Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(1): 1017–1037.
DOI: 10.3934/mbe.2024042
Received: 10 September 2023
Revised: 28 November 2023
Accepted: 04 December 2023
Published: 22 December 2023

Research article

A hybrid neural network-based intelligent body posture estimation system in sports scenes

Liguo Zhang[1], **Liangyu Zhao**[1] **and Yongtao Yan**[2,*]

[1] School of Physical Education, Shandong University, Jinan 250000, China

[2] Department of Physical Education, Shenzhen Polytechnic, Shenzhen 518055, China

* **Correspondence:** Email: sunbirdshang@163.com.

Abstract: Body posture estimation has been a hot branch in the field of computer vision. This work focuses on one of its typical applications: recognition of various body postures in sports scenes. Existing technical methods were mostly established on the basis of convolution neural network (CNN) structures, due to their strong visual information sensing ability. However, sports scenes are highly dynamic, and many valuable contextual features can be extracted from multimedia frame sequences. To handle the current challenge, this paper proposes a hybrid neural network-based intelligent body posture estimation system for sports scenes. Specifically, a CNN unit and a long short-term memory (LSTM) unit are employed as the backbone network in order to extract key-point information and temporal information from video frames, respectively. Then, a semi-supervised learning-based computing framework is developed to output estimation results. It can make training procedures using limited labeled samples. Finally, through extensive experiments, it is proved that the proposed body posture estimation method in this paper can achieve proper estimation effect in real-world frame samples of sports scenes.

Keywords: hybrid neural network; body posture estimation; intelligent systems; key-point extraction

1. Introduction

Multimedia feature extraction technology occupies an important position in computer vision tasks [1]. It is the focus on the difficulty of multimedia feature extraction technology to accurately describe, extract and analyze multimedia image features [2]. The multimedia feature extraction technology based on the deep learning method surpasses the traditional algorithm and becomes the mainstream method of the current multimedia data processing [3]. It is widely used in knowledge extraction, pose estimation, gait estimation, motion analysis and other tasks [4]. And it has potential value in sports effect evaluation and posture correction [5].

In sports effect evaluation and posture estimation and correction tasks, traditional methods mainly rely on human operations [6]. This approach requires a lot of manpower and material resources, and the experimental results are easily affected by human beings [7, 8]. With the continuous improvement and development of deep learning technology, it gradually replaces traditional methods and has achieved major breakthroughs on public datasets [9]. Therefore, in our paper, we adopt the 3D human pose estimation algorithm based on deep learning as the core method of multimedia feature extraction technology [10]. It is used to extract the movement trajectory and the 3D pose of the individuals in the video [11]. It is also used to analyze the effect and performance of sports through the extracted features and the effect of posture correction [12].

To analyze the effect evaluation and posture correction of sports, we propose a multimedia feature extraction technology based on deep learning. This can effectively estimate the three-dimensional pose and motion trajectory of the human body [13, 14]. First, we adopt the 3D pose estimation network model as the backbone network model [15, 16]. It is mainly composed of LSTM model and CNN feature extraction model, which are used to extract the temporal information and key-point feature information of the video, respectively [17]. Second, we adopt the framework of semi-supervised learning [18]. It uses the 2D mapping model trained on labeled data of the 3D pose for supervision, and continuously improves the model trained on unlabeled data to improve the generalization ability of the network model [19]. Finally, extensive experiments demonstrate that our method has a broad impact on sports performance evaluation and posture correction tasks [20, 21].

2. Related work

2.1. Multimedia feature extraction

The human eye can quickly sense two-dimensional images through previous knowledge and learning experience, thereby realizing image feature extraction [22]. With the continuous development of society, the continuous change of science and technology, and the continuous development of social networks, video has progressively increased to a huge volume [23]. Extracting feature information is a difficult task [24].

To process these data, multimedia feature extraction technology was born, which is widely used in image data processing tasks [25]. The multimedia feature extraction method based on the traditional algorithm has a clear principle and a reasonable design structure, which can meet people's needs for processing a small amount of data [26]. However, with the continuous popularization of big data technology, the data generated every day is in units of billions of records, and it is difficult for traditional algorithms to meet the demand [27]. At present, people's requirements for image data are constantly improving [28]. In the traditional image feature extraction algorithm, the recognition accuracy does not meet the requirements, the recognition time is long and the recognition results are quite different from the actual ones [29]. Therefore, the current image feature extraction method based on the deep learning method has gradually become the mainstream, which can effectively improve the accuracy of image feature extraction [30]. Oladipo et al. [31] developed a new age estimation system by combining genetic algorithms with back propagation-trained artificial neural networks (ANNs) and using local binary pattern feature extraction technology (LBGANN) for black faces. Gasmi et al. [32] proposed an optimal deep neural network model based on adaptive optimization algorithms, which takes medical images and natural language problems as inputs and provides accurate answers as outputs. Our model

first classifies medical problems according to the embedding stage.

Based on traditional computer image processing algorithms, combined with the information distribution of features in multimedia images, traditional multimedia image feature extraction methods can extract various image features [33]. The method adopts multi-scale feature analysis tools to transform the image data inputted by multimedia, and performs three-dimensional optimal approximation to the target features of the image, to facilitate the acquisition of various types and information in the image [34]. The traditional feature extraction method first classifies the multimedia image data by layered classification, filters out all the non-identified image windows under the condition of retaining the image window, and filters most of the non-identified areas on the base of ensuring sufficient detection rate [35]. Secondly, after the traditional multimedia feature extraction method completes the shearing transformation of the original 3D multimedia visual image, the sub-image is acquired, and the multi-directional local binary mode is used to obtain a circular area on the acquired sheared transformed image, and find the corresponding N feature points.

Based on completing image feature preprocessing, the multimedia feature extraction technology based on deep learning can perform feature extraction on images and subsequent classification or other tasks. In this process, deep learning methods can be used to generate corresponding sparse representations of the provided human pose motions, and an intelligent artificial neural network model is constructed. And it uses gradient descent to project the minimum value of the image objective function, and outputs the optimized minimum value. Finally, according to the gradient direction of the function, the optimization of the network model and the identification of features are carried out. Generally, deep learning technology relies on data and models, and needs to train a large amount of relevant data and design a better network model to achieve the required goals.

2.2. Sports effect evaluation

Adolescence is the most rapid stage of human development and learning awareness. There are various types of sports, including long-distance running, long jump and rope skipping. Cultivating the habit of sports and developing a healthy body during this stage is crucial for their overall well-being. However, current sports lack reasonable evaluation methods and evaluation standards, and the evaluation is usually performed manually, which is easily disturbed by human factors. Therefore, we have carried out research on sports effect evaluation, which has important theoretical significance for improving the research content and proposing constructive strategies to enrich the reform and innovation of sports models.

Youth sports contribute to the improvement of professional physical fitness and provide a strong guarantee for the development of society. As the basic sports quality of the human body, endurance can measure people's professional work ability and physical condition. Having excellent endurance is the foundation of maintaining a good working condition, and improving the endurance of young people can help improve their professional ability. Sports can effectively improve the human body's endurance. Being in the state of confrontation, running and chasing, especially for periodic sports, can improve the human body's ability to control rhythm. Sports cultivate people's coordination abilities and provide guarantee for the healthy development of society. At the same time, integrating professional skills with physical education teaching will not only enhance the dynamic nature of physical education but also contribute to the enhancement of human quality of life through sports.

