IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.Doi Number

Dynamic gesture recognition algorithm Combining Global Gesture Motion and Local Finger Motion for interactive teaching

LI Jiashan, LI Zhonghua*
Corresponding author: **LI Zhonghua** (e-mail: 568495506@qq.com)

ABSTRACT Adding dynamic gesture recognition to the new interactive teaching system is of great significance to improve the teaching efficiency. However, the features extracted by traditional dynamic gesture recognition methods are usually difficult to accurately represent the differences between dynamic gestures. Aiming at the problems of complex time sequence and spatial variability of dynamic gestures, this paper proposes a gesture recognition method combining global motion and local motion of fingers. Firstly, preprocessing of dynamic gesture data is carried out, including removing invalid gesture frames, completing gesture frame data and normalizing joint length. Then, according to the given hand joint coordinates, dynamic gesture key frames are extracted by using gesture distance function, and the global motion features of hand in space and the local motion features of fingers in hand are extracted based on gesture key frames. Secondly, the key frame gesture features of gesture global motion and finger local motion are fused, and linear discriminant analysis is used for feature dimensionality reduction. Finally, support vector machine with Gaussian kernel is used to realize dynamic gesture recognition and classification. Experimental results on DHG-14/28 and FPHA dynamic gesture dataset show that the accuracy of classification and recognition is 98.57% 88.29% and 97.31% respectively.

INDEX TERMS Dynamic gesture recognition; key frame; support vector machine; gesture motion.

I. INTRODUCTION

The new teaching system based on virtual simulation is promoting the reform of education. The combination of virtual simulation and traditional classroom is the trend of future development. But at present, there is no effective interactive way for teachers to operate the virtual simulation system conveniently in the context of traditional classroom. As an important interaction mode in the fields of computer graphics, virtual reality, human-computer interaction and sign language translation, gesture interaction provides a simple and convenient interaction experience [1]. According to the timing of gestures, they can be divided into static gestures and dynamic gestures [2]. Static gestures refer to a single frame of stationary gestures, while dynamic gestures refer to consecutive frames of gestures over a period of time. Compared with static gestures, dynamic gestures are difficult to be accurately recognized because they need to pay attention not only to the changes of hand shape, but also to the movements of fingers in time and space [3]. In general, the motion law of complex dynamic gestures has the following three obvious features :(1) variability of time. The motion speed of dynamic gestures is uncertain. For the same gesture, different people can complete it at different speeds. Even for the same person, the completion speed is different each time. (2) The variability of gesture integrity. In many cases, user/operator gestures are incomplete or redundant compared to system-defined gestures. (3) Spatial variability. The motion space and distance of the gesture are different, and the distance and range of the same gesture are always different for different people. These features will make it difficult to accurately represent the features of different dynamic gestures. The complex timing, spatial variability and inaccurate feature representation of dynamic gestures bring difficulties and challenges to the recognition and classification of dynamic gestures [2].

Many dynamic gesture recognition of the work is based on RGB images, depth image, smooth flow of information or gesture trajectory [4], Simonyan et al. [4] made use of dual data flow features to classify dynamic gestures. One data flow used static RGB images for classification, while the other data was fluent in light flow and trajectory information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access

The RGB image information contains the local feature information of the single frame gesture, and the optical flow and trajectory information contain the global feature information of the gesture. However, this method does not combine the characteristics of the two data streams, but only uses the two data streams separately. In this paper, global gesture motion features and local hand internal finger motion features are considered, and the two features are combined for dynamic gesture recognition and classification. Based on gesture image, On the basis of gesture images, Molchanov et al. [5] adopted connectionist temporal classification (CTC) to solve the dynamic gesture timing problem. However, this method is conditional independent, assuming that the outputs of different time frames are independent of each other. For the dynamic gesture sequence, the gesture sequence has the continuity of time and space, which does not conform to the dynamic gesture motion. Gongfa Li et al. proposed a hand gesture recognition method based on convolution neural network. And the convolution neural network is applied to the recognition of gestures, and the characteristics of convolution neural network are used to avoid the feature extraction process, reduce the number of parameters needs to be trained, and finally achieve the purpose of unsupervised learning [36]. Dinh-Son Tran et al. proposed a novel method for fingertip detection and hand gesture recognition in real-time using an RGB-D camera and a 3D convolution neural network (3DCNN). This system can accurately and robustly extract fingertip locations and recognize gestures in real-time

Recently, due to the wide use of hardware devices such as Intel RealSense, Microsoft Kinect, Open-pose and the development of high-precision hand tracking methods, it is easy for people to obtain high-precision hand skeleton data. In fact, the motion of the hand bone usually accurately reflects the feature differences of different dynamic gestures [3,6]. Based on the hand node coordinate input, this paper first proposes an effective method to extract the key frame of the dynamic gesture, aiming at the variability of the time and the integrity of the gesture. In this way, redundant frames in different dynamic gestures are removed and video with different length of dynamic gestures is unified to the same length. Then, based on the dynamic gesture key frame, the gesture motion features are represented as the global motion of the hand in space and the local motion of the fingers inside the hand, and LDA method in supervised dimension reduction is adopted here for feature dimension reduction. Finally, support vector machine (SVM) with gaussian kernel is used to realize effective dynamic gesture recognition. In this paper, a dynamic gesture feature representation is proposed, which can effectively represent the motion features of dynamic gestures and lay a foundation for the accurate recognition of gestures.

II. RELATED WORK

Spatial and temporal information feature processing of dynamic gesture is the key and difficult point of dynamic gesture recognition and classification [2]. Dynamic gesture recognition can be divided into traditional manual feature extraction methods and deep learning methods.

Most of the traditional manual feature extraction methods for dynamic gestures adopt dynamic time warping (DTW) [7-8], Fourier time pyramid [9], hidden Markov models (HMM) [10] and other methods to solve the spatiotemporal information processing problem of dynamic gestures. Among them, the DTW method [7-8] uses a pare-wise comparison strategy to regulate the time information. This method relies on a standard gesture version for comparison, but there is no standard version for comparison in the gesture data set, so the standard gesture can only be artificially set. The Fourier time pyramid method [9] deals with the spatiotemporal information features of dynamic gestures by segmental extraction of complete gesture frames. HMM believes that the next state of the dynamic event is only related to the previous state and has nothing to do with the previous state [10], which ignores the coherence of the dynamic gesture.

