Improved OpenPose and Attention Mechanism for Sports Start Correction

Bin Jiang
Xinxiang Institute of Engineering, Xinxiang, 453000, China
Email: jiangbin@xxgc.edu.cn

Keywords: human posture estimation, attention mechanism, sports, openpose algorithm, time mode

Received: July 23, 2024

The development of sports science and computer vision technology has made human pose recognition have great potential for application in sports. Considering the limitations of starting motion correction, the study proposes to use improved OpenPose posture estimation to achieve motion information analysis, and introduces attention mechanism to perform correlation analysis on athletes' limb posture information. Specifically, based on OpenPose's limitations in detecting subtle actions, inter frame information association and pose information classification processing are performed. Secondly, design a lightweight attention module and a network model under time attention mechanism for corrective analysis. The experimental results show that the improved algorithm has a starting posture recognition accuracy of over 0.85, and compared with the Dynamic Time Warping (DTW) model and the high-resolution human keypoint detection network (HRNet-w48), its recognition accuracy for the three stages of the starting action is over 93%. The ROC results indicate that the maximum accuracy of the research method's posture recognition can reach 0.97, which is superior to other baseline models, and its detection and recognition time for starting actions is less than 0.1 seconds. The methods proposed by the research institute have broad application prospects in improving sports performance and scientific training, and can provide suggestions and references for athletes' movement correction.

Povzetek: Predlagana metoda združuje izboljšano OpenPose knjižnico in mehanizem pozornosti za analizo športnih začetkov. Omogoča natančnejše popravljanje športnih gibov ter optimizacijo treningov.

1 Introduction

The sports start plays a crucial role in many competitive events, and a good start movement not only provides athletes with better starting speed and reaction time, but also directly affects the outcome of the whole game. On the contrary, inaccuracy and instability of the starting movement may lead to problems such as reduced speed, wasted energy and falls, which can have a negative impact on athletes' performance [1]. As there are large differences in physical fitness, starting posture habits and mental attention among athletes, improving incorrect starting posture and performance through start correction is of great value and significance for improving the level of athletic competition and improving athletes' performance [2]. Traditional start correction methods mainly rely on the experience and intuition of manual coaches, and lack objective and accurate assessment means. And due to the transient and complexity of the starting movement, it is difficult to capture and judge the small starting differences with the naked eye, so in order to avoid the subjectivity and limitation of judgement and corrective action, the analysis with the help of machine vision can effectively improve the accuracy of correction [3]. The study proposes to analyse the sports with the help of human posture estimation and attention mechanism to systematically capture the athletes' movement behaviour and acquire posture data. At the same time, the attention mechanism is introduced to improve the attention to the analysis of the starting movement, and to achieve the capture of key dynamic movement information and improve the accuracy of the analysis of the starting movement by simulating the human visual attention process. The guidance of the attention mechanism can help observers better understand the essential laws of the starting movement, and then achieve the target of corrective guidance [4]. Unlike some scholars who only combine human pose estimation with image data and multi-scale position networks, this approach can achieve good accuracy, but it inevitably lacks capture of dynamic details in human pose recognition. There are also some studies that combine attention mechanisms with neural network algorithms to solve the problem of information overfitting in human pose recognition, and the computational cost of algorithms. However, if applied in sports, it shows insufficient attention to the characteristics of each stage of posture. The innovation of the research lies in the combination of human pose recognition technology and lightweight attention mechanism, which can highlight weight information differences on the basis of grasping frame pose correlation analysis, improve information processing ability, increase the interpretability of the model, and more accurately focus on key poses and stages in running operations, improving recognition efficiency and accuracy. And this method can effectively avoid the limitations brought by subjective pose recognition, and can also reduce computational complexity while ensuring algorithm performance. In the starting motion, differences in the position and speed of certain joints can affect the starting effect of athletes. The introduction of attention mechanism can enable the model to focus on key features, reduce the interference of irrelevant information, and also reduce the overall resource consumption of the model, improving its applicability on different devices. The study analyses sports correction from four aspects, the first part is a review and discussion of the relevant literature on human posture correction and motion analysis algorithms, the second part is the design of the correction model of sports starting movement based on human posture recognition algorithms and the attention mechanism, the third part is the examination and analysis of the results of motion correction recognition under the fusion method, and the last part is an overview summary. overview summary of the full paper.

2 Literature review

Human posture estimation is a key core technology in the field of computer vision, which mainly refers to the detection of various parts of the human body, orientation and scale information from images or videos, and is used in human activity analysis and human-computer interaction. Tani Y et al. proposed a mobile weight-supporting walker, which used a colour-depth camera to measure the trainer's abdomen and completes ellipsoid fitting of the point cloud, by which the walker can estimate the position and orientation of the walking training object and track the training object based on proportional control. Experiments show that the researcher's proposed walking training system possesses high accuracy with a root-mean-square error of only 9 cm [5]. In order to increase the speed of data storage in human body segmentation, Le V and other scholars used 3D human posture estimation and human activity recognition in human body segmentation, firstly, generate human body 3D point cloud data and create action markers through colour-depth images, and evaluate the effect of video human body segmentation on the dataset with different deep learning methods, and the results show that the efficiency of human body segmentation is greatly improved [6]. In order to address the shortcomings of human gesture recognition in distinguishing small-scale keypoints and confusing them, Xu J and other scholars proposed a multi-scale location enhancement network based on convolutional neural network, which dynamically selects and fuses multi-dimensional features using multi-scale adaptive fusion units, and designs the location enhancement module to differentiate confusing semantics. Experimental validation shows that the research proposed method has higher accuracy in both single and multi-person pose estimation [7].Kumar P et al. in their in-depth study of human pose recognition in 2D and 3D, validate the accuracy of HPE in various aspects of activity recognition, animated games and video tracking, and a case study study on the use of Deep Learning algorithms in human pose estimation, and finally summarise the limitations and research directions of HPE [8].Huang L and other scholars applied human posture recognition technology in sports injury identification and correction, taking functional motion detection technology as the basis, calculating human key points through the detection system, selecting test indicators, and automatically detecting human health by calculating human posture [9]. Table 1 summarizes the relevant literature.

Table 1: Summary of previous research work

Literature	Method	Advantages	Insufficient
Tani Y et al. [5]	Measuring abdominal point cloud data using a color depth camera	The walking training system has high accuracy and small root mean square error	The application location of the human body is relatively limited
Le V et al. [6]	Combining 3D human pose estimation and human activity recognition to generate point cloud data	Improve human segmentation efficiency, enhance data storage speed and processing efficiency	Possible overfitting issues in action evaluation
Xu J et al. [7]	Propose a multi-scale position enhancement network based on convolutional neural networks to distinguish pose recognition	Can distinguish small-scale keypoints and improve pose estimation accuracy	Not considering the temporal nature of human pose recognition
Kumar P et al.[8]	Using deep learning algorithms to evaluate human posture	It is well applied in animated games and video tracking.	Not considering the deep features of human pose recognition
Huang L et al. [9]	Application of human pose recognition technology in sports injury recognition and correction	Automatic detection of key points in the human body can effectively analyze movement ability	Insufficient quantitative analysis of actual corrective training effects
Zhou K et al. [11]	Behavior recognition system based on spatiotemporal convolution and attention mechanism	The accuracy of behavior detection and recognition has significantly improved, solving the problem of overfitting of deep information	There are few dimensions for detecting behavioral actions

Li Y et al. [12]	Human Bone Behavior Recognition Algorithm Based on Spatiotemporal Attention Graph Convolutional Neural Network	The algorithm has high robustness and stability, meeting the needs of human motion recognition	The effectiveness of the application scenario needs further verification
Li X et al. [13]	Convolutional neural network based on dual attention mechanism for two-dimensional human action recognition	Reduce costs while maintaining recognition accuracy,	Environmental adaptability is limited, making it difficult to meet the requirements of dynamic video action recognition
Zhang Y et al. [14]	Building a lightweight fusion model for 2D and 3D action recognition	The streamlined action recognition framework achieves a recognition accuracy of 92.2%, saving computational resources	The fine feature performance of action recognition still needs to be improved

Different sports joints have their own functional bias, such as flexibility, stability, etc., and corrective training is of great practical value for the improvement of athletic ability and the activation of joint function. Among them, the attention mechanism in neural networks can better achieve the allocation of computational resources [10]. Zhou K and other scholars proposed a behaviour recognition system based on spatio-temporal convolution and attention mechanism, which fuses deep spatial information into segments to solve the problem of overfitting of depth information. The method was used in display surveillance data and the results showed that the framework has higher recognition accuracy in behaviour detection [11]. In order to fully explore the spatio-temporal features of human actions, Li Y and other scholars proposed a human skeleton action recognition algorithm based on spatio-temporal attention map convolutional neural network, which optimises the algorithm from both spatial and temporal aspects by using the dual attention mechanism, introduces the global average pooling and the auxiliary classification loss in spatial recognition, and extracts the temporal domain segments automatically from temporal recognition. It is shown through experiments that the algorithm has strong robustness and stability, and can better meet the human action recognition needs [12]. In order to reduce the cost of commonly used human action recognition algorithms and improve the recognition efficiency, Li X and other scholars proposed a convolutional neural network based on double attention mechanism. This method was used in 2-dimensional human action recognition of colour images and tested on public datasets, which proved that the cost of using this method is 70%~80% of the commonly used methods, and the accuracy of human action recognition is still maintained at a high level [13]. In order to improve the timeliness of video popularity recognition, Zhang Y and other scholars constructed a lightweight fusion model of 2D and 3D action recognition, using 2D convolutional network to extract feature maps and introducing 3D convolutional network to process the temporal information, which greatly reduces the complexity of action recognition. It was shown through experiments that the recognition accuracy of the streamlined action recognition framework proposed by the study reached 92.2% [14].