Sports help to form a good attitude towards life. Social competition is intensifying day by day,

which requires people to have a stronger sense of competition and awareness of the overall situation. By participating in sports learning, sports competitions, etc., people can fully develop their potential and fully demonstrate their abilities. Participating in sports and competitions can not only improve people's awareness of competition, but also cultivate people's fighting spirit. Sports help to cultivate people's organizational ability and promote the improvement of the public's professional ability. By participating in sports, people can cultivate the consciousness of collectivism, properly handle the relationship between the individual and the collective, between themselves and others, and enhance their organization and discipline. Generally, people's organizational management ability can be improved by participating in sports learning and sports training, thereby improving the communication ability between people.

Although active participation in sports can effectively improve people's professional ability and social communication methods, there is still a lack of effective methods in the current sports effect evaluation. Generally, there are corresponding rules and requirements for participating in physical education classes, sports events and sports competitions. By actively participating in sports, these rules can be learned and behaviors can be restrained through these rules. Those who violate these rules will be punished accordingly. These rules can effectively improve people's control ability and provide a good guarantee for the future. At the same time, our investigation found that the current evaluation methods for exercise effects are relatively lacking, so an active and effective exercise effect evaluation method is urgently needed to avoid the interference of human factors in the actual process.

In the field of sports science, there has been considerable research on methods for evaluating sports effectiveness. However, although there have been many studies dedicated to evaluating the effects of various sports, there are still some research gaps. For example, most existing studies mainly focus on the effectiveness of specific sports or training programs, with less attention paid to comparative studies between different sports programs. With the continuous development of sports science and health research, more and more researchers are paying attention to comparative research between different sports events. This comparative study can help us better understand the characteristics and effectiveness of different sports, and provide more comprehensive sports advice for sports enthusiasts. In studies comparing different sports, researchers usually choose common sports such as running, swimming, cycling, yoga, etc. They may compare the weight loss effects, degree of improvement in cardiovascular health and degree of muscle strength increase of these programs. Through these comparative studies, we can better understand the advantages and disadvantages of different sports and provide more scientific sports advice for sports enthusiasts [36].

In addition, existing research often only focuses on the short-term effects of exercise, while neglecting the impact of long-term exercise on individual health and development. In terms of effectiveness evaluation methods, many studies use traditional quantitative evaluation methods, such as performance tests and questionnaire surveys. Motion estimation is the process of estimating the motion of an object in a video sequence [37]. In our method, we use an optical flow-based method to estimate hand motion. Optical flow is a 2D vector field that describes the motion of objects in an image between consecutive frames. We can estimate the motion of the hand in the video sequence by calculating the optical flow [38].

However, the effectiveness and reliability of these methods are often questioned. For example, questionnaire surveys may be influenced by the subjective biases of respondents, while performance tests may be influenced by the testing environment and other factors. Therefore, we need more com-

prehensive and reliable evaluation methods to improve the accuracy and credibility of our research. In addition, existing research often overlooks the impact of exercise on mental health. Although the effects of exercise on physical health have been widely recognized, the positive effects of exercise on mental health have received less attention and research. This is undoubtedly an important research gap that requires further research and exploration.

2.3. Review and evaluation of existing posture correction methods

Poor sitting posture of the human body and prolonged sitting have a serious impact on people's physical and mental health. However, with the continuous progress of society and the continuous development of science and technology, people often cause more and more staff to sit for a long time due to work, and long-term poor sitting posture leads to more and more people suffering from physical diseases. At the same time, this issue also occurs in people who do not follow the level. The elderly and children are not immune. Young people are often in high-intensity work and often stay up late to work overtime and study. At the same time, without scientific sitting posture training and rehabilitation, it is difficult for patients to maintain or restore a good sitting posture, which eventually leads to the spread of diseases, such as for the human spine. How to scientifically and effectively maintain a good human body movement and prevent sedentary and incorrect body posture is particularly important.

Gesture recognition is the process of recognizing and understanding human gestures through computer vision technology. In our method, we use deep learning techniques to recognize gestures. Specifically, we use convolutional neural networks (CNN) to extract image features, and then use recurrent neural networks (RNN) to process temporal data to recognize gestures [39]. Therefore, it is necessary to carry out healthy training of limb functions for patients, and patients need to exercise limbs under the guidance of doctors. The functional recovery of the human body after limb injury is also very dependent on human movement. Through the movement of the human body, the damaged function of the human body can not only be restored, but also the patient's anxiety can be relieved by tender consolidation. Performing correct body posture exercises under the guidance of a doctor can improve the function of the human body and reduce the time and energy spent on body repair. Under normal circumstances, it is difficult for doctors to take care of many patients. Patients need to supervise themselves and carry out rehabilitation training. Moreover, problems are prone to occur in the process of rehabilitation training. There is an urgent need for a method that can truly provide comprehensive and correct guidance.

Human pose recognition has wide applications in fields such as healthcare, sports science and virtual reality. In order to achieve accurate and real-time pose recognition, researchers are constantly exploring new sensors and related technologies. Among them, the flexible integrated PAN/PVDF piezoelectric sensor, as an advanced sensing technology, has been widely used in the field of human pose recognition due to its good flexibility and sensitivity. This article will explore the synergistic enhancement characteristics of this flexible integrated PAN/PVDF piezoelectric sensor [40]. In the field of human posture correction, although there have been many studies dedicated to developing and applying various posture correction methods, there are still some research gaps. These research gaps are mainly manifested in the following aspects:

First, most of the existing posture correction methods are based on traditional physical therapy concepts and techniques, such as manual therapy, exercise therapy, etc. However, the effectiveness of these methods in correcting certain specific diseases or physical conditions has not been fully validated

and studied. For example, for certain neurological diseases or congenital body deformities, traditional physical therapy may not achieve significant corrective effects, which requires us to develop new and more targeted posture correction methods. Second, existing research often only focuses on static and instantaneous posture correction effects, while neglecting dynamic and long-term posture correction processes. In fact, human posture is a dynamic process that requires comprehensive evaluation and research from multiple time points and perspectives. Therefore, we need to strengthen research and evaluation of the effectiveness of long-term posture correction. Finally, existing research often only focuses on external correction methods of the body, while neglecting the influence of internal factors on posture. In fact, many internal factors of the body, such as muscle strength and neural control, can affect the posture and balance of the human body. Therefore, we need to conduct a more comprehensive study and explore the impact of internal factors on posture, and develop corresponding corrective methods.

In summary, although existing posture correction methods have achieved certain results, there are still some obvious research gaps. These research gaps provide us with space and direction for further research, which helps us to have a more comprehensive understanding and evaluation of the actual effects and application value of posture correction.

To achieve the task of correcting human posture motion, many predecessors have proposed various human posture correction techniques. But these methods consume a lot of manpower and material resources, and the effect of posture correction is difficult to achieve. Therefore, we use a deep learning method to achieve the goal of human posture correction. The method in our paper avoids the problem of human intervention and introduces some relevant quantitative indicators to provide reliable data support.

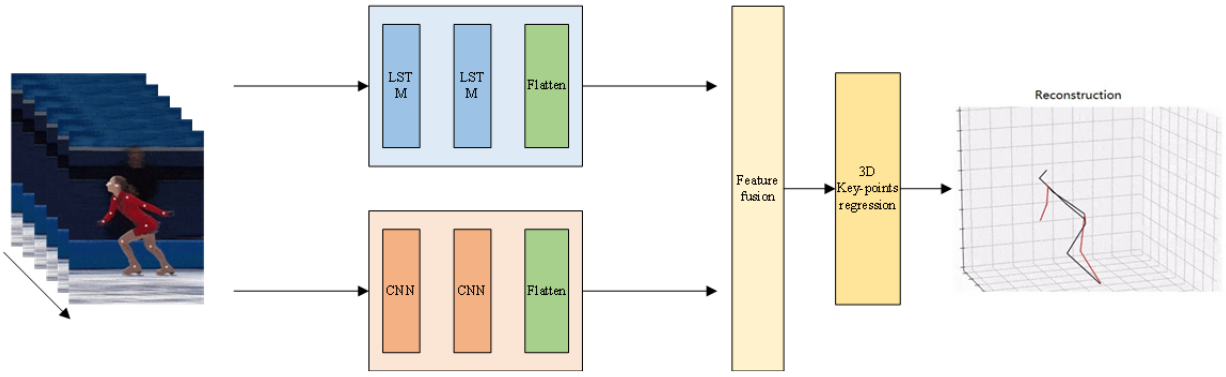

Figure 1. Structure diagram of the pose estimation network model.