For dynamic gesture recognition of deep learning methods often use HMMs [10], long short-term memory (LSTM) [11-12], the generalized time warping(GTW) [13], DTW [7-8], spatial pyramid pooling (SPP) [14] solve the problem of time and space information processing. Wu et al. [15] used HMMs, combined with deep confidence network and convolutional neural network, to extract the time dependence of skeleton features from RGB-D data. However, because the deep confidence network adopts unsupervised learning, it does not compress the data in combination with gesture categories. Nguyen, [6] proposed a hand joint point coordinates based on symmetrical positive determined(SPD). manifold learning method of neural network. The network consists of three parts: a convolutional layer, a spatiotemporal Gaussian aggregation layer, and the final SPD matrix learned from the skeleton data. This method is similar to the one presented in this paper, which uses physical link points between joints to extract features. However, this method is rough in time series processing. In order to capture the time sequence of skeleton sequence, the spatiotemporal gesture recognition network is used to construct many sub-sequences. Abavisani et al. [16] proposed a single modal dynamic gesture recognition method based on multi-modal training. The temporal and spatial semantic alignment loss is used to align the time and location information, which is closely related to the covariance matrix alignment. However, in the method of dynamic gesture recognition using neural network, it is difficult for network design to fully consider the specific gesture motion features of dynamic gesture. This paper proposed a new dynamic gesture recognition method. Method proposed divides the motion of dynamic gesture into two parts: the global motion of the hand in space and the local motion of the fingers inside the hand, and uses the key frame to extract the time information.

III. DYNAMIC GESTURE RECOGNITION METHOD

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access

In this paper, a new dynamic gesture recognition framework is proposed based on individual differences and space-time continuity of gestures. As shown in Figure 1, the frame input is 3D joint coordinates of dynamic gesture [29]. Considering that the feature extracted from RGB-d image has the influence of view angle, this paper extracts gesture feature from 3D joint position information of gesture. By inputting the depth map of static gesture RGB-d data, the position information of 22 joint points of hand is obtained according to the characteristics of hand model. According to the anatomical characteristics of the hand, the hand is regarded as a chain structure composed of 21 segments. Firstly, data preprocessing is carried out, including the removal of invalid gesture frames, the completion of gesture frame data and the normalization of joint length. Then key frame of the dynamic gesture is extracted, and based on the key frame, the global feature of hand movement in space and the local feature of fingers inside the hand are extracted. Then, the linear discriminant analysis (LDA) feature dimension reduction is performed after the fusion of the two features. Finally, SVM with Gaussian kernel is used for dynamic gesture recognition classification. Combined with the spatiotemporal continuity of dynamic gestures, the framework solves the timing problem of gestures and effectively extracts the global characteristics of hand movements and the local characteristics of finger movements.

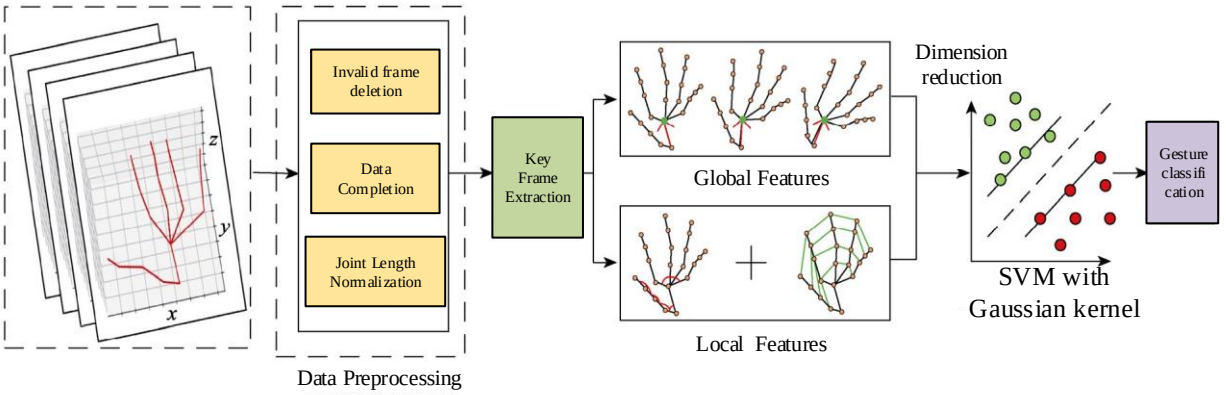


FIGURE 1. Dynamic gesture recognition framework

A. DYNAMIC GESTURE DATA PREPROCESSING

First of all, for the time variability of dynamic gestures, there are many redundant frames in the gesture video due to the fast or slow movements of the tester for the same gesture. In addition, in the process of gesture extraction, for the purpose of extracting joint position information, the tester usually needs to keep a static state for several seconds. This gesture frame is independent of the gesture category. In this paper, the gesture frame independent of the gesture category is defined as invalid gesture frame. In order to avoid interference in key frame extraction, invalid gesture frame needs to be removed first.

Secondly, for the integrity of dynamic gestures, the frame completion method is adopted for gestures dissatisfied with the frames of key frames, so that the frames of dynamic gestures meet the requirements of the frames of key frames.

Finally, in view of the spatial variability of dynamic gestures, when different people make the same gesture, different palm sizes and gesture amplitude will usually produce individual differences. In this paper, the joint length normalization method is used to eliminate the influence of individual differences so as to solve the spatial variability of dynamic gestures.

1. Gesture invalid frame deletion

Dynamic gesture is a series of gestures that change continuously over a period of time. The shape and position of the hand change with time. Dynamic gesture data sets are usually obtained through depth cameras or data gloves. The acquired dynamic gestures usually have the problem of how to define the start and end frames. In the data set sequence adopted in this paper, participants are required to fully open their whole hand in front of the camera within a few seconds before each sequence. This operation is mainly used to initialize the gesture estimation algorithm. Therefore, in each gesture sequence, there are some invalid gestures unrelated to gesture categories. In order to avoid the interference caused by invalid frames to gesture classification, invalid gestures should be deleted first. In addition, dynamic gesture starting frame extraction is also a difficult point in dynamic gesture classification. The dynamic gesture data set adopted in this paper has manually marked the valid start and end frames, so according to the number of start and end frames provided in the data set, the invalid frames before and after the start and end frames are deleted.

2. Gesture frame data completion

In order to facilitate processing and analysis, the number of key frames of each gesture should be unified. The number of key frames of some gestures in the collected video is less than the specified number. If the number of gestures less than the specified number of key frames is directly considered as invalid gestures, the number of gestures in the data set will be sharply reduced, which may lead to missing part of gestures. Therefore, this paper adopts the method of data completion

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access

for gesture data whose frame number is less than the key frame, and uses repeated gesture frames to complete the data. Repeat all existing frames in sequence from the start frame. In order to keep the characteristic of gesture motion, the repeated gesture frames are inserted directly after the repeated gesture frames until the video reaches the specified number of frames. Gesture frame completion can keep the number of samples in the training data set unchanged. Repeating the existing gesture frames can effectively maintain the integrity of dynamic gestures, which better illustrates the improvement of the accuracy of gesture recognition and the generalization of this method.

