Human motion recognition based on limit learning machine

Hong Chen¹,²,³, Hongdong Zhao¹, Baoqiang Qi⁴, Shi Wang², Nan Shen², and Yuxiang Li²

Abstract

With the development of technology, human motion capture data have been widely used in the fields of human–computer interaction, interactive entertainment, education, and medical treatment. As a problem in the field of computer vision, human motion recognition has become a key technology in somatosensory games, security protection, and multimedia information retrieval. Therefore, it is important to improve the recognition rate of human motion. Based on the above background, the purpose of this article is human motion recognition based on extreme learning machine. Based on the existing action feature descriptors, this article makes improvements to features and classifiers and performs experiments on the Microsoft model specific register (MSR)-Action3D data set and the Bonn University high density metal (HDM05) motion capture data set. Based on displacement covariance descriptor and direction histogram descriptor, this article described both combine to produce a new combination; the description can statically reflect the joint position relevant information and at the same time, the change information dynamically reflects the joint position, uses the extreme learning machine for classification, and gets better recognition result. The experimental results show that the combined descriptor and extreme learning machine recognition rate on these two data sets is significantly improved by about 3% compared with the existing methods.

Keywords

Human motion recognition, extreme learning machine, covariance description, direction displacement histogram, computer vision

Date received: 9 December 2019; accepted: 7 May 2020

Introduction

With the rise of big data on the Internet, the explosive growth of information, and the widespread application of machine learning algorithms such as deep learning in the Internet field, artificial intelligence has once again entered a period of rapid development.¹–³ At present, the main concern of the market is robot technology. Robot technology is a modern technology that integrates multiple technologies. With the development of big data technology, robot technology has become more and more widely used in the modern industry.⁴,⁵ Robot technology has the following characteristics: First, it has a high degree of freedom

¹ School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, China
² School of Mathematics and Information Technology, Hebei Normal University of Science and Technology, Qinhuangdao, China
³ Key Laboratory of Electro-Optical Information Control and Security Technology, Tianjin, China
⁴ Department of Information Engineering, Qinhuangdao Institute of Technology, Qinhuangdao, China

Corresponding author: Hongdong Zhao, School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300130, China. Email: zhaohd@hebut.edu.cn
and a strong ability to control the system. In the modern professional industry, to achieve high control of the system, the robot will show a high degree of freedom to meet the needs of different programs; the second is that the robot depends on the Internet of Things and big data in the process of operation technical support, through the use of Internet technology to achieve automatic control of robots; and the third is to have high computing and information feedback. Through the joint research of robot technology and computer vision to study human behavior recognition, analyzing and identifying different actions from video data are very important in many visual applications. For example, in the field of security monitoring, human motion recognition is used in video surveillance systems, which can automatically warn of emergencies, which is of great significance in urban law and order, fire protection, and traffic dispatching; in the field of intelligent monitoring, it can fall in the elderly or children or send out an alert when a dangerous behavior is made for timely rescue; to a certain extent, it can solve the problem that some contemporary people have no time to take care of the elderly and children; and in the field of health care, by studying the performance of human gait, it can hurt the legs of patients. The degree of analysis is used to develop the best treatment plan. In the field of virtual reality technology, the somatosensory games and virtual fitting rooms that are currently popular among young people are the application of human motion recognition technology. It can be seen that computers and smart devices are increasingly becoming an essential part of our lives. The growing demand for these smart devices has increased the need to use simple and practical smart systems. Therefore, systems based on visual interaction and control are becoming more and more common, and motion recognition is becoming more and more popular in the research field because of its wide application in various fields of human–computer interaction. However, in terms of calculation of exercise volume, the research is relatively simple. For example, recently popular step counting software and applications such as sports bracelets and WeChat sports are counting walking movements and calculating caloric calories. This kind of equipment and software does not recognize the motion accurately, which further leads to inaccurate step counting results, so the results obtained when calculating the calories consumed by exercise are also inaccurate. The research value of body motion behavior recognition is more and more important in the world today, and the number of people who pay attention to this research will be more and more. The research topics of motion recognition include image processing technology, machine learning, pattern recognition, artificial intelligence, and other fields. It is a research area combining various technologies. As a key technology, motion recognition technology has been paid more and more attention by scholars, and many schools, research institutions, and scholars have invested in this work and achieved some success, such as MIT, Carnegie Mellon University Colleges, and universities, and have established specialized research laboratories to carry out this research work.