In summary, the above-mentioned content uses image information data, point cloud data, neural network and other methods for human skeleton action recognition, which has good application effect, but it is difficult to control the error results of data processing and data redundancy problems caused by too much data information. The study, while leveraging the advantages of human posture estimation for motion analysis, differs from the above literature in focusing on the features extracted by posture recognition. The study realises the corrective analysis of sports from both spatial and temporal perspectives, focuses on the continuity and dynamics of the movement analysis, and provides a detailed delineation of the fine phases of the starting movement. And unlike the previous literature which only uses the attention mechanism to achieve the classification of feature weights, the study implements the attention mechanism to focus on the analysis of the starting movement, and applies the human visual attention to capture the information of the key dynamic movements, further enrich the theoretical foundation of exercise correction.

3 Research on starting correction in sports based on human posture estimation and attention mechanism

The variability of different sports makes the movement requirements of such forms as standing start and squatting start different, and the correction effect is often limited due to the transient and speedy nature of the starting movement [15]. In order to better correct the starting movement in sports, the study introduces the human posture estimation method for motion information analysis, in order to better detect and estimate the position of joints from the image information, and introduces the attention mechanism to better correlate and analyse the athlete's limb posture information, so as to better provide suggestions for the correction of the movement.

3.1 Sports analysis based on human posture estimation algorithm

Human posture estimation is a method to recognise and estimate human posture through computer vision and deep learning techniques, which can be used to obtain human posture information by tracking and estimating the key parts of the human body in real time using devices such as cameras or sensors. In order to improve the accuracy of human joint point detection, the study proposes a video-based human posture estimation algorithm, i.e., the Openpose algorithm establishes an association matrix implementation with frame posture relationships on the basis of static posture analysis, and generates and adequates key points in the image based on motion continuity and joint restorativeness, and then obtains the global human body posture [16].OpenPose In the detection of skeletal joints, OpenPose can make good use of its OpenCV and C programming advantages to achieve the tracking and identification of human joints, which mainly takes the input colour image and predicts the part confidence map and force vector field through the feed-forward network to encode and correlate the links between different skeletal joints [17-18]. The network architecture is shown schematically in Fig. 1.

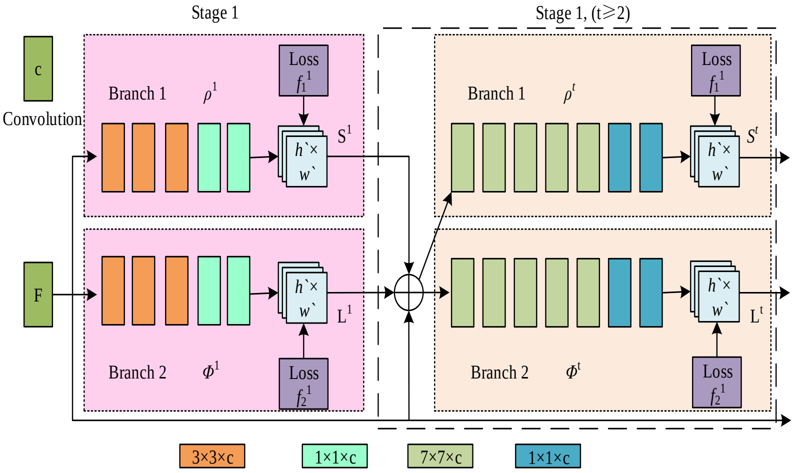


Figure 1: Schematic diagram of OpenPose network structure.

This network structure needs to extract the features and mapping relationships of its original image before starting, and subsequently outputs the confidence map and affinity vector field in one part of the stage, both of which can be iterated to achieve the prediction of the features, and Eq. (1) is the computational expression.

$$\begin{cases} S^t = p^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2 \\ L^t = \varphi^t(F,, S^{t-1}, L^{t-1}), \forall t \geq 2 \end{cases} \tag{1}$$

In Eq. (1), F denotes the feature mapping of the original graph, L^{t-1} is the set of vector fields under the prediction stage $t-1$, S^t is the confidence map, L^t is the affinity vector field, t is the node step, and p^t, φ^t denotes the intermediate results of S^t and L^t under the step. Meanwhile loss function exists in different network structures to achieve numerical prediction, which can be expressed as equation (2).

$$\begin{cases} f_S^t = \sum_j^J \sum_P W(P) \cdot \left\| S_j^t(p) - S_j^*(p) \right\|_2^2 \\ f_L^t = \sum_1^C \sum_P W(P) \cdot \left\| L_C^t(p) - L_C^*(p) \right\|_2^2 \end{cases} \tag{2}$$

In Eq. (2), S_j^* denotes the body part confidence of the real label, L_C^* is the vector field of the real label, W is the binary mask, P is the image position, and f_S^t, f_L^t denotes the upper and lower two branch networks. However, OpenPose inevitably exists misdetection and omission in detection, and the lighting conditions, shadow problems, and the habit of individual actions will make the difference analysis of subtle actions less effective. Therefore, the study is based on OpenPose processing static images to correlate the information between frames, and the inter-frame gesture distance is used to achieve the tracking and correlation of key gestures. The human body posture is smooth and coherent, and some jumping nodes can cause image jitter. The tracking and association ability of inter frame postures can effectively process adjacent frame information and better grasp the understanding and analysis of human behavior. The pose distance can be expressed as equation (3).

$$d_c(p_1, p_2) = \sum_i \frac{n_i}{m_i} \tag{3}$$

In Eq. (3), i denotes the skeletal joint points, p_1, p_2 are both poses, $\frac{n_i}{m_i}$ denotes the extracted bounding box similarity of the two poses, m_i denotes the matching points of p_2 , and n_i is the feature point of p_1. The correlation matrix can be established by bipartite graph maximum matching for the pose tracking information. The bipartite graph maximal matching algorithm for inter-frame human posture tracking can represent multiple key frames of human posture as a bipartite graph, where each vertex set represents a joint state at one point in time, and the edges represent the connections of the same joint at two points in time. The transformation formula of the classical bipartite graph can be expressed as equation (4).

$$H = (V^{+}, V^{-}, E) \tag{4}$$

In Eq. (4), V^{+}, V^{-} denotes the set of column nodes in the bipartite graph and E is the set of edges. The bipartite maximum matching algorithm calculates the similarity of human poses between different frames and performs maximum matching based on these similarities. Even if individual key point detection fails or has a large error in some frames, it still achieves pose processing through maximum matching with other frames. The algorithm has clear performance in structure and computation and is well applied in human pose tracking. The algorithm will consider the spatial prior knowledge and motion continuity between joints to determine the best matching relationship. Continuous joint changes can show its motion trajectory, so considering the shallow nature of information capture by traditional cameras, the study uses a bilinear interpolation algorithm to simulate the continuous action, see equation (5).

$$\begin{cases} x_{cg,k}^{id} = \dfrac{j-c}{j-i} * x_{i,k}^{id} + \dfrac{c-i}{j-i} * x_{j,k}^{id} \\ y_{cg,k}^{id} = \dfrac{j-c}{j-i} * y_{i,k}^{id} + \dfrac{c-i}{j-i} * y_{j,k}^{id} \end{cases} \tag{5}$$

In Eq. (5), $x_{i,k}^{id}, y_{i,k}^{id}$ denotes the horizontal and vertical coordinates of the better articulation point of forward search, $x_{j,k}^{id}, y_{j,k}^{id}$ denotes the horizontal and vertical coordinates of the better articulation point of backward search, i, j is the number of frames, id is the tracking marker, and k is the number of articulation points. The bilinear interpolation algorithm utilizes weighting to achieve image processing, which can effectively achieve inter frame interpolation and smooth transitions in simulating continuous actions. This interpolation method can effectively utilize the information of surrounding points to estimate the values of frame pixels, thereby achieving smoothing effects. When judging and repairing the wrong part of the skeletal joint points with the help of OpenPose, when the confidence map of the joint points is set to be less than the threshold, it needs to be repaired, and vice versa, the position prediction is done by interpolation, and the range of the diagonal length of the frame corresponding to the human body's posture is delineated. Fig. 2 shows the flowchart of repairing joint points.

