IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 5, 2020, accepted August 24, 2020, date of publication August 28, 2020, date of current version September 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3020141

Human Motion Gesture Recognition Algorithm in Video Based on Convolutional Neural Features of Training Images

XIANGUI BU
School of Sports and Physical Education, Shandong Sport University, Rizhao 276800, China
e-mail: buxiangui@sdpei.edu.cn

ABSTRACT The main work of human motion gesture recognition is to recognize and analyze the behavior of human objects in the video. Although the current research in the field of human motion gesture recognition has achieved certain results, the human motion gesture recognition in real life scenes has great effects due to factors such as camera movement, target scale transformation, dynamic background, viewing angle, and illumination. This article first proposes a new method of constructing human motion posture features to describe human behavior. This method is based on deep convolutional neural network features and topic models. Experiments have verified that compared with the traditional feature map extracted from the convolutional neural network fully connected layer, the feature map extracted from the convolutional neural network convolutional layer is not only lower in dimension but also has higher discrimination. Secondly, based on the feature map of the convolutional neural network, the training map downsampling strategy is used to overcome the interference caused by the object's scale change and shape change. Finally, based on the basketball gesture recognition method, the behavior performance of the legs and arms in 9 basketball actions of walking, running, jumping, standing dribbling, walking dribbling, running dribbling, shooting, passing and receiving is analyzed. As well as the corresponding signal waveform characteristics, a two-stage data division method for basketball is proposed. The unit action data is extracted for analysis to realize feature extraction. In order to select the most suitable classifier for basketball gesture recognition, the constructed feature vector uses four Different classifiers are trained to construct different classifiers to realize the division of actions.

INDEX TERMS Convolutional neural networks, training images, human motion gesture recognition, classifier, feature vector.

I. INTRODUCTION

Under the current technical background, video human motion gesture recognition is widely used in smart wearable devices, security monitoring, human-computer interaction, sports training and competition, military, medical care and other fields. Video human motion gesture recognition technology has broad application prospects and great commercial value, such as smart sports bracelets and watches, pedometers, fall prevention devices for the elderly, smart home appliances, VR glasses and other many devices such as series games have a certain market share. In the current era of big data and the background of the Internet era, with all kinds of ubiquitous electronic devices, especially smart phones that everyone can keep in their hands 24 hours a day, it can collect information from all walks of life, all ages, and even all classes. Video human body motion posture data will be very rich. Useful information can be obtained in such a huge "database", so as to play its value and make due contributions to all aspects of human life. In the wave of this new era, researchers are using video human motion posture data to analyze and extract information and ultimately improve the quality of human life, solve practical problems in human life, and meet human material and spiritual needs, which will give humans Life brings great convenience and a stronger sense of happiness. In summary, the research on the recognition of human motion gestures in video is very necessary.

Early video human motion gesture recognition mainly researched the recognition of simple single-target behaviors such as head motion, hand motion, body posture, and

The associate editor coordinating the review of this manuscript and approving it for publication was Zhihan Lv.

facial expressions. At present, everyone's research content is gradually turning to the research of multi-target, complex motion, and complex background video human motion gesture recognition. The classic video human motion gesture recognition methods are mainly as follows: (1) template matching method [1], [2]; (2) state space method [3]; (3) semantic-based method [4]. The template matching method is to match the test sequence with the reference sequence frame by frame or perform fusion matching. The template matching method is easy to implement, has low time complexity, and has a better effect on identifying behaviors with obvious differences. However, because the template matching method is easily affected by factors such as light and deformation, the recognition of some behaviors that are not very different will change It's quite difficult [5], [6]. The deep convolutional features of video frames are extracted through deep neural networks; then a pyramid-like structure of video feature expression is constructed, and then the feature expression of the entire video sequence is obtained; finally, the video is classified and recognized by the classifier [7], [8]. A multi-source deep neural network model is constructed, which non-linearly represents image features from different sources, and estimates the human body's action posture by extracting high-level human joint features from video information [9]. By fine-tuning the deep convolutional neural network trained on the Imagen net dataset, several feature fusion strategies are proposed based on the idea of fusing the features of the convolutional neural network. Compared with the classic traditional methods, this method has improved the recognition rate by 5 to 6 percentage points, and the feature extraction time has been greatly reduced [10], [11]. Hoping to improve the application of deep learning methods in video classification and video human motion posture, a 3D VGG-Net model is proposed. By adding time dimension convolution on the basis of traditional two-dimensional convolution, two-dimensional convolution is extended to three-dimensional [12], [13]. The researchers replaced the 3*3 convolution kernel in the VGG-Net model with 3*3*3, and then directly performed convolution operations on the video stream. This work has changed the recognition mode of the traditional video human motion gesture recognition method, and directly realized the video end-to-end classification and recognition, but the training of the entire network is very time-consuming and memory resources [14], [15]. Based on the VGG-Net model, combined with the traditional optical flow characteristics, the optical flow graph is also regarded as an image, and a dual data stream deep convolutional neural network is proposed [16], [17]. By using the trajectory features in the traditional method, by tracking the trajectory of the optical flow, the convolutional features are concentrated in the areas where the motion is more significant, and then the extracted features are subjected to a one-step down-sampling operation to obtain the final convolutional feature map, Using linear SVM as a classifier for action recognition [18], [19]. It is proposed to use Recurrent Neural Network (RNN) to establish a time series model for video sequences. First, the convolutional neural network is used to extract the deep convolutional features of each frame of the video; then the features extracted by the convolutional neural network are sent to the length Time memory network [20]; finally train the convolutional neural network and the long-short-term memory network at the same time. The recognition rate of this model on the standard database is much higher than the traditional method. Considering that the video sequence exists in segments, different videos Segments contain different semantic information. Based on this idea, a deep segmentation model is proposed. When building a sequence model, the video is divided into many segments, and then features are extracted for each segment. The importance of the features of different video segments is not the same, so when building the model, the researchers also took into account the importance of the features of different segments [21]–[23]. The DSM method has obtained very good recognition results on multiple behavioral data sets. Deep learning Some simple non-linear models can be used to transform the original data into more abstract, more complex, higher-level expressions, and more complex functions can be learned through enough model combinations [24], which can be substituted The feature extraction process of video human motion gesture recognition. The core of deep learning is that the features of each layer are not designed by manual engineering, but a process of autonomous learning from data [25]. In recent years, deep learning has focused on human faces. Recognition, image recognition, speech recognition, natural language processing and many other fields and directions show strong and excellent learning ability. In the field of video human motion gesture recognition, deep learning can improve video human motion gesture recognition through automatic feature extraction The efficiency of this greatly simplifies the cumbersome feature engineering [26], [27]. Convolutional neural networks are usually used to process data in the form of multiple arrays, such as color images composed of three two-dimensional arrays, which contain three color channels. Pixel intensity, which is commonly referred to as "R, G, B" three color channels [28]. Many data forms are multi-dimensional arrays, for example, language signals are one-dimensional signal sequences, and image audio signals are two-dimensional data signals. Video and three-dimensional images belong to three-dimensional data signals. Convolutional neural network is a deep feedforward artificial neural network, which has four key ideas, namely local connection, weight sharing, pooling sampling, and the superposition of multiple processing layers., It has significant performance in image processing [29]–[31]. In recent years, convolutional neural networks are often used in the research of video human motion gesture recognition, and its research results also prove that convolutional neural networks are very suitable for processing video human bodies. Motion gesture is recognition data [32]–[34]. Judging from the current development of video human motion gesture recognition, the feature extraction step may no longer have major innovations and changes, and choose to use convolutional neural networks to jump After this step [35], [36], letting the computer automatically learn the characteristics it needs is the development requirement and trend of behavior recognition. In the work, based on smart phone the human body's daily behavior recognition is the background, and the convolutional neural network is optimized to make it more suitable for the behavior recognition data to carry out detailed experiments, comparative studies and discussions. Through the above discussion, it can be found that the video human motion gesture recognition algorithm based on deep learning not only has excellent generalization ability than the traditional video human motion gesture recognition algorithm, but also avoids the cumbersome process of traditional manual feature extraction. The model performs feature learning instead of manual feature selection based on prior knowledge, and achieves better accuracy than traditional manual features. Therefore, it is very valuable to conduct research on human body recognition algorithms based on deep learning.

Firstly, extract the CNN features of the video frame; secondly, in order to better describe the object, the extracted CNN features are down-sampled in the training image to enhance their feature expression ability. Video sequence representation based on the LDA topic model: first generate a visual vocabulary for each frame of the video to construct a visual dictionary; then use the LDA topic model to describe the video topic, so as to obtain the topic histogram of each video, complete the coded representation of the video sequence. The preprocessing process is of the video human motion gesture recognition data set in the UCI public database. It explains in detail several experimental schemes designed, describes the CNN architecture in each experimental scheme, clarifies the reason and purpose of the design, and the principle of the designed CNN architecture and network parameter settings. The definition of basketball posture, the corresponding definition of the actions to be classified in basketball, to prepare for the recognition of basketball actions; next is the data division part, first analyze the basketball actions, and then summarize the data according to the analysis results Classification method: Finally, according to the characteristics of the different waveforms of the upper and lower limbs of the basketball action, the corresponding key data is extracted, and the classification method is designed to complete the gesture recognition of the basketball action.