3. Methodology

3.1. Overview

In order to realize the extraction of multimedia features, we propose a three-dimensional pose estimation network model, which can be effectively applied in sports effect evaluation and posture correction tasks. First, we propose a human pose estimation network model, which is mainly composed of LSTM units and CNN convolution units. The LSTM unit mainly learns the relationship between the image sequences in the video, while the CNN convolution unit mainly learns the key point features of the human body in each frame. Second, we focus on the structure and constituent elements of the

CNN convolution module. We adopted a standard LSTM unit to learn time series features. Finally, we propose a semi-supervised learning framework. We use the labeled 2D human pose key points to map to the 3D space to obtain 3D human key points, and perform supervised learning through the motion trajectory and human pose in the 3D space. At the same time, a trained label key-point prediction model is used to supervise 3D key-point learning without labels.

The overall framework of the hybrid neural network-based motion scene intelligent agent pose estimation model can be divided into two main parts: the pose estimation module and the self supervision module.

Input: The input of the model is image data from a moving scene.

Pose estimation module: This module is responsible for extracting key points from input images and mapping them into three-dimensional space. This module consists of one or more neural networks, including convolutional layers, pooling layers and fully connected layers. The output of this module is the 3D attitude estimation result.

Self-supervision module: This module is used to supervise the learning process of the pose estimation module. It calculates a loss function by comparing the predicted key point positions of the model with the actual annotated key point positions, and uses this loss function to optimize the parameters of the pose estimation module.

Loss function: The self-supervised module calculates the loss function based on the output of the pose estimation module and the actual annotated key point positions. Common loss functions include mean square error loss function, cross entropy loss function, etc.

Optimization algorithm: The model uses optimization algorithms to minimize the loss function and update the parameters of the pose estimation module. Common optimization algorithms include gradient descent algorithm, Adam algorithm, etc.

Output: The output of the model is the optimized 3D pose estimation result. Through this overall framework diagram, we can clearly see the relationship and interaction between pose estimation and self-monitoring modules. The self-supervised module supervises the learning process of the pose estimation module and optimizes the model parameters by calculating the loss function and optimizing algorithms, thereby improving the model's performance and generalization ability.

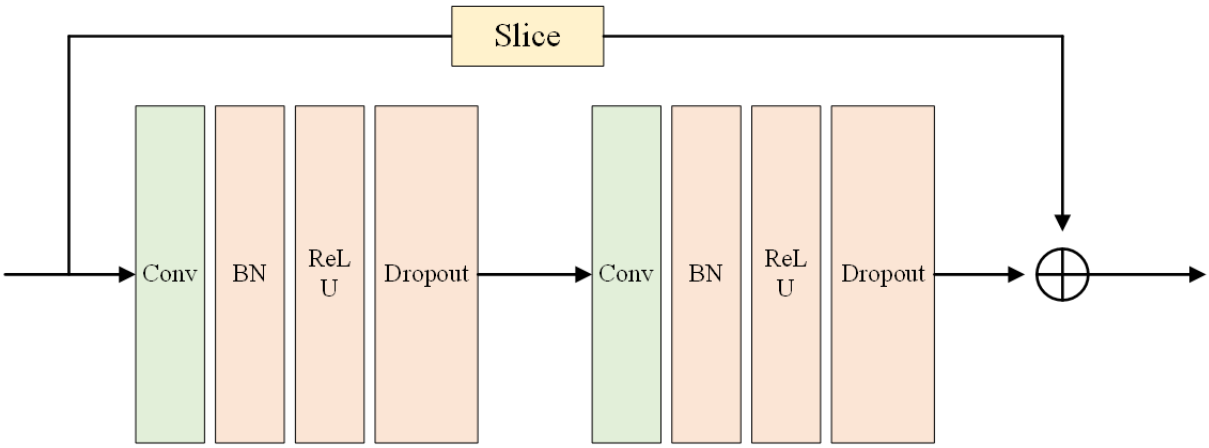

Figure 2. Structure diagram of convolution module.

3.2. 3D pose estimation network model

In order to learn the pose features of the human body from the multimedia video data and avoid the influence of factors such as perspective, and clothing on the 3D human pose estimation, we use a deep learning-based method to extract the human pose features in the video. This paper introduces long short-term memory network and convolutional neural network. Long short-term memory network is a temporal recurrent neural network. Its memory unit and gating mechanism make it have excellent performance in learning the temporal dependencies of long-sequence video data. It is suitable for learning the human key-point features of time series in this paper. Convolutional neural networks can extract local related features of data layer by layer through local connections, weight sharing, pooling mechanisms, etc. Since the three-dimensional human pose motion has local region correlation, it can be extracted with a convolutional neural network model. At the same time, the 3D human pose motion also has the characteristics of time series, so the LSTM unit is used to process the 3D human pose with time series.

The structure of the pose estimation network model is shown in Figure 1. It is mainly composed of an LSTM memory unit and a CNN convolutional layer. It forms specific constraints for the estimated 3D human pose, making the output human pose more reasonable. Firstly, the two-layer long and short-term memory network model is composed of two layers of LSTM units. The input image features are converted into a one-dimensional motion constraint vector through the Flatten layer after passing through the LSTM unit to limit the range of three-dimensional human motion. The convolutional neural network model is mainly composed of 4 convolutional layers, 4 pooling layers and 1 Flatten layer. After the input video features are extracted by the convolutional layer for key-point features, they are converted into a one-dimensional feature vector in the Flatten layer.

Finally, the features extracted from the two network models are fused, and the three-dimensional human motion skeleton is obtained after the 3D human body key point regressor. The three-dimensional human motion stock can be used for subsequent sports performance evaluation and posture correction and other related tasks. The pose similarity of human motion is high, and other shooting conditions and walking conditions will affect the motion of human pose. This paper uses the Softmax loss function and the center function to optimize the network model with multiple loss functions:

$$L = L_s + \lambda L_c, \tag{3.1}$$

$$L_s = -\sum_{i=1}^{m} \log \frac{e^{w_{y_i}^T x_i + b y_i}}{\sum_j^n e^{e^{W_{y_i}^T x_i + b_{y_i}}}}, \tag{3.2}$$

$$L_c = \frac{1}{2} \sum_{i=1}^{m} \left\| x_i - c_{y_i} \right\|_2^2. \tag{3.3}$$

Among them, *log* represents the natural logarithm with e as the base, L_s represents the model output, m represents each label, and only the label with the correct solution in e^w is 1, while the others are all 0. Therefore, cross entropy only calculates the corresponding "correct solution". It can be seen that the value of cross entropy error is determined by the output result corresponding to the correct label.

