Hindawi
Mobile Information Systems
Volume 2021, Article ID 2621691, 12 pages
https://doi.org/10.1155/2021/2621691

*Research Article*

# Gesture Recognition Algorithm of Human Motion Target Based on Deep Neural Network

**Zhonghua Xia,[1,2] Jinming Xing,[3] Changzai Wang,[4] and Xiaofeng Li[5]**

[1]*College of Physical Education and Training, Harbin Sport University, Harbin 150025, China*
[2]*Suan Sunan Rajabhat University, Bangkok 10300, Thailand*
[3]*School of Physical Education, Northeast Normal University, Changchun 130024, China*
[4]*Sun Yat-Sen Memorial Secondary School, Zhongshan 528454, Guangdong, China*
[5]*Department of Information Engineering, Heilongjiang International University, Harbin 150025, China*

Correspondence should be addressed to Xiaofeng Li; lixiaofeng@hiu.net.cn

Received 28 May 2021; Revised 5 July 2021; Accepted 16 July 2021; Published 23 July 2021

Academic Editor: Sang-Bing Tsai



There are some problems in the current human motion target gesture recognition algorithms, such as classification accuracy, overlap ratio, low recognition accuracy and recall, and long recognition time. A gesture recognition algorithm of human motion based on deep neural network was proposed. First, Kinect interface equipment was used to collect the coordinate information of human skeleton joints, extract the characteristics of motion gesture nodes, and construct the whole structure of key node network by using deep neural network. Second, the local recognition region was introduced to generate high-dimensional feature map, and the sampling kernel function was defined. The minimum space-time domain of node structure map was located by sampling in the space-time domain. Finally, the deep neural network classifier was constructed to integrate and classify the human motion target gesture data features to realize the recognition of human motion target. The results show that the proposed algorithm has high classification accuracy and overlap ratio of human motion target gesture, the recognition accuracy is as high as 93%, the recall rate is as high as 88%, and the recognition time is 17.8 s, which can effectively improve the human motion target attitude recognition effect.

## 1. Introduction

Humans' perception of information from the outside world is mainly obtained by vision. With the advancement of science and technology, the application of computer technology and related equipment to perceive and understand human behavior and actions has gradually formed, and a wealth of image and video information has been produced [1]. Therefore, how to combine algorithm and logical operations to make computers have human visual functions and perform analysis is a research hot-spot in computer simulation and application technology [2]. Among them, the gesture recognition of human motion targets is an important research direction of Computer Vision. Human motion gesture recognition technology is a technology that analyzes the relevant information of human motion behavior and judges the state of human motion behavior. It can provide user exercise status information, so it is widely used in sports health, user social behavior analysis, indoor positioning, and other fields.

Deep neural network is a fully connected neuron structure with multiple hidden layer structures. As a representative technology of deep learning, deep neural network has achieved great research results in visual research fields such as human behavior recognition. In literature [3], a human body gesture recognition algorithm based on CNN is proposed. By constructing a convolutional neural network model, a total of 11 layers of the model are set up, and five human gestures are collected in the human gesture data set, and the human gesture is operated by convolution and pooling. The fully connected layer of convolutional neural network is used to classify and process the human gesture

data set, and the data set is trained and recognized to realize the human gesture recognition. The algorithm's human gesture extraction feature recognition efficiency is high, but the algorithm's extraction feature recognition recall rate is low. In literature [4], a method for detecting the attitude of astronauts in a space capsule in a weightless environment based on a fast open attitude model is proposed. By constructing a fast open attitude model and using a lightweight deep neural network, the attitude features of the astronauts in a weightless environment are extracted. To ensure the accuracy of model recognition, three small convolution kernels are used to build a fast open attitude device; through the parameter sharing of the convolution process, the branch structure is changed; the residual network is used to suppress the hidden danger of gradient disappearance; and the astronauts work attitude detection is realized. The detection efficiency of the astronaut's operation gesture of this method is high, but there is a problem of low attitude detection accuracy. In literature [5], a deep neural network based on contextual long- and short-term memory architecture is proposed, which uses content and metadata to detect robot context features. It extracts from user metadata and uses it as an auxiliary input to process the tweet text in the contextual long- and short-term memory network, but the feature extraction effect of this method is poor. In literature [6], a new method for training deep neural networks to synthesize dynamic motion primitives is proposed. It can use a new loss function to measure the physical distance between motion trajectories, rather than between parameters that have no physical meaning. We evaluate the proposed method and show that the method's loss function minimization can get better results than using more traditional loss functions, but the algorithm recognition time is longer.

In response to the above problems, this paper proposes a human motion target gesture recognition algorithm based on deep neural network. The idea is as follows:

(1) According to the static gesture of the human, the distance between the key nodes is calculated. The Kinect interface equipment is used to collect the coordinate information of the human bone joints, calculate the difference in the feature value of the human motion gesture, extract the node characteristics of the motion gesture, and use the deep neural network to build the overall structure of the key node network and reduce the node position.

(2) The local recognition region is introduced to generate high-dimensional feature map, and the sampling kernel function is defined to determine the neighborhood of the central pixel.

(3) The depth neural network classifier is constructed to obtain the weighted value of the depth neural network classifier, calibrate the gesture features of the human motion target, fuse and classify the gesture data characteristics of the human motion target, obtain the result of the gesture recognition of the human motion target, and realize the gesture recognition of the human motion target.

## 2. Related Work

At present, a large number of scholars in academia at home have carried out extensive research on the gesture recognition of human motion targets and have achieved certain research results. In literature [7], a human action recognition framework with invariable depth and perspective is proposed, which encapsulated the motion content of the action as an RGB dynamic image, which was generated by an approximate rank pool. And the fine-tuned receiving model is used for processing, long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM) learning model sequence is used to learn the long-term view invariant shape dynamics of the action, and the view invariant features of the key deep human gesture frame based on the structural similarity index matrix are generated. The algorithm has a short recognition time, but the algorithm is affected by the complex changes in the position of key nodes, resulting in lower accuracy of human action recognition. Literature [8] proposed to learn human gesture model from synthetic data for robust RGB-D motion recognition. By analyzing the human gesture in a large amount of human motion targets human skeleton data, 3D human body assembly of different body shapes is synthesized and each gesture with 180 camera viewpoints is rendered. At the same time, the clothing texture, background, and lighting are randomly changed, and the generative confrontation network is used to calculate and minimize the gap between the synthesis and the real image distribution. The learning CNN model is used to transfer the shared human gesture. The CNN model invariant feature extractor is constructed. Pyramid models time changes and uses linear support vector machines to achieve classification. This algorithm has better performance in RGB and RGB-D action recognition, but there is a problem of longer recognition time. In literature [9], a new ellipse distribution coding method is proposed to understand the behavior of the human gesture under infrared imaging. First, the elliptical Gaussian coordinate coding is used to calculate the relationship between adjacent joint points, and then the prediction between the infrared image and the real image is measured. In the end, the infrared human gesture image recognition is completed, but the algorithm takes a long time to recognize.