II. VIDEO HUMAN MOTION GESTURE RECOGNITION ALGORITHM BASED ON TRAINING IMAGE CNN FEATURES AND LDA TOPIC MODEL

A video human motion gesture recognition algorithm framework based on CNN features of training images is constructed. The main steps of the framework include: the first step is to input each frame of the video into a pre-trained CNN model to extract the output feature map; the second step, in order to overcome the interference caused by the target's scale, it is proposed A training graph pooling strategy. Based on the extracted CNN feature graphs, a CNN feature of the training graph is constructed for each input image; the third step is to be able to dig out the potential semantic relationships in the video through the LDA topic model Obtain the "topic" of each visual feature, and then obtain the topic histogram of the entire video; the fourth step, based on the topic characteristics of the video, use the SVM classifier to classify. The convolution part of the pretrained model is drawn as two left and right blocks, and the feature map calculated on the left layer is separated, but the data used in the previous layer depends on the connected dotted line, as shown in the first layer and second layer after the input layer The border between them is separated, which means that the 128map on the right layer is calculated by the 48map on the layer. The frame diagram is shown in Figure 1:

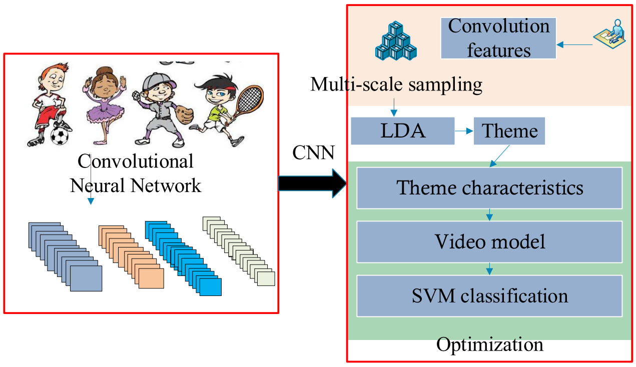

FIGURE 1. Video human motion gesture recognition algorithm framework based on CNN features of training images.

A. TRAINING IMAGE CNN FEATURE CONSTRUCTION

The VGG-16 deep convolutional neural network includes 13 convolutional layers, 6 maximum pooling layers and 3 fully connected layers. The 13 convolutional layers are divided into 6 convolutional blocks, and each convolutional block is followed by a maximum pooling layer. The first two fully connected layers have 4,098 channels. The third fully connected layer has 1000 channels for classification. The activation functions of all hidden layers use the rectified linear unit ReLU. Table 1 is the network configuration parameters of each layer of VGG-16.

After comparison experiments, the output of the last fully connected layer (4096 dimensions) of VGG-16 was not selected as the CNN feature map, but the output feature map of the last convolutional layer of VGG-16 after maximum pooling (7*7*512 dimensions) as the CNN feature of the input video frame. Figure 2 shows the CNN feature visualization heat map obtained by extracting some video frame sample pictures on the VGG-16 in the experimental data set. The darker part of the figure indicates that the feature response is relatively strong. It can be seen from the visualized heat map that a large number of activation areas of the feature map are concentrated on the human object, which

TABLE 1. Network configuration of VGG.

Input (224 x 224 RGB picture)
conv3-48
conv3-48
Maxpool
conv3-64
conv3-64
Maxpool
conv3-256
conv3-256
conv3-256
Maxpool
conv3-512
conv3-512
conv3-512
Maxpool
conv3-512
conv3-512
conv3-512
Maxpool
Fc-4098
Softmax

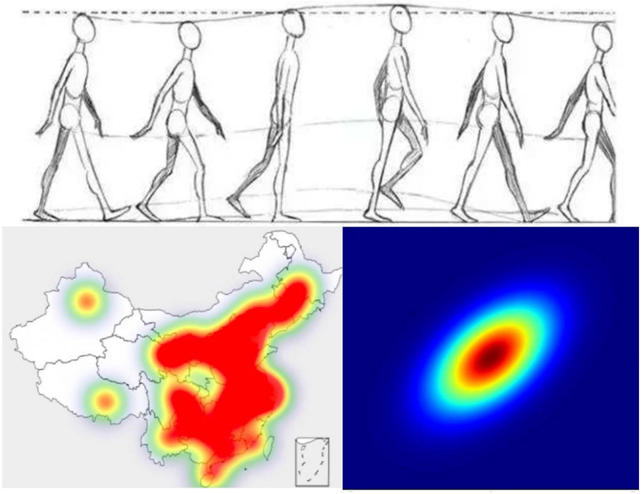

FIGURE 2. Visualized heat map of VGG-16 convolutional layer output.

proves that the VGG-16 model trained on ImageNet also has a good feature description ability on the video human motion gesture recognition data set.

Assume a convolutional neural network with K layers and graph J. The picture J is fed into the K-layer convolutional neural network, and feature maps obtained by convolution of different kernels can be obtained. Therefore, the convolution feature map can be expressed by the following formula:

$$G_j = \left\{G_{ji} : j = 1 \ldots M, i = 1 \ldots C_j\right\} \tag{1}$$

Among them, G_j represents the j-th feature map on the i-th convolutional layer, and C_j represents the number of convolution channels, that is, the number of convolution kernels. The global maximum pooling on the convolutional layer (seeking the maximum response value on the Jth convolutional feature map) can be expressed by the following formula:

$$N_{G_j} = \max(f_i(m_1.m_2)), m_1.m_2 \in G_j \tag{2}$$

The maximum pooling operation describes the local maximum response of the feature map, and the average pooling describes the overall response on the feature map. The research objects in the image generally appear in different shapes, spatial coordinates and scales. The pure pooling operation does not fully consider the spatial and scale information of the target object in the image, so this article proposes a multi-scale pool to overcome the changes in the scale space of the target. The definition of the multi-scale pooling strategy is as follows:

$$N_i = \left[N_{G_{i,j}} : j = 1 \ldots C_j\right] \tag{3}$$

B. VIDEO REPRESENTATION BASED ON LDA TOPIC MODEL

Generally, the length of each video in the behavioral video data set is different. Generally, classifiers require the same input dimensions, so the video length needs to be normalized. Among them, the most common approach is to use the bag-of-words model for normalization. Through the "bag of words" model, videos of different lengths can be converted into a word frequency vector of uniform length. However, the "bag of words" model does not take into account the relationship between vocabulary and vocabulary, that is, the semantic relationship. The LDA topic model is based on an improvement of the "bag of words" model, which can discover hidden topic information in documents or videos through semantic mining methods. So instead of just knowing which words appear, you can find those words related to the topic appearing in the document or video. The LDA topic model can overcome the interference caused by synonyms and polysemous words. For example, the word "apple" can represent both Apple mobile phones and fruits in daily life. If the word "apple" appears in an article about digital products with a high probability, the LDA topic model distinguish its meaning well.

Each frame of the video may be a certain basic action, such as kicking, raising a hand, or turning around. Complex actions are composed of a combination of these basic actions. In this article, these basic actions can be regarded as a "topic", using the LDA topic model to find the combination of basic actions (topics) of human behavior in the video. This is the topic-based video representation. The flowchart is shown in Figure 3. It can be seen from Figure 3 that in order to train the LDA topic model, two more steps are needed: first, building a visual dictionary of visual vocabulary; second, using Euclidean distance as a metric to match the multi-scale CNN features of each image with the closest match The visual words are mapped and quantified for the visual features of the pictures, and a word frequency matrix based on visual words is constructed for each video.

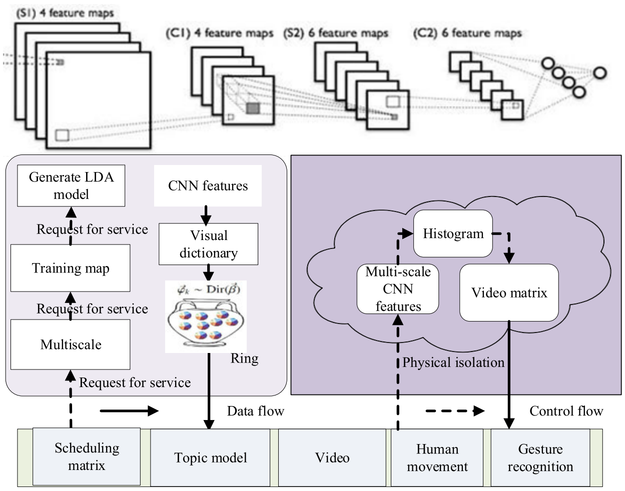

FIGURE 3. **Flow chart of topic-based video representation.**

When building a visual dictionary, you need to use the k-means algorithm. The general k-means algorithm is initialized randomly when initializing the cluster centers. There are two main problems with random initialization: it is easy to cause the algorithm to fall into a local optimal solution; improper initialization will cause the cluster centers to be too concentrated or too scattered, which will double the amount of calculation.