This requires extracting each frame of image from the video and then recognizing the human body from the image, further recognizing the human motion. Therefore, estimating a person’s complete three-dimensional (3-D) pose from an RGB image is one of the most challenging problems in computer vision. This involves dealing with two tasks that are inherently confusing. First, the two-dimensional (2-D) position of the human joint point or landmark must be found in the image, which is a blur due to the huge change in visual appearance caused by the viewpoints of different cameras, external and self-occlusion, or changes in clothing, body shape, and lighting. Secondly, it is still an uncertain problem to upgrade the coordinates of 2-D landmarks from a single image to 3-D. The number of 3-D spaces that can be formed consistent with the position of 2-D landmarks of humans may be infinite. Finding the correct 3-D pose that matches the image requires injecting additional information, which is usually in the form of 3-D geometric pose priors and time or structural constraints. In the field of computer vision, due to its complexity and variability, human movements and poses have always been the focus of attention. Although video retrieval and classification technology has been widely used, the problem of human motion recognition has not been solved well. First of all, there are too many human bones, and each bone has multiple degrees of freedom, which makes the performance of human movements extremely complicated. In addition, different activities of the human body may have similar movements, such as the two actions of shooting and waving. It can be easily identified, but if you only look at the limb movements, the existing recognition algorithms may be difficult to distinguish, and the above problems need to be solved urgently.

To solve the problem of human motion recognition based on hidden Markov model (HMM), Pichao solved the problem of reducing the dimension of feature space. First, they describe how to derive different features from human motion capture based on markers and define a total of 29 features and a total of 702 dimensions to describe human motion. They then proposed a strategy for systematically exploring the space of possible subsets of these features and identifying meaningful low-dimensional feature vectors for motion recognition. They evaluated our approach using a data set of 353 movements, which were divided into 23 different types of full-body movements. Their results show that the low-dimensional feature space is sufficient to achieve high motion recognition performance, and using only four dimensions, they can achieve 94.76% accuracy on the data set, which is quite high-dimensional compared to the feature vector considering many features, such as joint angles. In recent years, smartphone-based human motion recognition has received increasing attention in many fields such as mobile health, health tracking, and...
pervasive computing. However, changes in mobile phone orientation and position can easily affect motion recognition performance. Different users also have an impact on recognition accuracy. Most of the existing work focuses on one or two aspects of the above problems or training different models for different phone positions and directions. Peng proposed a universal framework for human motion recognition based on smartphones, which can effectively distinguish six daily movements, regardless of the position and orientation of the device. They chose a set of more powerful and effective features to solve the performance degradation caused by different phone locations, phone orientations, and users. In the experiment, they used data sets collected by three volunteers on Android smartphones to access our methods. Experimental results show that the proposed feature extraction algorithm is superior to most existing algorithms.27

Sevgi proposes a new human motion recognition framework based on HMM, namely hybrid event probability sequence (HEPS), which can identify unlabeled motion in video. First, the center of the moving object is used to extract the trajectory effectively. Second, HEPS is constructed using these trajectories, which represent different human behaviors. Finally, an improved particle swarm optimization (PSO) with inertial weights is introduced to recognize human actions using HMM. The proposed method was evaluated on the University of central Florida (UCF) human behavior data set with an accuracy rate of 76.67%. Comparative experimental results show that HMM has better effect than HEPS and PSO.28 Wang et al. proposed a variety of human motion pattern recognition algorithms based on microelectromechanical system (MEMS) inertial sensors, which solved the problem of low pattern recognition accuracy in various human motion pattern recognition. The time-domain features of the MEMS acceleration sensor are selected as the features of pattern recognition. The time-domain features of MEMS gyroscopes are used as features to aid recognition. The algorithm can accurately identify a variety of motion patterns, including walking, running, standing, going upstairs, going downstairs, falling forward, falling backward, and walking backward. A hierarchical recognition algorithm is used in the recognition process. Support vector machines (SVMs) are trained to recognize these two types of motion patterns because they are difficult to distinguish. Experiments based on the embedded firefighter positioning system platform show that the recognition algorithm can recognize multiple movement modes of firefighters with an average accuracy rate of more than 94%.29