There are also many studies in China. In literature [10], a human gesture recognition method is proposed based on a small number of key sequence frames. By preselecting the original motion sequence, using the motion trajectory to obtain the extreme value method, the primary key frame sequence was constructed, and the frame reduction algorithm was used to obtain the final key frame sequence. According to different human gestures, a hidden Markov model is constructed, the Baum–Welch algorithm is used to obtain the trained model, and the forward algorithm is used to recognize the human gesture. The algorithm's human gesture recognition accuracy is relatively high, but the algorithm has a large amount of calculation and has the problem of long recognition time. In literature [11], a multiperson gesture detection algorithm optimization based

on reinforcement learning is proposed and the SSD algorithm is used to construct a target detector, obtain the initial bounding box of the human body, and set it as an agent. Reinforcement learning is used to combine Markov decision process and $Q$ network to build a target fine model to train the agent and iteratively adjust its nine actions and four directions, and the stacked hourglass algorithm is used to build a gesture detector to detect the gesture of the adjusted bounding box. The human body detection accuracy of this algorithm is high, but the recall rate of human gesture detection is low. In literature [12], a human skeleton behavior recognition method is proposed based on temporal and spatial weighted gesture motion characteristics. The bilinear classifier is used to iterate to calculate and obtain the action weights of the joint points and static gesture categories and determine the joint points and gestures. Dynamic time warping and Fourier time pyramid algorithm are used to construct a long time sequence model of human skeleton behavior, and support vector machines are used to classify human skeleton behaviors to realize human skeleton behavior recognition. The algorithm has a good recognition effect, but the recognition time of the human skeleton behavior of the algorithm is longer.

For this reason, this paper proposes a human motion target gesture recognition algorithm based on deep neural network and uses MSCOCO data set and MPII data set as data sources to test the proposed algorithm, which verifies the superiority of the method proposed in this paper.

## 3. Algorithm for Target Human Motion Gesture Recognition Based on Deep Neural Network

*3.1. Extract Motion Gesture Node Features.* According to the static gesture of the human, the distance between the key nodes is calculated to extract the gesture characteristics of the human motion target. From a physiological point of view, there are a total of 20 human bone joints, which are used as the key nodes of the human movement target gesture [13]. By observing the movement of the human, the correlation of the bones and joints is obtained, and the key nodes of different movement gestures are selected in a targeted manner. For different human bodies, the size of the bones cannot be exactly the same, so according to the distance information of the key nodes, the movement gesture of a specific individual is represented. For different sports, the movement distance of the head, arms, and legs is not fixed. According to different gestures, the distance of joint points changes with the movement [14, 15]. In order to eliminate the influence of the change of the joint point spacing on the human gesture, the joint point spacing is set as a fixed distance value, and this distance value represents the static gesture characteristics of the human. First, the features of the static gesture of the human are extracted. The human skeleton in the static state is selected, the human head joint as the reference node is used, the distance from other joints to the head joint is calculated, and this distance is used as the element of the feature matrix [16, 17]. The static gesture feature can also characterize the motion gesture at the same time. Based on the static gesture feature extraction, the motion gesture feature is further extracted. The motion gesture objects are human bones at different moments, and $n$ frames of pictures are extracted equidistantly for each motion gesture, and the displacements $l$ of each joint point of the $i$ and $i+1$ frames of each motion gesture are calculated.

$$l = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2}, \quad (1)$$

where $x_i$, $y_i$, and $z_i$ refer to the position coordinate of a node at the image of frame $i$ and $x_{i+1}$, $y_{i+1}$, and $z_{i+1}$ are the position coordinate of the same node at the image of frame $i+1$. Images of frame $n$ are selected, and $s\,(s \le 20)$ nodes in each frame are extracted, and $n-1$ displacement at each node is obtained, which corresponds to $n-1$ coordinates. The characteristic matrix characterizes this movement gesture by coordinate distance $\mathbf{L}$.The calculation formula is as follows:

$$\mathbf{L} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1n-1} \\ l_{21} & l_{22} & \cdots & l_{2n-1} \\ \cdots & \cdots & \cdots & \cdots \\ l_{s1} & l_{s2} & \cdots & l_{sn-1} \end{bmatrix}, \quad (2)$$

where $l_{sn-1}$ is the distance of the node $s$'s frame $n$ and frame $n-1$. The aforementioned feature matrix is the feature vector of the motion gesture. The key information required for feature extraction is the coordinate position information of the key nodes of the human skeleton. Considering that the coordinate information needs to be relatively stable and have high accuracy, this study uses the Kinect interface device to collect the coordinate information of the human bone joints. According to the above process, the difference of the feature values of different human motion gestures is calculated, and the extraction of the motion gesture features is completed.

*3.2. Construction of the Node Structure Diagram.* Since the structure of the key nodes of the human gesture is a graph structure, it needs to be processed effectively. A deep neural network is used to construct a graph of the movement gesture node structure, and the key nodes are learned to realize regional positioning. The purpose of constructing the overall structure of the network is to learn the position of the node graph in the corresponding input image, and each position is divided into different regions to achieve the purpose of reducing the position of the node. The narrowed positioning range is a regional component, and a certain key point corresponds to a region category so that a key trajectory is established, expressed as a multilayer bone sequence, and each node corresponds to a specific multilayer bone sequence [18, 19]. According to the characteristics of the network, hierarchical representation and localized distinction are carried out, and the purpose of sequence classification is realized through the above conversion. Each graph of the original data has a corresponding node graph. In the key point coordinate data, the data are a series of frames, and each frame has node joint coordinates. The two frames of node vectors are constructed into a spatial

structure graph [20, 21]. The abovementioned spatial structure can be used as the input of the image in the neural network, and the adjacent grid becomes the specific area of the image pixel, and the high-latitude feature map is obtained after convolution processing. Classification processing on the feature map is performed to obtain the corresponding coordinate position [22]. The coordinates of key nodes undergo affine transformation and can be rotated within a certain range of angles to construct gestures of different spatial structures. It is shown in Figure 1.