In order to solve this problem, an optimized cluster center initialization method is adopted. The specific process has the following steps: the first step is to randomly select a feature point from the feature set as the initial cluster center; the second step is to calculate the probability that the feature sample point 10,000 EX can be selected as the initial cluster center by formula, Find the sample point most likely to be the cluster center, that is, the sample point that makes the largest; the third step, repeat the second step until k cluster centers are found.

$$\Pr ob = \frac{Dis(x)}{\sum\limits_{i} Dis(x)} \tag{4}$$

Therefore, the entire algorithm flow of building a visual dictionary is shown in Table 2. With the built visual dictionary, the multi-scale CNN features extracted from each video can be mapped to a unique visual word through Euclidean distance and nearest neighbor matching algorithm. In other words, every visual feature, whether useful or not, will be mapped to a visual word, which also contains interference features. In order to reduce the interference of noise in visual features, when the visual features are mapped to visual words using Euclidean distance, a value is set to filter out the interference features. When the Euclidean distance between the visual feature and the nearest cluster center is greater than this value, the feature can be discarded, and the feature will be retained only when the Euclidean distance is less than this threshold. The determination of the threshold will directly affect the generation of the word frequency matrix, which will have a great impact on the training of the subsequent LDA model. The threshold value in this article is obtained by summing the distances from all samples to the cluster center, and then calculating the average value. This method can well ensure that the obtained threshold is more appropriate, will not cause some useful visual features to be filtered out, and at the same time can effectively filter out some interference features.

TABLE 2. **Visual dictionary construction process.**

Input: Feature sample point collection

1. Randomly select a data sample point from the feature sample set as the first cluster center.
2. Use the formula to calculate the probability of each feature sample point as the cluster center, and select the highest probability as the cluster center.
3. Repeat step 2 until k initial cluster centers are found.
4. Calculate the Euclidean distance from each sample point to each cluster center, and divide the sample points into various categories.
5. Recalculate each cluster center.
6.Calculate the Euclidean distance between every two cluster centers
7. Determine the distance relationship between the data point and the original cluster center. If the conditions are met, the calculation of the distance between the data point and the center of this class can be ignored. If it is not met, continue to calculate the distance between the data sample and the center of class.
8. Repeat steps 5, 6 and 7 until the cluster centers no longer change or the maximum number of iterations is reached.
9. At this time, the calculated cluster centers are regarded as the visual words in the visual dictionary, and the construction of the visual dictionary is completed.

Output: cluster centers are visual words

The training process of the smooth LDA model has four main steps: the first step is to randomly give a topic to each visual word in each video in the video set; the second step is to rescan the entire video set for each word, Using the Gibbs sampling formula, calculate the probability of each subject of the currently sampled visual word; the third step, update the subject of the current sampled visual word according to the obtained probability distribution; the fourth step, repeat the second and third steps until The distribution of topic-visual words and the distribution of video-topics converge or reach a preset number of iterations, and these two distribution matrices are output to obtain a trained LDA model.

C. RESEARCH ON GESTURE RECOGNITION ALGORITHM

With the continuous development of the information age, human-computer interaction has become a hot topic of current research. Gesture recognition is a very important branch in the field of human-computer interaction. The enhancement of human-computer interaction through gesture recognition has broad application prospects. The basis of attitude recognition is attitude collection. The related data fusion and attitude calculation methods based on inertial sensors have been specifically introduced in the previous section. Compared with traditional optical collection methods, this method requires less equipment and is portable. It has good performance and can adapt to posture acquisition in complex scenes. Compatible with the posture analysis and recognition system based on inertial sensors described in this article, this chapter uses posture angle as the input for posture recognition for posture recognition, mainly to distinguish the angle between each bone, combined with the posture from the PC side The intuitive posture obtained by the host computer is reconstructed, the main movements and postures are respectively recognized, and the relevant parameters such as the number of steps in the movement are recognized.

According to the data characteristics of the attitude angle in the process of static and motion, static and motion can be separated. This is a two-classification problem, that is, one or the other. Using the earth is as a reference, the recognition of postures first needs to distinguish two types of static and motion. For the motion state, the main recognition is the "walking" state and its derived parameters, such as the number of steps and pace, while the static state includes "standing", "Lying", "Lying","Sit". After such division, it can be recognized for each gesture. The collected data of the moving state and the static state is shown in Figure 4, and it can be seen that the data trend is obviously different, and the degree of dispersion of the data in the moving state is greater.

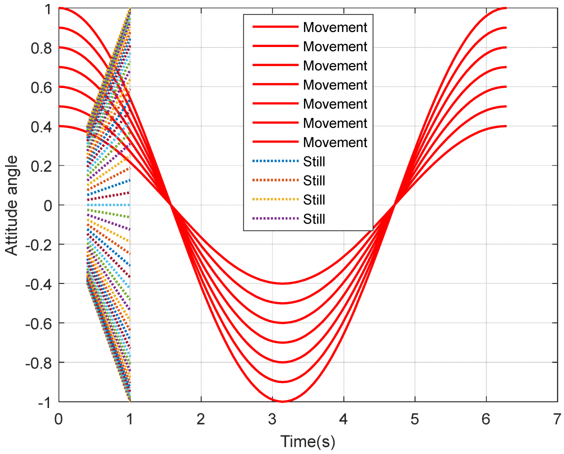

FIGURE 4. Comparison of data trends in motion and static state.

The data is windowed. The window length is 600 sampling points, and the time domain standard deviation is extracted. The following Figure 5 shows the standard deviation characteristics of the time domain data collection data. It can be seen from the table that as the posture changes, the corresponding the standard deviation presents a big difference, and this difference can be used to separate static and motion.

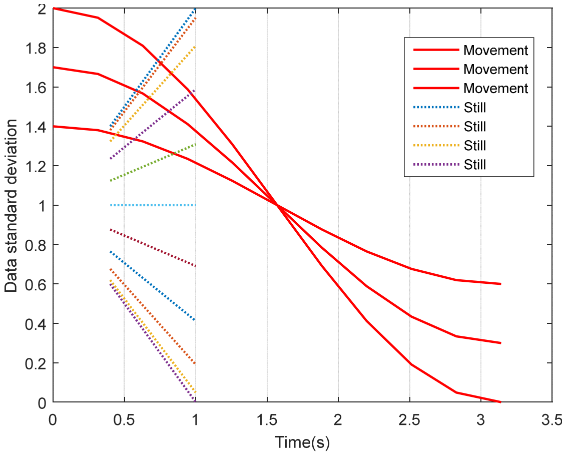

FIGURE 5. Time-domain standard deviation curve in motion and static state.

The recognition of the motion state is slightly different from the static state. The motion state presents higher dynamics and higher requirements for the attitude angle. By observing the user's posture during exercise and dividing the gait cycle, it can be obtained that the angle collected by the sensor of the leg changes the most. By collecting data from the sensors installed on the left and right thighs and calves, the walking process can be Statistics of the number of steps, step frequency, and pace.

III. RESEARCH ON RECOGNITION ALGORITHM OF VIDEO HUMAN MOTION GESTURE

Although the collected original CNN feature value signal value or preprocessed CNN feature value signal value can be used in the sample set of the classifier training and gesture recognition in the human motion gesture recognition system, the original CNN feature value signal value can reflect The physical meaning of human body movement is relatively narrow, and the recognition rate is not optimistic. Feature extraction and selection are the "bridge" of the two processes of data collection and motion gesture recognition. In the human motion gesture recognition system, the original CNN feature value signal value is generally extracted and selected. At the same time, feature extraction and selection are also one of the research focuses in human motion gesture recognition. The purpose of feature extraction and selection is to describe the physical meaning involved in human movement and the representation of the object. Its main task is to extract and select the most effective feature vector for human motion gesture recognition from the original CNN feature value signal, so as to provide effective support for improving the classification and recognition rate of the system, so feature extraction and selection have a direct impact to the recognition performance of the entire system. The feature extraction methods of CNN feature value signals in human motion gesture recognition can be summarized into three categories: time domain analysis method, frequency domain analysis method and time-frequency mixed analysis method. Table 3 lists the more widely used features.

TABLE 3. List of common features.