Depth motion maps (DMMs) have shown effectiveness for human motion recognition; however, they have lost time information and suffered intraclass changes due to changes in motion speed. To solve these challenges, Javid proposed a new method for human motion recognition. First, an adaptive layered DMM (AH-DMM) is calculated on the temporal layered window of the video sequence to capture temporal information. In addition, adaptive windows and steps are used to ensure that the AH-DMM is robust to changes in motion speed. Then, the Gabor filter is used to encode the texture information of the AH-DMM to generate a compact and differentiated motion representation. Finally, the representation is used as an input to the collaborative representation classifier. Experimental results on the public benchmark MSR-Action3D data set and DHA data set prove that the method is better than the latest depth-based action recognition method.21

The covariance descriptor and the direction displacement histogram descriptor (HOD) are combined to form a new descriptor that contains both the static dependency information of each joint between each joint and the dynamic displacement relationship of each joint between adjacent frames. And these three feature descriptors were, respectively, tested on the MSR-Action3D data set and the HDM05 data set with linear SVM, extreme learning machine (ELM), and voting-based ELM (V-ELM). The results show that the ELM classifier has better classification performance on human action data, and the recognition rate of the combined features on the MSR-Action3D data set is much higher than the two original features used alone.

**Proposed method**

**Basic extreme learning machine**

Feedforward neural networks, especially back propagation (BP) neural networks, have been widely used in recent years, but the traditional BP algorithm is essentially a one-step method parameter optimization problem, which has problems of slow convergence and convergence to a local minimum. Even though various improved feedforward networks have faster training speed and better generalization ability than BP algorithm, they still cannot get the global optimal solution.

To solve the above problems, an ELM method was proposed to train a single-hidden layer feedforward neural network (SLFN). In ELM, hidden nodes are randomly initialized and no longer adjusted during the entire training process. The only freedom to learn the parameters is the weights between the hidden layer and the output layer. In this way, ELM can be reduced to a parametric linear model for solving linear classification problems.

Traditional feedforward neural networks use gradient descent to train the network. This makes the training speed of the neural network slow, and each weight needs to be adjusted. A large number of network parameters need to be set, and it is easy to fall into a local optimal situation. Existing problems limit the development of feedforward neural networks. As a new type of single-hidden layer neural network, the ELM neural network solves the current
shortcomings of traditional feedforward neural networks. It has the advantages of extremely fast learning speed, simple and easy to learn, and good generalization ability. He has paid close attention and researched extensively and has been widely used in various fields such as image quality assessment, image classification, short-term wind speed prediction, fault diagnosis, and medical diagnosis traffic sign recognition.

The network structure of the basic ELM is the same as that of an SLFN, as shown in Figure 1, including an input layer, a hidden layer, and an output layer. The learning algorithm of an SLFN uses an error BP algorithm to update the weights, and the learning algorithm of the ELM neural network is based on randomly setting the input weights and hidden layer offsets. For adjustment, you only need to set the number of hidden neurons, and the output weight is obtained by solving a linear equation system. Its training process does not need iteration and is completed in one time. Compared with traditional neural networks, its training speed is significantly faster.

For a given training \( D = \{(x_i, t_i) | x_i \in \mathbb{R}^n, t_i \in \mathbb{R}^m\} \), the activation function \( f(x) \), the number of hidden layer nodes is \( L \), and the ELM regression model can be expressed as

\[
\sum_{i=1}^{L} \beta_i f(a_i x + b_i) = t_i
\]

where \( H_k \) is the neuron matrix and can be expressed as

\[
H_k = \begin{bmatrix}
  f(a_1 x_1 + b_1) & \cdots & f(a_L x_1 + b_L) \\
  f(a_1 x_2 + b_1) & \cdots & f(a_L x_2 + b_L) \\
  \vdots & \ddots & \vdots \\
  f(a_1 x_k + b_1) & \cdots & f(a_L x_k + b_L)
\end{bmatrix}
\]

\( \beta_k \) is the output layer matrix and can be expressed as

\[
\beta_k = \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_L \end{bmatrix}^T
\]

\( T_k \) is the output layer matrix and can be expressed as

\[
T_k = [t_1 \ t_2 \ \cdots \ t_k]^T
\]

The output weight can be obtained by solving equation (4)

\[
\beta_k = (H_k^T H_k)^{-1} H_k^T T_k
\]

Therefore, the prediction model of the time series after the basic ELM training can be expressed as

\[
t = \sum_{i=1}^{L} \beta_i f(a_i x + b_i)
\]

The algorithm steps of the basic ELM can be summarized as

\text{Step 1:} Convert the initial \( N \) sample sequences \( x_1, x_2, \ldots, x_N \) into the training set \( (x_i, t_i), (x_2, t_2), \ldots, (x_k, t_k) \), where \( x_i = [x_1 \ x_{i+1} \ \cdots \ x_{i+m-1}]^T \) is selected as the input, \( t_i = x_i + m \) is selected as the output. Where \( k = N - m > L \).