After a single frame of human gesture has been rotated through key nodes $(a_x, a_y)$, the multiframe state $(b_x, b_y)$ and $(c_x, c_y)$ are formed. The combination becomes the graph data, which are used as the input of the deep neural network. A key node matrix based on the graph data is established, which is a collection of all nodes in the gesture structure graph. The formula is defined as follows:

$$\mathbf{R}_{ns} = \begin{bmatrix} x_1 & x_2 & \dots & x_s \\ y_1 & y_2 & \dots & y_s \\ C_1 & C_2 & \dots & C_s \end{bmatrix}, \tag{3}$$

where $\mathbf{R}_{ns}$ is the graph structure gesture matrix at frame $n$ and of $s$ nodes; $x$, $y$ is the position function, and the set is $\{x_1, x_2, \dots, x_s\}$ and $\{y_1, y_2, \dots, y_s\}$; and $C$ is the node coordinate confidence, and the set is $\{C_1, C_2, \dots, C_s\}$. The confidence is used to judge whether the key node exists. Through the above conversion process, the overall structure of the key node network is constructed to provide a graph structure basis for subsequent network training.

*3.3. Locate the Local Recognition Area of the Node Structure.* In the gesture structure, the position changes of the key nodes of the human are more complicated. Compared with the number of pictures, the feature map constructed by the key nodes can express limited information. To accurately recognize the overall motion target gesture, a large number of training coordinate positions and classification label values need to be calculated, which not only increases the difficulty of recognition but also increases the amount of calculation [23]. Therefore, in this research, the overall positioning of key points is transformed into the positioning of the local recognition area to improve the computational efficiency of the gesture recognition algorithm. The target detection task mainly finds the target from an image, unifies the detection steps into the deep network, first inputs the original scale feature map, and then propagates forward to the shared convolutional layer to generate a higher-dimensional feature map. Classification and position regression are performed on multiple feature scales at the same time, each pixel of the feature map is taken as the center, and a default box is generated and mapped from the feature map to the original image location according to the center point coordinates. The sampling kernel function is determined, and the neighborhood of the center pixel is determined [24, 25]. In the single-frame image structure, the center position to the surrounding forms a grid so that the

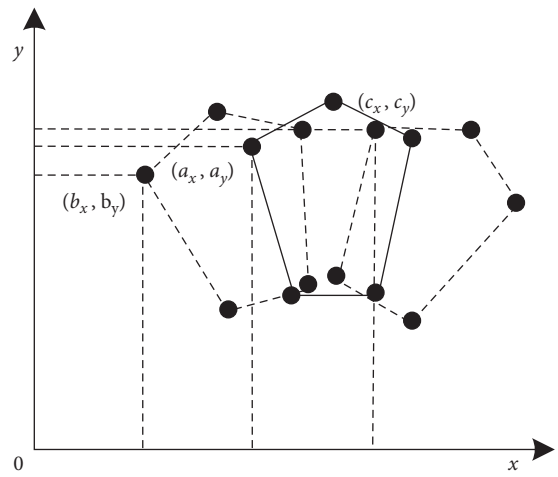



Figure 1: Space structure gesture.

neighborhood pixels form a fixed spatial order, and the retrieval is performed according to the dimension. Because the spatial structure graph is irregular, it is necessary to set a fixed node as the root node to mark the neighborhood set to realize the weight function. The weight function is iterative updated by the network's preset hyperparameters. Sampling function and weight function are converted into network form. The formula is defined as follows:

$$f(p_i) = \sum_{p_i \in D}^{k} \frac{1}{\phi(p_i)} f\left(c\left(p_i, p_j\right) \cdot w\left(p_i, p_j\right)\right), \tag{4}$$

where $f(p_i)$ is the network form of node $p_i$ and $p_i$ belongs to the neighboring domain $D$. $D$ is divided into $k$ subsets, and each subset corresponds to a weight value; $p_j$ is a node of neighboring domain, and $c(p_i, p_j)$ refers to the sampling function, $w(p_i, p_j)$ refers to the weight function, and $\phi(p_i)$ is a regular term and node neighborhood subset cardinality. The contribution $G_d$ of different subsets to the neural network is analyzed The formula is defined as follows:

$$G_d = \frac{Z_\varphi}{f(p_i)} \times 100\%, \tag{5}$$

where $Z_\varphi$ is the subset to label specific points. Formula (5) is similar to the standard network. It is used for a single frame of image and changes the pixel scale. Each subset corresponds to a single node. After the neural network is sampled, the state of the single node is learned to obtain the graph structure characteristics of the key nodes of the human. After classification, the areas where different key nodes are located are distinguished. Further, the minimum spatio-temporal neighborhood is divided into the multisequence space. The minimum time domain represents the minimum set of nodes in different frames, and the minimum space domain represents the minimum set of nodes in a single frame. Since the data structure is interconnected with the same number, it cannot be used to describe the same neighborhood, so the bone sequence is extended to the space-time domain. The formula is defined as follows:

$$T(p_i) = \left\{ p_i | d(p_i, p_j) \le D, |o - t| \le \lfloor \frac{\chi}{2} \rfloor \right\}, \quad (6)$$

where $T(p_i)$ is the minimum time and space domain of node $p_i$; $d(p_i, p_j)$ is the sampling distance from node $p_i$ to $p_j$; $|o - t|$ is the number of sampling frameworks, and $\chi$ is the time domain length. Through time-space sampling, the smallest time-space feature is detected so as to complete the location of the local recognition area of the feature map.

## 4. Establish the Algorithm for the Recognition of Target Motion Gestures of Humans

The gesture recognition algorithm is based on a large amount of data and establishes the relationship between the target output and the actual output by constructing an activation function. By continuously adjusting the weights and variation parameters, the optimal solution is obtained to realize the recognition of the human motion target gesture, and the algorithm of the human motion target gesture recognition is described as follows.

To sum up, the specific process of human motion target gesture recognition is shown in Figure 2.

## 5. Experimental Analysis and Results

In order to verify the effectiveness of the algorithm for the recognition of human motion target gestures based on deep neural networks, simulation experiments are carried out. The experiment uses MATLAB simulation software, combined with the Libsvm simulation toolbox, applies the human motion target gesture recognition algorithm in the actual operation simulation, and uses deep neural network technology to recognize the human motion target gesture.

