A ViT-DQN-Based Real-Time Martial Arts Training System with Multimodal Fusion for Action Recognition and Optimization

Huo Fang
Department of Physical Education, Zhengzhou University of Aeronautics, Zhengzhou, 450015, Henan, China
Email: hf@zua.edu.cn

Keywords: martial arts training, computer vision, reinforcement learning, action recognition, technique optimization

Received: March 14, 2025

This paper presents an intelligent martial arts training system that integrates computer vision and reinforcement learning to address the inefficiencies, lack of personalization, and delayed feedback in traditional martial arts instruction. The system employs a Vision Transformer (ViT) for real-time action recognition and a Deep Q-Network (DQN) for training strategy optimization, enabling precise, adaptive feedback for athletes. By combining deep learning with IoT sensor data, the system analyzes posture, movement accuracy, and exercise intensity in real-time to maximize training effectiveness. A large-scale experiment involving 200 martial arts practitioners across multiple age groups demonstrated that the system achieved high recognition accuracy for key movements—96.8% for chopping, 98.1% for kicking, and 96.8% for grappling—significantly outperforming traditional CNN- and LSTM-based models. In terms of fluency optimization, the DQN model surpassed PPO and A3C with near-perfect fluency scores for chopping and sidekick. Moreover, athletes using the system achieved notable improvements in competitive outcomes: the under-18 group's win rate rose from 65% to 85%, while the 23–27 age group improved from 75% to 90%. These findings validate the system's effectiveness in enhancing training efficiency and technical precision and demonstrate the potential of artificial intelligence for intelligent martial arts instruction and broader sports training applications.

Povzetek: ViT-DQN je sistem za pametno učenje borilnih veščin, ki v realnem času prepoznava gibe in optimizira tehnike z uporabo večmodalnih senzorjev in krepitvenega učenja.

1 Introduction

With the continuous development of artificial intelligence and deep learning technology, intelligent training systems have been widely used in various fields, especially sports training [1-2]. Computer vision and reinforcement learning have brought unprecedented innovation to athlete training. Traditional martial arts training often relies on coaches' guidance and athletes' autonomous learning. However, this training method is inefficient and difficult to customize, which easily leads to slow improvement of athletes' training effects. As a comprehensive competitive event, martial arts [3-4] not only has high requirements on athletes' strength, speed, and flexibility but also requires precise movement skills and high coordination. The movement correction in traditional training methods relies on the experience of coaches, cannot capture the subtle movement deviations of athletes in real-time, and lacks sufficient objectivity and data support.

In order to solve these problems, intelligent training systems based on computer vision [5-6] and reinforcement learning [7-8] have gradually become a new trend. Computer vision technology can identify athletes' movements in real-time and provide accurate feedback, thus providing strong support for movement analysis and evaluation, reducing errors caused by human intervention, and being able to work stably in a variety of complex environments. Through the automatic processing of image data by deep learning technology, the accurate recognition of martial arts movements has been greatly improved, allowing athletes to obtain real-time and accurate movement feedback. Reinforcement learning enables the system to automatically adjust training strategies through continuous interaction with the training environment, helping athletes find the best movement trajectory during training and promoting skill improvement. Combined with real-time data from IoT sensors [9-10], the intelligent system can detect the accuracy and fluency of movements, conduct a detailed analysis of multi-dimensional factors such as the strength and speed of athletes' movements, and further personalize training plans.

In response to the above problems, this paper proposes the following hypotheses and goals.H1. Improved action recognition accuracy

Using ViT, through its global attention mechanism, it can effectively capture the long-range dependencies and spatial structure information in the image, thereby achieving higher recognition accuracy and stability in complex martial arts action recognition tasks.H2. Optimization of real-time feedback speed

Combining the lightweight ViT structure with the efficient DQN reinforcement learning algorithm, an end-to-end closed-loop system is constructed. This system can control the overall delay from action detection to strategy feedback within 100ms without affecting the recognition quality, meeting the real-time requirements of martial arts training.

H3. Enhanced generalization ability under complex action patterns

The fusion of RGB-D images and IMU sensor data, and the joint modeling of ViT+DQN, can enhance the system's adaptability to unseen action patterns and improve the model's generalization performance under different lighting, backgrounds, and individual differences.

With the rapid development of artificial intelligence technology, the application of computer vision sports has received widespread attention. In particular, in terms of athlete motion analysis and real-time feedback, computer vision [11-12] has become an important tool for improving training effects and competitive performance. In existing research, many action recognition methods based on computer vision have achieved remarkable results. Traditional image classification models, such as CNN [13-14], are widely used in action recognition. Action recognition methods based on CNN [15-16] can accurately classify actions to a certain extent by analyzing athletes' frame-by-frame images. However, these methods rely on local features extracted by convolutional layers and may have specific limitations in dealing with complex actions and long-term dependent scenes. To address these problems, some recent studies have begun to explore Transformer-based models. The action recognition framework based on ViT has significantly improved by segmenting images into small blocks and using a self-attention mechanism to model global context. This method is particularly suitable for recognizing complex actions and can better handle long-term dependencies and spatial structure information. Therefore, ViT [17-18], as a cutting-edge technology, has important potential recognizing martial arts actions. The movements of athletes not only rely on image data, but the integration of sensor data also provides important support for improving the accuracy of movement recognition. Based on the method of fusing deep sensor data with visual data, the IMU sensor is combined with camera image data to achieve a multi-dimensional analysis of movement accuracy and posture. This multimodal fusion strategy effectively improves the system's ability to recognize complex movements, especially when athletes perform high-speed or highly complex movements. Sensor data can provide the system with more kinematic information and enhance the robustness of the model.

Although existing computer vision methods and sensor fusion technologies have achieved paticular success in motion recognition, their application in sports training, especially in personalization and real-time feedback, still faces many challenges. Recent advances in artificial intelligence have expanded the scope of intelligent training systems, particularly through integrating reinforcement learning and deep learning techniques.

Reinforcement learning [19–20] enables systems to adjust training strategies dynamically by interacting with the environment and optimizing behavior through reward signals, which is especially valuable in complex and uncertain training scenarios. For example, Ortiz [19] emphasized how AI-enhanced systems support psychomotor learning in martial arts by providing timely feedback and adaptive strategies. Wu and Zhou [20] further demonstrated how deep learning enables real-time recognition and positioning of martial arts actions from video input, offering strong support for intelligent feedback loops. Deep Q-Networks (DQN), a widely used reinforcement learning algorithm, have shown great promise in control and decision-making problems. Chen [21] proposed a composite CNN-GRU model for fine-grained motion modeling, enhancing the interpretability and effectiveness of martial arts action analysis. Similarly, Park et al. [22] highlighted how computer vision-based systems can assist in skill refinement and martial arts assessment, further validating AI's role in sports training.