\text{Step 2:} Determine the number of hidden layer neurons \( L \), the activation function is \( f(x) \), randomly generate input weight \( w_i \) and hidden layer threshold \( \beta, 1 \leq i \leq L \).

\text{Step 3:} Calculate the neuron matrix \( H_k \).

\text{Step 4:} Calculate the output weight \( \beta_k \) by formula (6).

**Motion capture technology**

Motion capture refers to the process of recording the movement of objects or people. It has been widely used in military, entertainment, sports, and medical fields and has also been used in computer vision and robotics. During the motion capture process, the actors’ actions are sampled multiple times per second, and the images obtained from multiple cameras are used to calculate the 3-D position. The experimental data used to verify the motion recognition algorithm in this article are obtained through motion capture technology. The two types of motion capture methods used in the data set used in this article are introduced below.

**High-precision motion capture system.** The emergence of motion capture as a photogrammetric analysis tool in biomechanical research can be roughly divided into two categories, namely optical systems and nonoptical systems:
Human motion recognition technology

Human motion recognition is one of several major problems in the field of computer vision and has many important applications in many fields. One of the most active applications is the interactive entertainment system. The recent popularity of contactless perception interoperators (such as Microsoft's Kinect sensor) has made it more widely used in these fields. The depth data collected by the sensor can be used to analyze the operator's body skeleton information in real-time. Each of these systems has three degrees of freedom for each marker, and rotation information requires at least three orientations between the markers to determine. Optical systems generally include passive marking, active marking, semiactive nonperceptual marking, and no marking.

In addition to optical systems, there are many motion capture systems that do not use visible light as markers and are collectively referred to herein as nonoptical systems. There are mainly the following: inertial motion capture systems, mechanical motion capture systems, and magnetic motion capture systems.

Kinect motion capture system. The Kinect sensor was developed by Microsoft and was originally applied to the Xbox game console. It allows players to control the game console by voice and gestures without the need for a handheld controller. The word Kinect is a self-innovating word combining the words kinetic and connection.

The shape of the Kinect sensor is similar to a webcam, with three lenses, an RGB color camera in the middle, and a 3-D structured light depth sensor composed of an infrared emitter and an infrared complementary metal oxide semiconductor (CMOS) camera on the left and right sides. This device can generate a 3-D image of the space in front of the sensor and use a random decision forest algorithm to automatically recognize the structure of the human body in near real time. The existing research results show that the depth sensor itself can accurately calculate the 3-D position of the working environment, and the joint center of the human body structure can be used to judge the human body posture. Compared with large motion capture systems, motion data captured by Kinect sensors are relatively noisy and have a low frame rate. However, due to its low price, it is still widely used in places where data accuracy is not high.

At present, there are mainly three motion recognition methods for visual information: the first is a motion recognition method based on spatiotemporal features. This method first converts human motion sequences into a set of static models and compares them with existing motion samples during the recognition phase. This method has low computational complexity but is more sensitive to noise. The second is an image-based statistical recognition method. This method statistically analyzes the low-level information of each frame of action to understand the entire sequence of actions, but the disadvantage of this method is that it requires a large amount of calculation. The third method is based on a skeleton model. This method first uses a 2-D or 3-D skeleton model of the human body to obtain the human pose of each frame and then describes different actions that change at any time. This method has high accuracy, but the requirements are also high, and the recognition method used in this article belongs to the third category.

Experiments have shown that using skeleton information alone to identify actions is more effective than using other low-level image data. This method is particularly effective when identifying simple user gestures, but its application in other fields, especially nonentertainment fields, has not been proven.

In human motion recognition, there are three main tasks: data capture, feature description, and motion modeling. This section will introduce the research progress of these tasks.