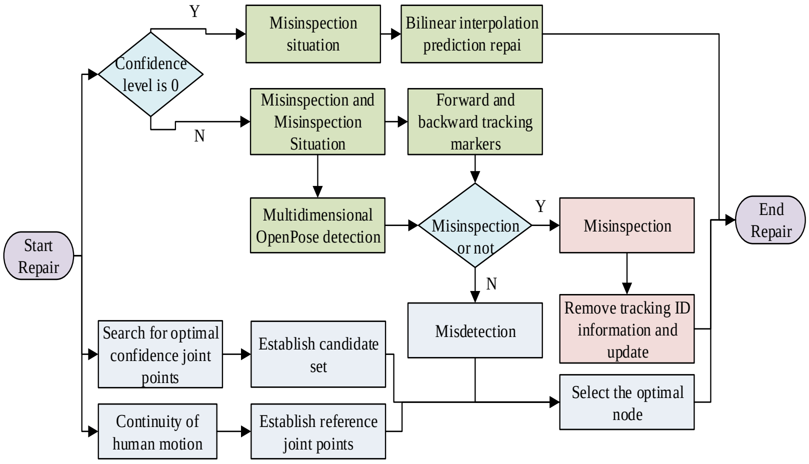

Start Repair
Confidence level is 0
Y
N
Misinspection situation
Bilinear interpolation prediction repai
Misinspection and Misinspection Situation
Forward and backward tracking markers
Multidimensional OpenPose detection
Misinspection or not
Y
N
Misinspection
Misdetection
Remove tracking ID information and update
Search for optimal confidence joint points
Establish candidate set
Continuity of human motion
Establish reference joint points
Select the optimal node
End Repair

Figure 2: Flowchart for the repair of joints.

In the section of candidate joints, the study does not process the skeletal joints with single pixel points, but chooses to select the skeletal joints and their intersection regions as the minimum strength after the image is hyperpixel segmented, and then realises the regional feature delineation of the image information. Simple linear iterative clustering is a hyperpixel generation method based on mean clustering and mainly sets the number of hyperpixels of the same size, then samples the initialised cluster centre and calculates the distance metric between the pixel and the cluster centre. The mathematical expression of the distance metric is shown in equation (6).

$$D = \sqrt{(\frac{d_c}{N_c})^2 + (\frac{d_s}{N_s})^2} \tag{6}$$

In Eq. (6), d_c, d_s represents the distance measurement of pixel points c, s , and N_c, N_s are the number of pixels in its image. When more than half of the human body poses are detected, the human body tracking information needs to be fine phosphorus labelled with the help of forward and backward search and its pose similarity is calculated, the mathematical expression of which is shown in Eq. (7).

$$PS_{i,j}^{id,k} = \alpha * \sum_k \frac{n_k}{m_k} + (1-\alpha) * \|H_i - H_j\|_2 \qquad (7)$$

In Eq. (7), i, j denotes the number of frames, k denotes the number of joints, id is the markers, n_k, m_k is the number of optical flow feature points in the bounding box extraction, $\sum_k \frac{n_k}{m_k}$ denotes the gesture distance, H_i, H_j denotes the colour histogram of the gesture frames in the i and j frames, $\| \ \|_2$ denotes the Euclidean distance, and α is the value of weights. The classification process of the tracking marker information of the gesture can better identify the discrimination error and omission, and if the similarity of the two gestures is higher than the threshold value, there is no identity problem of the gesture.

3.2 Analysis of starting movements with improved attention mechanisms

In sports start correction research, the attention mechanism model can help identify the key movements and postures of athletes, and by combining with human posture estimation techniques, useful features can be extracted from the key points and skeleton information of athletes, and focus attention on information such as the starting line, starting position, and power point [19]. The more common attention mechanisms are the spatial attention mechanism and the channel attention mechanism, which emphasise certain features of the image by focusing attention on specific regions or weighting out the channels, respectively [20]. A Convolutional Block Attention Module (CBAM) designed for network models is proposed for feature analysis. This CBAM module is formed by two modules of channel attention and spatial attention in series, and its structure is schematically shown in Fig. 3.

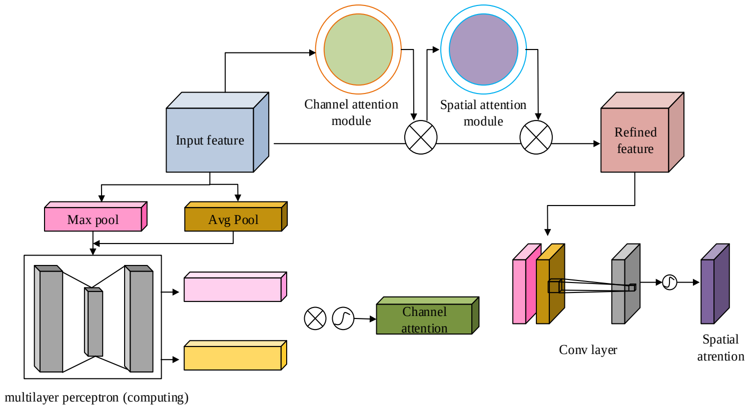


Figure 3: Schematic diagram of CBAM module structure.

CBAM takes the input feature maps based on channel and spatial dimensions and computes the attention maps separately, and completes the feature adjustment by pixel multiplication of the attention maps with the original feature maps. The channel attention module mainly adds a global maximum pooling layer that pays more attention to the details of the features, and its parallel approach also makes the information more prominent. The input feature maps in the channel attention are processed by global maximum pooling and global average pooling respectively, and then the force map is generated by the multilayer perceptron with shared weights. Its computational formula can be expressed as equation (8).

$$M_c(F) = \sigma(MLP(AP(F)) + MLP(MP(F)))$$
(8)

In Eq. (8), $MLP(AP(F)$ represents the global average pooling, $MLP(MP(F)$ is the global maximum pooling, and σ is the feature weight parameter. The spatial attention module pays more attention to the importance of spatial information to the attention mechanism, and its connected feature maps will be processed by convolution, function activation to get the feature maps that can be multiplied by pixels with the input features, which is mathematically expressed as Eq. (9).

$$M_s(F) = \sigma(f^{7*7}[(AP(F)); MP(F)])) \qquad (9)$$

In Eq. (9), it represents the convolution layer with convolution of 7*7. Sports starting action includes different phases, such as the pre-positioning phase, preparation phase and starting phase, etc., and the action essentials of different action time periods are different, adding the Temporal Pattern Attention Mechanism can analyse the key moments of the starting action during the model training process, and better focus on the details of the starting action, to avoid lifting the head too early, failing to push up with force, etc. [21]. Therefore, the study uses Temporal Pattern Attention Mechanism (TPAM) to analyse the action sequence, the schematic diagram of which is shown in Fig. 4.

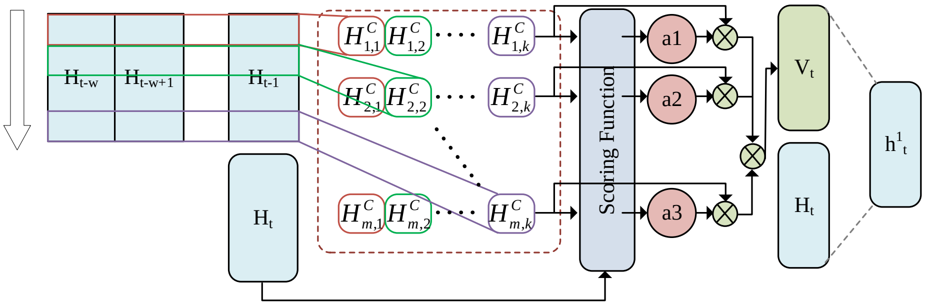

H_{t-w}
H_{t-w+1}
H_{t-1}
H_t
$H^C_{1,1}$ $H^C_{1,2}$ ···· $H^C_{1,k}$
$H^C_{2,1}$ $H^C_{2,2}$ ···· $H^C_{2,k}$
$H^C_{m,1}$ $H^C_{m,2}$ ···· $H^C_{m,k}$
Scoring Function
a1
a2
a3
V_t
H_t
h^1_t

Figure 4: Temporal pattern attention mechanism.