In particular, DQN-based intelligent fitness systems can simulate realistic training environments and deliver reward feedback, enabling athletes to adjust their movements and intensity in real-time based on sensor input and performance feedback. This adaptive capability is especially critical for skill-based sports such as martial arts and gymnastics, where precision and timing are key.

Moreover, as Gordon [23] pointed out, AI-enhanced training platforms promote self-regulated learning, empowering athletes to take charge of their progress. The combination of DQN and sensor data [24] represents a promising direction for enhancing personalization and real-time responsiveness in training systems. In these systems, sensors capture biomechanical and physiological data, which DQN algorithms use to continuously update and refine training plans, improving training outcomes and optimizing movement accuracy.

Additionally, the optimization techniques in deep learning architectures, as discussed by Wang and Song [25], offer insights into how convolutional neural networks can be fine-tuned for improved performance in data-intensive environments such as real-time motion tracking. Meanwhile, the recommendation algorithms proposed by Chen [26] using collaborative filtering and short-text similarity demonstrate the potential for integrating personalized recommendation systems into intelligent training platforms — allowing the system to suggest tailored training modules or corrective feedback based on an athlete’s history and progress.

In general, martial arts movements are highly complex and spatiotemporal. Local receptive fields limit traditional CNNs and are insufficient in modeling long-distance dependencies and global motion structures. ViT uses a self-attention mechanism to model global features of images effectively and can more accurately capture the spatial structure and dynamic changes of movements. ViT and DQN have good modal fusion capabilities. They can jointly model visual features with sensor time series data to form a unified optimization framework, achieving end-to-end closed-loop learning from action recognition to policy feedback.

2 Model design

2.1 Data collection

In order to carry out real-time recognition and training optimization of martial arts movements, data collection is the basis of system design. The accuracy and comprehensiveness of data collection directly determine the performance of subsequent training optimization and movement recognition algorithms.

Image data is the core input data in martial arts action recognition, and its quality and diversity will directly affect the performance of the action recognition model. To ensure the comprehensiveness and accuracy of the data, this system uses an RGB-D camera to take full-body photos of athletes. The RGB-D camera can not only provide standard RGB image data but also provide depth information to help the system capture the relative position and movement trajectory between the athlete and the background, thereby improving the accuracy of action recognition.

Choose high-resolution video equipment and use two cameras to shoot at different angles to ensure the athlete's entire action process can be covered. The camera's position is reasonably arranged according to the layout of the training scene to ensure that every detail of the athlete's action can be captured and possible occlusion problems can be eliminated. In order to capture the complete picture of the athlete performing martial arts movements, image data should be collected from multiple perspectives, including front, side, and oblique angles.

Martial arts training is conducted under different lighting conditions. To enhance data diversity, the data collection process should cover a variety of lighting environments, including natural light, indoor light, intense light, and weak light. In addition, data should be collected under different backgrounds to avoid excessive reliance on background factors in the action recognition model, thereby improving the model's generalization ability.

The illumination conditions for collecting images are shown in Table 1.

Table 1: Lighting conditions

Lighting environment	Minimum light (Lux)	Maximum light (Lux)
Natural Light	100	1000
Interior lighting	50	500
Glare	500	10000
Low light	1	50

Each collected image data is annotated in detail. The annotation content is the action category. Martial arts include: chopping fist, kicking, grappling, hook punch, side kick, turning and kicking.

The description of martial arts moves is shown in Table 2.

Table 2: Description of martial arts movements

Action	Describe	Application
Chop fist	Use straight arms to exert force and punch out diagonally from the chest to hit the target quickly.	Used to attack targets above or from the side, to train arm strength and speed.
Kick	Use your feet as the force point, kick from low to high, with your toes pointing toward the target, and your movements should be smooth and explosive.	Train leg flexibility and explosive power, suitable for close combat defense and attack.
Catch	By controlling the opponent's arms or neck, pressure is applied to make the opponent lose resistance.	It is a key technique used in confrontation to control and subdue the opponent.
Uppercut	Bend your wrist and strike the target with your fist from bottom to top, concentrating the force on the fist.	Suitable for hitting the opponent's head or chin to increase the hitting angle and power.
Side kick	Lift your legs sideways and strike the target with the sole of your foot horizontally, exerting force quickly and keeping your center of gravity stable.	Trains leg control and attack range, often used for defensive kicks.
Turn and kick	Turn around and kick your legs, and exert force quickly with your legs. The main target is the opponent's upper body or head.	Improves turning speed and leg explosiveness, often used to counterattack enemies.

Figure 1. Patch-based segmentation and tokenization process for ViT input: The figure shows a temporal sequence of grappling actions in martial arts. The background (green), mat edge (orange), and athletes (red and yellow) are segmented into distinct regions. Each frame is divided into 16×16-pixel patches (not visually outlined here for clarity), which are flattened and linearly projected into embedding tokens. These tokens are then passed into the Vision Transformer for spatial feature extraction. The color segmentation shown here was used for visual preprocessing and token consistency across frames. The image segmentation of martial arts movements is shown in Figure 1.

Figure 1: Demonstration of martial arts movements

Images are processed by denoising, enhancement, and standardization to ensure the high quality and diversity of training data. Gaussian filtering is a commonly used image denoising technology that aims to reduce noise in images, especially high-frequency noise generated during camera shooting. It removes noise in images by performing convolution operations on them and smoothing them with Gaussian kernels.

The mathematical representation of the Gaussian filter is as follows:

$$G(x,y)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)(1)$$

In formula 1, $G(x,y)$is the weight of the Gaussian kernel, $\sigma$is the standard deviation of the Gaussian function, and determines the size and blur of the filter. x and y are the positions of the filter on the image.

The Gaussian filter processes the image using the following convolution formula:

$$I'(x,y)=\sum_{m=-k}^{k}\sum_{n=-k}^{n}G(m,n)\cdot I(x+m,y+n)(2)$$

$I'(x,y)$is the image after Gaussian filtering, $G(m,n)$is the Gaussian kernel value at each pixel, and k is the radius of the Gaussian filter.

Image enhancement increases the diversity of the data set. It improves the generalization ability of the model by performing a series of transformations on the original image, including cropping, rotation, scaling, and color adjustment. These methods can simulate different environmental conditions and changes, allowing the model to learn more features during training and improve its adaptability to unknown data.