Feature type	Feature name
Time domain characteristics	average value
	Variance or standard deviation
	Root mean square
	Signal amplitude area
	Signal amplitude vector
	Correlation coefficient between two axes
	Energy
	Mean absolute deviation
	Time domain integration
	Posture direction
Frequency domain characteristics	FFT coefficient
	Frequency domain
	Energy spectral density
	Spectral coefficient
Time-frequency characteristics	Wavelet coefficient

The time domain analysis method is to analyze the CNN feature value signal from the signal time domain, and extract the corresponding valuable features. This method is relatively simple to extract. As shown in Table 3, the most commonly used time-domain features are: mean, root mean square, variance or standard deviation, signal amplitude area, correlation coefficient between two axes, energy, etc. Among them, many time-domain features are extracted based on the physical meaning of the CNN feature value signal. For example, the correlation coefficient between the axes reflects the correlation between the signals of each two coordinate axes, and the root mean square reflects the value of the resultant force of the three-dimensional CNN feature values. The frequency domain analysis method is to analyze the CNN feature value signal from the angle of the signal frequency domain, and extract the frequency domain feature from the CNN feature value signal after the fast Fourier transform. As shown in Table 3, the commonly used frequency domain features include FFT coefficients, frequency domain direct and so on.

There are many kinds of classification algorithms for human motion gesture recognition, and the data acquisition devices and acquisition schemes are also very different. Therefore, the extraction of CNN feature value signals cannot be standardized by a unified standard. Generally, in response to different types of motion postures and recognition requirements, researchers must design features that are more compatible with the relevant recognition model in order to achieve better recognition performance of the system.

A. GESTURE RECOGNITION CLASSIFICATION ALGORITHM

The quality of the classification algorithm is directly related to whether the recognition ability of the entire recognition system can meet the requirements, and whether the recognition of human motion posture can be correctly realized. For video human motion gesture recognition based on the CNN features of the training image, a statistical-based pattern recognition method is usually used. This method generally requires a large number of samples for training and learning to establish a corresponding recognition model. The recognition model is used to recognize the test samples. For human motion gesture recognition, classification algorithms are mainly used. Simple and effective classification algorithms include: decision trees, k nearest neighbors.

Decision tree is a hierarchical structure composed of nodes and directed edges. The tree contains three kinds of nodes: root node, internal node, leaf node, there is one and only one root node, which is a collection of all training data. Each internal node in the tree divides the data set reaching the node into 2 or several blocks according to a specific attribute. Each end node is a clearly classified data collection. The specific process is as follows:

(1) The tree starts from a single node N, where N represents the training tuple in D.

(2) If the tuples in D is all of the same class, then node N is set as a leaf node and marked with this class. Otherwise, calculate the information gain of each attribute according to the training set D, select the attribute with the largest information gain as the split attribute, and calculate its corresponding split point (split condition).

(3) Test node N with D to generate different branches.

(4) For the tuples on each branch of the classification result of D, the algorithm uses the same process to recursively form a decision tree.

(5) The recursive division step only stops when one of the following termination conditions is established:

a. All tuples of partition D (provided at node N) belong to the same category;

b. There are no remaining attributes available for further division of tuples. At this time, using the majority voting method, convert N into a leaf node, and mark it with the majority class in D.

c. The given branch has no tuples, that is, the number of branches is empty. At this point, create a leaf node with the majority class in D.

(6) Return the result decision tree.

The k-nearest neighbor method is a supervised pattern recognition method. When a new sample appears, first calculate the distance between the new sample and the sample data in each category, and then find out the k nearest neighbor sample data. In the k data, which category of samples account for the vast majority, the new sample is judged as which kind. The most commonly used measure of distance between samples is Euclidean distance. Any instance x is composed of n feature vectors: $(b_1(x), b_2(x) \ldots b_n(x))$ represents the r-th attribute value of instance x. The distance between two instances is defined as:

$$d(x_1, x_2) = \sqrt{\sum_{r=1}^{n} b_r(x) - b_r(x_j)} \tag{5}$$

Then, the nearest sample among the k nearest neighbor samples is selected as the category of the unknown sample, and the k nearest neighbor method can also be used to guess a true value of the unknown sample.

The general steps of the k-nearest neighbor method are as follows:

(1) Construct training set D.

(2) Choose the value of k. The k value can be completely artificially selected. Generally, an optimal k value can be obtained through multiple tests. The k value plays a vital role, and the k value may directly affect the classification effect.

(3) Calculate the distance between the sample to be classified and each sample in the training set D.

(4) The calculated distance values are arranged in ascending order, and the k nearest training samples are selected as the approximate samples to be classified. The value of k directly determines how many samples are selected for category statistics.

(5) Calculate the proportion of each category in the selected k samples, and use the majority voting method to search for the optimal attribution category. The class with the largest ratio will be selected as the ideal class, that is, the class of the sample to be classified.

B. IMPROVED RANDOM FOREST'S RECOGNITION OF HUMAN MOTION GESTURE

Recognition algorithm plays a vital role in video human gesture recognition, and the quality of the algorithm directly affects the recognition effect. As a combination algorithm, the random forest algorithm has better results than the base classifier, but it has problems such as static and easy to fall into local optimality. A random forest algorithm for bee mating optimization is proposed and applied to the algorithm based on Video human motion gesture recognition based on CNN features in training images. Using this algorithm to recognize 5 daily behaviors and 1 abnormal behavior of the human body has achieved very good results.

Due to the introduction of two random ideas, the random forest algorithm avoids the problem of overfitting. The classification accuracy is higher than its base classifier, but at the same time it makes the algorithm less stable. On the other hand, because the random forest algorithm is a static algorithm, the classification model cannot be adjusted once it is trained. Therefore, the algorithm is easy to fall into a local optimal solution, and the global optimal solution cannot be obtained. In view of this, this chapter proposes a random forest algorithm for bee mating optimization. The algorithm introduces the intelligent optimization algorithm of bee mating to dynamically change the decision tree in the random forest to enhance the diversity of the random forest and improve the classification ability and stability of the random forest. In order to prevent the algorithm from overfitting the data set, this chapter divides the original data set X into three equal parts: training set D, validation set M, and test set T. The training set D is used to train the decision tree to form a random forest (bees), and then the initial bee colony is generated; the validation set M is used to calculate the fitness value of the initial bee colony, and is used to raise young bees by the worker bees; the test set T is used to calculate the bees The fitness during the mating process and the performance of the queen bee (the optimal random forest) in each iteration. The young bee inherits the excellent genes of the queen bee, so the probability of the offspring population evolving to the optimal solution becomes greater. On the other hand, in order to avoid premature maturity, the mutation operator is used to represent the worker bees, and the worker bees are used to generate new solutions for each young bee, so as to maintain the diversity of the evolutionary population and strengthen the algorithm's global optimization ability. Due to the introduction of multiple random forests, and each iteration will adjust the decision tree between the forests and introduce a new decision tree, the performance of the new queen bee is better than or equal to the previous generation queen bee, so after mfN iterations, the best performance queen bee will be produced, And the stability of the algorithm is strong. The specific generational evolution process of the algorithm is shown in Figure 6.

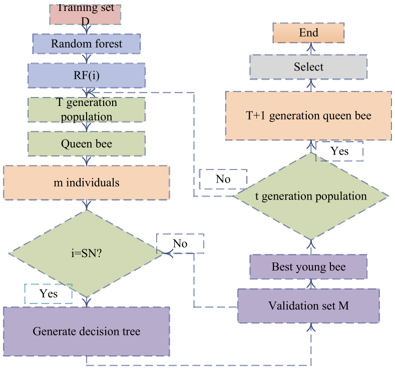

FIGURE 6. Schematic diagram of HBMORF algorithm.

The human body gesture recognition process is as follows: first collect the CNN feature value signals generated by the tester during exercise, identify behavior markers, check the collected data, and select valid data; then perform feature extraction of these CNN feature value signals to form feature vectors; Finally, use these feature vectors to train the model and classify and recognize.

Feature selection uses a series of rules to obtain the relative relationship of the importance of features. Random forest implements feature ranking by measuring the importance of features. Commonly used techniques include: 1) Count the number of times each feature is used as a segmentation feature, and use this frequency to indicate its importance; 2) When building a decision tree, the Gini index method is used to measure the segmentation effect of nodes, that is, the importance of features is indicated by calculating features.

The random forest algorithm is used to conduct feature vector use frequency experiments on the data set DataB. When the accuracy rate is guaranteed to reach more than 90%, the use frequency of random forest attributes is shown

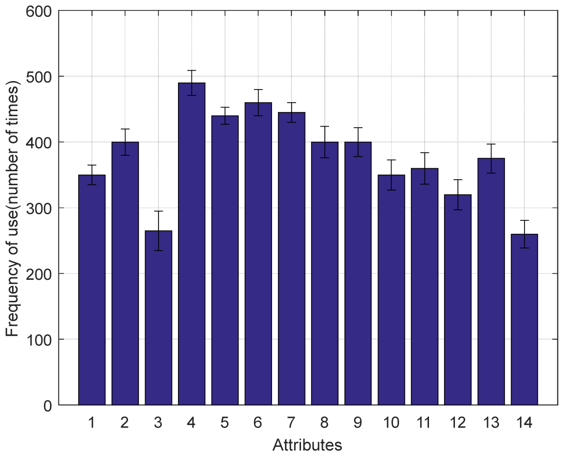

FIGURE 7. Attribute usage frequency chart.