This attention mechanism is a convolutional processing and scoring function operation on the implicit layer with the help of a convolutional kernel to better evaluate the correlation between sequences, the mathematical expression of which is shown in Equation (10).

$$H^C_{i,j} = \sum_{l=1}^{w} H_{i(t-w-1+l)} \times C_{j,T-w+l} \qquad (10)$$

Where w denotes the length of the sliding window, $H^C_{i,j}$ is the result of i row vector and j convolution kernel computation, t is the time, T denotes the range covered by the attention mechanism, and l is the length of the implicit layer. The weight matrix of the attention mechanism can be expressed as equation (11).

$$\alpha_i = \mathrm{si}\,gmoid(f(H^C_i, h_t)) \qquad (11)$$

In Eq. (11), h_t denotes the hidden layer vector. The attention weight matrix and multiply the hidden vector and the attention weight matrix, and then get the context vector. The study combines the human body key detection network with the attention mechanism, which can better highlight the limb poses in the motion images and reduce the amount of model computation, and the combination of the two methods can extract the important features and achieve the union of local and overall information when performing the starting movement correction analysis, which in turn improves the target detection and expression ability of the network model. In other words, the results of the CBAM attention mechanism module are processed with a convolutional kernel size of 1 * 1 to obtain the final results, and the network can also achieve the extraction of features at different scales.

The study was conducted on the collection of athletes' starting take-off manoeuvres on various video websites and the example of squatting take-off manoeuvre including the data of different phases of this manoeuvre. The pose images of the starting action are imported into the network model and the positional data of the key points are obtained by prediction for analysis, and the image data are subsequently processed to highlight their target key information and reduce the interference of redundant information on the image detection accuracy. Considering the uncertainty of the range of values of the data to be processed for the study, standard deviation standardisation is used, i.e. by calculating the arithmetic mean and standard deviation of the original data variables. The standardisation process is carried out to make its processed variable values based on up and down around 0. The mathematical expression is shown in equation (12).

$$x* = \frac{x - m}{n} \qquad (12)$$

In Eq. (12), x represents the original data, m is the calculated mean, n represents the standard deviation, and $x*$ is the normalised data value. In sports training, in order to better correct the starting posture of athletes, the study establishes a motion image dataset that includes three stages of different postures, namely, positioning, preparation, and starting, and the network model based on the improvement of the attention mechanism is used for the correction analysis of the starting movement, and the system design is carried out by using the browser and server model (B / S mode). The requirement analysis of the correction system can be seen in Fig. 5.

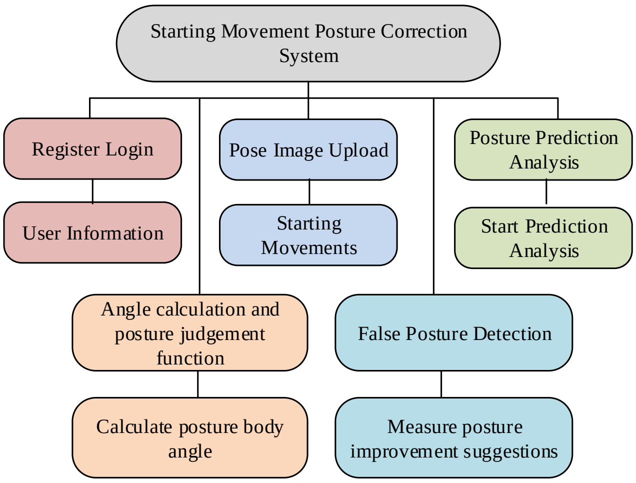

Starting Movement Posture Correction System
Register Login
User Information
Pose Image Upload
Starting Movements
Posture Prediction Analysis
Start Prediction Analysis
Angle calculation and posture judgement function
Calculate posture body angle
False Posture Detection
Measure posture improvement suggestions

Figure 5: Requirement's analysis of the starting movement correction system.

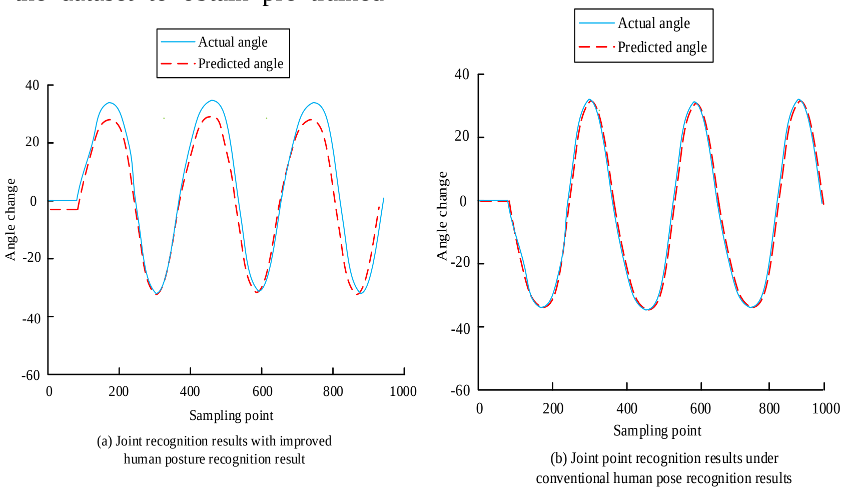
The starting movement correction system consists of five aspects: login and registration, postural image upload, postural image prediction, postural limb angle calculation and judgement, and result prediction and suggestion. After logging in and registering with the system, the user can upload the starting posture image, which will then be transmitted to the network model for prediction and analysis, and will be given the prediction results and related improvement suggestions. Accuracy is a fundamental indicator for measuring the overall performance of a classification model, representing the proportion of correctly predicted samples in the total sample size. In starting posture recognition, higher accuracy indicates that the model can effectively distinguish between correct and incorrect postures, which helps evaluate the effectiveness of posture correction. The recall rate reflects the model's ability to recognize positive examples (such as correct starting posture), and the higher the recall rate, the more real positive examples the model can find. The F1 score is the harmonic mean of accuracy and recall, used to comprehensively evaluate the performance of the model in positive and negative class recognition. In scenarios, especially when there is class imbalance (such as fewer samples with correct motion posture than incorrect ones), F1 score can effectively reflect the balance of the model between the two, improving the comprehensiveness of the evaluation. Equation (13) is the mathematical expression for accuracy, recall, and F1 value.

$$\begin{cases} P = (C / T) \times 100\% \\ R = (C / N) \times 100\% \\ F = 2 / (\frac{1}{P} + \frac{1}{R}) \end{cases} \tag{13}$$

In Eq. (13), P is the accuracy, R is the recall rate, F is the F-value, C represents the number of instances of a certain class correctly identified and classified, N is the total number of instances of a certain class in the test data, and T is the total number of instances of a certain class predicted by the classifier. These indicators focus on both the accuracy of overall predictions and the ability to identify specific categories, which helps to analyze and correct athletes' starting postures.

4 Analysis of the application of the corrective results of the starting movement

Research in the experimental process, the system front-end technology and back-end technology used are HTM, Javascript, and Java language, respectively, in the development language Python for library function calls. The deep learning framework structure is Pytho3.6 + PyTorch, the operating system in the hardware environment is Windows 10, CPU, GPU, memory hard disk size of Intel Core i5-9300H, NVIDIA RTX 3090, 24G / 1T, the operating system and the database in the software environment is Ubuntu19.10 and MySql. Research on collecting the starting movements of athletes on various video websites and expanding the dataset through data augmentation methods. This involves randomly reducing or enlarging the starting motion image data of athletes at a certain ratio, flipping or translating it according to a certain angle or distance transformation, and transforming the motion data image through color space transformation. Subsequently, the collected data will be standardized using deviation normalization, and the raw data will be transformed into linear transformation data in the (0,1) interval to ensure dimensional consistency. Then, a dataset of starting action posture images will be established. The image dataset includes three categories of images: positioning posture, preparation posture, and mid run posture. Each category contains 800 images of starting motion posture, for a total of 2400 images. Divide the starting posture dataset into training and testing sets in a 7:3 ratio. The training of the model requires a large amount of data, and using 70% of the data for the training set can help the model learn fully on diverse data and reduce the risk of overfitting. A ratio of 7: 3 can provide sufficient sample size for effective training while ensuring that the test set has sufficient size for statistical analysis and model evaluation. The study performed key point criteria on the captured image information of the starting action, including 27 key points and a human body detection frame. After combining the attention module with the network model, the motor pose of the starting action is trained and analysed, and Adam is selected as the training optimiser with a total of 300 rounds of training and an initial learning rate of 0.0001. Meanwhile for the calculation of the font joint angle the cosine calculation is performed based on the known joint coordinates. During the training process, data processing is carried out first. The collected images and videos are annotated using the Labelme software tool, and the motion image keypoints are manually annotated according to the dataset annotation standards. Each image includes 17 human keypoints and a human detection box. Each annotated image will generate a. json file. After merging the standard completed file with the dataset file, it can be used for training and testing the network model. Subsequently, the collected images will be rotated clockwise and counterclockwise to achieve data augmentation. A high-resolution detection network will be trained on the dataset to obtain pre trained parameters. The attention model will be added to the network model to detect and estimate joint positions from image information. Feature data acquisition and action sequence analysis can be achieved through a lightweight attention module and time attention mechanism, which can then be used to train the posture training set for starting movements and associate limb posture details.When conducting result detection, considering that the superpixels of some joint points are very small and the body pixels around the joint points have significant differences, this study directly takes the superpixels of the bone joint points as the minimum granularity, establishes a box (30 * 30) centered on the joint points, and establishes a candidate joint point set for the current frame's joint points to achieve joint point repair. Setting the search length to 5 during the forward and backward search processes ensures both the continuity of motion and the accuracy of joint repair. Firstly, the research proposed human pose estimation algorithm is analysed for image effect detection, its joint point detection results are analysed and compared with the traditional joint point recognition methods, and the results are shown in Fig. 6.