3. Normalized hand joint length

Gesture data sets usually need to be collected by different participants, and keep gesture universality. However, different participants had different hand sizes and joint lengths. In order to eliminate individual differences of the hands, this paper normalized the hand joint length to the same length, but did not change the Angle between the joints. DeSmedt et al. [3] normalized the hand joint length as the average length of the data set, but increased the calculation amount. Based on the standard finger length, the hand joint length is normalized in this paper.

As shown in Figure 2, the location of the joint point is shown, W_i represents the *i-th* joint point, and W_0 represents the root joint point. Take a frame as an example to briefly describe the normalization process. W_{ij} is used to represent the position of the *i-th* node in *j-th* frame. For convenience, the subscript *j* is omitted in the normalization process, and is simplified as W_i.

Where, i = 0, 1, 2, …, 21. Vector is used to represent the joint pair formed by 22 joint nodes:

$$V_i = \begin{cases} W_i - W_{i-1}, & 1\text{锟斤拷} \quad 21,\ i \neq 6,10,14,18 \\ W_i - W_5, & i = 6,10,14,18 \end{cases} \tag{1}$$

The normalization process is:

$$\overline{V}_i = L_i \frac{V_i}{\|V_i\|} \tag{2}$$

$$\overline{W}_i = \begin{cases} W_0, & i = 0 \\ \overline{V}_i + W_{i-1}, & 1\text{锟斤拷} \quad 21,\ i \neq 6,10,14,18 \\ \overline{V}_i + W_5, & i = 6,10,14,18 \end{cases} \tag{3}$$

It should be pointed out that the normalization of hand joint length in this paper is based on a standard finger length, and the standard finger length is established by referring to ACT hand joint segment [17], where L_i is the corresponding standard length of section *i* joint segment

B. DYNAMIC GESTURE FEATURE REPRESENTATION

First of all, from a global perspective, dynamic gesture is a series of spatial changes of the hand with the passage of time, which can be divided into translational motion and rotational motion according to the motion characteristics of the object. Translational motion is represented by the moving distance of the center point of the hand. According to the motion characteristics of the hand, the position of the center point of the palm can uniquely determine the position of the hand in space. The rotation motion is characterized by the change of the main direction vector of the hand. In this paper, the main direction of the hand is defined as: the vector that the elbow points to the center of the palm. Considering the features of interactive gestures, it does not include the straight line of the hand around the middle finger root joint and the elbow, so this paper does not consider the features of the self-rotation motion.

Secondly, locally, in addition to the spatial changes of the hand, the hand shape changes caused by the local movement of the fingers inside the hand. This paper equates the hand joint with the 21-segment chain structure. The local motion of the finger is caused by the joint bending of the finger, which can be understood as the overall change of the chain structure caused by the Angle change between the chain segments. Considering that as many as 16 elements are used in the rotation matrix, and the universal joint deadlock phenomenon occurs in Euler Angle, the rotation quaternion is used to represent the change in this paper. However, for the chain segment structure, subtle Angle errors will be accumulated, which will easily cause a large distance error after passing through multiple chain segments [3]. Therefore, in order to eliminate the distance error caused by the accumulation of Angle error, the relative distance feature of the finger is added to the local motion feature of the finger. Similarly, considering the physical characteristics of the hand, there is no self-rotating motion of the finger around the straight line where the finger root joint is connected with the elbow.

To sum up, based on the geometric characteristics of gestures and the perspective of time-space continuity, this paper proposes four characteristic representations of dynamic gestures. The process of dynamic gesture movement includes the global movement of the whole hand in space (i.e. the translational movement and rotational movement of the hand in space) and the local movement of the fingers inside the hand (i.e. the translational movement and rotational movement of the fingers inside the hand). The details are as follows:

(1) The translational motion of the hand in space

The movement process of the hand in space is depicted by the distance between the hand center point (connection node 1) in two frames before and after, that is:

$$T_j - T_{j-1} = \left\| \overline{W}_{1,j} - \overline{W}_{1,j-1} \right\| \tag{4}$$

(2) The rotational motion of the hand in space

The flip information of the hand in space is depicted by the distance of the main direction vector of the hand between the two frames. In this paper, the main direction of the hand is defined as $\overline{W}_1 - \overline{W}_0$, and the flip information is expressed as:

$$P_j - P_{j-1} = \| \ \| \overline{W}_{1,j} - \overline{W}_{0,j} \| - \| \overline{W}_{1,j-1} - \overline{W}_{0,j-1} \| \ \| \tag{5}$$

(3) Translational motion of the fingers inside the hand

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access

The translational motion of the finger is characterized by the relative distance between the fingertips. In order to avoid the accumulation of rotation errors due to the rotation Angle information between joint segments. In this paper, the distance between the adjacent fingertips and the distance between the fingertips and the fingertips relative to the wrist are extracted as the features of finger translation. Specifically, it is the distance between the adjacent fingertips:

$$\begin{aligned} &D_0 = \left\|\bar{W}_9 - \bar{W}_4\right\|, \quad D_1 = \left\|\bar{W}_{13} - \bar{W}_9\right\|, \\ &D_2 = \left\|\bar{W}_{17} - \bar{W}_{13}\right\|, \quad D_3 = \left\|\bar{W}_{21} - \bar{W}_{17}\right\| \end{aligned} \tag{6}$$

And the distance between the tip of the finger and the wrist:

$$\begin{aligned} &D_4 = \left\|\bar{W}_4 - \bar{W}_0\right\|, \quad D_5 = \left\|\bar{W}_9 - \bar{W}_0\right\|, \\ &D_6 = \left\|\bar{W}_{13} - \bar{W}_0\right\|, \quad D_7 = \left\|\bar{W}_{17} - \bar{W}_0\right\|, \\ &D_8 = \left\|\bar{W}_{21} - \bar{W}_0\right\| \end{aligned} \tag{7}$$

(4) The rotational motion of the fingers within the hand

The three common representations of rotation in space are directional cosine rotation matrix, Euler rotation, and quaternion. Directional cosine matrix uses a 4*4 matrix to represent the transformation matrix rotated around any axis. The data is redundant and it is difficult to track and recognize gestures with multiple restrictions. Euler rotation is a combination of a series of coordinate axis rotation transformations in a certain order of coordinate axes. This method needs to rotate in a fixed axis order, which will result in different recognition results.

In this paper, quaternion is used to characterize the rotation between pairs of hand joints. Quaternions include information about the rotation axis and rotation angle, and can represent rotation along any axis with four-dimensional data, which can be represented as: $q = q_0 + q_1 \cdot i + q_2 \cdot j + q_3 \cdot k \cdot$ Where q_0 is a real number, i, j, k can be interpreted as a geometrically rotated orthogonal coordinate (X, Y, Z). In order to represent the uniqueness of joint-to-joint rotation and gesture representation, quaternion is used to represent the rotation between hand-to-joint pairs.