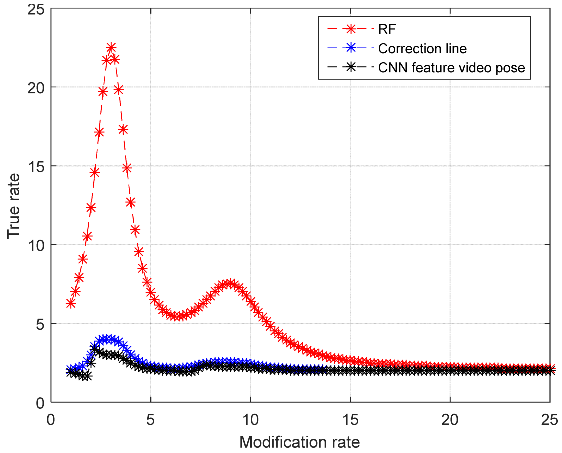

FIGURE 8. ROC curve of RF and CNN feature video attitude algorithm.

in Figure 7. As can be seen from the figure, APS and FRS account for a relatively large proportion. According to statistics, the total frequency of use of all attributes is 5,312, and the frequency of use of APS and FRS is 2,620, accounting for 49.3. Among them, “Sx is used most frequently, APSz is the second, and both APS and FRS are used more frequently than other features. When a feature is used as a segmentation feature, the more important the feature is. Therefore, APS and FRS Group feature vector is of great significance to human body gesture recognition.

The receiver operating characteristic (ROC) curve is an effective tool for comparing two classification models. The curve shows the trade-off between the true rate (TPR) and false positive rate (FPR) of a given model. The RF and CNN feature video pose algorithms calculate the class prediction probability when performing optimal classification voting, and the model gives the class prediction probability of each test tuple. Randomly select 50 test tuples (including 25 positive tuples (falling) and 25 negative tuples (walking)) from the data set DataA for classification, sort the test tuples in descending order of probability, and calculate the tuples TPR and FPR, the ROC curve obtained after 5th-order polynomial curve fitting is performed on the requested data, as shown in Figure 8. The figure shows the ROC curve of the RF and CNN feature video pose algorithms, and the correction line represents random guessing. The closer the ROC curve of the model is to the correction line, the lower the accuracy of the model. It can be seen from the figure that the CNN feature video attitude algorithm is far from the correction line than the RF algorithm, and the CNN feature video attitude algorithm curve begins to encounter real example tuples, and as the tuples move to higher numbers, the curve rises steeply. Later, the real there are fewer and fewer case tuples, and more and more false positive tuples, and the curve is flat and becomes more horizontal. It can be seen that compared with the RF algorithm, the CNN feature video pose algorithm is more accurate and the classification prediction effect is better.

In order to test the effect of bee colony size on the performance of the algorithm, we used the glass dataset to do ten experiments on each bee colony size, and obtained the maximum accuracy of the algorithm and the relationship between the average accuracy and the colony size, as shown in Figure 9.

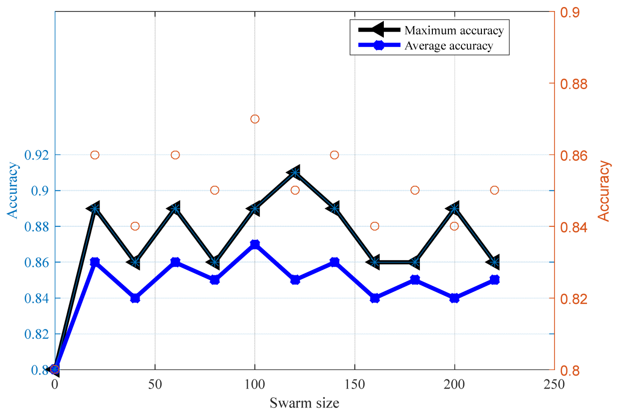

FIGURE 9. Relationship between population size and accuracy.

It can be seen from Figure 9 that the accuracy of the CNN feature video attitude algorithm fluctuates with the change of the colony size. Before the colony size reaches 100, as the colony size increases, the overall accuracy rate shows an upward trend; After it is 100, as the colony size increases, the accuracy rate decreases as a whole; when the colony size is 100, the maximum accuracy and average accuracy of the algorithm are both maximum. So this chapter sets the population size to 100.

In order to further verify the performance of the CNN feature video pose algorithm, 10 independent experiments were carried out according to Experiment 1 and Experiment 2, each time with a different random seed to obtain the training set and attribute subsets to generate different initial bee colonies. The average accuracy rate is obtained from the results of 10 experiments, and the experimental results

TABLE 4. Comparison of classification accuracy of various classification algorithms.

Accuracy	Experiment one			Experiment two		
	Smallest	Average	Maximum	Smallest	Average	Maximum
SVM	89.792	92.331	93.563	82.423	84.188	85.14
RF	94.4642	94.9512	95.2746	85.0756	85.9929	87.0125
CNN feature video pose algorithm	95.8965	96.2862	96.5786	92.5383	92.941	33.409

are compared with the RF and SVM algorithms, as shown in Table 4.

From the experimental results in Table 4, we can see that the improved random forest algorithm has a great improvement in accuracy. In experiment 1, the maximum accuracy rate reaches 96.5786%, which is 1.36% higher than the maximum accuracy rate of the RF algorithm, and is better than SVM 3.24% higher. In the second experiment, the maximum accuracy of the CNN feature video pose algorithm is 7.37% higher than that of the RF algorithm. And compared with the maximum accuracy of Experiment 1, the difference is only 3.18%, while the maximum accuracy of the RF algorithm is 8.26%. This shows that the CNN feature video gesture algorithm is not only accurate, but also more robust for human gesture recognition. In Experiment 1 and Experiment 2, the minimum accuracy of the CNN feature video attitude algorithm is much higher than that of SVM and RF calculation, and the difference between the minimum accuracy rate and the maximum accuracy rate is very small, indicating that the CNN feature video attitude algorithm has strong stability. The human body gesture recognition of different testers has strong adaptability.

In order to further analyze the classification results and obtain the recognition rate of each class, the confusion matrix related to the maximum accuracy of the RF and CNN feature video pose algorithms in the experiment is analyzed, as shown in Table 5, Table 6, Table 7, and Table 8. Among them, the row represents the category of the original data, and the column represents the predicted category.

TABLE 5. Confusion matrix of the classification results of the RF algorithm in experiment 1.

Class	Walking	Falling	Lying	Sitting	Standing up	Sitting down	Total	Recognition rate%
Walk	406	16	0	13	0	3	424	96.2185
fall	7	0	8	5	0	0	31	46.8752
Lie flat	2	0	731	2	0	0	733	99.7269
Sit still	11	0	4	362	6	0	378	96.0423
stand up	8	0	1	3	3	10	18	36.8422
Sit down	8	2	4	1	1	2	24	39.1305
Total	442	19	748	387	10	15	1609	95.2737

TABLE 6. Experiment 1 confusion matrix of CNN feature video pose algorithm classification results.

Class	Walking	Falling	Lying	Sitting	Standing up	Sitting down	Total	Recognition rate%
Walk	411	1	0	12	0	0	424	96.9257
fall	4	21	5	2	0	0	31	68.7501
Lie flat	1	0	731	2	0	0	733	99.7269
Sit still	6	3	0	368	7	1	378	97.8893
stand up	6	0	0	3	0	4	18	36.8422
Sit down	2	0	3	4	0	12	24	56.5216
Total	432	25	739	392	7	17	1609	96.5795

TABLE 7. Confusion matrix of the classification results of the RF algorithm in experiment 2.

Class	Walking	Falling	Lying	Sitting	Standing up	Sitting down	Total	Recognition rate%
Walk	467	12	0	4	0	0	424	96.6815
fall	13	10	7	2	0	0	31	36.6657
Lie flat	4	0	635	0	2	0	733	99.3731
Sit still	132	8	0	283	0	2	378	67.3758
stand up	3	0	6	0	2	0	18	8.3334
Sit down	9	0	4	9	0	1	24	8.6958
Total	624	32	652	298	4	3	1609	87.0015

From the comparison results of Table 5 and Table 6, Table 7 and Table 8, it can be seen that for the recognition rates of the six behaviors, the CNN feature video pose algorithm is higher than the RF algorithm. Experiment 1 and Experiment 2 used different sources of data sets for testing models, and physical differences between testers and other reasons will cause the CNN feature rules of the collected training images to be different. It can be seen from Table 5 and Table 7 that the recognition rate of the RF algorithm for falling behavior is less than 51%, while the recognition rate of the CNN feature video gesture algorithm can reach about 61%; the recognition rate for standing and sitting is lower than 41%, while the recognition rate of the CNN feature video gesture algorithm can reach about 51%. The recognition rate of falling is low, mainly because it is

TABLE 8. Experiment 2 confusion matrix of CNN feature video pose algorithm classification results.