After data enhancement, the image is normalized. Standardization and normalization can ensure that the image data is suitable for the deep learning network, especially when training the neural network. The data is processed with zero mean and unit variance to accelerate convergence. The standardization formula is as follows:

$$I_{norm}(x,y)=\frac{I(x,y)-\mu}{\sigma}(3)$$

The selection of sample quantity is mainly based on the following points: First, high-frequency combat actions (such as kicking and chopping) appear frequently in training and have more technical variations, so the number of samples is relatively sufficient to enhance the model's recognition robustness of these actions; second, complex combination actions (such as turning kicks and uppercuts) are more challenging to execute and have more delicate action structures, so it is relatively chanllenging to obtain samples during the collection process, so the sample size is slightly smaller but still maintains a sufficient training scale; finally, all categories are divided in a 3:1 ratio between the training set and the test set to ensure that the data distribution of each category is balanced and to avoid overfitting or recognition bias due to sample bias. The dataset constructed by the above strategy can reflect the true distribution characteristics of martial arts actions and provide a data basis for the model's generalization ability in complex action recognition. The images are divided into the dataset in a 3:1 ratio, and the dataset distribution is shown in Table 3.

Table 3: Dataset distribution

Category	Training set	Test Set	Total
Chop fist	669	223	892
Kick	633	211	844
Catch	594	198	792
Uppercut	537	179	716
Side kick	630	210	840
Turn and kick	558	186	744

2.2 Action recognition model

ViT is a computer vision model based on the Transformer architecture, which aims to address the limitations of traditional convolutional neural networks in image recognition tasks. ViT takes advantage of the Transformer's advantages in natural language processing and uses a global attention mechanism to model long-distance dependencies in images, thereby better capturing the global context information of the image, which is suitable for real-time recognition tasks of martial arts movements.

ViT model structure is shown in Figure 2.

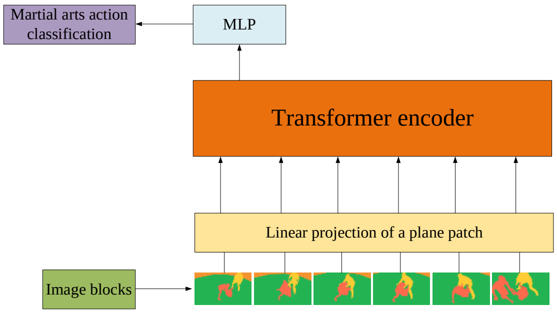
Martial arts action classification
MLP
Transformer encoder
Linear projection of a plane patch
Image blocks

Figure 2: ViT model structure

The input image is divided into fixed-size image blocks. Each image block is expanded into a one-dimensional vector and mapped to a high-dimensional embedding space through a linear transformation. To introduce spatial information, the ViT model retains the relative position information between image blocks by adding position encoding. These position encoding vectors are obtained through learning and added to the embedding vector of each image block. The addition of position encoding ensures the model can understand the spatial relationship between image blocks.

ViT uses multiple Transformer encoder layers to process the input sequence of image patches. Each encoder layer consists of a self-attention mechanism and a feed-forward neural network. The self-attention mechanism allows the model to pay attention to other image patches in the image when processing each patch, thereby capturing the global dependencies of the image.

For an input embedding sequence, compute the query, key, and value matrices:

$$Q = XW_Q (4)$$
$$K = XW_K (5)$$
$$V = XW_V (6)$$

W_Q, W_K, $W_V$are learnable weight matrices corresponding to query, key, and value, respectively.

Calculate the attention score :

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V (7)$$

In Formula 7, $d_k$is the dimension of the key vector, and softmax is used to normalize the scores to ensure that their sum is 1.

During the training process, ViT optimizes the model parameters through backpropagation and adjusts the network weights so that the network achieves the best performance on the classification task. The training process involves loss functions, optimization algorithms, and training strategies.

ViT uses the cross-entropy loss function to measure the difference between the model's predicted category and the actual label. For a multi-category classification problem, the formula is:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) (8)$$

In Formula 8, C is the number of categories and $\hat{y}_i$is the probability of category i predicted by the model. The cross-entropy loss function can effectively drive the model to update weights in each training iteration to improve classification accuracy.

Adam optimizer is a commonly used optimization algorithm in deep learning. It combines the advantages of Momentum and RMSprop and can adjust the learning rate adaptively. The updated formula of the Adam optimizer is as follows:

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{v_t + \epsilon}} \cdot m_t (9)$$

$\theta_t$are the parameters of $\eta$the model at time step t, and are the learning rate. $m_t$and $v_t$are the first and second moments of the gradient, which are calculated by recursive update:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t (10)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 (11)$$

In formula 10-11, $\beta_1$and β_2 are the decay rates of the first-order moment and the second-order moment.

In order to improve computational efficiency and reduce memory consumption, ViT adopts a batch training strategy. Each training uses a small batch of data for forward and back propagation, and the small batch size is 32. In order to speed up training and prevent overfitting, ViT uses step decay, that is, gradually reducing the learning rate during training.

2.3 Action optimization model

DQN is a method that combines deep learning and reinforcement learning. It uses DNN to approximate the Q-value function and optimize the strategy selection during training. In the intelligent martial arts training system, the DQN model helps athletes adjust their action strategies during training and gradually optimize their action performance to achieve the best results.

The input to the network is the athlete's current state, including movement posture and sensor data (IMU data, speed, angle, etc.). These inputs are represented as vectors and fed into the neural network for processing. The state space contains the various possible movement postures of the athlete during training and the corresponding sensor feedback.

The output is a Q-value function, representing the expected reward when taking an action in a given state. Based on the Q-value, the model selects the optimal action strategy for each input state.

DQN uses a deep neural network to approximate the Q-value function. The input layer receives the athlete's current state, extracts feature through multiple hidden layers, and finally calculates the Q-value of each action through the output layer. The core goal of DQN is to learn an optimal Q-value function so that the selected action can bring the maximum long-term return in each state. The model gradually learns the best strategy by continuously updating the Q-value function.

The DQN model is shown in Figure 3.