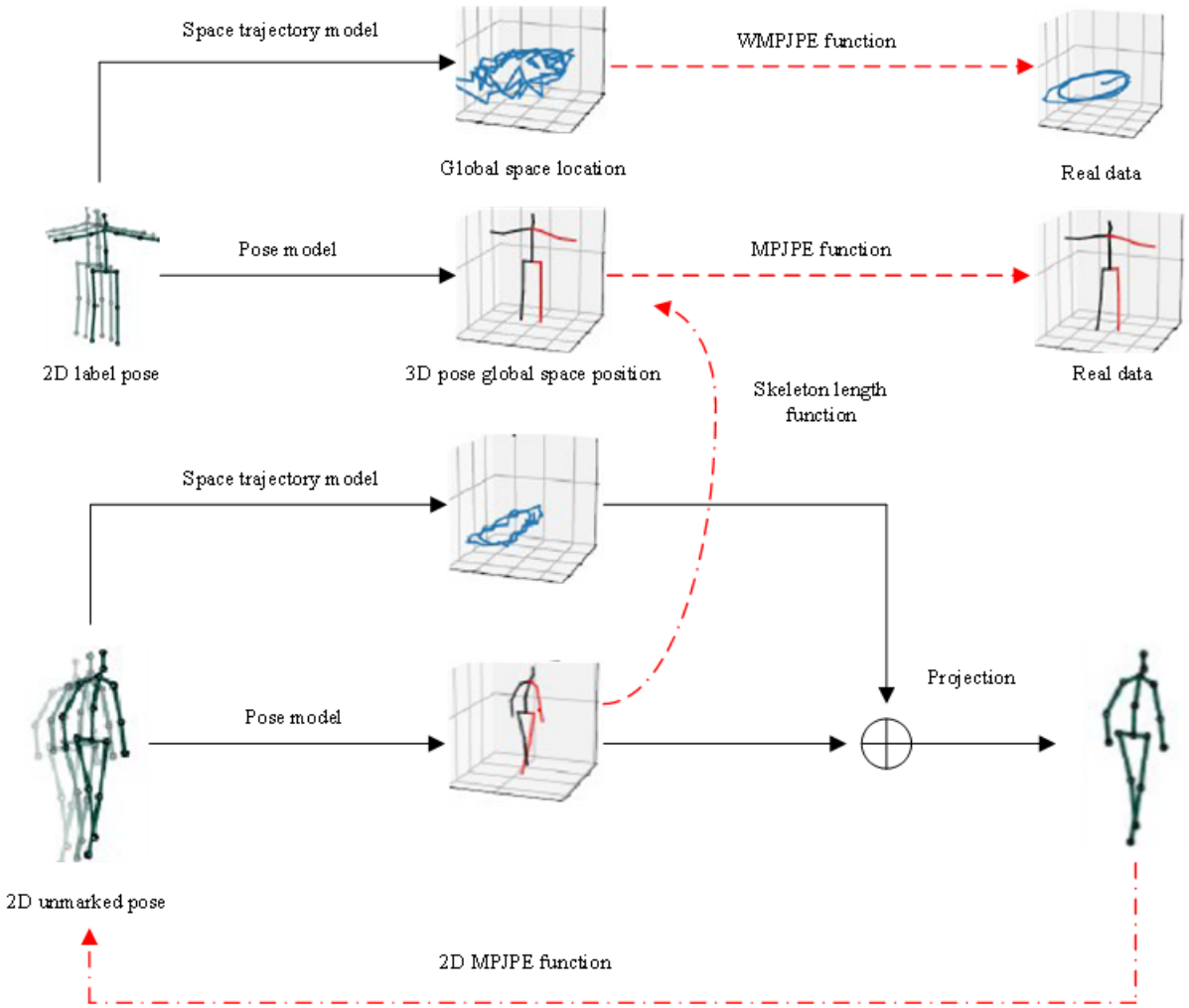

Figure 3. Framework diagram of semi-supervised learning method.

The key to LSTM is to control the flow of information through gating mechanisms. Specifically, LSTM uses three gates: forget gate, input gate and output gate. These gates are determined by a series of learnable weight parameters. First, we define the output of the forgetting gate as I_t, define the candidate value for memory update as F_t, and define the output of the output gate as O_t. The calculation process is as follows:

$$I_t = \sigma\left(X_t W_{xi} + H_{t-1} W_{hi} + b_i\right), \tag{3.4}$$

$$F_t = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right), \tag{3.5}$$

$$O_t = \sigma\left(X_t W_{xo} + H_{t-1} W_{ho} + b_o\right). \tag{3.6}$$

Among them, W_f, W_i, W_c, W_o are weight matrices, bf, bi, bc, bo are bias terms. $[h_{t-1}, x_t]$ indicates that the hidden state t-1 and input x_t are concatenated horizontally. $\sigma(\cdot)$ represents the sigmoid function. When training a general neural network model, $S = f(W_T \cdot X + b)$ is usually used, where W is the weight, X is the input and b is a constant. Assume we have input data $X = (X_1 + X_2 + X_3 + ...X_t)$ for a time series, where x_t represents the input of the t-th time step. At the same time, LSTM also has a set of internal states, including memory unit F_t and hidden state O_t.

3.3. CNN feature extraction module

In this paper, the convolutional neural network model is mainly used to process the image data of each frame in the video, and the position and motion status of the key points of the human body are learned from it. The convolutional neural network in this paper is mainly composed of 4 convolutional layers, 4 pooling layers and Flatten layers. The structure of the convolution module is shown in Figure 2, in which the convolution layer, the pooling layer, the ReLU activation function and the Dropout layer constitute the convolution module. These convolution modules are arranged according to a specific experimental sequence to obtain the convolutional neural network structure.

In these cases, the term "experimental sequence" refers to the arrangement and combination of convolutional modules in the design and construction process of convolutional neural networks (CNN). This arrangement and combination is based on specific task requirements and goals to optimize network performance and solve specific machine learning problems. The process of determining and selecting a CNN network structure typically includes the following steps:

- First, it is necessary to clarify what the problem is to be solved. This may involve various tasks such as image classification, object detection, speech recognition and natural language processing.
- According to the defined problem, corresponding datasets need to be collected. The size and complexity of the dataset also affect the design of the network structure.
- Network design is a key step in determining the network structure. At this stage, the designer will design a suitable network structure based on the nature of the problem, the characteristics of the dataset and existing knowledge and experience.
- The designed network structure needs to be verified through experiments.
- The network structure is designed and optimized based on experimental results to improve its performance.

Strude is the step size, which is the sliding interval. The calculation method for convolutional layers is sliding window dot multiplication, so the effect of step size on sliding windows is still easy to understand, and is also calculated as 5×5. For example, in the convolution operation shown in Figure 2, with a step size of $s = 1$, there is no interval. When $s = 2$, when the convolution window moves, there needs to be an interval of 1 lattice. In this way, when $s = 2$, the scale of the convolutional layer output is changed from 3×3 and has become 2×2. Finally, when the feature input size is $H \times H$ and the convolution kernel parameters are F (convolution kernel size), S (stripe), P (padding), the output feature size is:

$$N = \frac{W - F + 2P}{S} + 1, \tag{3.7}$$

$$k' = d \times (k - 1) + 1, \tag{3.8}$$

$$o = (i - 1) \times s + k - 2p, \tag{3.9}$$

where N represents the output feature map size, k represents the dilated kernel size and o represents the deconvolution output size.

3.4. Semi-supervised learning method framework

The semi-supervised learning method framework proposed in this paper mainly relies on the basic convolutional neural network model, which needs to predict the two-dimensional key-points of the

input video. Then it will be projected to the three-dimensional space through the network model, and finally obtain the three-dimensional human pose estimation. In the two-dimensional key-point prediction method, mature two-dimensional human pose estimation methods such as HRNet and HigherHRNet with relatively high accuracy can be used to obtain the positions of human key points in the two-dimensional image. At the same time, the coordinates of the human body key-points with labels are used for supervised learning, and these key-points need to be accurately marked on the two-dimensional image.