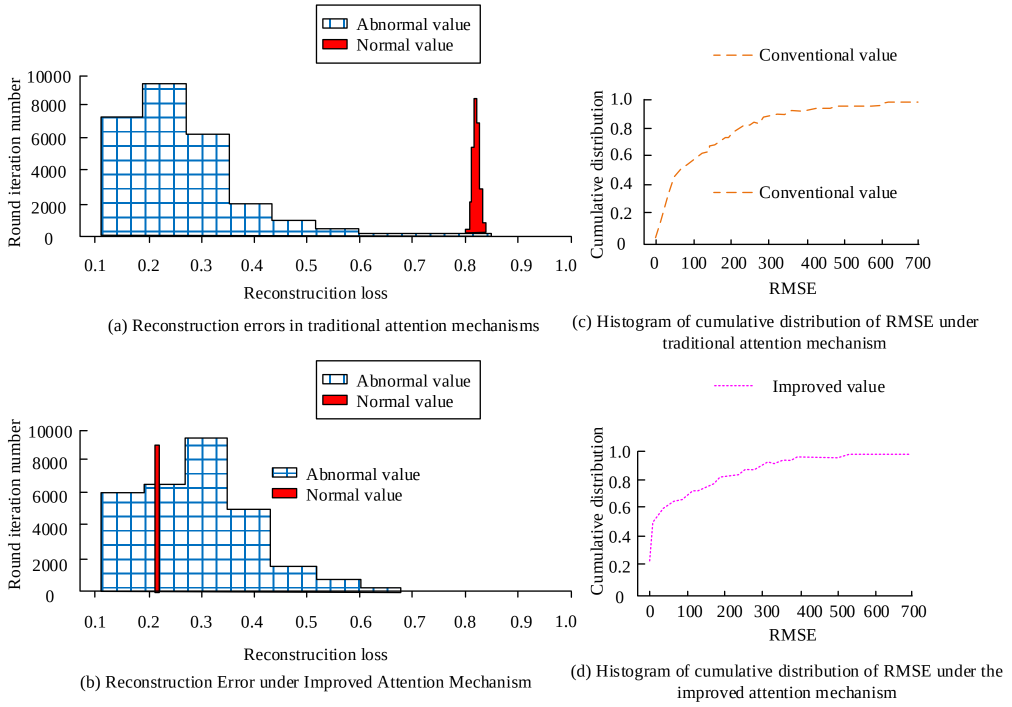


(a) Joint recognition results with improved human posture recognition result

(b) Joint point recognition results under conventional human pose recognition results

Figure 6: Posture recognition effect.

In Fig. 6, compared with the traditional key point recognition method, the improved human gesture recognition shows a smaller error result of the action gesture curve, and its recognition accuracy is basically above 90%, while the traditional recognition method has a large fluctuation situation when the sample points are 180, 450, and 720, and the maximum key point extraction deviation reaches 3.22%. Subsequently, the recognition error results of the improved attention mechanism proposed in the study are analysed and the results are shown in Fig. 7.



(a) Reconstruction errors in traditional attention mechanisms

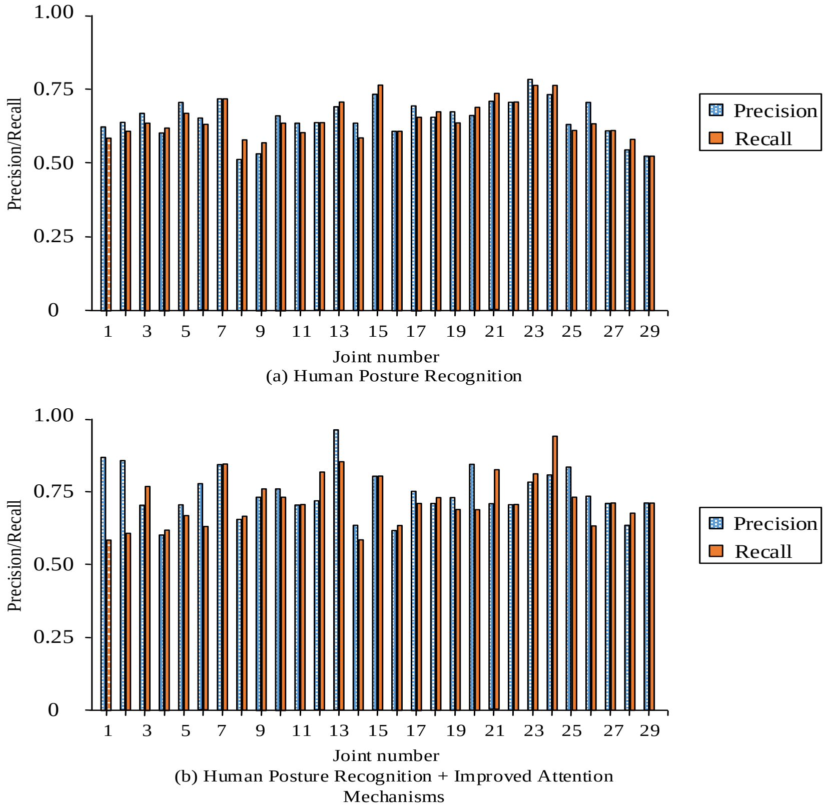
(c) Histogram of cumulative distribution of RMSE under traditional attention mechanism

(b) Reconstruction Error under Improved Attention Mechanism

(d) Histogram of cumulative distribution of RMSE under the improved attention mechanism

Figure 7: Error results before and after the improvement of the attention mechanism.

In Fig. 7, the addition of the improved attention mechanism makes the model show lower error results in the processing of the dataset, and its anomaly loss situation improves from 0.8 to about 0.2, and its cumulative root mean square error value also decreases with the addition of the improved attention mechanism, and the curve convergence is better. The pose recognition algorithm with the addition of the attention mechanism is subsequently analysed and its results are shown in Fig. 8.

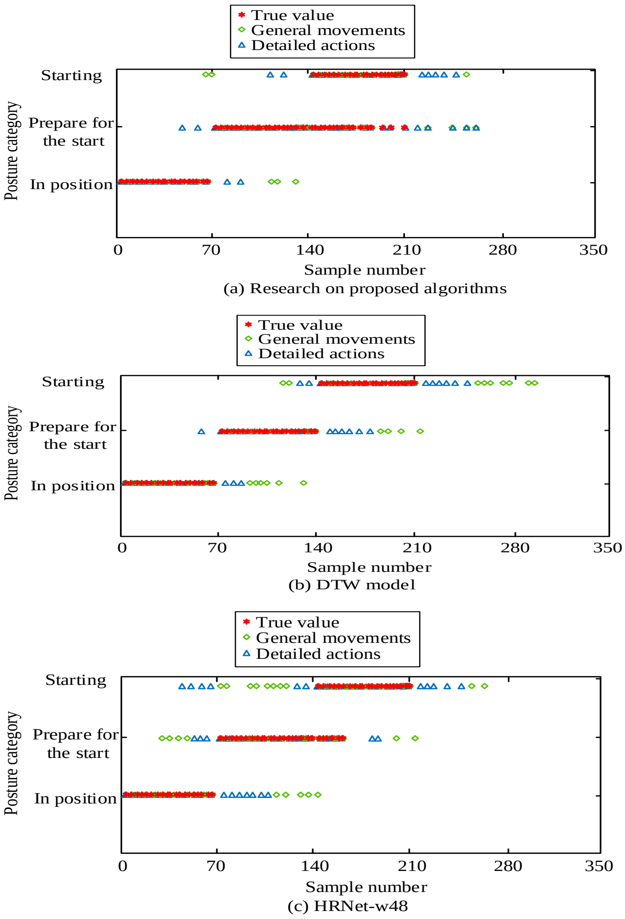


(a) Human Posture Recognition

(b) Human Posture Recognition + Improved Attention Mechanisms

Figure 8: Recognition results under improved pose recognition algorithm.