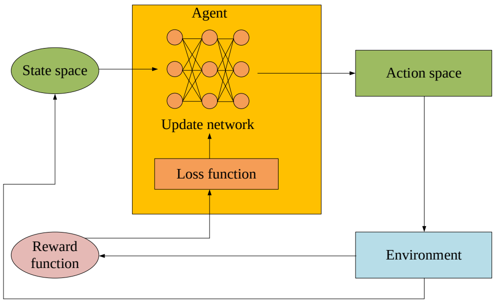

Figure 3: DQN model structure

In order to optimize the training process, the reward function is designed to ensure that the system can effectively learn and provide feedback based on the athlete's performance. The accuracy reward measures the athlete’s accuracy when performing martial arts moves. When the athlete performs a move, the system will give a reward based on the move’s accuracy. The accuracy reward is defined as:

$$R_{accuracy} = -\left|\theta_{target} - \theta_{current}\right| (12)$$

A fluency bonus can be defined by calculating the speed between actions:

$$R_{smoothness} = -\left|v_{current} - v_{previous}\right| (13)$$

Strength rewards can be measured using force data collected by sensors:

$$R_{strength} = -\left|F_{target} - F_{current}\right| (14)$$

Real-time feedback can be achieved through real-time updates of sensor data:

$$R_{feedback} = -|Angle_{target} - Angle_{current}| (15)$$

The total reward function of the system is the weighted sum of the above reward terms:

$$R_{total} = w_1 R_{accuracy} + w_2 R_{smoothness} + w_3 R_{strength} + w_4 R_{feedback} (16)$$

In Formula 16, w_1, w_2, w_3, and are $w_4$the weights of the corresponding reward items respectively. Among them, $w_1 = 0.4, w_2 = 0.3, w_3 = 0.2, w_4 = 0.1$.

In each training process, DQN updates the Q value through the Q-learning algorithm. The Bellman equation is used to update the Q value function:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left(R_t + \gamma \max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t)\right) (17)$$

In Equation 17, $\gamma$is the discount factor, which determines the importance of future rewards. $\alpha$is the learning rate, which determines the step size of the update.
DQN uses the ε-greedy strategy to balance exploration and exploitation. In each training step, the exploration and exploitation strategies are:

$$A_t = \begin{cases} \text{random action}, & \text{with probability}\epsilon \\ \arg\max_{a'} Q(S_t, a'), & \text{with probability} 1 - \epsilon \end{cases} \quad (18)$$

As training progresses, ϵ it will gradually decrease, prompting the model to rely more on the best strategy learned in the later stage. To avoid high variance in Q value updates during training, DQN uses an experience replay mechanism, which stores past states, actions, rewards, and following states in a replay buffer. During each training, a small batch of experience is randomly sampled from the buffer to update the Q value, preventing the model from overfitting a specific training sequence and improving the stability of the training.
To ensure that the intelligent martial arts training system based on ViT and DQN proposed in this study has good reproducibility and experimental transparency, this paper sets the key hyperparameters of the model training stage in detail. The ViT model is initialized with ImageNet pre-trained weights to accelerate convergence and improve recognition performance under small samples. The input image is uniformly cropped to 224×224 pixels and divided into 196 image patches with a block size 16×16. Each patch is linearly mapped to a 768-dimensional embedding vector. The Transformer backbone structure of ViT contains 12 encoder layers, 12 attention heads, and a 3072-dimensional feedforward network. The optimizer uses AdamW, the initial learning rate is 3e-4, the weight decay is set to 0.01, and the batch size is set to 32. The training process is carried out for 100 epochs, and the step decay strategy is used to reduce the learning rate to 1/10 of the original value at the 60th and 85th epochs, respectively. All model training was completed on an NVIDIA RTX 3090 GPU, and a full training session took about 18 hours. In the action optimization module, the DQN model takes the player's current state as input and outputs the Q value of each action category. The network structure contains three fully connected layers with dimensions of 256, 128, and action space size (6), respectively. DQN uses an experience replay buffer (size of 10,000) for training and adopts an ε-greedy strategy for exploration, with an initial exploration rate of ε=0.9 and gradually decaying to 0.1. The optimizer is Adam, the learning rate is set to 1e-4, the discount factor γ is set to 0.95, and the target network update frequency is updated every 100 steps. During training, the maximum number of interaction steps is set to 50,000; the batch size is 64, and the average reward for every 200 steps is used to evaluate the training effect. DQN training is also run on the RTX 3090 platform, with an average training time of about 12 hours. Considering that the feedback of the reinforcement learning environment has high noise, the sliding window average strategy is used to monitor the training process. In

addition, in terms of data input, the video frame rate is 30fps, and the duration of each action sample is controlled at about 1.5 seconds; that is, each input sequence contains about 45 frames of images in order to take into account the integrity of the action and the efficiency of temporal modeling. In order to enhance the model's understanding of the action timing information, ViT uses an overlapping sliding window to extract the image block sequence when processing action video clips. None of the models used independent validation sets for parameter adjustment throughout the training process. The main reason is that the system emphasizes end-to-end closed-loop training and online feedback mechanisms, and its performance evaluation depends more on the performance of the final test set and the actual training feedback effect.

3 Model fusion and system implementation

3.1 Fusion strategy

In the intelligent martial arts training system, the fusion of image data and sensor is crucial to improving the training effect. Image data can provide spatial posture information of athletes in action, while sensor data can provide accurate dynamic feedback. Data fusion technology combines data from different modalities to provide richer input information for the model.

ViT is used to extract visual features from images. ViT can cut the input image into small pieces then capture the long-distance dependencies in the image through a multi-layer Transformer structure to generate an embedded representation of the image. These visual features can reflect the posture and movement of the athlete to a certain extent, but they are not enough to capture the dynamic changes of the movement.

Sensor data provides information about the timing changes during the action. A time series convolutional network is used to process these time series data. TCN can effectively learn the long-term and short-term dependencies in time series data and extract the dynamic features of the athlete's action process. These features help the model understand the timing of the action and reflect the athlete's execution stability and fluency.

ViT and the sensor features processed by TCN will be fused according to a certain weight to form a comprehensive feature vector. This fused feature vector will be used as the input of the DQN model to optimize the training process. In this way, DQN can make more accurate action optimization decisions after receiving richer information.

To improve the synchronization between image recognition and action optimization, a joint training strategy is adopted. That is, the ViT and DQN models are optimized simultaneously during the training process. The core of the joint training strategy is that extracting image features and processing sensor data can be optimized under the same training framework. ViT is responsible for extracting the spatial information of the action from the image, while DQN optimizes the athlete's action strategy based on this visual information and sensor data.