*5.1. Experimental Environment and Data Set.* The experiment uses MSCOCO data set and MPII data set and conducts training and testing on this data set. The MSCOCO data set is a human gesture estimation data set, which contains about 30,000 sample images of human images and camera-collected images, and the number of joint points is 18. The MPII data set is a state-of-the-art articulated human gesture estimation benchmark. It contains about 25,000 sample images, including more than 40,000 people with annotated body joints, and the number of joint points is 16.

Sixteen human joint points in MSCOCO data set and MPII data set were selected and numbered, and 10,000 sample images data were selected from each of the two data sets, with a total of 20,000 sample images for experimental analysis. Randomly 10,000 sample images are selected as the training data, and the rest 10,000 sample images as the test data. In the MATLAB simulation software for 100 groups of training, algorithm training is used to test the human skeleton characteristics. The training parameters of the gesture recognition algorithm are shown in Table 1.

*5.2. Evaluation Criteria*

(i) *Gesture Classification Accuracy*. This refers to the correctness of the classification of the human motion target's gesture, which is used to reflect the accuracy of the human motion target's gesture classification. The calculation formula for the accuracy of gesture classification is as follows:

$$S_z = \frac{\lambda_z}{\gamma_d} \times 100\%, \quad (7)$$

where $\lambda_z$ is the number of correctly classified human motion target gestures and $\gamma_d$ is the total number of human motion target gestures to be classified.

(ii) *Gesture Recognition Time*. The gesture recognition time is used as an indicator to compare the proposed algorithm with Literature [7]–Literature [11] algorithm to verify the performance of the proposed algorithm.

(iii) *Gesture Recognition Recall Rate*. It refers to the degree of success in recognizing the gesture of the relevant human motion target in the data set for measuring the human gesture estimation. The calculation formula of the recall rate of gesture recognition is as follows:

$$S_q = \frac{\delta_z}{\gamma_d} \times 100\%, \quad (8)$$

where $\delta_z$ refers to the number of related human motion target gestures recognized and $\gamma_d$ is the total number of human motion target gestures to be recognized.

(iv) *Overlap Ratio*. The overlap ratio describes the overlap between the output of the algorithm and the calibration range, and the overlap ratio can be used as a key index to judge the effectiveness of gesture recognition.

(v) *Gesture Recognition Precision*. Taking the gesture recognition precision as the index, the advantages of the proposed algorithm are verified.

*5.3. Results and Discussion.* According to the examples of human motion gesture on MSCOCO data set and MPII data set, the human motion gesture classification accuracy of the proposed algorithm and the algorithms in literature [7]–literature [11] are compared and analyzed by using different data sets. It is shown in Figure 3.

Based on the example of the depth map of human motion gesture in Figure 3, the accuracy of human motion gesture classification of different algorithms is calculated, and the comparison results are shown in Table 2.

According to Table 2, the proposed algorithm has higher accuracy rates of human motion gesture classification on the MSCOCO data set and MPII data set, respectively, as 0.82 and 0.86. The highest accuracy rate of other literature

(i) Input: the original data of human motion, as well as the positioning results in the minimum space-time domain of the motion gesture node structure diagram.
(ii) Output: recognition results of human motion target gesture of $S_b(x)$.
(iii) According to the test samples in the gesture database of the human motion target to be recognized $C_S$ and the sample training set $X_S$, obtain the human gesture feature distribution set as $R_z = [1/u, 1/u, \ldots, 1/u]^T$.
(iv) Where $u$ is the number of human motion target gestures in the training set.
(v) Construct a deep neural network classifier and obtain the weighted value of the deep neural network classifier as $F_J = 1/1 + \exp((d_{ik})^2 - \varphi_u)$.
(vi) where $d_{ik}$ is the initialized eigenvalues and $\varphi_u$ is binarized fitting results. Through the feature extraction results of motion gesture nodes, the deep neural network is introduced to obtain the input and output iterative equations of the deep neural network classifier: $\kappa(\eta) = R_z - \mathbf{L}(\partial\chi_w/\partial\vartheta_q)$.
(vii) where $\chi_w$ is the learning pace length and $\vartheta_q$ is the maximum iteration times of the training. Using the structural similarity algorithm, the weighted coefficient of the human motion target gesture classifier is obtained and expressed as $\nu = C_S(\partial\chi_w/\partial X_S)$.
(viii) Through the deep neural network classifier, the gesture characteristics of the human motion target are calibrated, and the recognition statistics are obtained as $T_J = (\sum_{l=1}^{\nu} F_J - X_S/\kappa(\eta))\vartheta_q$.
(ix) According to the classification method of the video image, the data are fused and classified and recognized. The image pixels after feature extraction are traversed through the window sliding to traverse the entire image, and the calculation process can be expressed as $B(a,b) = \sum_{i=1}^{i}\sum_{j=1}^{j} u_{a+i,b+j} \cdot u_{i,j} + \sigma$.
(x) where $B(a,b)$ is the traverse result, $(a,b)$ is the position of the output of the feature map of the previous layer of the node; $u_{a+i,b+j}$ is the value of the feature diagram at line $a+i$ and column $b+j$, $u_{i,j}$is the value at line $i$ and column $j$; and $\sigma$ is the derivative error.
(xi) As the number of traversal results deepens, a connection and sharing relationship is formed, and the activation function is used to transform the linear transformation into a nonlinear transformation [19]. After introducing the nonlinear activation function, the deep network can simulate any function. The PReLU activation function was selected for this study
$$F(m) = \begin{cases} m, & m > 0, \\ 0.01m, & m \le 0 \end{cases}.$$
(xii) where $m$ is the input node. This function is a piecewise function. When $m \le 0$, the gradient is not 0, which solves the dead zone problem of the disappearance of the gradient. The sliding window is used with the same size and step size to calculate the sliding matrix and feature map. From the perspective of the amount of data and the number of parameters, the amount of calculation is reduced. It can reduce dimension and abstract results at the same time and improve the fault tolerance of the algorithm [20].
(xiii) The fully connected method is used to connect the network nodes, and the output formula of each neuron is $S_{\omega,\theta}(m) = F(\varpi^T m + \theta)$.
(xiv) where $S_{\omega,\theta}(m)$ is the input value of the node; $F$ is the activated function; $\varpi$ is the weight vector, and $\theta$ is the deviation. $T$ is the transpose symbol. Through the abovementioned full connection method, the output information characteristics can be gathered.
(xv) The aggregated human motion target gesture information features are extracted, and the gesture recognition result of the human motion target based on the difference of biological characteristics is obtained: $S_b(m) = (S_{\omega,\theta}(m)/\varpi)F(m)$.
(xvi) End

Algorithm 1: Human motion target gesture recognition algorithm.

algorithms does not exceed 0.65, and the highest accuracy rate of literature [7] is 0.63. The highest accuracy rate of literature [8] is 0.65, the highest accuracy rate of literature [9] is 0.62, the highest accuracy rate of literature [10] is 0.58, and the highest accuracy rate of literature [11] is 0.50. The deep neural network used in this paper has strong representation ability and the best classification effect.