In Fig. 8, the precision curve of the model before the addition of the attention mechanism is basically lower than 0.80 in performing the starting gesture recognition, and the average values of its accuracy and recall are basically located at 0.75 and 0.73, with the overall recognition effect being poor. On the other hand, the networks under the fusion model exhibit recognition accuracies above 0.85, and the maximum recognition accuracy reaches 0.93, and the overall curve of the recall rate is also on the average value of 0.87. The addition of the attention mechanism resulted in an improvement in the precision of pose recognition. The matching accuracies of different algorithmic models in performing human pose recognition for the starting action were subsequently analysed, and the comparison results are shown in Fig. 9.



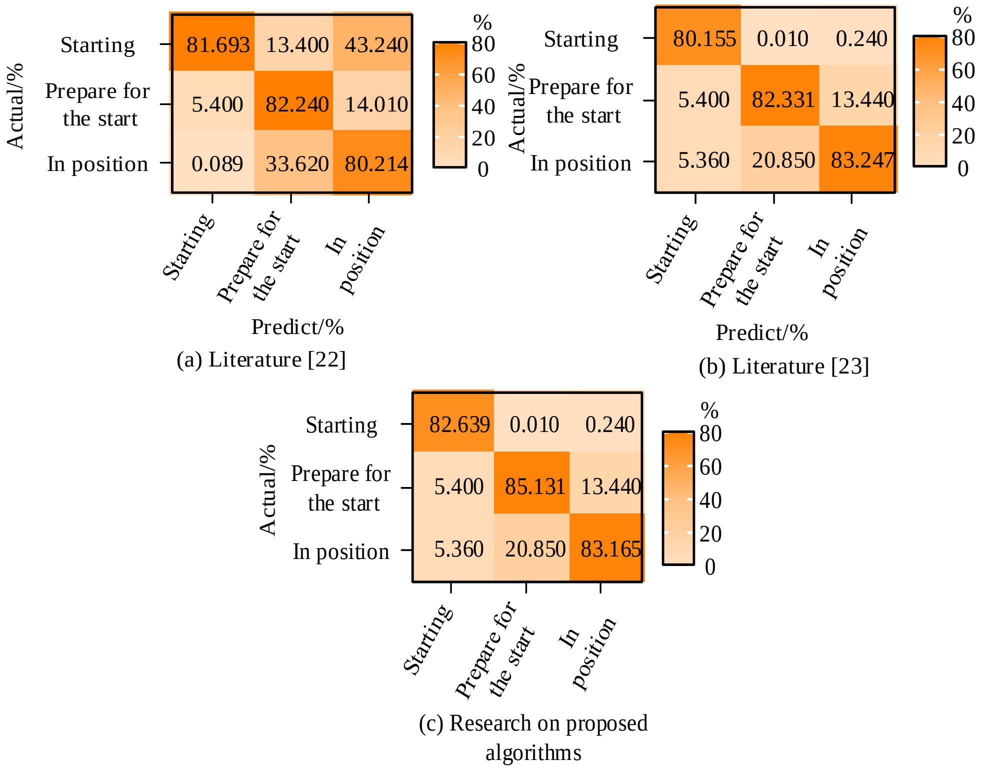
(a) Research on proposed algorithms
(b) DTW model
(c) HRNet-w48

Figure 9: Recognition results of the three algorithms for the starting movement poses.

In Fig. 9, the proposed fusion recognition algorithm exhibits recognition accuracies of 96.12%, 94.37%, and 93.54% in the three phases compared to the Dynamic Time Warping (DTW) model of the dynamic regularisation pose algorithm (DTW) and the high-resolution human body key-point detection network (HRNet-w48), with the matching effect on the starting pose being the best. The DTW model, on the other hand, showed pose recognition accuracies of 90.05%, 88.12% and 87.14%. The HRNet-w48 model was slightly better than the DTW model, but its accuracy did not exceed 93% and had a maximum difference magnitude of 3.47% with the algorithm proposed in the study. Subsequently, the algorithm proposed in the study will be compared with the posture action recognition technology that enhances the classification observation model [22], and the posture recognition technology that combines KNN and DTW [23], in order to better test and analyze its effectiveness. The result is shown in Fig. 10.

Fig. 10 shows the identification confusion matrix of different algorithms for the starting action stage, where the horizontal and vertical coordinates represent the predicted and true values, respectively. Specifically, the recognition and prediction results of different algorithms for starting actions are all above 80%, but there are still some differences. Among them, the recognition accuracy of literature [22] for the three actions is 81.693%, 82.240%, and 80.214%, respectively, and the recognition accuracy of literature [23] for the three actions is 80.155%, 82.231%, and 83.247%, respectively. The recognition accuracy of the algorithms proposed in the study in the starting action is generally above 82%, with an accuracy of 85.131% in the preparatory stage. The algorithm proposed in the study can perform correlation analysis on different frame pose relationships and perform targeted recognition while grasping the allocation of different pose data information, thus greatly improving the accuracy of pose recognition. Then, the training results of the three algorithms in Fig. 10 in the pre positioning stage, the preparation stage and the starting stage are compared, and the results are shown in Table 2.

The results in Table 2 show that the proposed algorithm can better learn the posture of the starting action, and the F1 value of each action is more than 0.9, which is much higher than the other two comparison algorithms. The pose recognition results of the different algorithms are analysed with the help of ablation experiments and the results are shown in Fig. 11.



(a) Literature [22]

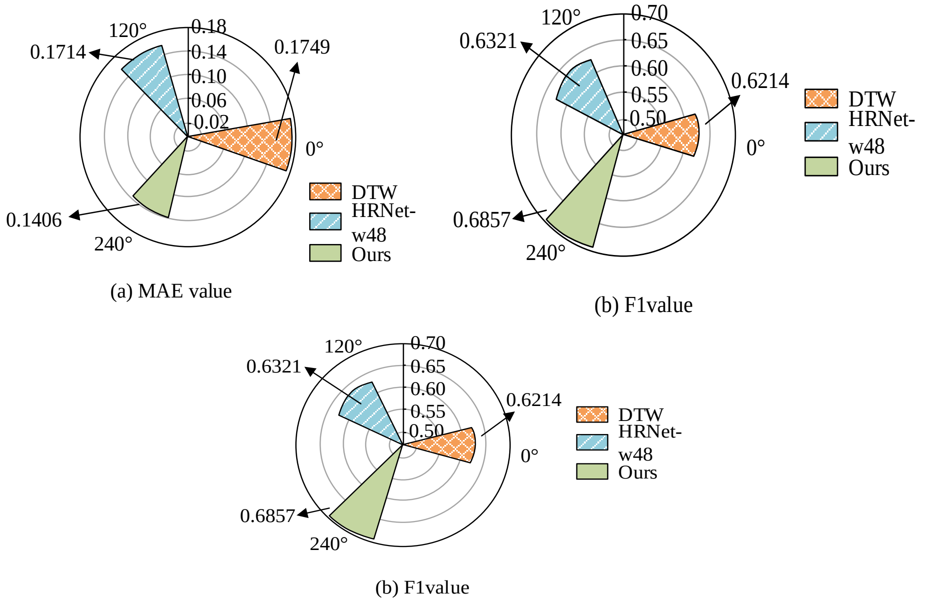
(b) Literature [23]

(c) Research on proposed algorithms

Figure 10: Starting movement recognition results for three algorithms.

Table 2: Training specificity and F1 value distribution results of three algorithms.

Stage	**Algorithm**	**Specificity**	**F1value**
Pre preparation stage	Research algorithm	0.996	0.975
	Literature [22]	0.895	0.951
	Literature [23]	0.897	0.953
Preparation stage	Research algorithm	0.999	0.987
	Literature [22]	0.886	0.967
	Literature [23]	0.874	0.958
Starting stage	Research algorithm	0.999	0.997
	Literature [22]	0.891	0.962
	Literature [23]	0.882	0.961



(a) MAE value

(b) F1value

(b) F1value

Figure 11: Ablation results of different algorithms.

In Fig. 11, the MAE error value (0.1406) exhibited by the proposed algorithm is smaller than that of the DTW model (0.1749) and the HRNet-w48 model (0.1714), and its F1 value (0.6857) is larger than that of the DTW model (0.6214) and the HRNet-w48 model (0.6321). The HRNet-w48 model performed well in the ablation experiment. The reason for the above results is that studying inter frame information correlation for OpenPose pose estimation can effectively capture dynamic changes in action sequences. Compared to DTW's fixed similarity matching and HRNet-w48's single frame prediction, inter frame information association provides richer contextual information and enhances the understanding of motion dynamics. And the lightweight attention module and temporal attention mechanism designed in the research can reduce the interference of redundant information, improve the matching degree of posture features, and enhance the attention mechanism to emphasize the information weight of important features compared to DTW. The other two comparison methods have limited ability in handling temporal features and complex behaviors. The above results indicate that the improved posture recognition algorithm proposed in the study can better recognize the starting action. Further utilizing Receiver Operating Characteristic (ROC) to analyze the application of the model, the closer the ROC curve is to the upper left corner, the larger the area under the curve, and the higher the accuracy of the experiment. Compare the Kinect human pose recognition, long short-term memory neural network with keyframe attention for motion pose recognition, KF-LSTM, and improved YOLOv7-POSE algorithm models. The results are shown in Fig. 12.