The semi-supervised learning method group is mainly used to improve the mapping relationship of the neural network model, so part of the real 2D key-point detection data and the corresponding 3D estimation model need to be used for pre-training in the whole experiment. The structure of the semi-supervised learning method is shown in Figure 3. The upper part represents supervised training and the lower part represents unsupervised training. The real 2D label data is used in supervised training, and the trained model is used for supervision in semi-supervised training. In the whole semi-supervised training process, a spatial trajectory model and a pose estimation model are introduced, and the distance between the parameters learned by these models and the real label data is used as a method to measure the learning effect of the model.

For the calculation of the center of the knee and ankle joints, we often need to construct virtual tracking markers to help predict the center of the knee and ankle joints. We have three constraint conditions for calculating the constraint points of human joints.

1) The line connecting the known joint center and the anatomical tracking marker point on the outer side of the current joint should be perpendicular to each other.

2) It must be located in the plane determined by the other three tracking markers of the lower limb (including the known joint center).

3) The distance between the anatomical tracking marker points on the outer side of the current joint is the diameter of the joint. Based on the above three constraints, three position coordinate points can be obtained.

$$R^m_{\text{joint}} = \alpha = \arccos\left(\frac{\overline{N_{k_s}N_{k_1}} \cdot \bar{b}}{\left|\overline{N_{k_S}N_{k_t}}\right|}\right), \tag{3.10}$$

$$R^m_{\text{form}} = \frac{z_0}{\left[(x_0 - x_{k_0})^2 + (y_0 - y_{k_0})^2\right]^{\frac{1}{2}}}, \tag{3.11}$$

$$M_r = \left[R^m_{\text{joint}}, R^m_{\text{root}}, R^m_{\text{gra}}, R^m_{\text{step}}, R^m_{\text{forw}}\right]^T, \tag{3.12}$$

where joint represents the knee key point constraint, root represents the glue part key point constraint, gra represents the center of gravity deviation constraint, step represents the pedestrian stride information constraint, forw represents the human body structure motion constraint and M represents the human body motion constraint matrix, which maintains the joint points Timing features of motion constraints.

4. Experimental results and analysis

In this article, in order to train and test a hybrid neural network-based intelligent agent pose estimation model for motion scenes, the following steps were used to prepare training and testing data:

Image data was collected from various motion scenes, including images of different people under different postures, clothing, lighting and occlusion conditions. For each image, this paper manually marked human key points such as hips, shoulders, elbows, knees and ankles. These marker points will be used as target key points in training and testing data.

This article collected a total of 80 sets of data for testing and analysis. For these data, it is first necessary to shuffle or randomize the overall dataset to ensure the randomness and unbiased partitioning. This step can ensure the uniformity of the distribution of each sample in the training and testing sets, avoiding model bias caused by sample selection. After completing data shuffling or randomization, it is necessary to determine the partition ratio between the training and testing sets. Usually, the number of training sets accounts for the majority of the overall dataset, such as 60 to 80%. The remaining data is used for the test set to evaluate the model's generalization ability. The specific partition ratio can be adjusted based on the size of the dataset and the needs of model training. After determining the partition ratio, different methods can be used to divide the dataset into training and testing sets. A common method is to use stratified sampling, which involves layering based on certain attributes of the samples (such as motion type, difficulty, etc.), and then extracting a certain proportion of samples from each level as a training or testing set. Another method is direct random partitioning, which involves randomly selecting a certain number of samples from the overall dataset as training and testing sets.

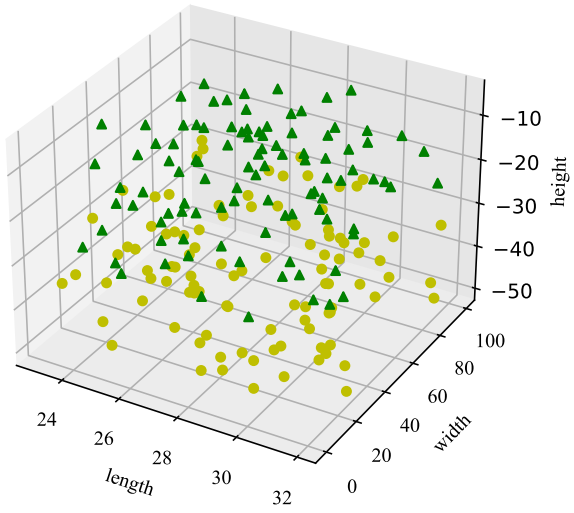

Figure 4. Sampling diagram of human motion trajectory.

In order to map 2D key points into 3D space, the authors used camera calibration to obtain internal parameters of the camera (such as focal length, optical center and lens distortion). These parameters are used to convert key points in 2D images into coordinates in 3D space. The authors divided the collected image data into a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate the performance of the model. The author uses a hybrid neural network to train the model, which can automatically extract key points from images. The author used a large amount of image data to train the model and used techniques such as cross validation to optimize

the model parameters. On the test set, the author evaluated the performance of the model and used standard attitude estimation metrics to measure its performance. This article collected a large number of human posture images from public datasets in academia and industry. These datasets include images of various human postures, clothing, lighting and occlusion conditions, providing rich training samples for our model. In order to better adapt to our model, we also collected some custom image data in our own experimental environment. We create this data by capturing videos in various sports scenes and manually labeling human key points.

Figure 4 shows the location sampling of the trajectory of a specific task motion in three-dimensional space. The green triangle represents the head of the human body, and the yellow circle represents the buttocks of the human body. The movement trajectory of the human body can be roughly obtained through the movement of these two types of human key points. The length, width and height in the figure represent the size of the sampling space, and negative values indicate downward. This sampling space is relatively dense, so that the movement of the human body in the three-dimensional space can be clearly analyzed.

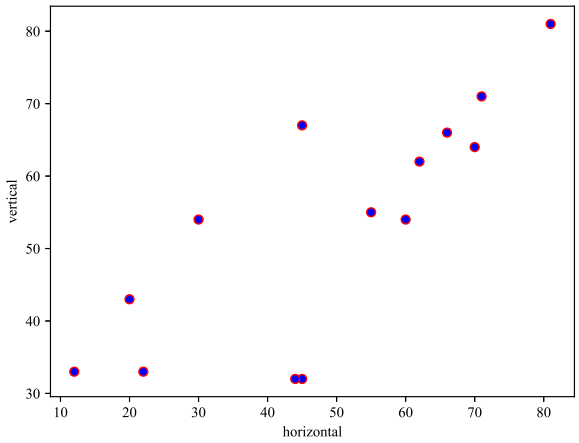

Figure 5. Human body posture correction diagram.

Figure 5 shows a rectification diagram of the human body posture, wherein the coordinate system represents the image coordinates of the three-dimensional space mapped to the two-dimensional space. Judging from the movement of the key points of the human body, the two-dimensional spatial positions of the corresponding key points of the head and buttocks are in a straight line distribution, which can be expressed in the form of $y = ax + b$. From Figure 5, we found that the positions of these key points do not show a linear distribution, and there is obviously a turning point in the middle, so it is necessary to correct the specific character pose.