In order to comprehensively evaluate the performance of this gesture recognition algorithm, the index of overlap ratio is proposed, and different algorithms are compared and analyzed. The higher the overlap ratio is, the closer the algorithm output value is to the calibration value and the better the algorithm performance is. The comparison results of the overlap ratio of different algorithms are shown in Figure 4.

Literature [7] algorithm has a maximum overlap ratio of about 60% of the human motion target gesture recognition results, literature [8] algorithm has a maximum overlap ratio of approximately 58%, and literature [9] algorithm has a maximum overlap ratio of approximately 50%. In literature [10], the highest overlap ratio of the algorithm is about 58%, the highest overlap ratio of the algorithm in literature [11] is about 56%, and the highest overlap ratio of the proposed algorithm is about 80%. It can be clearly seen that the algorithm recognition results in this paper have a higher overlap rate with the calibration range, and the recognition effect is better.

In order to verify the gesture recognition precision of the human motion target gesture recognition algorithm based on the deep neural network, the Literature [7]–Literature [11] algorithm and the proposed algorithm are used to test the recognition precision of the human motion target gesture recognition algorithm. In this way, the comparison results of the recognition precision of human motion target gestures of different algorithms are obtained. It is shown in Figure 5.

According to Figure 5, when the number of iterations is 500, the average human motion target gesture recognition precision rate of the algorithm in Literature [7] is 80%, and the average human motion target gesture recognition

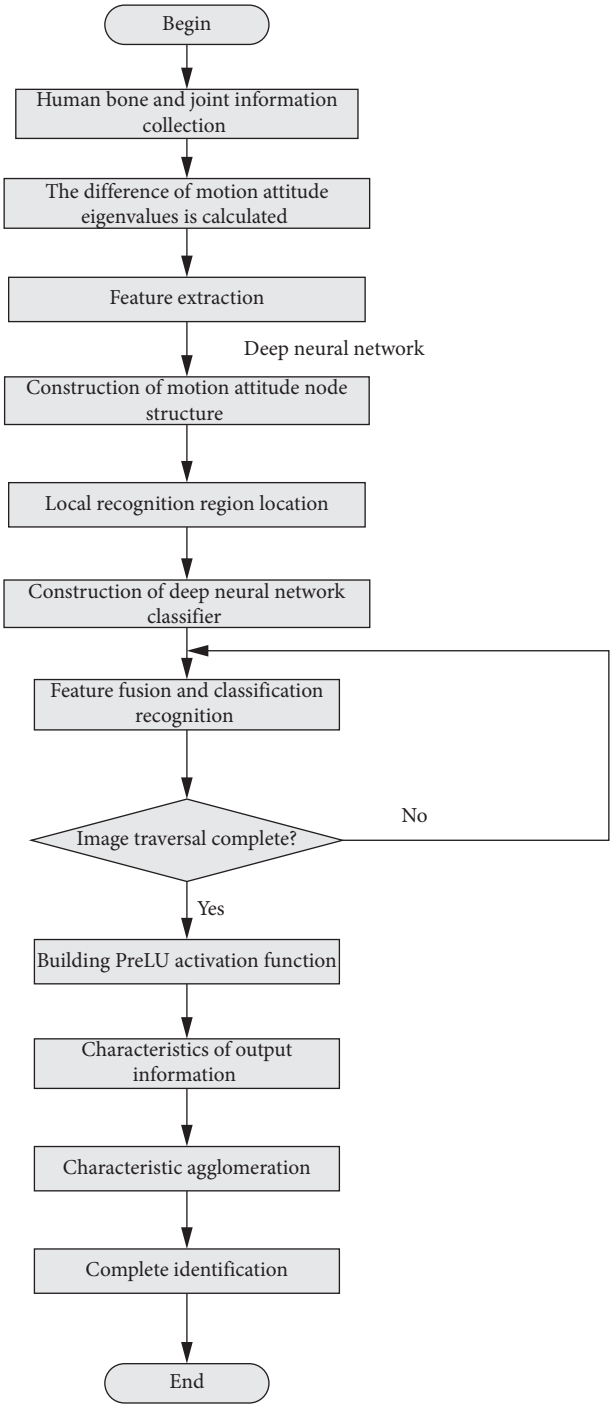
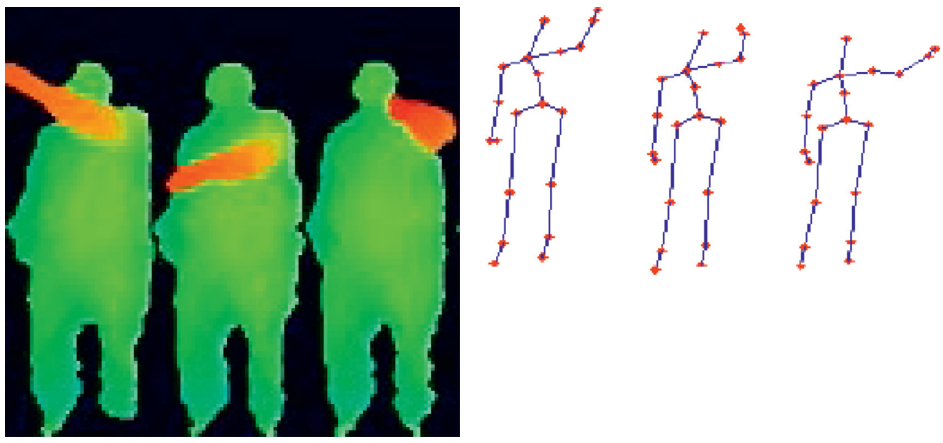



Figure 2: Gesture recognition algorithm of human motion target based on deep neural network.

Table 1: Training parameters of the gesture recognition algorithm.

| Parameters | Value (or running state) |
|---|---|
| Number of input images | 20,000 sample images |
| Iterations | 500 |
| Optimizer | Random gradient descent |
| Initial value of weight correction | 0.01 |



(a) (b)

Figure 3: Example of a depth map of human motion gesture: (a) MSCOCO data set; (b) MPII data set.