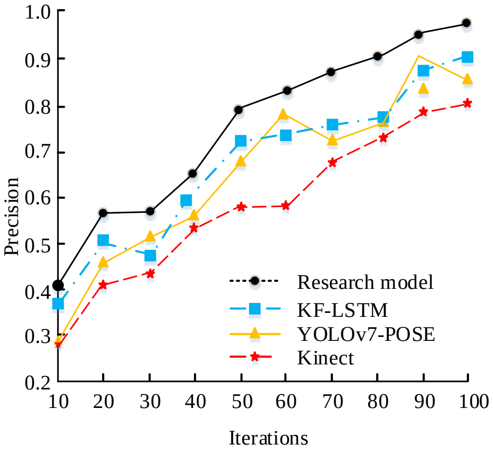


Figure 12: Baseline model comparison results.

The Fig. 12 shows the comparison results of ROC curves for different baseline models in motion posture recognition. It can be seen from the figure that the posture recognition model based on Kinect exhibits the worst ROC results, with a maximum value not exceeding 0.85, and its accuracy shows significant node fluctuations with increasing iteration times. The accurate variation curves of the KF-LSTM model and YOLOv7-POSE model are slightly better than those of the Kinect method, but their maximum values still do not exceed 0.9, and their curves fluctuate significantly in the later stage. Compared with other models, the fusion model proposed in the study has the largest area under the ROC curve and a maximum accuracy of 0.97, indicating better application performance in attitude recognition. The classification effect of the algorithm in performing limb gesture recognition is analysed with the help of Receiver Operator Characteristic Curve (ROC) and the results are shown in Fig. 13.

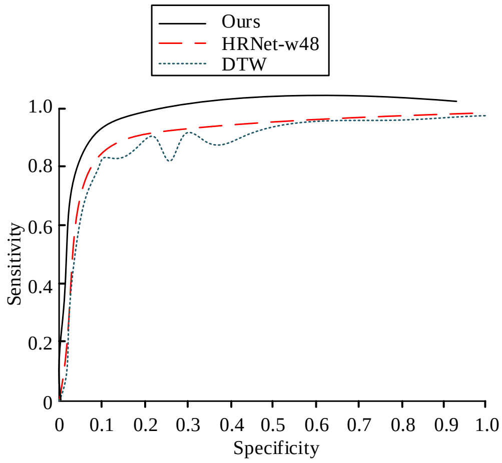
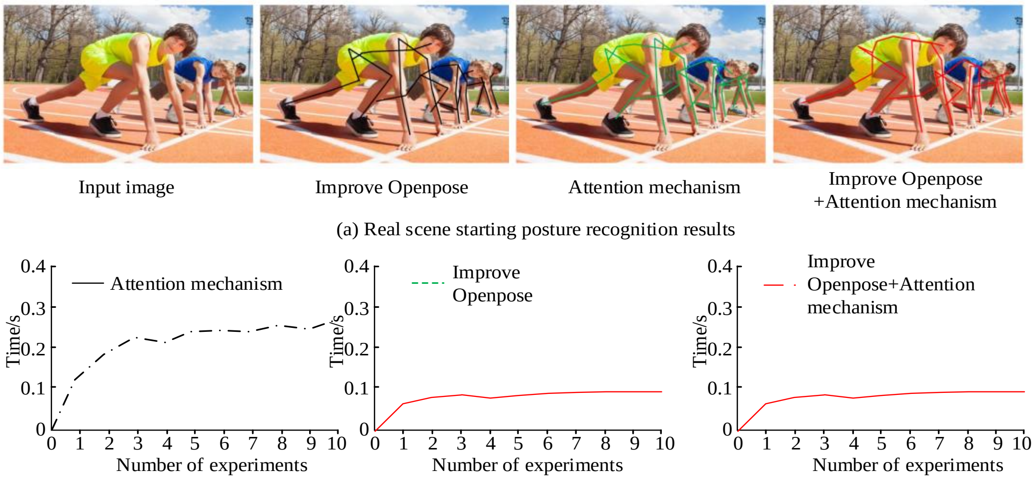


Figure 13: ROC results for different algorithms.

The Receiver Operating Characteristic Curve (ROC) is a tool used to evaluate the performance of binary classification models, mainly used to demonstrate the performance of classification models under different thresholds. The ROC curve is related to the true positive rate and false positive rate. It is a graph drawn with false positive rate as the horizontal axis and true positive rate as the vertical axis. Different threshold parameters can affect the performance of the model. To ensure the rationality and stability of the data results, a threshold of 0.5 is set. AUC (Area Under Curve) represents the area under the ROC curve. The AUC value is between 0 and 1, with a perfect classifier having an AUC of 1 and a purely random classifier having an AUC of 0.5. The larger the AUC value, the stronger the model's ability (sensitivity) to correctly identify true examples, and the lower the probability (1-specificity) of mislabeling negative examples as positive examples. The area under the ROC curve reflects the ability of different algorithms to distinguish classification samples. The closer the ROC curve is to the upper left corner, the higher the true rate can be achieved at a lower false positive rate, which means the model can more effectively distinguish between positive and negative classes, and the better the performance. The smoothness of the curve direction can also reflect whether it will be affected by abnormal data samples. Overall, the algorithm proposed in the study has the largest area under the curve, and the overall direction is smooth without obvious fluctuations. In Fig. 13, the area under the curve exhibited by the study's proposed algorithm is the largest and significantly better than the two comparative algorithms, and its ROC curve is smoother. The ROC curve of the

DTW model among them shows significant fluctuations. The fluctuation of the ROC curve is usually caused by the imbalance of the dataset (such as significant differences in the proportion of positive and negative samples) or the inconsistent performance of the model at different thresholds. The Dynamic Time Warping (DTW) algorithm is a technique used to measure the similarity between two time series of different lengths. The time series data involved in posture and action recognition is usually very complex, including the different movement patterns of athletes, the speed of movements, and the continuity of movements. When calculating the length of fine sequences, DTW may need to rely on specific constraints to improve efficiency for longer sequences, and this algorithm is more prone to biased results when processing attitude data due to imbalanced data distribution. The DTW algorithm relies on features in time series for matching and recognition. The accuracy of feature extraction, similarity between actions, parameter settings, data outliers, and other factors can all affect the performance of the algorithm, leading to fluctuations in the ROC curve. The key detection analysis of the research-proposed pose recognition algorithm is visualised in Fig. 14.



(a) Real scene starting posture recognition results

(b) Detection time change results

Figure 14: Posture recognition visualisation results and detection time.

In Fig. 14, the proposed fusion algorithm combines the human posture recognition with the attention module, which shows better recognition of the starting movement and the overall detection time is less than 0.1 s. The other algorithms show a certain joint error in the recognition results, and the detection and recognition time of the movement is basically in the range of 0.25~0.28. The above results show that the proposed fusion algorithm can stably analyse the starting posture movements and capture the subtle movements better.

5 Discussion

Human pose recognition techniques are widely used in human segmentation, video tracking, and sports action recognition and correction. And the attention mechanism is also commonly used in human action recognition, which usually unites the point cloud data of colour-depth images to identify human skeleton actions by spatio-temporal convolution. Specifically, scholars in the literature [6] achieved the creation of 3D point cloud data of human body with the help of colour-depth images and deep learning methods. The scholars in literature [7] take the help of multiscale location augmentation network for human body estimation. In literature [12] scholars use spatio-temporal convolutional attention for human skeleton work recognition and in literature [14] 2D and 3D action recognition models are fused. Unlike previous research, this study proposes a frame pose relationship correlation analysis based on the OpenPose human pose estimation recognition algorithm, which improves the accuracy of recognition through continuous action design and joint skeleton point repair. And based on the consideration of differences in the allocation of posture data information, a lightweight attention mechanism that focuses on time patterns is introduced to improve the accuracy of posture recognition. The experimental results show that the addition of improved attention mechanism results in lower error performance of the model in dataset processing, with abnormal loss improved from 0.8 to around 0.2, and the starting posture recognition accuracy is above 0.85. The proposed model performs well in pose classification and recognition, and the MAE error value (0.1406) shown in the ablation experiment results is smaller than that of the DTW model (0.1749) and HRNet-w48 model (0.1714). The accuracy of action recognition in references [22] and [23] does not exceed 85%, while the algorithm proposed in the study can improve recognition accuracy by obtaining different posture data information allocation. The starting motion is a fast and constantly changing process, and studying the inter frame information correlation of OpenPose posture estimation can effectively capture the posture changes of athletes at different time points during the starting process. Compared to DTW's fixed similarity matching and HRNet-w48's single frame prediction, inter frame information association provides richer contextual information and enhances the understanding of motion dynamics [24]. And the lightweight attention module and temporal attention mechanism designed in the research can reduce the interference of redundant information and improve the matching degree of posture features. In physical education teaching and training, Song Z and other scholars proposed a 3D skeleton keypoint detection algorithm based on a pose estimation model and visual background extractor. By analyzing continuous action information on a time series, the detection is performed, and the results show that the method has good pose estimation detection accuracy (over 90%) [25]. This approach is similar to studying attention mechanisms that consider temporal information, but it differs from research that focuses on decomposing and analyzing action postures. The detection efficiency of the research algorithm for motion recognition is 0.1 seconds. Although it performs better than other models, it still needs further improvement, including enriching the data selection of dynamic motion images and correcting the accuracy of detection analysis. At the same time, the adaptability and application stability of the algorithm under different hardware platform deployment content, resource requirements, and collaborative situations should also be analyzed.