Taking the quaternion between the joint segments of $\bar{V}_0(x_0, y_0, z_0)\bar{V}_1(x_1, y_1, z_1)$ as an example. The calculation is as follows:

$$\begin{cases} q_0 = \cos\left(\cos^{-1}\left(V_0 \times V_1\right)/2\right) \\ q_1 = \dfrac{k_1}{\sqrt{k_1^2 + k_2^2 + k_3^2}} \cdot \sin\left(\cos^{-1}\left(\bar{V}_0 \times V_1\right)/2\right) \\ q_2 = \dfrac{k_2}{\sqrt{k_1^2 + k_2^2 + k_3^2}} \cdot \sin\left(\cos^{-1}\left(\bar{V}_0 \times V_1\right)/2\right) \\ q_3 = \dfrac{k_3}{\sqrt{k_1^2 + k_2^2 + k_3^2}} \cdot \sin\left(\cos^{-1}\left(\bar{V}_0 \times V_1\right)/2\right) \end{cases} \tag{8}$$

Where:

$$\begin{cases} k_1 = (y_1 - y_0)(z_2 - z_1) - (y_2 - y_1)(z_1 - z_0) \\ k_2 = (z_1 - z_0)(x_2 - x_1) - (z_2 - z_1)(x_1 - x_0) \\ k_3 = (x_1 - x_0)(y_2 - y_1) - (x_2 - x_1)(y_1 - y_0) \end{cases} \tag{9}$$

In this paper, q_0 is used as the rotation feature to represent the rotation angle information. The remaining three numbers in the quaternion represent the information of the rotation axis, which depends too much on the position information of the joint points. The same gesture can vary greatly depending on the position information of the joint points. Therefore, in this paper, using the quaternion angle information as a feature, Q_i *(i=0, 1, 2, … 13)* is expressed for the corresponding 14 rotation characteristics of 22 joints. As shown in Figure 2, the arc marks the rotation angle between the joint segments.

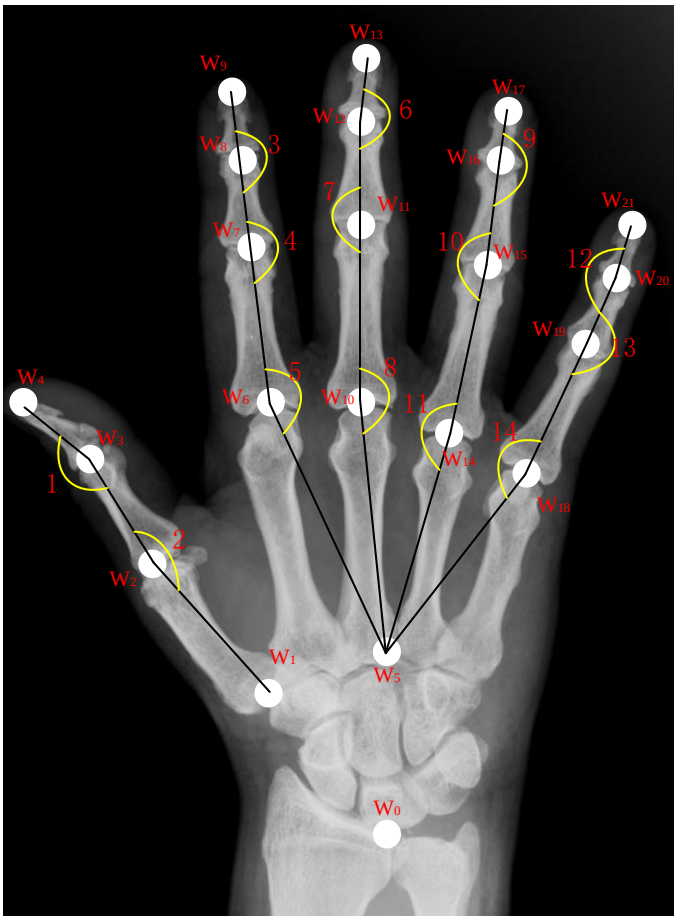
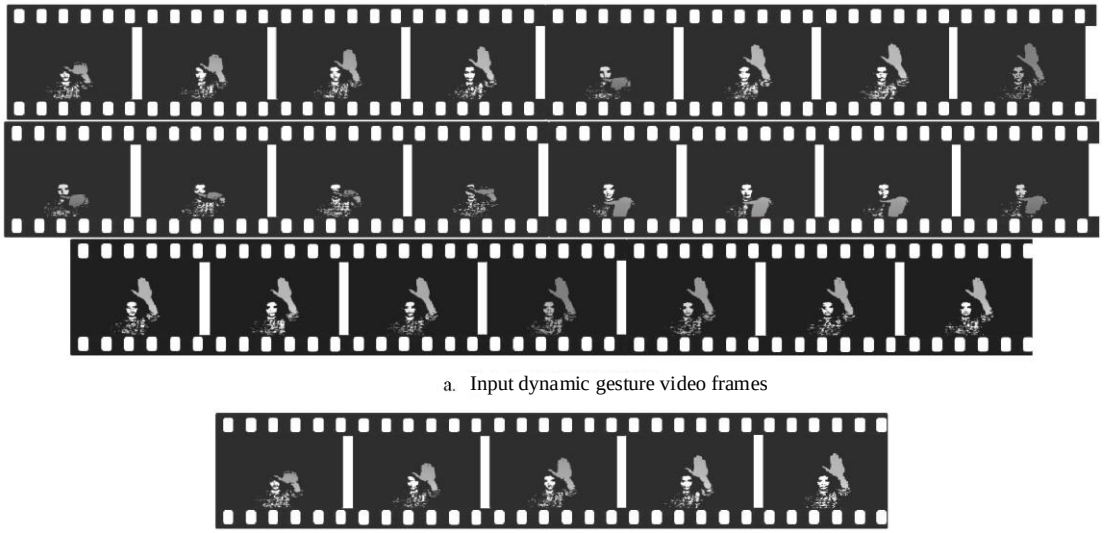


FIGURE 2. Extracting key frames without segments

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access



a. Input dynamic gesture video frames

b. Extracted dynamic gesture key frames

FIGURE 3. Extracting key frames without segments

C. DYNAMIC GESTURE KEY FRAMES EXTRACTION

1. Gesture distance function

In order to extract the key frame of dynamic gesture effectively and fuse the four characteristic representations of global gesture motion and local finger motion, a gesture distance function is proposed in this paper. By sorting the gesture distance, the gesture frame with significant feature change in dynamic gesture is selected as the key frame. In other words, the frame that produces the motion mutation is used as the gesture key frame. Define the distance between the first and second frames of a dynamic hand potential as follows:

$$H_j = \lambda_1 \sum_{i=0}^{13} \left(Q_{i,j} - Q_{i,j-1} \right) + \lambda_2 \sum_{k=0}^{8} \left(D_{k,j} - D_{k,j-1} \right) + \lambda_3 \left(P_j - P_{j-1} \right) + \lambda_4 \left(T_j - T_{j-1} \right), j = S+1, \cdots, E \quad (10)$$

Where, the starting frame number of the dynamic gesture is S, and the ending frame number is E. λ is the balance parameter, which is to ensure that the gesture distance values calculated from different features are in the same quantity level. In the experiment, $\lambda_1=100$, $\lambda_2=1$, $\lambda_3=1$, $\lambda_4=1$.