The first point is the usability and quality of motion capture data. Using high-precision motion capture systems to obtain accurate skeleton data usually requires a higher price, but many organizations provide open databases, such as the Carnegie Mellon University (CMU) Mocap database at Carnegie Mellon University and the HDM05 database at the University of Bonn. On the other hand, Kinect like Microsoft and some other relatively cheap sensors make it much easier to obtain motion data. Although there will be a loss in accuracy, these losses are still within acceptable limits, plus these sensors, the depth information can be obtained, and the position of the joint can be calculated, it can be used in general. And these acquisition systems are widely used due to their low price. Many such data can be found at present, the most typical of which are the MSR-Action3D data set and MSRC12 Kinect Gesture Data set released by Microsoft Corporation.

The second point is to find reliable feature descriptors for the action sequence. There are three main types of action descriptors: full sequence descriptors, single frame descriptors, and point of interest descriptors. The latter two require subsequent steps such as descriptor aggregation and timing modeling to complete the goal of action recognition.

The third point is to dynamically model the action. Generating models such as HMM or discriminative models such as conditional random field (CRF) can be used to analyze motion sequences. In these methods, joint positions and joint position histograms are used. And because these complex models cannot find enough training data, it is very easy to cause overfitting problems. In addition, the 3-D joint positions calculated from the depth data are much more noisy than the motion capture data, which make these
data from it is very difficult to obtain an accurate classification, especially among similar actions. There are other methods that use recurrent neural networks to describe complex nonlinear actions, but the results are not ideal. Another neural network method is a conditionally restricted Boltzmann machine. Because a large number of parameters need to be calculated, these models require a large number of data samples and long-term training to obtain more accurate estimates of the model parameters.

**Experiments**

**Experimental background**

To evaluate the combined new descriptors and verify the classification ability of the ELM algorithm on human action data, a comparative experiment was performed on the data sets MSR-Action3D and HDM05. To compare experimental results, this chapter uses the same training and test set partitions as in the above references. To avoid the randomness brought by one partition method, this chapter and subsequent chapters also perform a variety of different data set partitions. In the experiments, linear SVM and ELM classifiers and V-ELM classifiers based on voting were used. Among them, support vector machine (LIBSVM) software (version 3.1.7) was used for SVM.

**Experimental data**

1. MSR-Action3D data set

   The MSR-Action3D data set contains 20 different actions of 10 actors. Each action is repeated two to three times. There are a total of 567 motion sequences. This article uses the same 544 as the reference [7]. This data set was collected by Microsoft Kinect sensor and recorded depth information and skeletal joint position information. All the methods in this article only use the skeletal joint position information. Each skeleton of this data set contains 20 joint points, as shown in Figure 2. This data set collects 30 frames of data per second and adopts a data recording method that directly records the 3-D spatial position of joint points in each frame.

   In the use of data, the entire data set is divided into three subsets, each of which contains eight types of actions, and there are intersections between the subsets. Each subset is tested separately, and the average training accuracy of the three subsets is finally calculated. In all experiments of this data set, the actions of five performers are selected as the training set, and the actions of the remaining five performers are used as the test set.

2. HDM05 data set

   The HDM05 data set is recorded in advanced streaming file/advanced MP3 catalog (ASF/AMC) format. The main differences from the MSR-Action3D data set are (i) the data set uses the professional motion capture device Vicon Mx system to obtain data with less noise. (ii) There are 31 skeleton nodes in this data set, which records more details than MSR-Action3D. (iii) The number of motion frames collected per second is much higher, reaching 120 fps.

   In each experiment in this article, 11 moves performed by 5 people were used; these 11 moves were deposit floor, elbow to knee, grab high song, hop both legs, jog, kick forward, lie down, floor, rotate both arms backward, sneak, squat, and throw basketball. In this article, 277 motion sequences are selected in this data set. The movements of three actors are trained and the movements of the other two actors are tested.

   The human skeleton structure recorded by ASF is shown in Figure 3, which consists of 31 joint points including the root node and a total of 56 degrees of freedom.

   The whole skeleton can be recorded as a tree with the waist node as the root node and the remaining nodes according to the hierarchical relationship. Among them, the root node is a special node that determines the position and orientation of the entire skeleton. The remaining joint points are displaced relative to the parent node. According to the actual situation of the human body, each joint point has different degrees of freedom.

**Discussion**

**Experimental analysis on MSR-Action3D data set**

In the case of changing the descriptor hierarchy parameters, different combinations of features and classifiers are used for experiments. The SVM uses a linear kernel function, the
The number of ELM hidden layer units is set to 10000, and V-ELM is voted for 10 ELM, each of the number of hidden layer units of ELM is 1000, and the same data subset division method is used. The average recognition rate of each data subset is presented in Table 1.