When training the model, a large amount of image data is required to train the model. These images can come from different sports scenes. During the training process, common optimization algorithms are used to optimize the parameters of the model. In addition, techniques such as cross validation can be used to prevent overfitting and improve the generalization ability of the model. Hyperparameters refer to parameters that need to be manually set during the training process, such as learning rate, batch size, number of training rounds, etc. The selection of these parameters will have a significant impact on the performance of the model. To find the optimal hyperparameter combination, some common grid

hyperparameter search methods are used. These methods can help us find the optimal combination of hyperparameters, thereby improving the performance and generalization ability of the model.

When training and testing motion scene intelligent agent attitude estimation models based on hybrid neural networks, hyperparameter tuning is an important step, as it can affect the performance and generalization ability of the model. During hyperparameter tuning, the hyperparameters that need to be adjusted and their search space may include the following aspects:

1) Learning rate: Learning rate is the parameter used to update weights during model training. It controls the learning speed of the model in each iteration. The search space for learning rate can be set to a range containing multiple different values, selecting values between 0.001 and 0.1.

2) Batch size: Batch size is the number of samples used in each training iteration. It will affect the convergence speed and training stability of the model. Usually, the search space for batch size can be set to a range containing multiple different values, and this article uses values between 32 and 256.

3) Optimizer: The optimizer is an algorithm used to update model weights. Common optimizers include random gradient descent (SGD), Adam, and others. During hyperparameter tuning, it is necessary to select a suitable optimizer and adjust its relevant parameters, such as learning rate, momentum, etc.

4) Loss function: The loss function is a function used to measure the difference between the predicted results of a model and the actual results. During hyperparameter tuning, it is necessary to select a suitable loss function and adjust its parameters, regularization coefficients, etc.

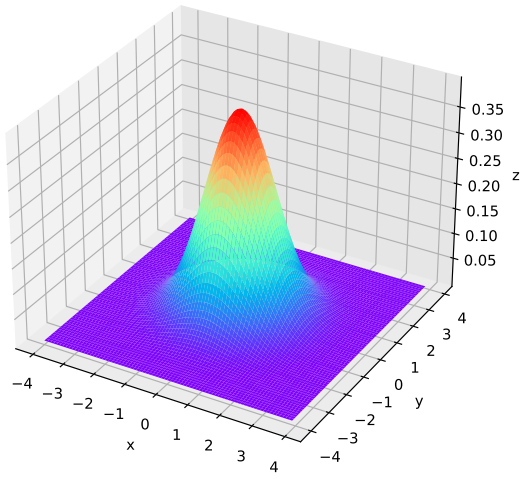

Figure 6. Heatmap of key points.

Figure 6 shows a heatmap of the total human body key points in a 2D image. In this paper, the trained two-dimensional pose estimation model needs to be used to extract the key points of the image, and then map the two-dimensional key points into the three-dimensional space. In this method, the first step is to construct a two-dimensional human body template that represents the contour of the human body in a specific posture. Then, by matching the input image with the template, the most similar pose is found to estimate the posture of the human body. The advantage of this method is its simplicity and ease of use, but its disadvantage is its low accuracy and sensitivity to interference from factors such as lighting and clothing. In order to improve the accuracy of attitude estimation,

traditional methods include models based on hybrid neural networks. This model combines convolutional neural networks (CNN) and recurrent neural networks (RNN) to extract image features using CNN, and process temporal data using RNN to obtain more accurate human pose estimation results. Therefore, the position of the two-dimensional key points plays a decisive role in the prediction of the three-dimensional space key points, so it is necessary to ensure that the predicted two-dimensional key points are accurate enough to avoid mapping to the wrong spatial positions. The bottom axes x and y of the figure represent the 2D image surface, and the z represents the confidence of the predicted key points.

Figure 7 shows the optimization of the human pose estimation model proposed in this paper. In order to evaluate the performance of a hybrid neural network-based intelligent pose estimation model for motion scenes, the author used accuracy metrics. Accuracy refers to the ratio of the number of samples correctly predicted by the model to the total number of samples. In pose estimation tasks, accuracy can be used to measure the model's recognition ability for different poses. The bottom coordinate system represents the learning rate and step size setting of the influencing factors of the network model, and the negative values represents the multiple of dividing by the setting. From the figure, we can see that setting a smaller learning rate can effectively improve the overall performance of the network model.

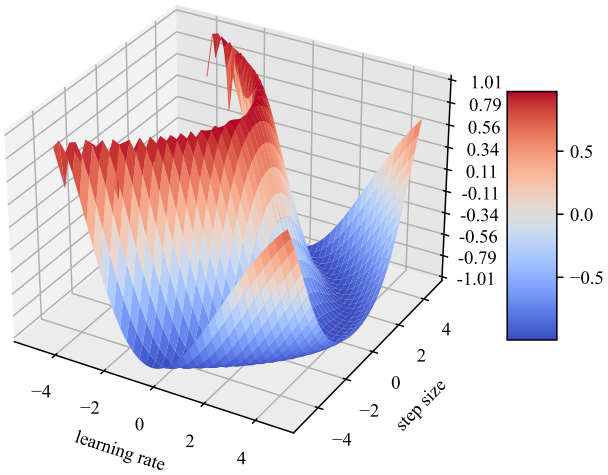

Figure 7. Network model optimization diagram.

Figure 8 shows the human pose output by the 3D human pose estimation network model proposed in this paper. The obtained three-dimensional human body pose is more reasonable, and the human body motion description is more realistic. In the figure, x, y and z represent the three-dimensional space of the network model. Figure 9 shows the matching situation of the 3D human pose output by this paper and the 2D key points, in which the coordinate axis still represents the 3D space position. Since the method used in this paper is to map the obtained 2D key points into the 3D space, there will be a problem that one 3D pose corresponds to the 2D human pose from multiple perspectives. Therefore, this paper is convenient for comparison, and the generated 3D human pose is matched with the key-points of the 2D human body under a specific perspective, and the final effect is shown in the

figure. From the figure, we can clearly see the matching status of the three-dimensional pose of the human body and the two-dimensional key points.

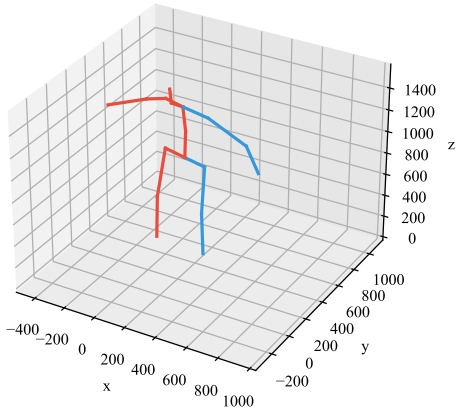

Figure 8. 3D human body pose diagram.

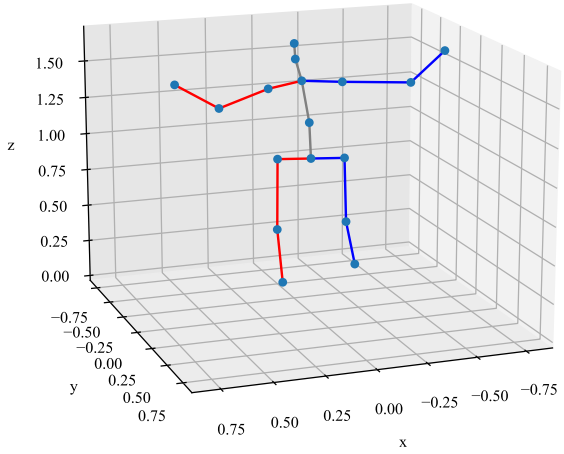

Figure 9. Matching diagram of 3D human pose and 2D key points.