2. Extracting key frames segmentally

For dynamic gesture video sequences, if the first k frame with the maximum gesture distance function is directly selected as the key frame, it is easy to see that all the key frames are adjacent to each other. As for the upward sliding gesture dynamic sequence whose effective frame is frame 44~66 as shown in Figure 3, the key frame extracted directly by using the gesture distance function is frame 52~56. These key frames are all adjacent frames, which cannot effectively represent the whole gesture process. To facilitate observation, take the depth chart as an example, as shown in Figure 3. The figure shows that when the key frame of the gesture is extracted directly without segmentation, there is serious information redundancy, and it does not include the starting gesture, and the complete information of the dynamic gesture is lost. In order to avoid information redundancy and keep gesture integrity, it is necessary to consider segmental extraction of dynamic gesture key frames.

In the segmentation extraction dynamic gesture key frame, assuming that the start frame of the gesture is F_S and the end frame is F_E, the whole valid gesture can be represented as $\{F_S, \ldots, F_E\}$. If k key frames are extracted, the whole gesture can be evenly divided into k segments. After segmentation, the gesture segment I is

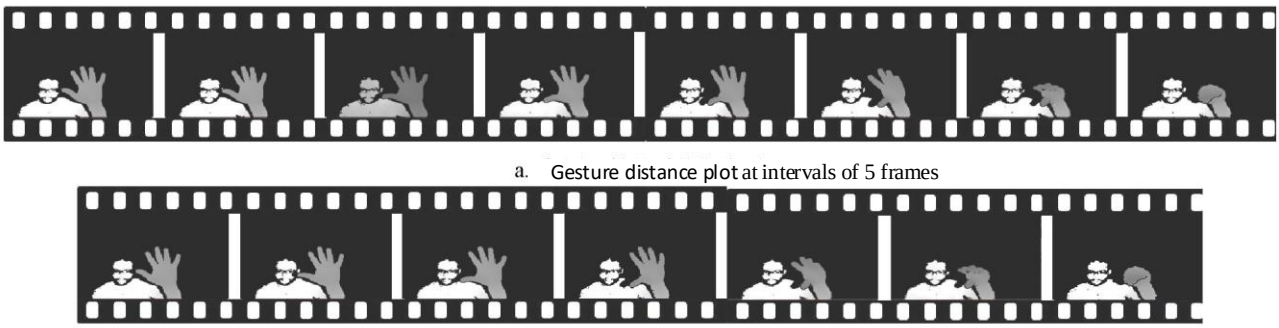
$$I = \left\{ \{F_S, \cdots, F_{S+d-1}\}, \cdots, \left\{ F_{S+(k-1)\cdot d}, \cdots, F_E \right\} \right\} \quad (11)$$

Where, $d=[(E-S+1)/k]$. Then select the frame with the maximum distance within each gesture segment as the key frame of the segment.

The frame range after deleting invalid frames in the data set of this paper is 7~149 frames, which can be selected as a wide range of key frames. Considering the difference between the length of video sequence of human motion recognition and gesture motion, the key frames are selected as 31 frames by comparing the accuracy of gesture recognition. Finally, to ensure the integrity of the gesture, the start and end frame of the gesture is added as the key frame.

If the start frame (end frame) is already included in the key frame, then the adjacent frame (the next or the previous frame) is selected to replace it, and then the start and end frames are added. The algorithm steps are as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access



a. Gesture distance plot at intervals of 5 frames

b. Segmented extraction of gesture keyframes

FIGURE 4. Segmented extraction of gesture key frames

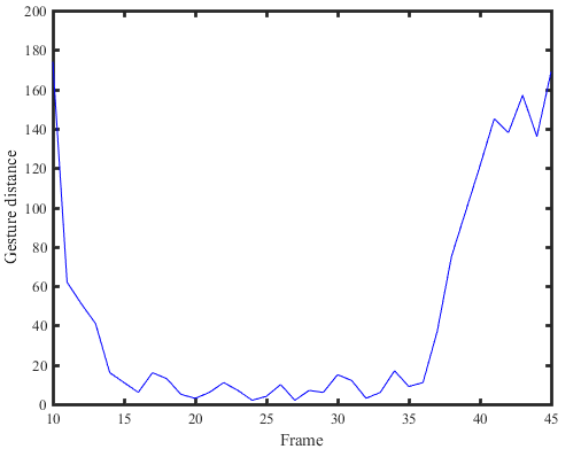


FIGURE 5. Gesture distance diagram of frame 10~45 in grab gesture

Algorithm 1: Key frame extraction algorithm.

Input: 3D coordinate information of 22 key nodes of dynamic gesture.

Output: The k-frame key frames of the dynamic gesture.

Step 1: According to the start and end frame, delete the invalid frame of the gesture.

$$\left\{F_s^{'},\cdots,F_E^{'}\right\}\leftarrow\left\{F_1,\cdots F_N\right\}$$

Step 2: Complete the gesture frames.

$$\left\{F_s,\cdots,F_E\right\}\leftarrow\left\{F_s^{'},\cdots,F_E^{'}\right\}$$

Step 3: Eq. (3) is used to normalize the joint length of each frame, and the position information of the joint node after normalization is obtained.

$$\overline{W}_{i,j},i=0,1,\cdots,21,S\text{数}\quad E$$

Step 4: The dynamic gesture is segmented according to Eq. (9).

Step 5: According to Eq. (10), the distance H_j between two adjacent frames in a video segment is calculated.

Step 6: The frame with the maximum distance is selected as the key frame in each video segment. $\{F_{m1}, F_{m2},\ldots, F_{mk}\}$

Step 7: Add the start and end frames of the gesture (F_S and F_E), and finally get the key frame of the dynamic gesture as $\{F_S, F_{m1}, F_{m2},\ldots, F_{mk}, F_E\}$.