HODs do not overlap, so the last two cells of the table are empty. Mix stands for combined feature descriptors. The overlap of combined feature descriptors refers to the combination of covariance descriptors with overlap and HODs without overlap.

To easily see the difference in data, the highest recognition rate of each combination of features and classifiers in different layers is shown in Figure 4.

It can be seen from Table 1 and Figure 4 that the combined feature descriptor combined with the ELM classifier has obvious advantages in this division of the data set. In addition, the classification effect of the ELM algorithm on covariance descriptors and combined feature descriptor is better than SVM. There has also been a noticeable improvement.

### Experimental analysis of different training and test sets

To avoid the randomness caused by a partitioning method of the data set, this article conducted the above experiments 10 times under different training and test set partitioning methods. The average results are presented in Table 2 and Figure 5:

The experimental results show that the combined feature descriptor combined with the ELM classifier has obvious advantages in this data set under many different partitioning methods of the data set, and the classification effect of the ELM algorithm on covariance descriptors and
Combined feature descriptors is better significantly improved than that of SVM.

**Experimental analysis of high density metal oxide semiconductor (HDM05) data set**

To verify the effectiveness of the method on high-quality data, a similar experiment was performed on the motion capture data set HDM05. The main differences between this data set and the MSR-Action3D data set are: (i) HDM05 data set uses professional motion capture equipment to obtain data with less noise; (ii) the skeleton joint points of this data set are 31, which will make all kinds of feature descriptors longer. (iii) The number of frames collected per second is also much higher, reaching 120 fps instead of 30 fps in the previous data set.

In the experiment, 11 actions performed by 5 actors were used. In the case of ensuring that the action categories are the same, this article randomly selects 277 motion sequences in this data set, using the actions of 3 actors for training and the action of the remaining 2 actors for testing; as in the previous data set, exhaustively divided all 10 types of training and test sets, eliminated one of the obvious errors, and finally counted the experimental results of the remaining 9 types of partitions. The average results of nine experiments are presented in Table 3 and Figure 6, where Figure 6 shows the highest recognition rate under the combination of each feature and classifier in Table 3.

It can be seen from the experimental results that the classification results of the ELM and V-ELM algorithms for various feature descriptors are better than SVM. Although the recognition rate of V-ELM is slightly lower than that of ELM, the time spent during the experiment is significantly reduced compared to ELM. Comparing Table 2 and Table 3, it can be found that although the number of samples used for training is similar, the classification accuracy of the same feature descriptor and classifier on the HDM05 data set is much higher than that of the MSR-Action3D data set. Analysis shows that this is determined by HDM05 data set of low noise, high frame rate, and more joints.

The experiment used 11 actions performed by 5 people. Under the condition of ensuring that the action categories are the same, 277 motion sequences were randomly selected under this data set, and the actions of 3 actors were used for training and the actions of the remaining 2 actors were the test; like the previous data set, exhaustively divided all 10 training and test set division methods and eliminated one of the obvious errors. Finally, the remaining nine types were counted. The average results of nine experiments are presented in Table 4 and Figure 7. It needs to be mentioned that, because the data recorded by the HDM05 data set is high in accuracy, the feature descriptors also increase accordingly. When the number of hidden layer nodes reaches 100,000, the computer memory used in the experiment can no longer meet the experimental requirements. Considering this situation, currently there is no actual use value, so no experiments with a higher
Table 4. Average recognition rate of HDM05 data set under nine partitioning methods.

| Method      | L = 1 | L = 2 | L = 3 | L = 2, OL | L = 3, OL |
|-------------|-------|-------|-------|-----------|-----------|
| CV + ELM    | 94.91 | 96.58 | 96.88 | 97.15     | 97.11     |
| HOD + ELM   | 96.98 | 97.29 | 97.98 |           |           |
| MIX + ELM   | 94.72 | 96.48 | 97.33 | 97.04     | 97.24     |
| CV + LR-ELM | 95.71 | 96.21 | 96.84 | 96.27     | 97.17     |
| HOD + LR-ELM| 97.17 | 97.80 | 97.94 |           |           |
| MIX + LR-ELM| 94.60 | 96.57 | 97.25 | 97.01     | 97.38     |

L: number of layers of feature descriptors; OL: overlap between window layers of each layer; CV: covariance descriptor; HOD: direction displacement histogram descriptor; ELM: extreme learning machine.