Table 2: Comparison of the accuracy of human motion gesture classification with different algorithms.

| Algorithm | MSCOCO data set | MPII data set |
|---|---|---|
| The proposed algorithm | 0.82 | 0.86 |
| Literature [7] | 0.63 | 0.62 |
| Literature [8] | 0.65 | 0.60 |
| Literature [9] | 0.56 | 0.62 |
| Literature [10] | 0.58 | 0.50 |
| Literature [11] | 0.50 | 0.46 |

accuracy rate of the algorithm in Literature [8] is 60%, and the average human motion target gesture recognition precision rate of the algorithm in Literature [9] is 70%, the average human motion target gesture recognition precision rate of the algorithm in Literature [10] is 78%, and the average human motion target gesture recognition precision rate of the algorithm in Literature [11] is 60%, and the average human motion target gesture recognition precision of the proposed algorithm is as high as 93%. It can be seen that the proposed algorithm has a high precision in the recognition of human motion target gestures and can effectively improve the precision of human motion target gesture recognition. Because the proposed algorithm calculates the distance between key nodes according to the static gesture of the human, the Kinect interface device is used to collect the coordinate information of the human bone joints, the difference in the feature value of the human motion gesture is calculated, and the node feature of the motion gesture is extracted, thereby reducing the actual feature value. The deviation improves the recognition precision of the human motion target gesture.

In order to verify the gesture recognition time of the proposed algorithm, the Literature [7] algorithm, the Literature [8] algorithm, the Literature [9] algorithm, the Literature [10] algorithm, the Literature [11] algorithm, and the proposed algorithm are used to compare the recognition time of each joint point of the human motion target gesture of different algorithms. In this way, the comparison results of the recognition time of human motion target gestures of different algorithms are obtained. It is shown in Table 3.

According to Table 3, with the increase in the joint points of the human motion target gesture, the recognition time of each joint point of the human motion target gesture of different algorithms increases. When the human motion target gesture has 16 joint points, the recognition time of each joint point of the human motion target gesture of the algorithm in Literature [7] is 21.3 s, and the recognition time of each joint point of the human motion target gesture of the algorithm in Literature [8] is 27.3 s, the recognition time of each joint point of the human motion target gesture of the algorithm in Literature [9] is 24.6 s, the recognition time of each joint point of the human motion target gesture of the algorithm in Literature [10] is 28.4 s, the recognition time of each joint point of the human motion target gesture of the algorithm in Literature [11] is 29.7 s, while the recognition time of each joint point of the human motion target gesture of the proposed algorithm is only 17.8 s. It can be seen that the recognition time of each joint point of the human motion target gesture of the proposed algorithm is shorter, and the human motion gesture can be recognized more quickly. Therefore, the proposed algorithm uses a deep neural network to build the overall structure of the key node network, reduces the position of the node, and converts the overall positioning of the key point into the local recognition area to improve the calculation efficiency of the gesture recognition algorithm, thereby shortening the recognition time of each joint point of the gesture. On this basis, we further verify the recall rate of human motion target gesture recognition of the proposed algorithm and obtain the comparison results of the recall rate of human motion target gesture recognition of different algorithms. It is shown in Figure 6.

According to Figure 6, when the number of input images is 10000, the average recall rate of human motion target gesture recognition in algorithm [7] is 50%, and the average recall rate of human motion target gesture recognition in