Take grab gesture as an example to illustrate the effectiveness of key frame extraction in this paper. Figure 4a shows the depth graph of every 5 frames in the grab gesture, corresponding to the gesture graph of 10, 15, 20, 25, 30, 35, 40, and 45 respectively. Figure 4b shows the key frames of gesture capture extracted by algorithm 1, corresponding to the gestures of frame 10, 17, 22, 30, 34, 40 and 45 respectively. Figure 5 is the gesture distance diagram of frame 10~45, showing the gesture distance corresponding to each frame. It can be seen that the segmentation of dynamic gesture key frame can effectively represent the whole process of gesture change.

D. DYNAMIC GESTURE RECOGNITION AND CLASSIFICATION

1. Gesture feature fusion

In this paper, the fusion of the global motion of gestures and the local motion of fingers will jointly represent a dynamic gesture, similar to Luvizon et al. [18] 's idea of integrating features. The feature is fused into an m-dimensional eigenvector $\{y_1, y_2,\ldots, y_m\}$ of a single gesture. In the data set containing N samples, N gesture feature vectors are obtained respectively:

$$\mathrm{Y_i}=\left[\mathrm{y_{i,1}},\cdots,\mathrm{y_{i,m}}\right],\mathrm{i}=1,2,\cdots,\mathrm{N} \tag{12}$$

The feature of each dimension in the eigenvector is normalized respectively:

$$\overline{\mathrm{f_{i,j}}}=\frac{\mathrm{f_{i,j}}-\mathrm{f_j}}{\sigma_\mathrm{j}} \tag{13}$$

Where, $\mathrm{f_j}=\sum_{\mathrm{i=1}}^{\mathrm{N}}\mathrm{f_{i,j}}/\mathrm{N}$, $\sigma_\mathrm{j}=\sqrt{\sum_{\mathrm{i=1}}^{\mathrm{N}}\left(\mathrm{f_{i,j}}-\mathrm{f_j}\right)^2/\mathrm{N}}$.

Thus, the normalized eigenvector of N gestures is:

$$\overline{\mathrm{F_i}}=\left[\overline{\mathrm{f}}_{\mathrm{i,1}},\cdots,\overline{\mathrm{f}}_{\mathrm{i,m}}\right],\mathrm{j}=1,2,\cdots,\mathrm{N} \tag{14}$$

2. Gesture feature dimension reduction

For SVM, the sample eigenvector dimension in this paper is too much, and there is information redundancy in the dynamic gesture key frame. In order to make variables independent from each other and remove noise in gesture features, and at the same time consider category labels existing in samples, LDA method in supervised dimension

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access

reduction is adopted here for feature dimension reduction. The dimension of the feature vector after LDA is c/2, c is the number of gesture categories that need to be classified. Especially note that when c is odd, the dimension after LDA is (c-1)/2. The dimensionality reduction principle of this method is as follows: The data of the same kind should be as close as possible, and the data of different categories should be as far away as possible. In other words, after projection, the intra-class variance is the smallest, while the inter-class variance is the largest. In the process of dimension reduction, the priori knowledge of gesture categories is fully utilized. The process of mapping gesture features into a low-dimensional space makes full use of the information of gestur e categories. In this way, the characteristic variance between different types of gestures is maximized, while the characteristic variance between the same type of gestures is minimized, which is convenient for gesture recognition and classification.

3. Gesture recognition and classification based on SVM with Gaussian kernel

Support vector machine (SVM) improves the generalization ability through the theory of structural risk minimization, and transforms the optimal classification surface problem into a convex quadratic programming problem. It solves the practical problems of small sample, nonlinear, high dimension and local minimum. In the solution of the problem, the symmetric and positive semidefinite kernel function is used to map the input samples to the high-dimensional feature space, so that the linear inseparable problem can be transformed into a linear separable problem. The kernel function must satisfy the condition in mathematics. The commonly used kernel functions are linear, polynomial, Gauss, Sigmod. Kernel function is not often used at present, but through our analysis of its properties, we can better serve our life applications. Gaussian kernel function is one of the most commonly used kernel functions in solving practical problems, which can provide satisfactory results for practical problems. Therefore, the application and characteristics of them have become a very meaningful research content.

The general form of Gaussian kernel function is as follows:

$$k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\delta^2}\right),\delta>0 \qquad (15)$$

Gaussian kernel function has separability and locality. The separability of kernel functions refers to the ability of feature transformations derived from kernel functions to linearly separate training samples in feature space for a given training sample. Normally, we don't know if the sample is actually being acted upon to become linear separable in the middle until we use it. When selecting a Gaussian kernel function, training samples can almost always be linearly separable by selecting the appropriate parameters, which is guaranteed by its nature.

For the Gaussian kernel function, when the value of the kernel radius δ is very small, although the training sample is separable linearly, it is easy to produce over-fitting, which results in poor generalization of the hyperplane. The reason is that when the value of δ is very small, it only affects the samples in a small field whose distance is comparable to that of δ. When the distance between samples is much greater than δ, its value gradually tends to zero. Therefore, the Gaussian kernel function has strong interpolation ability and is good at extracting the local properties of samples. In order to determine the best value of δ, we take 20 values evenly between 0.01-1 and 100 values evenly between 10-1000. It is found that the best δ appears between 0.6 and 0.7, so we choose 0.65 as the value of δ.

Compared with other machine learning classification methods, SVM avoids the complexity of high-dimensional space and directly uses kernel function to map to high-dimensional space, and then directly solves the corresponding high-dimensional space decision problem with the solution method in the case of linear divisible. When the kernel function is known, the difficulty of solving high dimensional space problems can be simplified. At the same time, SVM has a good theoretical basis and does not involve probability measure, and the final decision function is only determined by a small number of support vectors. Computational complexity depends on the number of support vectors, not the dimensionality of the sample space, thus avoiding the dimensionality disaster.

In this paper, SVM with Gaussian kernel is used to realize the recognition and classification of dynamic gestures. This method can find the compromise between the learning accuracy and learning ability of the specific training sample according to the limited sample information, and has advantages in solving the small sample, nonlinear and high dimensional recognition. And the loss function used in the training process is shown in Eq. (16).

$$Loss_1=\frac{1}{n}\sum_{i}^{n}\left[p_i\log(q_i)+(1-p_i)\log(1-q_i)\right] \qquad (16)$$

Where n is he number of gesture categories that need to be classified. p_i is ground truth and q_i is the probability of prediction.