ViT is used to extract visual features from images, while TCN is used to process time series data collected by IMU sensors. In order to achieve effective coordination of multimodal information, a weighted fusion mechanism in the channel dimension is adopted. The specific process is as follows:

Let be the temporal visual features output by AAA, where T represents the number of time steps, Dv is the feature dimension; let BBB be the sensor features output by TCN. First, the two sets of features are mapped to a unified dimensional space through independent linear transformations:

Let $V \in R^{T*D_v}$ be the temporal visual features output by ViT, where T represents the number of time steps and Dv is the feature dimension; let $S \in R^{T*D_s}$ be the sensor features output by TCN. First, the two sets of features are mapped to a unified dimensional space through independent linear transformations.

$$V' = W_v V + b_{v,} S' = W_s S + b_s \ (19)$$

Among them, $W_v \in R^{D*D_v}$, $W_s \in R^{D*D_s}$ are learnable parameter matrices, and b_v, b_s are bias terms.

Subsequently, the aligned features are concatenated in the channel dimension and linearly combined by a trainable weight vector $\alpha = [\alpha_1, \alpha_2]$:

$$F = \alpha_1 V' + \alpha_2 S' (20)$$

$F \in R^{T*D}$ is the final fusion feature representation. This fusion operation is located in the later stage of the entire network architecture, that is, it is integrated after ViT and TCN complete the extraction of high-level semantic features to retain the modeling depth within the modality and enhance the cross-modal semantic consistency.

In order to demonstrate the overall architecture and operation process of the intelligent martial arts training system based on ViT and DQN proposed in this paper, Figure 4 is used to describe the process.

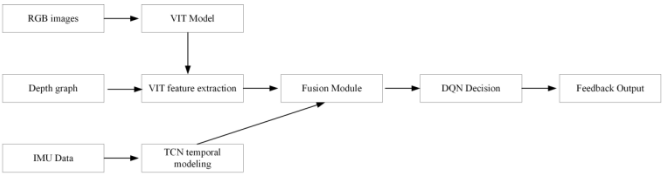

Figure 4: System operation process

3.2 Experiments and evaluation

In the real-time recognition task of martial arts movements, it is crucial to evaluate the model’s performance. In order to comprehensively measure the accuracy of the model in martial arts movement recognition, the following commonly used evaluation indicators are used:

Accuracy is a basic indicator used to measure the overall performance of a classification model, and it is defined as the ratio of all correctly classified samples to the total number of samples. The formula is:

$$A = \frac{TP+TN}{N}(21)$$

In Formula 19, N is the total number of samples.

Precision measures the proportion of positive cases when the model predicts positive cases. The formula for precision is:

$$P = \frac{TP}{TP+FP}(22)$$

The recall rate measures the proportion of samples that are positive and are correctly identified as positive by the model. The formula is:

$$R = \frac{TP}{TP+FN}(23)$$

The F1 score is the harmonic mean of precision and recall, which can comprehensively evaluate the classification performance of the model :

$$F1 = 2 \times \frac{P\times R}{P+R}(24)$$

Classification metrics analyze each action (chop, kick, grab, hook, side kick, turn kick) individually, thereby comprehensively evaluating the accuracy and robustness of the action recognition model.

In martial arts training, the fluency and continuity of movements are crucial to improving technical level. A group of martial arts experts with professional backgrounds was invited to give scores based on the coherence, rhythm, and accuracy of the athletes' movements. A 5-point system (1 point is very unsmooth, and 5 points are very smooth) is used to evaluate the fluency of each training movement.

During the system training process, various reinforcement learning optimization algorithms were used to optimize the athletes' training strategies. The best performance in improving training efficiency and movement accuracy was determined by comparing the effects of these optimization algorithms.

In order to evaluate the actual effect of the intelligent training system, 200 martial arts athletes were selected and divided into groups to compare the effects of traditional training methods and the intelligent training system in this paper. The athletes were divided into groups according to their age: <18, 18-22, 23-27, 28-32, 33-37, 38-42, >42. The effectiveness of the intelligent training system was evaluated by comparing the winning rate of the matches after training. The winning rate of the match is the ratio of the number of wins of each athlete in the match to the total number of matches, calculated as:

$$WR = \frac{WN}{TN}(25)$$

In Formula 23, WNand are TNthe number of wins and the total number of games, respectively.

This system uses an RGB-D camera to collect the athlete's visual information and combines it with an IMU sensor to obtain the temporal motion data of the action. The acquisition resolution of RGB images and depth images is 640×480, and the frame rate is 30fps; the sampling frequency of the IMU sensor is 100Hz, and it can synchronously provide three-axis acceleration and angular velocity information. In order to ensure the temporal consistency of multimodal data, the system uses a hardware trigger mechanism to perform synchronous calibration between sensors. In the preprocessing stage, the IMU data is denoised by a second-order Butterworth low-pass filter (cutoff frequency is 20Hz) to remove high-frequency noise interference; the image data is smoothed and denoised using Gaussian filtering, and background modeling technology is combined to reduce environmental interference. The above enhancement and preprocessing methods effectively improve the quality of the input data, providing a reliable guarantee for the stable training and accurate recognition optimization of the subsequent ViT and DQN models.

4 Results

4.1 Action recognition performance

The martial arts action recognition performance is shown in Table 4.

Table 4: Martial arts action recognition performance

Action	Accuracy	Precision	Recall	F1
Chop fist	0.968	0.968	0.958	0.963
Kick	0.981	0.964	0.961	0.962
Catch	0.968	0.951	0.974	0.962
Uppercut	0.984	0.971	0.971	0.971
Side kick	0.971	0.984	0.964	0.974
Turn and kick	0.961	0.964	0.977	0.970

Table 4 shows the performance evaluation results of martial arts action recognition using ViT. The accuracy, precision, recall, and F1 score of each action are at a high level, especially in the chop, kick, hook, and other actions; the recognition accuracy is perfect. The advantage of ViT in visual tasks mainly comes from its ability to capture long-distance dependencies and global context information in the image. Unlike traditional convolutional neural networks that extract features through local receptive fields, ViT cuts the image into multiple patches of fixed size and treats them as a sequence to input into the Transformer model so that the model can directly learn the relationship between various regions in the image. This global modeling capability is particularly suitable for complex dynamic action recognition tasks, such as martial arts actions, where the coherence and details of the action are essential, and the various parts are closely related. ViT 's self-attention mechanism can help the model capture the relationship between the various joints and parts of the athlete's body, allowing the model to accurately model the key details in the action, thereby improving recognition accuracy.

ViT 's input processing method makes it more robust. After the image data is cut into patches, ViT can automatically adjust the image patches of different actions, reducing the impact of different lighting and background changes on image recognition. In traditional CNN models, changes in lighting, posture, or camera angle often significantly affect the stability of feature extraction. In contrast, ViT 's global feature processing method can effectively reduce the interference of these factors, making the model more adaptable to various environmental changes.