Class	Walking	Falling	Lying	Sitting	Standing up	Sitting down	Total	Recognition rate%
Walk	468	7	0	4	1	1	424	96.6815
fall	8	15	5	0	0	0	31	36.6657
Lie flat	3	0	635	0	3	2	733	99.3731
Sit still	50	0	0	367	0	3	378	67.3758
stand up	1	0	1	2	4	0	18	8.3334
Sit down	7	3	0	6	0	11	24	8.6958
Total	537	25	641	380	8	17	1609	87.0015

wrongly recognized as walking and lying down behaviors, because there are some noise data in the data collection process; standing up and sitting down are mainly wrongly recognized as walking and sitting down behaviors, because these behaviors are some It is close, but the comparison of the results shows that the CNN feature video pose algorithm has better processing capabilities for noise data.

IV. CURVE FITTING CNN FEATURES AND K-NEAREST NEIGHBOR VIDEO HUMAN MOTION POSE RECOGNITION

In the research of human body motion gesture recognition, researchers generally tend to study the CNN feature data of the human behavior training map at each moment, and judge the type of human body motion pose at this moment. But the posture of human body movement is not a task that can be completed in an instant, but a coherent task with the nature of time series. Although the human body movement posture can be expressed as a frame in an animation, it can also be defined as a kind of movement posture, but each frame can more accurately reflect the transition state or final state of a movement posture. In this chapter, it is proposed to use polynomial least squares curve fitting to fit the CNN feature data of the training map, and then calculate the curve similarity between each curve, and finally combine the k-nearest neighbor algorithm to recognize the human body movement posture. Experimental results show that the method achieves a good recognition effect.

The k-nearest neighbor classification algorithm is one of the simplest machine learning algorithms. In this chapter, the CNN feature data of the training image of human motion posture is fitted into a curve, the curve is used as the research object of the classifier, the concept of curve similarity is introduced, and a k-nearest neighbor algorithm based on curve similarity is proposed as the classification The device classifies the curve and judges the category of the human body movement posture.

The steps of the k-nearest neighbor algorithm based on curve similarity are as follows:

(1) Given s groups of equal-length known classes (the number of categories is l) training image CNN feature data (including X, Y, Z axis three directions), and curve fitting each group of data, a total of 3*s fitting curve, the fitting curve is coded by axis direction and group number. And given a set of CNN feature data (that is, the data to be classified) of the training image of an unknown class and perform curve fitting, three fitting curves are obtained.

(2) Select the initial k value ((k value is optimized through multiple tests).

(3) Calculate the curve similarity between the fitting curve in a certain axis direction of the data to be classified and the fitting curve in the corresponding direction of each known class.

(4) Arrange the required similarity values in ascending order, and select the first k curves.

(5) Use the k curves selected in step 4 to calculate the selected ratio of each category in each axis direction.

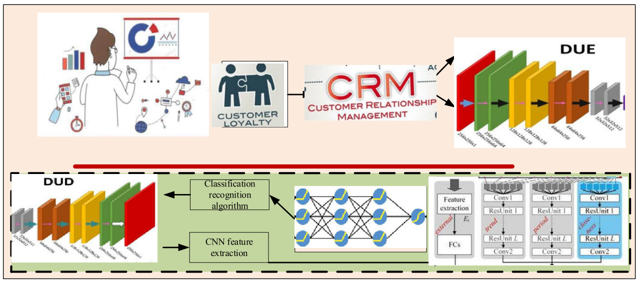

FIGURE 10. Flow chart of human pose recognition based on CNN features.

As shown in Figure 10, the main steps of human posture recognition based on CNN features are divided into three parts: First, preprocess the image data collected by the CNN feature sensing device, then extract the corresponding human posture image features, and finally use appropriate The classification algorithm performs posture classification and recognition. The application field of human gesture recognition technology based on CNN features is very wide. For example, in the field of games, using somatosensory technology to obtain the player's gestures and actions for human-computer interaction can free game players from the constraints of traditional game interactive devices, thereby greatly improving the player's gaming experience. In the medical field, natural limb movement detection through CNN features can effectively reduce the workload of medical staff, and the effect of patient rehabilitation training has also been effectively improved. In the field of education, by integrating human-computer interaction technology based on CNN features into classroom teaching, it provides teachers and students with a more natural way of human-computer interaction.

Extracting the effective features that describe the human posture from the image sequence is an important prerequisite to ensure accurate recognition of actions. The effects of different features will be greatly related to the characteristics of the target and the environment. The description ability of the same feature for different types of actions will be somewhat different, and the description ability of different features for the same category of actions is also uneven. Feature descriptions can be classified into global feature descriptions and local feature descriptions according to different characteristics, so how to use existing data Information and recognition targets to select appropriate feature descriptors are the key to human gesture recognition.

Global feature description is to adopt a top-down description method, which treats the recognition target as a whole. Global features cover comprehensive human body information. Because they are easily affected by image preprocessing such as the accuracy of recognition target positioning and background removal, global feature descriptions also have certain limitations, such as noise, occlusion, and camera viewing angle changes.

Local feature description is a bottom-up description method, that is, only useful parts of the moving target are extracted, and the observation target is regarded as a local descriptor or a collection of local image blocks. Compared with global features, local features will not change with changes in environmental background noise, object occlusion, or human motion, and they are also more stable to actions such as scale, translation, and rotation. Local feature extraction is generally divided into two parts: local feature area detection and local feature area description.

Dynamic time planning was first used in speech recognition, mainly for the recognition of isolated words. In gesture recognition, it can solve the problem of different time lengths for different targets to complete actions. DTW is a template matching algorithm. Through a given distance matrix, it finds a path from the upper left corner to the lower right corner so that the sum of the element values passed by the path is the smallest. In view of the problems of the DTW algorithm that has to plan a path every time it runs, there is a huge amount of calculation and a lot of space, etc., the DTW algorithm is improved, and a new global path window is proposed, which reduces the amount of calculations without reducing Correct rate, but has limitations for the problem of body occlusion. Aiming at the shortcomings of dynamic time warping in action recognition such as sudden changes in time structure and sensitivity to changes in illumination, this algorithm uses DTW to repeatedly randomly sample the random time warping formed by image sequence sampling, extracts the time elastic TE characteristics of sequence data, and then uses principal components Analyze dimensionality reduction to generate sequence subspace, and finally use linear discriminant analysis to complete gesture recognition, but the recognition rate for motion gestures is not high.

V. EXPERIMENTAL VERIFICATION

Experimental environment: software: operating system Windows XP, training image CNN feature video motion gesture recognition algorithm simulation environment: Matlab 7.1, compared classic algorithm simulation environment: Clementine 12.0. Hardware: Intel(R) Core(TM) 2 Duo CPU T6400@2.OOGHz, 2GB of memory. In order to verify the feasibility and effectiveness of the training image CNN feature video motion gesture recognition algorithm in human motion gesture recognition, a variety of recognition experiments are carried out in this chapter. The data involved in the experiment include the known data set series DataK01, DataK02, DataK03, the to-be-identified data set series DataR01, DataR02, DataR03 and the data set DataT. The related attributes are shown in Table 9.

TABLE 9. Description of the data set used in the experiment.

Data set	Number of data points	Number of categories	Containment behavior	Source
DataK01	4837	4	Walking, lying down, sitting still	Partial data of ABCD testers
DataR01	1809	4	Walking, lying down, sitting still	Part of E tester data
DataK02	1248	4	Fall, stand up, sit down	Partial data of ABCD testers
DataR02	304	5	Fall, stand up, sit down	Part of E tester data
DataK03	2378	7	Walk, fall, lie flat, sit still, stand up, sit down	Partial data of ABCD testers
DataR03	1000	6	Walk, fall, lie flat, sit still, stand up, sit down	Part of E tester data
DataT	202	5	Walk, fall, lie flat, sit still, stand up, sit down	Part A tester data

In order to verify the influence of the degree of the polynomial on the error of the fitting curve, the CNN feature data of the training map in the X-axis direction in the data set DataT is used to conduct experiments on the influence of the degree of different polynomials on the curve fitting error. Figure 11 shows the influence of the degree of the polynomial on R-square.

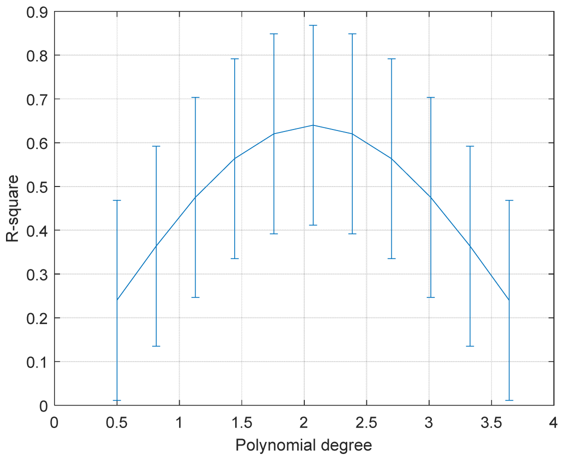

FIGURE 11. Influence of the degree of the polynomial on R-square.