Figure 10 shows an evaluation chart of human motion. We obtained the 3D pose and 2D key points of the human body from previous experiments, which can be effectively applied to motion effect evaluation and posture correction tasks. Figure 10 shows the evaluation of the sports effect, which is divided into various categories from 0 to 10, indicating the type of the sports effect. The corresponding value ranges from 0 to 1 to indicate the evaluation of the exercise effect, and the larger the value, the better the evaluation effect. In Figure 10, the ten categories from 0 to 10 may represent different types of exercise or intensity. Specifically, each category may represent a specific type of exercise or a specific range of exercise intensity. Category 0 represents a situation where there is no exercise, categories 1–3 represent mild walking exercise, categories 4–6 represent moderate running exercise, categories 7–10 represent high-intensity fitness exercise, etc. By comparing the actual collected sensor

data with known categories, the current type or intensity of human movement can be determined.

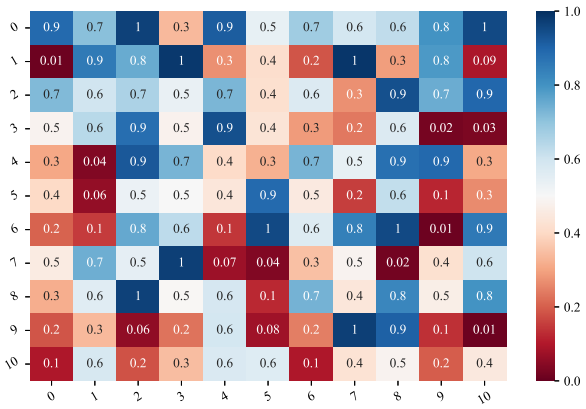

Figure 10. Human motion evaluation diagram.

To evaluate the performance of the proposed method in this article, we compared it with traditional two-dimensional attitude estimation methods. In the experiment, we used common public datasets such as MPII Human Pose and LSP Jitter for testing. To evaluate the performance of our proposed method, we conducted experiments on a public dataset and compared the results with traditional two-dimensional attitude estimation methods. In the experiment, we used common public datasets such as the MPII human posture dataset and the LSP jitter dataset for testing. For new related methods, we propose a method based on flexible integrated piezoelectric sensors for human pose estimation. This method collects information such as acceleration, velocity and angle during human motion, and utilizes signal processing and analysis techniques to extract features of human posture. Then, we use machine learning or deep learning algorithms for attitude estimation and motion type recognition. In the evaluation process, we used indicators such as accuracy and recall to measure the performance of different methods. The results show that the method based on flexible integrated piezoelectric sensors outperforms traditional two-dimensional attitude estimation methods in terms of accuracy and recall. Specifically, we used accuracy metrics to measure the accuracy of the model in estimating human posture, as well as recall metrics to measure the model's recognition ability for different types of motion.

The results show that the accuracy and recall of traditional methods are 85.3 and 83.1%, respectively. The accuracy and recall rates found in this article are 91.7 and 89.5%, respectively. It was found that the method proposed in this article outperforms traditional two-dimensional pose estimation methods in terms of accuracy and recall, and performs more evenly in different motion scenes and key point types. This indicates that the method proposed in this article can achieve more accurate and reliable attitude estimation, providing strong support for related applications such as motion analysis and behavior recognition.

5. Conclusions

Multimedia feature extraction methods based on deep learning have been widely used in various image feature extraction tasks. In this paper, multimedia feature extraction techniques are applied to

the tasks of sports performance evaluation and posture correction. In our paper, the 3D human pose estimation network is used to extract human motion information from video data, and the 3D human pose is obtained through the LSTM sequential unit and CNN feature extraction unit to analyze the motion and posture of the human body. In addition, we also propose a semi-supervised learning method that can use a small amount of labeled data for supervision, and then apply it to massive unlabeled data, which enhances the performance of the network model for practical applications. Although the method in our paper can achieve better analysis of human body posture, it still lacks effective data pairs in practical applications. The correspondence between the two-dimensional human posture and the three-dimensional human posture leads to a certain obstacle in the performance of the network model. Therefore, more attention should be paid to data processing in the follow-up work, and the multimedia feature extraction method can be continuously improved from the data-driven aspect.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgement

This work was supported by Shandong Province Special Task Project under grant 20CLYJ34.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Y. Zon, G. Huang, A feature dimension reduction technology for predicting ddos intrusion behavior in multimedia internet of things, *Multimedia Tools Appl.*, **80** (2021), 22671–22684. https://doi.org/10.1007/s11042-019-7591-7
2. T. Vijayakumar, R. Vinothkanna, M. Duraipandian, Fusion based feature extraction analysis of ecg signal interpretation–a systematic approach, *J. Artif. Intell.*, **3** (2021), 1–16. https://doi.org/10.36548/jaicn.2021.1.001
3. X. Zhu, F. Ma, F. Ding, Z. Guo, J. Yang, K. Yu, A low-latency edge computation offloading scheme for trust evaluation in finance-level artificial intelligence of things, *IEEE Internet Things J.*, 2023. https://doi.org/10.1109/JIOT.2023.3297834
4. D. Meng, Y. Xiao, Z. Guo, A. Jolfaei, L. Qin, X. Lu, et al., A data-driven intelligent planning model for uavs routing networks in mobile internet of things, *Comput. Commun.*, **179** (2021), 231–241. https://doi.org/10.1016/j.comcom.2021.08.014
5. Y. Zhu, W. Lu, R. Zhang, R. Wang, D. Robbins, Dual-channel cascade pose estimation network trained on infrared thermal image and groundtruth annotation for real-time gait measurement, *Med. Image Anal.*, **79** (2022), 102435. https://doi.org/10.1016/j.media.2022.102435

6. S. K. Prabhakar, S. W. Lee, Holistic approaches to music genre classification using efficient transfer and deep learning techniques, *Expert Syst. Appl.*, **211** (2023), 118636. https://doi.org/10.1016/j.eswa.2022.118636
7. Z. Guo, Q. Zhang, F. Ding, X. Zhu, K. Yu, A novel fake news detection model for context of mixed languages through multiscale transformer, *IEEE Trans. Comput. Social Syst.*, 2023. https://doi.org/10.1109/TCSS.2023.3298480
8. M. P. van Dijk, M. Kok, M. A. Berger, M. J. Hoozemans, D. H. Veeger, Machine learning to improve orientation estimation in sports situations challenging for inertial sensor use, *Front. Sports Active Living*, **3** (2021), 670263. https://doi.org/10.3389/fspor.2021.670263
9. S. Jang, J. Jang, Deep learning image processing technology for vehicle occupancy detection, *J. Korea Inst. Inf. Commun. Eng.*, **25** (2021), 1026–1031. https://doi.org/10.6109/jkiice.2021.25.8.1026
10. I. Akhter, A. Jalal, K. Kim, Adaptive pose estimation for gait event detection using context-aware model and hierarchical optimization, *J. Electr. Eng. Technol.*, **16** (2021), 2721–2729. https://doi.org/10.1007/s42835-021-00756-y
11. J. Yang, L. Jia, Z. Guo, Y. Shen, X. Li, Z. Mou, et al., Prediction and control of water quality in recirculating aquaculture system based on hybrid neural network, *Eng. Appl. Artif. Intell.*, **121** (2023), 106002. https://doi.org/10.1016/j.engappai.2023.106002
12. P. Pareek, A. Thakkar, A survey on video-based human action recognition: recent updates, datasets, challenges, and applications, *Artif. Intell. Rev.*, **54** (2021), 2259–2322. https://doi.org/10.1007/s10462-020-09904-8
13. L. H. Palucci Vieira, P. R. Santiago, A. Pinto, R. Aquino, R. d. S. Torres, F. A. Barbieri, Automatic markerless motion detector method against traditional digitisation for 3-dimensional movement kinematic analysis of ball kicking in soccer field context, *Int. J. Environ. Res. Public Health*, **19** (2022), 1179. https://doi.org/10.3390/ijerph19031179
14. G. Lin, Y. Tang, X. Zou, C. Wang, Three-dimensional reconstruction of guava fruits and branches using instance segmentation and geometry analysis, *Comput. Electron. Agric.*, **184** (2021), 106107. https://doi.org/10.1016/j.compag.2021.106107
15. Z. Luo, R. Hachiuma, Y. Yuan, K. Kitani, Dynamics-regulated kinematic policy for egocentric pose estimation, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 25019–25032.
16. C. Fang, T. Zhang, H. Zheng, J. Huang, K. Cuan, Pose estimation and behavior classification of broiler chickens based on deep neural networks, *Comput. Electron. Agric.*, **180** (2021), 105863. https://doi.org/10.1016/j.compag.2020.105863
17. T. Huang, Q. Zhang, X. Tang, S. Zhao, X. Lu, A novel fault diagnosis method based on cnn and lstm and its application in fault diagnosis for complex systems, *Artif. Intell. Rev.*, **55** (2022), 1289–1315. https://doi.org/10.1007/s10462-021-09993-z
18. B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, et al., Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 18408–18419.