In the chopping fist action, the precision and recall are 0.968 and 0.958, respectively, and the F1 score is 0.963. This shows that ViT can accurately judge the categories of various actions when performing action recognition, avoiding the occurrence of misclassification and missed judgment. High precision means that the model can effectively exclude most negative samples, while high recall means capturing all positive samples. For action recognition tasks, the balance between precision and recall is essential, because the diversity and complexity of martial arts actions make it crucial to judge each action accurately. ViT can accurately model action details through its deep self-attention mechanism, thereby achieving higher precision and recall. ViT can effectively extract rich features from the image data of martial arts actions, accurately identify different types of actions, and achieve a good balance between precision and recall. Therefore, ViT 's high accuracy in martial arts action recognition is due to its powerful global modeling ability, self-attention mechanism, and effective processing of complex visual information.

4.2 Movement fluency

In order to effectively measure the performance of action optimization, three optimization algorithms, DQN, PPO, and A3C, are compared. The fluency of the action is evaluated through expert scoring; in order to reduce the subjectivity of scoring and improve the objectivity of action fluency assessment, this paper invited five experts with more than ten years of experience in martial arts teaching and competition refereeing to form a scoring panel, and used structured scoring criteria to score the quality of athletes' action execution independently. The scoring criteria revolve around three core dimensions: action accuracy (accounting for 40%) - evaluating the consistency of the action path, force angle, and target position; action fluency (accounting for 40%) - evaluating the natural connection, rhythm control, and coherence between actions; body coordination (accounting for 20%) - examining the coordination and center of gravity control ability of the trunk and limbs. Each expert quantitatively scored each action sample on a 5-point scale based on the visual motion trajectory playback, sensor feedback data, and video recording provided by the system, with 1 point for "severe deviation/incoherence" and 5 points for "high accuracy/fluency." The final score is the weighted average of the scores of the five experts to ensure that the scoring results have good reliability and repeatability. The results are shown in Figure 5.

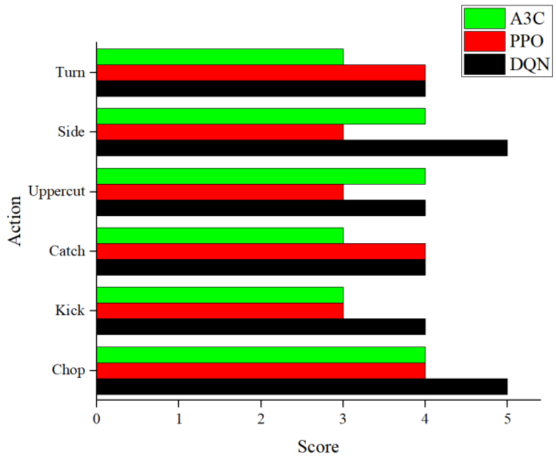

Figure 5. Movement smoothness

Figure 5 shows the performance of three reinforcement learning algorithms, DQN, PPO, and A3C, in the martial arts action optimization task. The expert rating mainly evaluates the fluency of the action. DQN obtained high scores in most actions, especially in the chopping punch and sidekick, while PPO and A3C were slightly inferior in the fluency scores of most actions. DQN is a reinforcement learning algorithm based on value iteration. It selects the optimal action strategy by learning a state-action value function. In the scenario of martial arts action optimization, DQN is particularly suitable for actions that require a long time to execute stably, such as chopping punches and side kicks. DQN optimizes the strategy through experience replay and target network so that the randomness of exploration can be effectively reduced during the training process, and the accuracy and fluency of the action can be steadily improved. Especially executing complex actions, DQN can gradually learn more efficient action execution strategies in multiple rounds of training, thereby significantly improving the fluency and stability of the action. In addition, DQN adopts offline learning and experience replay mechanisms so that the model can fully use historical experience in each training, reduce overfitting and the occurrence of local optimality, and further improve the optimization effect of the action. Therefore, DQN showed better fluency optimizing actions such as chop and sidekick, which require higher stability and accuracy.

Unlike DQN, PPO and A3C are reinforcement learning algorithms based on policy gradients. These algorithms improve the execution of actions by directly optimizing policies. However, the training process of PPO and A3C is more complicated than that of DQN, and may face greater training difficulties in high-dimensional state spaces. Although PPO can avoid gradient explosion and overfitting problems by limiting the update amplitude of the policy, its improvement in action fluency is not as apparent as DQN. The policy update performed by the PPO at each time step may make it difficult to significantly improve the fluency and accuracy of actions in some complex actions (such as kicking and grappling), so the expert score is relatively low. A3C uses multi-threaded parallel training to improve the robustness of the policy through multi-policy updates. However, its asynchronous update method may sometimes cause unstable training, especially when the details of the action are high, and the fluency may be affected. Therefore, although A3C and PPO have optimized the training process to a certain extent in actions such as turning kicks and hooks, the final action fluency score is slightly low due to unstable training and large policy update amplitude.

DQN shows obvious advantages in action optimization, especially when fine control and action stability are required. This is mainly attributed to DQN 's experience replay and target network mechanism, which enables the model to gradually learn the optimal strategy, reduce unnecessary exploration and over-optimization, and ensure the smoothness of action execution. In contrast, PPO and A3C are slightly inferior in fluency scores, mainly because their strategy update methods are more aggressive. This may lead to instability and local optimality during training, especially in martial arts movements that require fine control. The optimization effect is not as significant as DQN. Therefore, DQN based on value iteration is stronger in martial arts action optimization tasks, especially in action fluency and stability.

4.3 Training effect

The traditional training system is compared with the training system proposed in this paper to evaluate its effectiveness in improving the skills of athletes. 200 martial arts athletes were selected and divided into groups according to age, namely <18, 18-22, 23-27, 28-32, 33-37, 38-42, and >42. The results of the winning rate of martial arts athletes are shown in Table 5.