Figure 7. Comparison of the average highest recognition rate of the HDM05 data set under nine partitioning methods.

Conclusions

With the development of economy and science and technology, robot technology is becoming more and more powerful in terms of applications and functions. The recognition of human motion through robotics has a wide range of applications in many fields and has become a research hotspot in the field of computer vision in recent years. Human motion recognition based on ELM aims to determine the category to which the action belongs by extracting features and modeling from human motion capture data. In this article, the covariance descriptor and the HOD are combined to form a new descriptor that contains both the static dependency relationship of each joint in each frame and the dynamic change of the position of each joint between adjacent frames. These two descriptors were used to recognize human motion on the MSR-Action3D data set and the HDM05 data set in combination with linear SVM, ELM, and V-ELM. The results show that (i) ELM and V-ELM based on voting have better covariance features and combined features than SVM, and combined with combined features on low-quality MSR-Action3D data sets have a large recognition rate. This proves that the ELM algorithm has advantages in the application of human motion classification; (ii) the combined feature descriptors can complement the features on low-quality data and improve the classification accuracy.

This article studies a human motion recognition algorithm based on ELMs. The research analyzes the existing algorithms to recognize human movements and analyzes their shortcomings in real time. By extracting the histogram of oriented gradient (HOG) and local binary patterns (LBP) features of the action and constructing a joint feature vector, the grayscale, deformation, and rotation invariance of the image can be effectively guaranteed. The precise positioning of markers based on human motion characteristics, making full use of human motion shape characteristics, helps to remove suspected unrelated areas. Compared with other methods to detect human motion, experimental comparison proves that the algorithm has improved real-time performance and robustness. Using the ELM network to classify and recognize human movements, compared with traditional machine learning methods, the ELM network avoids iterative adjustment of its weights, saves computing time, and improves computing efficiency. The experimental comparison proves that the algorithm has good real-time performance.

The depth of research in this article is limited. It is just a single human motion recognition based on ELM. In the later work, I hope to achieve the following goals: extract features from the rotation angle of the subjoints to the subjoints to achieve rotation, scaling, translation, and invariance of scale to further improve the classification effect. For low-precision human motion capture data, we can first use spherical or spline differences for each joint point in the motion sequence to reduce noise in this way and then perform subsequent feature extraction and classification. At present, the recognition of human movements using ELM is based on a large number of hidden layer nodes, and the memory requirements of the experimental machine are very high. In later work, I hope to find new methods to reduce the demand for hidden layer nodes to improve practicality of the algorithm. The classification method in this article is applied to more fields to verify its universality.
Declarations of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was financially supported by Planning Subject of “the thirteenth Five-Year-Plan” key laboratory of National Defense Science and Technology of China (grant number 61421070104) and the National Natural Science Youth Foundation of China (grant number 61805067).

