific Reports, 15(1), 31700.

[14] Wei, B., & Gu, B. (2023, December). Research on Wushu Movement Recognition and Optimization Method Based on Deep Learning Algorithm. In International Conference on Big Data Analytics for Cyber-Physical System in Smart City (pp. 787-798). Singapore: Springer Nature Singapore.

[15] Thanh, N. T., & Công, P. T. (2019). An evaluation of pose estimation in video of traditional martial arts presentation. Journal of Research and Development on Information and Communication Technology, 2019(2), 114-126.

[16] Labintsev, A., Khasanshin, I., Balashov, D., Bocharov, M., & Bublikov, K. (2021). Recognition punches in karate using acceleration sensors and convolution neural networks. IEEE Access, 9, 138106-138119.

[17] Yao, S., Ping, Y., Yue, X., & Chen, H. (2025). Graph Convolutional Networks for multi-

modal robotic martial arts leg pose recognition. Frontiers in Neurorobotics, 18, 1520983.

[18] Shang, Y. (2024). Advancing Martial Arts Training: Neural Network-Based Recognition and Assistance Systems in Biotechnological Applications. Journal of Commercial Biotechnology, 29(5), 84-94.

[19] Yan, S., Chen, J., & Huang, H. (2022). Biomechanical analysis of martial arts movements based on improved pso optimized neural network. Mobile Information Systems, 2022(1), 8189426.

[20] Chen, G. (2024). An interpretable composite CNN and GRU for fine-grained martial arts motion modeling using big data analytics and machine learning. Soft Computing, 28(3), 2223-2243.

[21] Yu, K. (2025, March). Optimization Study of Traditional Martial Arts Movement Recognition Algorithm Incorporating OpenPose. In 2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE) (pp. 373-379). IEEE.

[22] Liu, M., & Zhang, J. (2022). Gesture estimation for 3D martial arts based on neural network. Displays, 72, 102138.