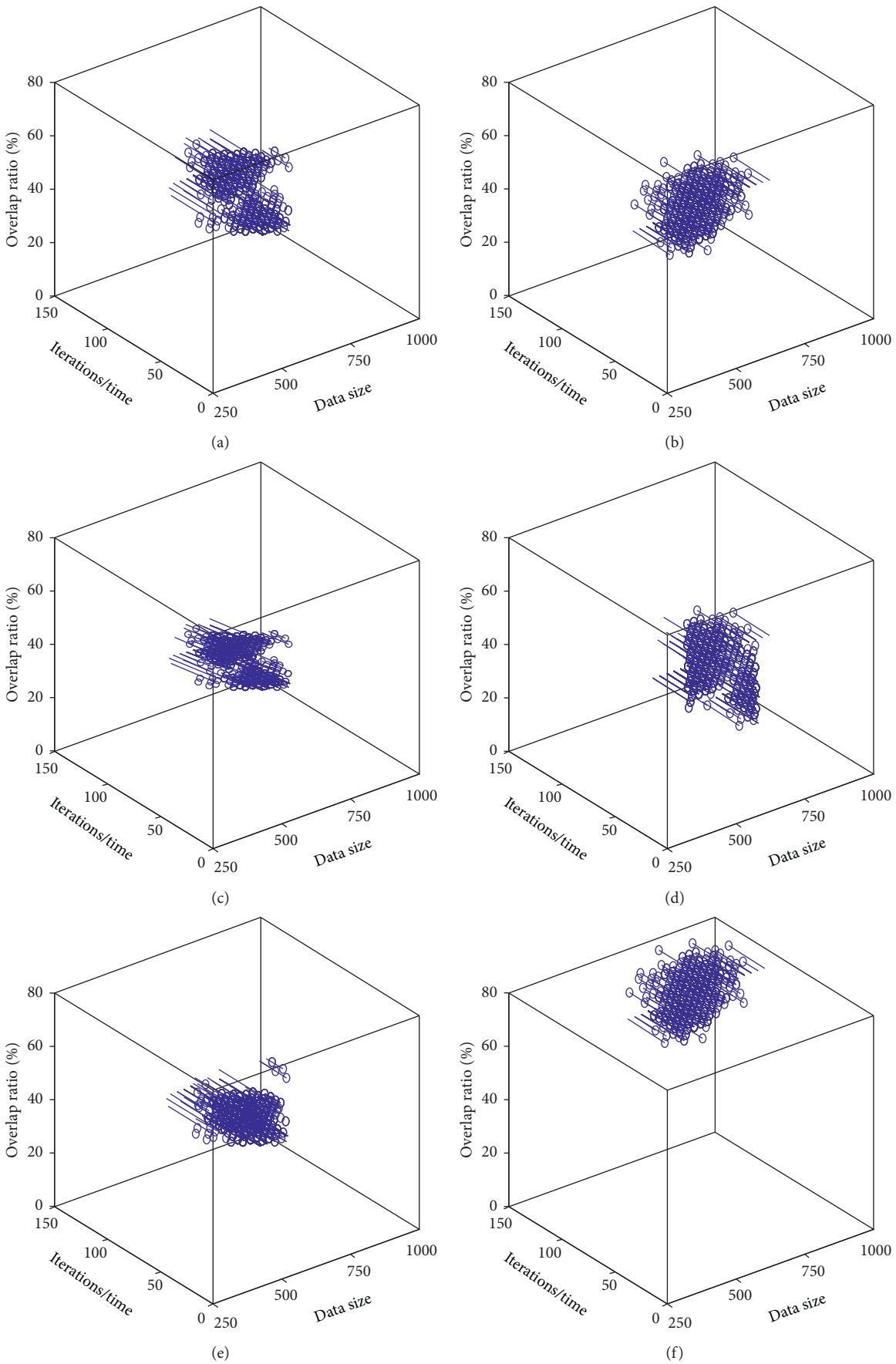



Figure 4: Comparison of the overlap ratio of human motion target gesture recognition results with different algorithms: (a) Literature [7] algorithm; (b) Literature [8] algorithm; (c) Literature [9] algorithm; (d) Literature [10] algorithm; (e) Literature [11] algorithm; (f) the proposed algorithm.

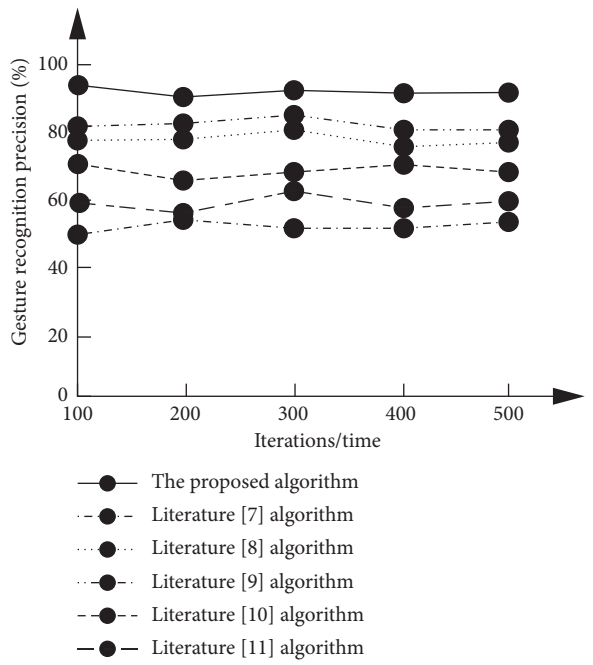



Figure 5: Comparison results of gesture recognition precision of human motion target with different algorithms.

Table 3: Time comparison results of human motion target gesture recognition based on different algorithms.

| Joint point/number | Identification time (s) | | | | | |
|---|---|---|---|---|---|---|
| | The proposed algorithm | Literature [7] algorithm | Literature [8] algorithm | Literature [9] algorithm | Literature [10] algorithm | Literature [11] algorithm |
| 1 | 11.9 | 13.6 | 18.2 | 17.0 | 18.0 | 17.5 |
| 2 | 12.6 | 14.8 | 19.3 | 17.9 | 18.8 | 18.0 |
| 3 | 12.8 | 15.2 | 19.9 | 19.2 | 20.2 | 19.9 |
| 4 | 12.9 | 15.7 | 20.9 | 19.6 | 20.5 | 19.3 |
| 5 | 13.1 | 16.4 | 20.9 | 19.8 | 21.4 | 20.8 |
| 6 | 13.8 | 16.6 | 21.6 | 20.4 | 21.0 | 21.5 |
| 7 | 14.6 | 16.9 | 21.8 | 20.9 | 22.7 | 22.0 |
| 8 | 15.1 | 17.0 | 22.5 | 21.4 | 22.8 | 22.1 |
| 9 | 15.2 | 17.9 | 22.5 | 21.0 | 23.4 | 23.7 |
| 10 | 15.6 | 18.6 | 23.4 | 21.6 | 23.7 | 23.5 |
| 11 | 15.8 | 19.2 | 23.8 | 22.7 | 24.6 | 28.8 |
| 12 | 16.4 | 19.7 | 24.1 | 22.7 | 24.1 | 25.6 |
| 13 | 16.6 | 20.2 | 24.2 | 23.3 | 25.0 | 25.9 |
| 14 | 17.0 | 20.9 | 25.6 | 24.8 | 25.7 | 26.4 |
| 15 | 17.2 | 21.0 | 25.4 | 24.6 | 26.5 | 27.5 |
| 16 | 17.8 | 21.3 | 27.3 | 24.6 | 28.4 | 29.7 |

algorithm [8] is 68%. The average recall rate of human motion target gesture recognition in algorithm [9] is 75%, the average recall rate of human motion target gesture recognition in algorithm [10] is 59%, and the average recall rate of human motion target in algorithm [11] is 79%. The gesture recognition recall rate is 88%, and the average human motion target gesture recognition recall rate of the proposed algorithm is 88%. It can be seen that the proposed algorithm has a higher recall rate of human motion target gesture recognition.

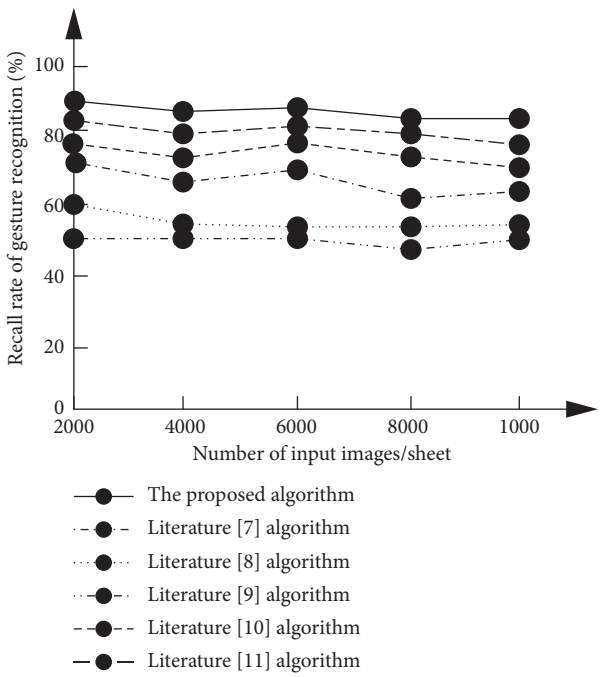



Figure 6: Comparison results of recall rate of human motion target gesture recognition with different algorithms.