The judgment criterion of the error parameter is: the closer the R-square is to 1, the better the curve fitting effect. In the experimental results, when the degree of the polynomial is 8, the value of R-square has reached 0.878, and then it grows slowly and close to steady, until the curve appears twists and turns at 2 o'clock, and finally begins to decline. Based on the judgment criterion, the preferred range of the polynomial degree in this chapter is. Analyzing the data in the optimal range, when the degree of the polynomial is about 2, R-squaxe is the closest to 1.

In this experiment, set the degree of the polynomial to 15, k to 3, set the number of data points gn included in each group of behaviors to 20, and set the X, Y, and Z axis directions to 1. Take the DataK01 data set as the known class, and DataT01 as the unknown class, and group each group of 20 data. The grouping results are shown in Table 10.

TABLE 10. Data set is grouped by GN = 20.

Data set	Number of groups	Number of each type of behavior		
		Walk	Lie flat	Sit still
DataK01	206	52	103	51
DataT01	83	23	38	20

Using the training image CNN feature video motion gesture recognition algorithm to classify the above behavior, the behavior recognition rate is 96.3415%. Further analyze the classification results to obtain the recognition rate of each classification, as shown in Table 11.

TABLE 11. Confusion matrix of long-term behavior classification time results.

	Walking	Lying down	Sitting still	Number of behaviors	Recognition rate%
Walk	22	38	1	3	91.6677%
Lie flat	1	1	21	38	99%
Sit still	2	2	2	22	95.2383%
Total	25	41	24	83	96.3425%

It can be seen from Table 11 that there are two groups of walking behaviors identified as lying down and sitting still. It can be inferred that the fluctuations of these two groups of behaviors are relatively stable, and are similar to lying down and sitting still, which leads to misjudgment. Another type of sitting behavior is judged to be walking. It may be that the tester has other actions during sitting, which may cause some deviations in the data.

In the training image CNN feature video motion gesture recognition algorithm, the parameters that affect the recognition effect mainly include the k value, the degree of the polynomial, and the number of data points contained in each group of behaviors gn. In this experiment, the DataK03 data set is used as the known class, and DataT03 is the unknown class. The data sets are grouped into groups of 10, 15 and 20. The grouping results are shown in Table 12.

Taking into account that the degree of the polynomial cannot exceed the number of data points in the group, the degree of the polynomial is 8 and gn is 10 and 15 respectively. The experiment is carried out. The experimental results are shown in Figure 12 Shown. It can be seen from the figure that when gn is 15, the behavior recognition rate is higher than the result when gn is 10. Considering the size of gn alone, it can be seen that a higher gn value is conducive to the recognition of human motion gestures, and the same The more CNN feature signals of the training image acquired by the action in the same time period, the more favorable the recognition of the human body posture, thereby improving the posture recognition rate. From the grouping results in Table 12, it can be seen that the number of discarded data points when gn is 15 is more than when gn is 10. In other words, if you want to reduce the proportion of discarded data, you need to increase the number of data points generated by each behavior. The number of CNN feature signals in the training image. However, due to the limitations of hardware devices and other factors, one behavior can only obtain a certain amount of training image CNN characteristic signals in a fixed period of time.

The parameter k value is an important parameter in the selection of k nearest neighbor algorithm, and the choice of k value will have a significant impact on the classification result. In order to verify the influence of the k value on the algorithm of this article, this article carried out three sets of experiments, respectively taking gn = 10, degree = 8; gn = 15, degree = 8; gn = 20, degree = 15. The choice of k value to the behavior recognition rate the results of

TABLE 12. Multiple grouping results of the data set.

Data set	gn	Number of groups	Number of each type of behavior					
			Walking	Falling down	Lying down	Sit still	Standing up	Sit down
DataK02	12	193	33	58	41	28	13	19
	16	124	21	38	29	18	4	11
	18	88	16	34	20	15	2	6
DataT01	11	90	21	14	25	21	4	4
	16	56	14	8	16	15	2	4
	23	40	11	8	11	10	1	1

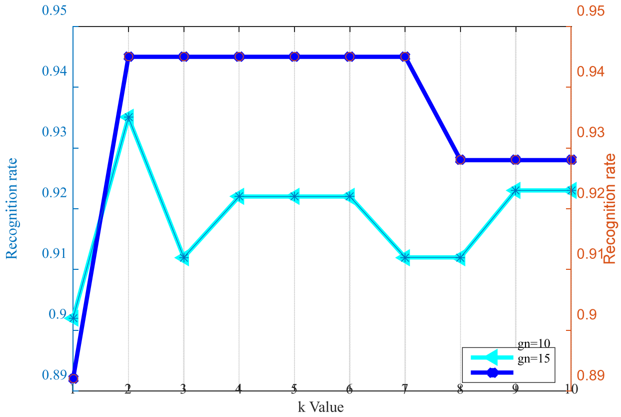

FIGURE 12. Influence of different gn values on behavior recognition rate.

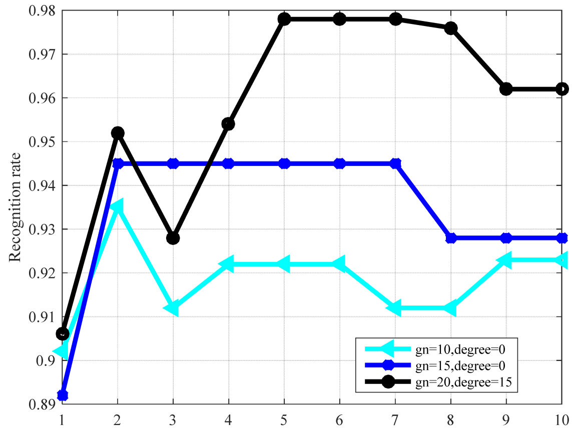

FIGURE 13. Influence of different k values on behavior recognition rate.

the impact are shown in Figure 13. It can be seen from the figure that although the recognition rate fluctuates as the value of k changes, the overall trend is that the recognition rate increases and then decreases as the value of k increases. The recognition rate changes with different gn and degree values. The behavior recognition rate of the three groups of experiments reaches the highest value when k = 5,6, but when gn = 15, degree = 8, it has reached the highest value when k = 3 Recognition rate.

Taking the CNN feature data of the fixed grouping training graph for each behavior as the research starting point, the polynomial least squares method is used for curve fitting, and the curve is used as the research object of human motion posture recognition. Each group of behaviors is fitted with three curves on the X, Y. Z axis., Using k-nearest neighbor algorithm as the classifier, introducing curve similarity as the criterion for distance judgment in the algorithm, closely combining human motion gesture recognition and analyzing the recognition method in all aspects, and finally determining the optimal k value and polynomial for parameters such as the number of times and the number of data points contained in each group of behaviors, the recognition rate can reach over 96%.

Each time a different random seed is used to obtain a training set and a subset of attributes to generate a different initial bee colony. Calculate the average accuracy rate of 10 test results, and compare the test results with the RF and SVM algorithms, as shown in Table 13.

TABLE 13. Comparison of accuracy rates.

Accuracy	Experiment one			Experiment two		
	Min	Ave	Max	Min	Ave	Max
SVM	89.82	92.34	93.57	82.43	84.19	85.19
RF	94.43	94.92	95.23	85.13	85.98	87.12
Optimize random forest algorithm	95.92	96.38	06.45	92.54	93.45	94.3

VI. CONCLUSION

In this article, only the CNN method is used to extract the local features of human behavior data segments and recognize them. Because CNN extracts the local features of data information, and the artificial feature method extracts global features, it is considered that the features extracted by CNN can be combined with artificial statistical features, and then this combined feature is used for classification Recognition. The six simple human behaviors are classified and studied, namely walking, going up stairs, going down stairs, sitting still, standing still, and lying still. Their behaviors are not only simple but also single. It hoped to recognize some more complex behaviors, such as running, cycling, squatting and tying shoelaces, climbing mountains and even various sports. It is hoped that human behavior recognition can give full play to its characteristics and be better applied in various fields. In the human motion gesture recognition system using the improved random forest as the classifier, the object of study is the data obtained by extracting the feature vector from the feature signal value of the training image CNN. In real life, the human body's behavior and posture will exist in the form of time period, including multiple training image CNN characteristic signal values. Taking this as the starting point of the research, the polynomial least squares method is used for curve fitting, and the curve is used as the research object of human motion gesture recognition, and the curve training set and the curve test set are established. The simple and effective k-nearest neighbor algorithm is used as the classifier, the curve similarity is introduced as the criterion for the distance between the two curves in the algorithm, and the optimal classification is selected by voting. The recognition method closely combines the research direction of human motion gesture recognition, and designs multiple experiments to analyze the recognition method in all aspects, and determine the optimal k value, the number of polynomials, and the number of data points contained in each group of behaviors, etc. Parameters, the recognition rate can reach more than 96%. And compared with a variety of classic classification algorithms, the recognition rate is significantly higher than these algorithms, indicating that the recognition method is feasible and very effective.