Table 5: Training effect

Age	Traditional systems	This article’s system
<18	65%	85%
18-22	70%	88%
23-27	75%	90%
28-32	68%	84%
33-37	60%	80%
38-42	55%	72%
>42	50%	68%

From the data in Table 5, there is a significant difference in the winning rate of martial arts athletes of different age groups between the traditional training system and the intelligent training system proposed in this paper. Whether it is young athletes or older athletes, the improvement effect of the system in this paper on the winning rate of the game is relatively significant, especially in the under 18 and 23-27 age range; the winning rate of the traditional training system is 65% and 75%, respectively, while the winning rate of the intelligent training system is increased to 85% and 90%, respectively. This improvement shows that the intelligent training system can better help athletes make significant progress in skills through real-time feedback and an optimized training process. The advantages of the intelligent training system in this paper are more obvious in young athletes, which may be related to the faster speed of young athletes accepting new technologies and adapting to innovative training methods. As age increases, the physical condition of athletes gradually declines, and the winning rate gap between the traditional training system and the intelligent training system gradually narrows. In the age groups of 33-37 and 38-42, the winning rates of the traditional training system were 60% and 55%, respectively. In comparison, the winning rates of the proposed system were 80% and 72%, respectively, indicating that the intelligent training system can still significantly improve in middle-aged and elderly athletes. However, with the increase of age, the effect of training may be limited by physical conditions and adaptability. In addition, the winning rate difference in the group >42 years old is more obvious, with the winning rate of the traditional training system being only 50%. In contrst, the winning rate of the proposed system increased to 68%, indicating that the intelligent training system still has a positive impact on the adaptation and skill improvement of older athletes, primarily through technical means to optimize movements and real-time feedback, which can effectively make up for the limitations of traditional training. Therefore, the performance of the proposed system is better than that of the traditional system in all age groups, especially in young and middle-aged athletes, which proves the effectiveness and potential of the intelligent training system in improving the skills of martial arts athletes.

In order to more intuitively demonstrate the advantages of the proposed intelligent martial arts training system in action recognition performance, we conducted comparative experiments with traditional methods on multiple key indicators. The experiment selected the current mainstream CNN-based action recognition models, such as Two-Stream CNN and 3D-CNN + LSTM, and compared their core performance parameters, such as accuracy, F1 score, and real-time feedback delay on the same test set. The results are shown in Table 6:

Table 6: Key indicator performance comparison

	Average recognition accuracy (%)	Average F1 score	Real-time feedback delay (ms)	Standard Deviation	95% confidence interval
Two-Stream CNN	92.3	0.918	145	±1.2	(90.1%,94.5%)
3D-CNN + LSTM	93.7	0.929	138	±1.0	(91.7%,95.7%)
ViT without DQN	96.1	0.954	115	±0.8	(94.5%,97.7%)
ViT-DQN	96.8	0.963	98	±0.6	(95.6%,98.0%)

As can be seen from Table 6, the ViT + DQN system proposed in this paper outperforms existing methods in all key indicators. First, regarding average recognition accuracy, the system reached 96.8%, which is 4.5 percentage points higher than Two-Stream CNN and 3.1% higher than 3D-CNN + LSTM. Second, the F1 score of the system is 0.963, which is significantly higher than other models, indicating that it has good recall ability while maintaining high precision, effectively reducing misjudgment and missed detection. In addition, in terms of real-time feedback delay, the system is only 98ms, much lower than the 138–145ms of traditional methods, meeting the strict requirements of martial arts training for fast feedback. Further analysis of standard deviation and confidence interval shows that the system performance fluctuates little, and the results are stable and reliable, verifying its generalization ability under different sample subsets.

5 Discussion

The intelligent martial arts training system based on ViT and DQN proposed in this paper shows significant advantages in multiple core performance indicators. First, in terms of action recognition accuracy, the system achieved an average recognition accuracy of more than 96.8% on a large-scale martial arts action dataset containing 8,658 annotated samples, and the kicking action even reached a high accuracy level of 98.1%. This performance is better than the traditional CNN architecture and some LSTM methods for time series modeling.

Second, in terms of response speed and real-time feedback capability, this system uses a lightweight ViT model and an efficient DQN strategy update mechanism to ensure that the end-to-end delay from image input to action optimization suggestions is controlled within 100ms, meeting the strict requirements for instant feedback in martial arts training. In contrast, traditional CNN+RL solutions often cause high inference delays due to complex model structures and unstable training, making it difficult to achieve proper "real-time guidance". In addition, combined with the kinematic parameters collected by IoT sensors, the system further improves the accuracy of action evaluation. It reduces the risk of misjudgment caused by a single source of visual information, reflecting the value of multimodal input in improving response quality.

In terms of generalization and environmental adaptability, the system significantly improves the model’s applicability in real training grounds by constructing a data set covering a variety of lighting conditions and different backgrounds. Experimental results show that even under extreme conditions such as low light or reflection, the recognition accuracy of the system remains above 94%, indicating that it has good environmental fault tolerance. In contrast, many existing methods only verify their performance in ideal laboratory environments and lack comprehensive support for real-world training scenarios. In addition, the system supports the generation of personalized training paths for athletes across age groups, reflecting a strong ability to adapt to individual differences, especially in young groups. The training effect is most significantly improved, fully demonstrating the system's application potential in teaching and youth sports training.

Although this system is superior to existing work in many indicators, it still has some limitations. On the one hand, the current version mainly models 6 types of typical martial arts movements. It has not yet covered more complex combination techniques and tactics, which, to a certain extent, limits its promotion in high-level competitive training. On the other hand, although multimodal data fusion enhances system stability, it also brings higher hardware deployment costs and data synchronization challenges, mainly when used in mobile training or remote areas, where there may be a problem of insufficient equipment. In addition, as a reinforcement learning algorithm based on the Q-value function, DQN may have limited strategy exploration efficiency when facing extremely high-dimensional state spaces. In the future, consider introducing the Actor-criticism framework to improve the stability of the strategy during long-term training.

In summary, the intelligent martial arts training system constructed in this study has shown superior performance to traditional CNN+RL solutions in terms of accuracy, response speed, and generalization ability, especially in complex action recognition and personalized training optimization, providing a new technical paradigm for intelligent sports training. At the same time, the system also exposed certain limitations in practical applications, suggesting that future research can be further optimized from the perspectives of action category expansion, lightweight deployment, algorithm iteration, and upgrade to promote the in-depth application and sustainable development of artificial intelligence in martial arts training and even the entire sports field.

6 Conclusion

This study proposes an intelligent martial arts training system based on computer vision and reinforcement learning. By applying ViT for action recognition and combining DQN to optimize the training process, real-time recognition of martial arts actions and optimization of techniques are achieved. Experimental results show that the system is significantly better than traditional training methods in terms of training effect, especially in young athletes. In addition, the intelligent training system can provide personalized training feedback and adjust training strategies according to the real-time performance of athletes, thereby effectively improving the competitive level of athletes. The contribution of this study is to promote the innovation of martial arts training technology by using advanced deep learning models and reinforcement learning algorithms. Its practical significance is reflected in improving athlete skills, acceleratingefficiency, and reducing training costs. It plays an important role in promoting the application of intelligent training and personalized feedback.