## 6. Conclusions

In order to improve the accuracy and recall rate of human motion target gesture recognition and shorten the time of human motion target gesture recognition, a deep neural network-based human motion target gesture recognition algorithm is proposed. According to the static gesture of the human, the distance between the key nodes is calculated, and the Kinect interface device is used to collect the coordinate information of the human bone joints, calculate the difference in the feature value of the human motion gesture, and extract the node characteristics of the motion gesture to improve the recognition accuracy of the human motion target gesture. Deep neural network is used to build the overall structure of the key node network, reduce the position of the node, locate the smallest space-time domain of the node structure diagram through time-space sampling, improve algorithm computing efficiency, and shorten recognition time. This algorithm can distinguish different human gestures and has certain validity and feasibility.

However, due to the complexity of the human motion target gesture recognition process, there is still something to be improved in this research. Subsequent research can be conducted from the aspect of gesture similarity to evaluate the difference in similarity between the human gesture and the standard gesture so as to measure individual differences to achieve priority matching.

## Data Availability

The data used to support the findings of this study are included within the article. Readers can access the data supporting the conclusions of the study from MSCOCO data set and MPII data set.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by Ministry of Education Science and Technology Development Center Fund under grant no. 2020A050116.

## References

[1] N. Wu and S. Haruyama, "Real-time audio detection and regeneration of moving sound source based on optical flow algorithm of laser speckle images," *Optics Express*, vol. 28, no. 4, pp. 1364–1370, 2020.

[2] S. Sun, M. Jiang, D. He, Y. Long, and H. Song, "Recognition of green apples in an orchard environment by combining the GrabCut model and Ncut algorithm," *Biosystems Engineering*, vol. 187, pp. 201–213, 2019.

[3] Y. Zhou, Y. Wang, Y. Zhao et al., "Human posture recognition based on CNN," *Computer and Modernization*, vol. 2, pp. 49–54, 2019.

[4] E. Qwu, Z. Tang, P. Xiong, C.-f. Wei, A. Song, and L. Zhu, "ROpenPose: A rapider OpenPose model for astronaut operation attitude detection," *IEEE Transactions on Industrial Electronics*, p. 1, 2021.

[5] K. Sneha and F. Emilio, "Deep neural networks for bot detection," *Information Sciences*, vol. 467, pp. 312–322, 2018.

[6] C. Rpa, B. Bra, A. Ag et al., "Training of deep neural networks for the generation of dynamic movement primitives," *Neural Networks*, vol. 127, no. 3, pp. 121–131, 2020.

[7] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Transactions on Image Processing*, vol. 29, pp. 3835–3844, 2020.

[8] J. Liu, H. Rahmani, N. Akhtar, and A. Mian, "Learning human pose models from synthesized data for robust RGB-D action recognition," *International Journal of Computer Vision*, vol. 127, no. 10, pp. 1545–1564, 2019.

[9] H. Liu, Y. Chen, W. Zhao, S. Zhang, and Z. Zhang, "Human pose recognition via adaptive distribution encoding for action perception in the self-regulated learning process," *Infrared Physics & Technology*, vol. 114, Article ID 103660, 2021.

[10] Q. Caixing, Y. Tu, Y. Yu et al., "Human posture recognition method based on a few key sequence frames," *Journal of Graphics*, vol. 40, no. 03, pp. 532–538, 2019.

[11] d. Huang and Z. Yingna Cai, "Optimization of multi-human posture detection algorithm based on intensive learning," *Computer Application and Software*, vol. 36, no. 04, pp. 186–191, 2019.

[12] C. Ding, K. Liu, Li Guang et al., "Research on human skeleton behavior recognition based on temporal and spatial weighted gesture motion characteristics," *Chinese Journal of Computers*, vol. 43, no. 1, pp. 29–40, 2020.

[13] X. Zhang, "Noise-robust target recognition of SAR images based on attribute scattering center matching," *Remote Sensing Letters*, vol. 10, no. 2, pp. 186–194, 2019.

[14] J. Wan, S. Xu, and W. Zou, "High-accuracy automatic target recognition scheme based on photonic analog to digital converter and convolutional neural network," *Optics Letters*, vol. 45, no. 24, Article ID 411214, 2020.

[15] C. Wu, "Assisting target recognition through strong turbulence with the help of neural networks," *Applied Optics*, vol. 59, no. 30, pp. 9434–9442, 2020.
[16] S. HongQ. Wang et al., "Target recognition method with frequency features on retina-like laser detection and range images," *Applied Optics*, vol. 58, no. 35, pp. 9532–9539, 2019.
[17] L. Qi, Q. Hu, Q. Kang, Y. Bi, Y. Jiang, and L. Yu, "Detection of biomarkers in blood using liquid crystals assisted with aptamer-target recognition triggered in situ rolling circle amplification on magnetic beads," *Analytical Chemistry*, vol. 91, no. 18, pp. 11653–11660, 2019.
[18] J. Xu, P. Bi, X. Du, and J. Li, "Robust PCANet on target recognition via the UUV optical vision system," *Optik*, vol. 181, pp. 588–597, 2019.
[19] M. Liu and W. D. Zhu, "Design and analysis of nonlinear-transformation-based broadband cloaking for acoustic wave propagation," *Wave Motion*, vol. 92, Article ID 102421, 2019.
[20] T. Huang, Y. Chen, B. Yao, B. Yang, X. Wang, and Y. Li, "Adversarial attacks on deep-learning-based radar range profile target recognition," *Information Sciences*, vol. 531, no. 8, pp. 159–176, 2020.
[21] S. Pereira, P. Canhoto, R. Salgado, and M. J. Costa, "Development of an ANN based corrective algorithm of the operational ECMWF global horizontal irradiation forecasts," *Solar Energy*, vol. 185, no. 6, pp. 387–405, 2019.
[22] A. Li, Z. Zhao, X. Liu, and Z. Deng, "Risley-prism-based tracking model for fast locating a target using imaging feedback," *Optics Express*, vol. 28, no. 4, pp. 5378–5392, 2020.
[23] S. Fan, Y. Jia, and J. Liu, "Feature selection for human posture recognition based on three-axis accelerometer," *Journal of Applied Sciences*, vol. 37, no. 3, pp. 427–436, 2019.
[24] R. Liao, S. Yu, W. An et al., "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, Article ID 107069, 2019.
[25] R. D. Bem, A. Ghosh, T. Ajanthan et al., "DGPose: Deep generative models for human body analysis," *International Journal of Computer Vision*, vol. 9, pp. 1–27, 2020.