6 Conclusion

During the running action, there are more errors in the starting action postures that the athletes are exposed to, such as the distance between the legs is too large, and the position of the centre of gravity of the body is too far back, which in turn will make the athletes fall or start errors and so on. The study proposes to combine the human posture recognition algorithm with the attention mechanism for the starting movement correction analysis, and conducts experimental analysis on it. The results show that the the fusion recognition algorithm proposed in the study exhibits recognition accuracies of 96.12%, 94.37% and 93.54% in the three phases of the starting action compared to the DTW algorithm and HRNet-w48, with the DTW model having accuracies of 90.05%, 88.12% and 87.14%, respectively. The results of the ablation experiments show that the research-proposed algorithm exhibits a MAE error value (0.1406) that is smaller than the DTW model (0.1749) and the HRNet-w48 model (0.1714), and its F1 value (0.6857) is larger than that of the DTW model (0.6214) and the HRNet-w48 model (0.6321). The algorithm proposed in the study provides better recognition of the starting movement, with an area under the ROC curve greater than the other comparative algorithms, and with a detection time (less than 0.1s) less than the detection values of the other two algorithms (0.25s to 0.28s). Enhancing the richness of data selection and corrective detection analysis of dynamic motion images are important improvements that the study needs to focus on in the future.

Data availability statement

All data generated or analysed during this study are included in this article.

Conflict of interest

The authors declare that they have no competing interests.

Funding statement

There is no funding in this article.

References

[1] Ma N, Wu Z, Cheung Y, Guo Y, Gao Y, Li J, Jiang B (2022). A survey of human action recognition and posture prediction. *Tsinghua Science and Technology*, 27(6), pp. 973-1001. https://doi.org/10.26599/TST.2021.9010068.

[2] Petz P, Eibensteiner F, Langer J (2021). Sensor shirt as universal platform for real-time monitoring of posture and movements for occupational health and ergonomics. *Procedia Computer Science*, 180, pp. 200-207. https://doi.org/10.1016/j.procs.2021.01.157.

[3] Li Y, Guo T, Liu X, Luo W, Xie W (2022). Action status based novel relative feature representations for interaction recognition. *Acta Electronica Sinica*, 31(1), pp. 168-180. https://doi.org/10.1049/cje.2020.00.088.

[4] Zhang J, Ye G, Tu Z, Qin Y, Qin Q, Zhang J, Liu J (2022). A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *Journal of Intelligent Technology*, 7(1), pp. 46-55. https://doi.org/10.1049/cit2.12012.

[5] Tani Y, Tasaki R, Terashima K (2019). Walking tracking system based on estimation of human posture in omni-directional mobile walker. *Journal of Meteorological Society of Japan*, 37(2), pp. 161-167. https://doi.org/10.17521/cjpe.2007.0050

[6] Le V (2023). Deep learning-based for human segmentation and tracking, 3D human pose estimation and action recognition on monocular video of MADS dataset. *Multimedia Tools and Applications*, 82(14), pp. 20771-20818. https://doi.org/10.1007/s11042-022-13921-w.

[7] Xu J, Liu W, Xing W, Wei X (2022). MSPENet: multi-scale adaptive fusion and position enhancement network for human pose estimation. *Visual Computing*, 39(5), pp. 2005-2019. https://doi.org/10.1007/s00371-022-02460-y.

[8] Kumar P, Chauhan S, Awasthi L (2022). Human pose estimation using deep learning: review, methodologies, progress and future research directions. *International Journal of Multimedia Information Retrieval*, 11(4), pp. 489-521. https://doi.org/10.1007/s13735-022-00261-6.

[9] Huang L, Liu G (2022). Functional motion detection based on artificial intelligence. *Journal of Supercomputing*, 78(3), pp. 4290-4329. https://doi.org/10.1007/s11227-021-04037-3.

[10] Wang X, Wu Y, Zhu L, Yang Y (2020). Symbiotic attention with privileged information for egocentric action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7), pp. 12249-12256. https://doi.org/10.1609/aaai.v34i07.6907.

[11] Zhou K, Wu T, Wang C, Wang J, Li C (2020). Skeleton based abnormal behavior recognition using spatio-temporal convolution and attention-based LSTM. *Procedia Computer Science*, 174(8), pp. 424-432. https://doi.org/10.1016/j.procs.2020.06.110.

[12] Li Y, Yuan J, Liu H (2021). Human skeleton-based action recognition algorithm based on spatiotemporal attention graph convolutional network model. *Journal of Computer Applications*, 41(7), pp. 1915-1921. https://doi.org/10.11772/j.issn.1001-9081.2020091515.

[13] Li X, Xie M, Zhang Y, Ding G, Tong W (2020). Dual attention convolutional network for action recognition. *IET Image Processing*, 14(6), pp. 1059-1065. https://doi.org/10.1049/iet-ipr.2019.0963.

[14] Zhang Y (2020). Fused behaviour recognition model based on attention mechanism. *Visual Computing in Industrial and Medical Arts*, 3(1), pp. 71-80. https://doi.org/10.1186/s42492-020-00045-x.

[15] Wu J, Xue Y, Meng X, Wan X (2021). Research on behaviour recognition algorithm based on SE-I3D-GRU network. *High Technology Letters*, 27(2), pp. 163-172. https://doi.org/10.3772/j.issn.1006-6748.2021.02.007.

[16] Wang Z (2021). Spatial-Temporal feature-based sports video classification. *International Journal of Applied Computer Science and Information Technology*, 12(4), pp. 79-97. https://doi.org/10.4018/IJACI.2021100105.

[17] Afza F, Khan M A, Sharif M, Kadry S, Damasevicius R (2021). A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. *Image and Vision Computing*, 106, pp. 1-20. https://doi.org/10.1016/j.imavis.2020.104090.

[18] Agethen S, Hsu W (2020). Deep multi-kernel convolutional LSTM networks and an attention-based mechanism for videos. *IEEE*, 22(3), pp. 819-829. https://doi.org/10.1109/TMM.2019.2932564.

[19] Liu J, Che Y (2021). Action recognition for sports video analysis using part-attention spatio-temporal graph convolutional network. *Journal of Electronic Imaging*, 30(3), pp. 33017.1-33017.16. https://doi.org/10.1117/1.JEI.30.3.033017.

[20] Fang L, Sun M (2021). Motion recognition technology of badminton players in sports video images. *Future Generation Computer Systems*, 124(9), pp. 381-389. https://doi.org/10.1016/j.future.2021.05.036.

[21] Chen Z (2022). Research on internet security situation awareness prediction technology based on improved RBF neural network algorithm. *Journal of Computer and Communications Engineering*, 1(3), pp. 103-108. https://doi.org/10.47852/bonviewJCCE149145205514.

[22] Cai X (2022). WSN-driven posture recognition and correction towards basketball exercise. *International Journal of Information System Modeling and Design*. https://doi.org/10.4018/IJISMD.300777.

[23] Fu D (2023). Research on intelligent recognition technology of gymnastics posture based on KNN fusion DTW algorithm based on sensor technology. *International Journal of Wireless and Mobile Computing: IJWMC*. https://doi.org/10.1504/ijwmc.2023.10057672.

[24] Le T, Huynh-Duc N, Nguyen C T, Tran M T (2023). Motion embedded images: an approach to capture spatial and temporal features for action recognition. *Informatica*, 47(3). https://doi.org/10.31449/inf.v47i3.4755.

[25] Song Z, Chen Z (2024). Sports action detection and counting algorithm based on pose estimation and its application in physical education teaching. *Informatica*, 48(10). https://doi.org/10.31449/inf.v48i10.5918.