Acknowledgment

This work was supported by the Henan Provincial Government Decision-Making Research Project (Project No. E305, 2010): Research on the Countermeasures for the Sustainable Development of the Martial Arts Industry in Henan Province, and the Henan Provincial Government Tendered Decision-Making Research Project (Project No. 2102B717, 2012): Research on the Inheritance and Development of Folk Sports in the Construction of New Rural Areas.

References:

[1] Chen G. An interpretable composite CNN and GRU for fine-grained martial arts motion modeling using big data analytics and machine learning[J]. Soft Computing, 2024, 28(3): 2223-2243. https://doi.org/10.1007/s00500-023-09565-z

[2] Tropin Y, Podrigalo L, Boychenko N, et al. Analyzing predictive approaches in martial arts research[J]. Pedagogy of Physical Culture and Sports, 2023, 27(4): 321-330.https://doi.org/10.15561/26649837.2023.0408

[3] Hui B. Visualization system of martial arts training action based on artificial intelligence algorithm[J]. Soft Computing, 2023: 1-12. https://doi.org/10.1007/s00500-023-08711-x

[4] Ueda T, Shima K, Mutoh A, et al. Martial Arts Demonstration Evaluation System Using Machine Learning to Reflect the Actual Evaluation Methods of Instructors[J]. Procedia Computer Science, 2024, 246: 312-319.https://doi.org/10.1016/j.procs.2024.09.410

[5] Al-Faris M, Chiverton J, Ndzi D, et al. A review on computer vision-based methods for human action recognition[J]. Journal of imaging, 2020, 6(6): 46.https://doi.org/10.3390/jimaging6060046

[6] Kong Y, Fu Y. Human action recognition and prediction: A survey[J]. International Journal of Computer Vision, 2022, 130(5): 1366-1401.https://doi.org/10.1007/s11263-022-01594-9

[7] Zhang K, Li Y, Wang J, et al. Real-time video emotion recognition based on reinforcement learning and domain knowledge[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(3): 1034-1047.https://doi.org/10.1109/TCSVT.2021.3072412

[8] Weng J, Jiang X, Zheng WL, et al. Early action recognition with category exclusion using policy-based reinforcement learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(12): 4626-4638.https://doi.org/10.1109/TCSVT.2020.2976789

[9] Meng Z, Zhang M, Guo C, et al. Recent progress in sensing and computing techniques for human activity recognition and motion analysis[J]. Electronics, 2020, 9(9): 1357.https://doi.org/10.3390/electronics9091357

[10] Babangida L, Perumal T, Mustapha N, et al. Internet of things (IoT) based activity recognition strategies in smart homes: a review[J]. IEEE sensors journal, 2022, 22(9): 8327-8336.https://doi.org/10.1109/JSEN.2022.3161797

[11] Bird JJ, Ek a rt A, Faria D R. British sign language recognition via late fusion of computer vision and leap motion with transfer learning to american sign language[J]. Sensors, 2020, 20(18): 5151.https://doi.org/10.3390/s20185151

[12] Cao J, Tanjo Y. High-Accuracy Human Motion Recognition Independent of Motion Direction Using A Single Camera[J]. International Journal of Innovative Computing, Information and Control, 2024, 20(4): 1093-1103.https://doi.org/10.24507/ijicic.20.04.1093

[13] Xin W, Liu R, Liu Y, et al. Transformer for skeleton-based action recognition: A review of recent advances[J]. Neurocomputing, 2023, 537: 164-186.https://doi.org/10.1016/j.neucom.2023.03.001

[14] Li J, Liu X, Zhang W, et al. Spatio-temporal attention networks for action recognition and detection[J]. IEEE Transactions on Multimedia, 2020, 22(11): 2990-3001.https://doi.org/10.1109/TMM.2020.2965434

[15] Lv Shuping, Huang Yi, Wang Yingying. Research on human action recognition based on two-stream convolutional neural network[J]. Experimental Technology & Management, 2021, 38(8).

[16] Ding C, Zhang L, Chen H, et al. Human motion recognition with spatial-temporal-convLSTM network using dynamic range-doppler frames based on portable FMCW radar[J]. IEEE Transactions on Microwave Theory and Techniques, 2022, 70(11): 5029-5038.https://doi.org/10.1109/TMTT.2022.3200097

[17] Zhang H, Yang K, Cao G, et al. ViT-LLMR: Vision Transformer-based lower limb motion recognition

from fusion signals of MMG and IMU[J]. Biomedical Signal Processing and Control, 2023, 82: 104508.https://doi.org/10.1016/j.bspc.2022.104508

[18] Wensel J, Ullah H, Munir A. Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos[J]. IEEE Access, 2023.https://doi.org/10.1109/ACCESS.2023.3293813

[19] Ortiz A C. Capturing, Modelling, Analyzing and providing Feedback in Martial Arts with Artificial Intelligence to support Psychomotor Learning Activities[D]. Master's thesis. Universidad Nacional de Educación a Distancia, 2020.From: https://oai.e-spacio.uned.es/server/api/core/bitstreams/4a43f35e-d961-4552-8ade-fa466eed7ae2/content

[20] Wu B, Zhou J. Video-Based Martial Arts Combat Action Recognition and Position Detection Using Deep Learning[J]. IEEE Access, 2024.https://doi.org/10.1109/ACCESS.2024.3487289

[21] Chen G. An interpretable composite CNN and GRU for fine-grained martial arts motion modeling using big data analytics and machine learning[J]. Soft Computing, 2024, 28(3): 2223-2243.https://doi.org/10.1007/s00500-023-09565-z

[22] Park J, Shin S, Kim T, et al. Applying Computer Vision to Martial Arts[J]. Journal of Asian Society for Health & Exercise, 2022, 4(2): 45-54. From: https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE11396109

[23] Gordon J. Empowered Learners: How Martial Arts Foster Self-Regulated Learning[J]. Mantle: The Annual Review of Interdisciplinary Research, 2023, 1(1): 29-39. https://doi.org/10.14288/mantle.v1i1.198385

[24] Yan, P., Zhang, Y. Retraction Note: Application of light sensors based on reinforcement learning in martial arts action optimization and image defect detection. Opt Quant Electron 56, 1638 (2024). https://doi.org/10.1007/s11082-024-07607-w

[25] Wang Y, Song L. Application and Optimization of Convolutional Neural Networks Based on Deep Learning in Network Traffic Classification and Anomaly Detection[J]. Informatica, 2025, 49(14). https://doi.org/10.31449/inf.v49i14.7602

[26] Chen Y. Human Resource Recommendation Based on CBCF-BAC and Short Text Similarity Algorithm[J]. Informatica, 2024, 48(22). https://doi.org/10.31449/inf.v48i22.6853