References
1. Rashid SA and Jafar FA. Experimental analysis on human haptic emotion for robot sensing technology development. Adv Sci Lett 2017; 23(5): 4414–4418.
2. Wu Y, Wang T, Li L, et al. Automatic path planning technology of stitching robot for composite fabric with curved surface. Mater Sci Technol 2017; 25(2): 16–21.
3. Sugihara T, Yasunaga H, Matsui H, et al. Accessibility to surgical robot technology and prostate-cancer patient behavior for prostatectomy. Jpn J Clin Oncol 2017; 47(7): 1–5.
4. Song A. Research on human-robot interaction telerobot with force telepresence. Sci Technol Rev 2015; 33(23): 100–109.
5. Malle BF. Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethic Inform Technol 2016; 18(4): 243–256.
6. Cao D, Zhuang L, Su K, et al. A wearable olfactory animal-robot system based on Wi-Fi technology. Chin J Sensor Actuat 2015; 28(3): 303–309.
7. Nyholm S and Smids J. Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic. Ethic Inform Technol 2018; 27(2): 1–10.
8. Das PK, Behera HS, Jena PK, et al. Multi-robot path planning in a dynamic environment using improved gravitational search algorithm. J Electr Syst Inform Technol 2016; 3(2): 295–313.
9. Huang J, Fan X, and Dong W. Current status of laparoscopic and robot-assisted nerve-sparing radical cystectomy in male patients. Asian J Urol 2016; 3(3): 150–155.
10. Wang X-M, Liu M-C, Wang H-R, et al. Visual identity system for automatic feeding robot. J Shenyang Univ Technol 2018; 40(5): 564–570.
11. Li W, Chen C, Su H, et al. Local binary patterns and extreme learning machine for hyperspectral imagery classification. IEEE T Geosci Remote Sens 2015; 53(7): 1–13.
12. Yu L, Dai W, and Tang L. A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. Eng Appl Artif Intell 2015; 47: 110–121.
13. Lei Z and David Z. Evolutionary cost-sensitive extreme learning machine. IEEE T Neurl Netw Learn Syst 2015; 28(12): 3045–3060.
14. Yang Y and Wu QMJ. Multilayer extreme learning machine with subnetwork nodes for representation learning. IEEE Trans Cybern 2015; 46(11): 2570–2583.
15. Wu B, Xi L, Fan S, et al. Fault diagnosis for wind turbine based on improved extreme learning machine. J Shanghai Jiaotong Univ 2017; 22(4): 1–8.
16. Al-Yaseen WL, Othman ZA, and Nazri MZA. Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Syst Appl 2017; 67: 296–303.
17. Bencherif MA, Bazi Y, Guessoum A, et al. Fusion of extreme learning machine and graph-based optimization methods for active classification of remote sensing images. IEEE Geosci Remote Sens Lett 2015; 12(3): 527–531.
18. Sovij D, EIrola E, Miehe Y, et al. Extreme learning machine for missing data using multiple imputations. Neurocomputing 2015; 174(PA): 220–231.
19. Yi C, Enyi Y, and Arindam B. A 128 Channel extreme learning machine based neural decoder for brain machine interfaces. IEEE T Biomed Circuit Syst 2015; 10(3): 679–692.
20. Wang Y, Xie Z, Xu K, et al. An efficient and effective convolutional auto-encoder extreme learning machine network for 3D feature learning. Neurocomputing 2015; 174(PB): 988–998.
21. Ullah J and Jaffar MA. Object and motion cues based collaborative approach for human activity localization and recognition in unconstrained videos. Clust Comput 2018; 21(2): 1–12.
22. Liu H and Wang L. Human motion prediction for human-robot collaboration. J Manuf Syst 2017; 44(Part 2): 287–294.
23. Wang W, Hou ZG, Cheng L, et al. Toward patients’ motion intention recognition: dynamics modeling and identification of iLeg—an LLRR under motion constraints. IEEE T Syst Man Cybern Sys 2016; 46(7): 1–13.
24. Yuting Y, Dingna D, Huan Z, et al. Motion recognition based on hidden Markov model with improved brain storm optimization. Space Med Med Eng 2015; 48(7): 1274–1287.
25. Qiu-Ju G, Xiao-Mu L, Xue-Mei G, et al. Compressive infrared classification of human motion using HMM. Acta Automatica Sinica 2017; 43(433): 398–406.
26. Wang P, Li W, Ogunbona P, et al. RGB-D-based human motion recognition with deep learning: a survey. Comput Vision Imag Understand 2018; 171: 118–139.
27. Wang P, Liu H, Wang L, et al. Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. Cirp Ann 2018; 67(1): 17–20.
28. Gurbuz SZ and Amin MG. Radar-based human-motion recognition with deep learning: promising applications for indoor monitoring. IEEE Signal Process Mag 2019; 36(4): 16–28.
29. Wang X-M, Liu Z, and Ni W-B. Motion measurement and recognition method of human arm based on MEMS sensors. J Chin Inertial Technol 2017; 25(6): 701–707.
30. Vinayak KG and Hemant PV. A study of vision based human motion recognition and analysis. Int J Ambient Comput Intell 2016; 7(2): 75–92.
31. Guo M and Wang Z. Segmentation and recognition of human motion sequences using wearable inertial sensors. Multimed Tool Appl 2018; 77(2): 1–20.
32. Guo Y, Li Y, and Shao Z. RRV: a spatiotemporal descriptor for rigid body motion recognition. *IEEE Trans Cybern* 2018, 48(5):1513–1525.

33. Wu D-H, Wang Z-L, and Chen Y. Research on segmentation and recognition of human daily action sequence based on HMMs and SVM. *Dalian Ligong Daxue Xuebao/J Dalian Univ Technol* 2015; 55(4): 411–416.

34. Motiian S, Siyahjani F, and Almohsen R. Online human interaction detection and recognition with multiple cameras. *IEEE Trans Circuit Syst Video Technol* 2017; 27(3): 649–663.