IV. EXPERIMENTS AND ANALYSIS

The experimental platform of this paper is Intel Core i7-9700, with 16 GB RAM and Windows10 64-bit operating system. Based on the 3D coordinate information of the hand node, this paper determines the starting and ending frames and deletes invalid frames of the gesture. Then, the joint length was normalized to eliminate individual differences, and the key frame of the gesture was extracted. Then, the global motion of the hand in space and the local motion of the fingers inside the hand were extracted respectively, and the feature fusion and LDA dimension reduction were carried out. Finally, SVM with Gaussian kernel is used for dynamic gesture recognition and classification.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access

A. DATASETS

TABLE I
THE GESTURE CATEGORIES CONTAINED IN THE DATASET

NO.	Gestures	Category
1	Grab	Fine
2	Expand	Fine
3	Pinch	Fine
4	Rotation CW	Fine
5	Rotation CCW	Fine
6	Tap	Coarse
7	Swipe right	Coarse
8	Swipe left	Coarse
9	Swipe up	Coarse
10	Swipe down	Coarse
11	Swipe X	Coarse
12	Swipe V	Coarse
13	Swipe +	Coarse
14	Shake	Coarse

The data set used in this paper is DHG-14/28 dynamic gesture data set [3]. The dataset contains 14 dynamic gesture categories, as shown in Table 1, and performs gestures in two ways: with just one finger and with the entire hand. Each gesture was completed by 20 participants in the above two ways, each executed five times, with a total of 2800 dynamic gesture sequences. Among the 14 gestures, 5 are Fine gestures and 9 are Coarse gestures. At the same time, the data set contains not only the dynamic gesture video frame depth image, but also 22 hand joint coordinates in 2D depth image and 3D space, in which the depth image resolution is 640×480, and the depth image and hand skeleton are captured at the speed of 30 frames /s.

FPHA dataset. This dataset contains 1175 action videos belonging to 45 different action categories, in 3 different scenarios, and performed by 6 actors. Action sequences present high inter-subject and intra-subject variability of style, speed, scale, and viewpoint. The dataset provides the3D coordinates of 21 hand joints as DHG dataset except for the palm joint. We used the 1:1 setting proposed in [31] with 600 action sequences for training and 575 for testing.

B. THE DETERMINATION OF FRAMES OF KEYS FOR GESTURES

In dynamic gesture key frames extraction, it is necessary to determine the frames of key frames, which will affect the recognition accuracy of gestures. In this paper, we compare the accuracy of gesture recognition and analyze the frame k value of different key frames. As shown in Figure 6, as the number of frames of key frames increases, the accuracy of gesture recognition increases and tends to be stable. The accuracy of gesture recognition tends to decrease when the key frames are larger than 31. It can be seen from Figure 7 that for 28 kinds of gestures in DHG-14/28 dynamic gesture data set [3], when the number of key frames is k=31, the gesture recognition accuracy is 88.29%, reaching the highest. The experiment shows that if the key frames are small, the key frames of the same gesture may be different, which leads to the low accuracy of gesture recognition. Therefore, in order to improve the accuracy of gesture recognition, a relatively large number of gesture key frames should be reserved to avoid the above problems. At the same time, even if some key frames of the same gesture differ greatly, there are still enough other key frames to shorten the distance between the same gesture and play a role in widening the gap between different gestures. In the experiment of this paper, k=31 key frames are selected for each dynamic gesture video. The global motion of the hand in space and the local motion of the fingers inside the hand are extracted according to the key frame, and the recognition and classification of dynamic gestures are realized based on gesture feature representation.

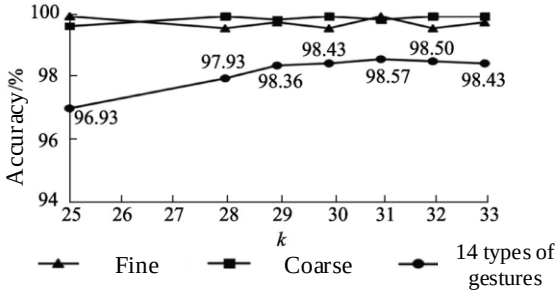


FIGURE 6. The gesture recognition accuracy of 14 gestures with different k values in DHG-14/28

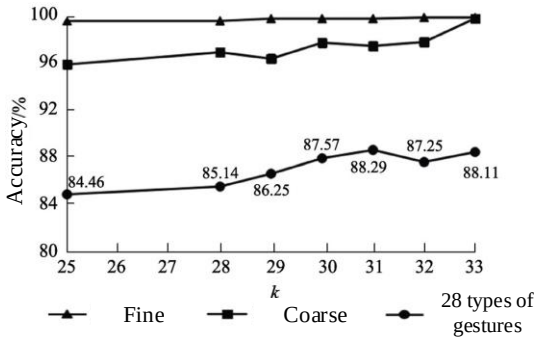


FIGURE 7. The gesture recognition accuracy of 28 gestures with different k values in DHG-14/28

C. COMPARISON WITH DIFFERENT RECOGNITION METHODS

In order to demonstrate the effectiveness of this method, it is compared with 14 existing gesture recognition methods [3,6,19-28,29-31]. The same experimental setup was followed in comparison, and the experiment was carried out by cross validation. The dynamic gesture data of 19 subjects were trained and the recognition test was conducted on the gesture data of the remaining 1 subject. The experiment was conducted with different test subjects' gesture data, and the experiment was repeated 20 times. In this paper, DHG-14, DHG-28 and FPHA of gestures were verified by experiments.

For DHG-14, the method presented in this paper was compared with 12 other dynamic gesture recognition methods. As can be seen from Table 2, the highest

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3065849, IEEE Access

recognition accuracy of the existing methods is 87.25%, while the method in this paper improves by 11.27% on this basis, and the gesture recognition accuracy reaches 98.46%. Meanwhile, this article experimented with the Fine class data and Coarse class data respectively. The recognition accuracy of existing methods [3,20-23,27-28] on Coarse gestures generally exceeds that of Fine gestures by 10.00%. However, the recognition accuracy of the method in this paper is relatively similar in these two categories of data, indicating that the gesture recognition method proposed in this paper has strong generalization ability. The recognition accuracy of Fine and Coarse gesture data is over 99.50%. Compared with the existing best methods, the recognition accuracy of this method is 18.10% and 4.69% higher respectively.

For DHG-28, 28 gesture recognition methods are compared with 12 other dynamic gesture recognition methods. As can be seen from Table 3, the highest recognition accuracy of the existing method is 83.58%, while the method in this paper improves by 4.89% on this basis. Experiments are carried out on Fine data and Coarse data respectively. compared with Coarse data, the recognition accuracy of Fine data by the methods in literature [3,20] is generally about 15.00% lower. In addition, the recognition accuracy of the method in this paper is only 2.37% lower than that in the 2 types of data, which indicates that the method in this paper has strong generalization ability for the Fine class data and Coarse class data in 28 gestures. The recognition accuracy of the method in this paper is more than 88.29% on both types of data, which is 27.90% and 11.03% higher than that of the existing best method. Meanwhile, the recognition accuracy of the method in this paper is higher than that of the data in Coarse class. It shows that the method in this paper pays more attention to the changes of hand details in dynamic gesture feature representation and extraction. The gesture feature integrates the global motion feature of the hand in space and the local motion feature of the fingers inside the hand.