19. Z. Guo, K. Yu, K. Konstantin, S. Mumtaz, W. Wei, P. Shi, et al., Deep collaborative intelligence-driven traffic forecasting in green internet of vehicles, *IEEE Trans. Green Commun. Networking*, **7** (2023), 1023–1035. https://doi.org/10.1109/TGCN.2022.3193849

20. P. Jarvis, A. Turner, P. Read, C. Bishop, Reactive strength index and its associations with measures of physical and sports performance: A systematic review with meta-analysis, *Sports Med.*, **52** (2022), 301–330. https://doi.org/10.1007/s40279-021-01566-y

21. Ž. Kozinc, G. Marković, V. Hadžić, N. Šarabon, Relationship between force-velocity-power profiles and inter-limb asymmetries obtained during unilateral vertical jumping and singe-joint isokinetic tasks, *J. Sports Sci.*, **39** (2021), 248–258. https://doi.org/10.1080/02640414.2020.1816271

22. L. K. Topham, W. Khan, D. Al-Jumeily, A. Hussain, Human body pose estimation for gait identification: A comprehensive survey of datasets and models, *ACM Comput. Surv.*, **55** (2022), 1–42. https://doi.org/10.1145/3533384

23. J. Yang, F. Lin, C. Chakraborty, K. Yu, Z. Guo, A. T. Nguyen, et al., A parallel intelligence-driven resource scheduling scheme for digital twins-based intelligent vehicular systems, *IEEE Trans. Intell. Veh.*, **8** (2023), 2770–2785. https://doi.org/10.1109/TIV.2023.3237960

24. T. W. Dunn, J. D. Marshall, K. S. Severson, D. E. Aldarondo, D. G. Hildebrand, S. N. Chettih, et al., Geometric deep learning enables 3d kinematic profiling across species and environments, *Nat. Methods*, **18** (2021), 564–573. https://doi.org/10.1038/s41592-021-01106-6

25. Z. Guo, D. Meng, C. Chakraborty, X. R. Fan, A. Bhardwaj, K. Yu, Autonomous behavioral decision for vehicular agents based on cyber-physical social intelligence, *IEEE Trans. Comput. Social Syst.*, **10** (2022), 2111–2122. https://doi.org/10.1109/TCSS.2022.3212864

26. J. Huang, F. Yang, C. Chakraborty, Z. Guo, H. Zhang, L. Zhen, et al., Opportunistic capacity based resource allocation for 6g wireless systems with network slicing, *Future Gener. Comput. Syst.*, **140** (2023), 390–401. https://doi.org/10.1016/j.future.2022.10.032

27. Q. Li, L. Liu, Z. Guo, P. Vijayakumar, F. Taghizadeh-Hesary, K. Yu, Smart assessment and forecasting framework for healthy development index in urban cities, *Cities*, **131** (2022), 103971. https://doi.org/10.1016/j.cities.2022.103971

28. D. Ravishyam, D. Samiappan, Comparative study of machine learning with novel feature extraction and transfer learning to perform detection of glaucoma in fundus retinal images, in *Soft Computing: Theories and Applications: Proceedings of SoCTA 2020, Volume 2*, Springer, (2021), 419–429.

29. S. Wan, Y. Zhan, L. Liu, B. Yu, S. Pan, C. Gong, Contrastive graph poisson networks: Semi-supervised learning with extremely limited labels, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 6316–6327.

30. Y. Chen, L. Liu, V. Phonevilay, K. Gu, R. Xia, J. Xie, et al., Image super-resolution reconstruction based on feature map attention mechanism, *Appl. Intell.*, **51** (2021), 4367–4380. https://doi.org/10.1007/s10489-020-02116-1

31. O. Oladipo, E. O. Omidiora, V. C. Osamor, A novel genetic-artificial neural network based age estimation system, *Sci. Rep.*, **12** (2022), 19290. https://doi.org/10.1038/s41598-022-23242-5

32. K. Gasmi, I. B. Ltaifa, G. Lejeune, H. Alshammari, L. B. Ammar, M. A. Mahmood, Optimal deep neural network-based model for answering visual medical question, *Cybern. Syst.*, **53** (2022), 403–424. https://doi.org/10.1080/01969722.2021.2018543
33. Z. Guo, Y. Shen, A. K. Bashir, M. Imran, N. Kumar, D. Zhang, et al., Robust spammer detection using collaborative neural network in internet-of-things applications, *IEEE Internet Things J.*, **8** (2021), 9549–9558, https://doi.org/10.1109/JIOT.2020.3003802
34. B. Artacho, A. Savakis, Unipose+: A unified framework for 2d and 3d human pose estimation in images and videos, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 9641–9653. https://doi.org/10.1109/TPAMI.2021.3124736
35. Z. Guo, L. Tang, T. Guo, K. Yu, M. Alazab, A. Shalaginov, Deep graph neural network-based spammer detection under the perspective of heterogeneous cyberspace, *Future Gener. Comput. Syst.*, **117** (2021), 205–218, https://doi.org/10.1016/j.future.2020.11.028
36. M. Yoo, Y. Na, H. Song, G. Kim, J. Yun, S. Kim, et al., Motion estimation and hand gesture recognition-based human-uav interaction approach in real time, *Sensors*, **22** (2022), 2513, https://doi.org/10.3390/s22072513
37. J. P. Sahoo, A. J. Prakash, P. Plawiak, S. Samantray, Real-time hand gesture recognition using fine-tuned convolutional neural network, *Sensors*, **22** (2022), 706, https://doi.org/10.3390/s22030706
38. L. Mo, W. Liancheng, Design of sports competition aided evaluation system based on big data and motion recognition algorithm, *ElectronicDesign Eng.*, **27** (2019), 6–10.
39. J. Kim, S. Yang, B. Koo, S. Lee, S. Park, S. Kim, et al., semg-based hand posture recognition and visual feedback training for the forearm amputee, *Sensors*, **22** (2022), 7984, https://doi.org/10.3390/s22207984
40. J. Mu, S. Xian, J. Yu, J. Zhao, J. Song, Z. Li, et al., Synergistic enhancement properties of a flexible integrated pan/pvdf piezoelectric sensor for human posture recognition, *Nanomaterials*, **12** (2022), 1155. https://doi.org/10.3390/nano12071155

AIMS Press