REFERENCES

[1] M. Zadghorban and M. Nahvi, "An algorithm on sign words extraction and recognition of continuous Persian sign language based on motion and shape features of hands," *Pattern Anal. Appl.*, vol. 21, no. 2, pp. 323–335, May 2018.

[2] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[3] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognit.*, vol. 76, pp. 80–94, Apr. 2018.

[4] P. Ji, C. Wu, H. Li, and X. Xu, "Vision-based posture recognition using an ensemble classifier and a vote filter," *Proc. SPIE*, vol. 157, pp. 101571J–101584J, Oct. 2016.

[5] Q. Chen, Y. H. Chen, and W. G. Jiang, "The change detection of high spatial resolution remotely sensed imagery based on OB-HMAD algorithm and spectral features," *Guang Pu Xue Yu Guang Pu Fen Xi= Guang Pu*, vol. 35, no. 6, pp. 1709–1714, Jun. 2015.

[6] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, Z. Ma, and J. Song, "Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model," *Pattern Recognit. Lett.*, vol. 119, pp. 187–194, Mar. 2019.

[7] B. Xie, X. He, and Y. Li, "RGB-D static gesture recognition based on convolutional neural network," *J. Eng.*, vol. 2018, no. 16, pp. 1515–1520, Nov. 2018.

[8] M. Arsalan and A. Santra, "Character recognition in air-writing based on network of radars for human-machine interface," *IEEE Sensors J.*, vol. 19, no. 19, pp. 8855–8864, Oct. 2019.

[9] S. Huo, T. Hu, and C. Li, "Improved collaborative representation classifier based on l2-regularized for human action recognition," *J. Electr. Comput. Eng.*, vol. 2017, pp. 1–6, Nov. 2017.

[10] J. Wang, T. Zheng, and P. Lei, "Hand gesture recognition method by radar based on convolutional neural network," *J. Bjing Univ. Aeronaut. Astronaut.*, vol. 44, no. 6, pp. 1117–1123, Jun. 2018.

[11] E. Tsironi, P. Barros, C. Weber, and S. Wermter, "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition," *Neurocomputing*, vol. 268, pp. 76–86, Dec. 2017.

[12] S. Y. Kim, H. G. Han, J. W. Kim, S. Lee, and T. W. Kim, "A hand gesture recognition sensor using reflected impulses," *IEEE Sensors J.*, vol. 17, no. 10, pp. 2975–2976, May 2017.

[13] C. C. D. Santos, J. L. A. Samatelo, and R. F. Vassallo, "Dynamic gesture recognition by using CNNs and star RGB: A temporal information condensation," *Neurocomputing*, vol. 400, pp. 238–254, Aug. 2020.

[14] L. Chen, J. Fu, H. Li, B. Zheng, and Y. Wu, "Hand gesture recognition using compact CNN via surface electromyography signals," *Sensors*, vol. 20, no. 3, pp. 672–683, Jan. 2020.

[15] S. K. Leem, F. Khan, and S. H. Cho, "Detecting mid-air gestures for digit writing with radio sensors and a CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1066–1081, Apr. 2020.

[16] X. Zhang and X. Li, "Dynamic gesture recognition based on," *Future Internet*, vol. 11, no. 4, pp. 91–102, Apr. 2019.

[17] N. Dawar, S. Ostadabbas, and N. Kehtarnavaz, "Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition," *IEEE Sensors Lett.*, vol. 3, no. 1, Jan. 2019, Art. no. 7101004.

[18] G. Zhu, L. Zhang, P. Shen, J. Song, S. A. A. Shah, and M. Bennamoun, "Continuous gesture segmentation and recognition using 3DCNN and convolutional LSTM," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1011–1021, Apr. 2019.

[19] J. Zhang, K. Shao, and X. Luo, "Small sample image recognition using improved convolutional neural network," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 640–647, Aug. 2018.

[20] Y. Hu, Y. Wong, Y. Du, M. Kankanhalli, W. Geng, and W. Wei, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PLoS ONE*, vol. 13, no. 10, pp. 206049–206056, Oct. 2018.

[21] W. Zhuang, Y. Chen, B. Wang, C. Gao, and J. Su, "Design of human activity recognition algorithms based on a single wearable IMU sensor," *Int. J. Sensor Netw.*, vol. 30, no. 3, pp. 193–203, Jan. 2019.

[22] J.-M. Liu and M.-H. Yang, "Recognition on images from Internet street view based on hierarchical features learning with CNNs," *J. Inf. Technol. Res.*, vol. 11, no. 3, pp. 62–74, Jul. 2018.

[23] J. H. Kim, G. Batchuluun, and K. R. Park, "Pedestrian detection based on faster R-CNN in nighttime by fusing deep convolutional features of successive images," *Expert Syst. Appl.*, vol. 114, pp. 15–33, Dec. 2018.

[24] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Comput. Vis.*, vol. 12, no. 1, pp. 3–15, Feb. 2018.

[25] M. J. Dong and Y. Yuan, "3D human feature recognition based on speeded-up robust features algorithm," *J. Beijing Inst. Clothing Technol.*, vol. 38, no. 4, pp. 37–44, Apr. 2018.

[26] N. Oukrich, C. E. Bouazzaoui, and A. Maach, "Human activities recognition based on autoencoder pre-training and back-propagation algorithm," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 19, pp. 5194–5202 Oct. 2017.

[27] A. A. P. Cattaneo and E. Boldrini, "You learn by your mistakes. Effective training strategies based on the analysis of video-recorded worked-out examples," *Vocations Learn.*, vol. 10, no. 1, pp. 1–26, May 2016.

[28] A. Scharl, A. Hubmann-Haidvogel, A. Jones, D. Fischl, R. Kamolov, A. Weichselbraun, and W. Rafelsberger, "Analyzing the public discourse on works of fiction–detection and visualization of emotion in online coverage about HBO's game of thrones," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 129–138, Jan. 2016.

[29] H. Cho and S. Yoon, "Divide and conquer-based 1D CNN human activity recognition using test data sharpening," *Sensors*, vol. 18, no. 4, pp. 1055–1067, Apr. 2018.

[30] T. Y. Zhong, W. P. Liu, and P. J. Liu, "A forest fire smoke detection algorithm based on fractional-calculus video fusion," *J. Bjing Forestry Univ.*, vol. 39, no. 3, pp. 24–31, Mar. 2017.

[31] L. K. Larkey, D. J. Roe, K. L. Weihs, R. Jahnke, A. M. Lopez, C. E. Rogers, B. Oh, and J. Guillen-Rodriguez, "Randomized controlled trial of qigong/tai chi easy on cancer-related fatigue in breast cancer survivors," *Ann. Behav. Med.*, vol. 49, no. 2, pp. 165–176, Apr. 2015.

[32] R. Rosas-Romero, "Remote detection of forest fires from video signals with classifiers based on K-SVD learned dictionaries," *Eng. Appl. Artif. Intell.*, vol. 33, pp. 1–11, Aug. 2014.

[33] M. Siddiqi, R. Ali, M. Rana, E.-K. Hong, E. Kim, and S. Lee, "Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis," *Sensors*, vol. 14, no. 4, pp. 6370–6392, Apr. 2014.

[34] M. Tsuchiya, Y. Yamauchi, and H. Fujiyoshi, "Efficient learning method for human detection based on automatic generation of training samples with the negative-bag MILBoost," *IEEJ Trans. Electron., Inf. Syst.*, vol. 134, no. 3, pp. 450–458, Jan. 2014.

[35] N. C. Hawley, M. L. Wieland, J. A. Weis, and I. G. Sia, "Perceived impact of human subjects protection training on community partners in community-based participatory research," *Prog. Community Health Partnerships, Res., Edu., Action*, vol. 8, no. 2, pp. 241–248, 2014.

[36] L. Morente, J. M. Morales-Asencio, and F. J. Veredas, "Effectiveness of an e-learning tool for education on pressure ulcer evaluation," *J. Clin. Nursing*, vol. 23, no. 13, pp. 2043–2452, Oct. 2014.

XIANGUI BU was born in Weishan, China, in 1981. He is currently a Master of physical education and education doctor, an Associate Professor, a Master's Supervisor, the Head of boxing course with Shandong Sport University and the first famous Teacher of Shandong Sport University. He is a Visiting Scholar with Russian St. Petersburg National Sports University and the Moscow National Sports Institute. In 2003, he joined Shandong Sport University and devoted to physical education training and management research.