For FPHA, the method presented in this paper was compared with 6 other dynamic gesture recognition methods. As can be seen from Table 4, the highest recognition accuracy of the existing methods is 93.77 %, while the method in this paper improves by 3.78% on this basis, and the gesture recognition accuracy reaches 97.31%. Meanwhile, we calculated the average recall rates and error rates for gestures using different methods. The method proposed in this paper has the highest recall rate for gestures, and the extraction of gesture key frames can capture gestures to the maximum extent and reduce miss detection.

In addition, the method in this article is less recognizable in the case of a combination of the two broad categories of gesture data than it was when you used the Fine or Coarse class data alone. This is because SVM is a convex optimization problem in nature. If the added dynamic gesture sample data is only an invalid constraint (non-support vector), the gesture classifier will not be affected. If there are enough support vectors in SVM, the increase of gesture sample data may lead to overfitting, which often performs poorly on larger data sets. Therefore, SVM is suitable for small sample dynamic gesture data set, and the increase of gesture sample data will slow down the SVM training speed, and may lead to skewed data. Therefore, the error of the learned gesture classifier results leads to the decrease of the recognition accuracy.

TABLE II
COMPARISON OF RECOGNITION ACCURACY OF 14 GESTURES (%)

Methods	Category		
	Fine	Coarse	Both
Skeleton-based[3]	73.61	88.32	83.10
CNN LSTM[20]	78.02	89.63	85.22
MFA RNN[21]	76.95	89.44	84.65
Skeletal Quads[22]	70.58	92.13	84.28
D-Pose TC[23]	81.91	95.03	85.29
SPD-ML[6]			87.25
HON4D[24]			70.83
HOG2[25]			85.24
A SoJT[19]			82.74
NIUKF-LSTM[26]			84.27
SL-fusion-Average[27]	76.33	90.68	85.58
MFA-Net[28]	75.42	91.52	85.84
Ours(k=31)	100	99.87	98.46

TABLE III
COMPARISON OF RECOGNITION ACCURACY OF 28 GESTURES (%)

Methods	Category		
	Fine	Coarse	Both
Skeleton-based[3]	68.32	86.56	79.83
CNN LSTM[20]	71.79	86.26	81.04
MFA RNN[21]			80.57
Skeletal Quads[22]			79.34
D-Pose TC[23]			80.75
SPD-ML[6]			83.58
HON4D[24]			74.58
HOG2[25]			76.74
A SoJT[19]			68.34
NIUKF-LSTM[26]			80.34
SL-fusion-Average[27]			74.65
MFA-Net[28]			81.47
Ours(k=31)	99.71	97.53	88.29

TABLE IV
COMPARISON OF RECOGNITION PERFORMANCE OF FPHA (%)

Methods	Category		
	mRecall	Error	Accuracy
SPD-ML[6]	85.71	7.33	84.16
3DCNN+LSTM[12]	88.59	5.32	86.29
MSF[30]	82.67	9.50	83.54
Hierarchical Net[31]	93.14	2.41	93.77
DS Evidence Theory 32]	82.53	7.63	83.91
MFA-Net [28]	93.27	3.42	92.24
Ours(k=31)	96.36	1.58	97.31

V. CONCLUSION

A dynamic gesture recognition framework is proposed based on the fusion of global gesture motion and local finger motion. The experimental results show that this method can effectively extract the key frames in the dynamic gesture, and use the key frames to replace all the dynamic gesture frames, avoiding data redundancy. At the same time, this paper proposes a dynamic gesture feature representation, which is characterized by the global motion feature of the hand in space and the local motion feature of the fingers inside the hand, and each motion is further divided into rotation motion

and translation motion. The experimental results show that the static gesture model and dynamic gesture model are effective, but there are still some shortcomings and improvements. Due to the limitation of gesture data set, this method does not consider the self-rotation of dynamic gesture.

The dynamic gesture recognition algorithm in this paper can be applied to the interactive teaching system. In addition, it will have a broad application prospect combined with virtual simulation. Virtual simulation technology has become popular in teaching practice, but the current virtual simulation teaching system lacks a convenient way of interaction. Dynamic gesture recognition combined with virtual simulation technology realizes a new generation of teaching system. Teachers can use gestures to control the virtual simulation system and switch the courseware. Students can also use gestures to respond to the teacher's interaction. In the future, our research work will focus on the combination of virtual simulation and hand recognition to build a new teaching system suitable for the classroom. In addition, we found that redundancy still exists in the selected key frames, so how to optimize the extraction method of feature frames to enhance the real-time performance of the network is the focus of our future research.

REFERENCES

[1] M. J. Reale, S. Canavan, L. Yin, K. Hu and T. Hung, "A Multi-Gesture Interaction System Using a 3-D Iris Disk Model for Gaze Estimation and an Active Appearance Model for 3-D Hand Pointing," in IEEE Transactions on Multimedia, vol. 13, no. 3, pp. 474-486, June 2011, doi: 10.1109/TMM.2011.2120600.

[2] R. R. Itkarkar and A. V. Nandi, "A survey of 2D and 3D imaging used in hand gesture recognition for human-computer interaction (HCI)," 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Pune, 2016, pp. 188-193, doi: 10.1109/WIECON-ECE.2016.8009115.

[3] Q. De Smedt, H. Wannous and J. Vandeborre, "Skeleton-Based Dynamic Hand Gesture Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, 2016, pp. 1206-1214, doi: 10.1109/CVPRW.2016.153.

[4] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 1933-1941, doi: 10.1109/CVPR.2016.213.

[5] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree and J. Kautz, "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 4207-4215, doi: 10.1109/CVPR.2016.456.

[6] X. S. Nguyen, L. Brun, O. Lézoray and S. Bougleux, "A Neural Network Based on SPD Manifold Learning for Skeleton-Based Hand Gesture Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 12028-12037, doi: 10.1109/CVPR.2019.01231.

[7] Z. Huang, C. Wan, T. Probst and L. Van Gool, "Deep Learning on Lie Groups for Skeleton-Based Action Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 1243-1252, doi: 10.1109/CVPR.2017.137.

[8] R. Vemulapalli, F. Arrate and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 588-595, doi: 10.1109/CVPR.2014.82.

[9] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012, pp. 1290-1297, doi: 10.1109/CVPR.2012.6247813.

[10] B. G. Celler, P. N. Le, A. Argha and E. Ambikairajah, "GMM-HMM-Based Blood Pressure Estimation Using Time-Domain Features," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 6, pp. 3631-3641, June 2020, doi: 10.1109/TIM.2019.2937074.

[11] C. Jian, J. Li and M. Zhang, "LSTM-based dynamic probability continuous hand gesture trajectory recognition," in IET Image Processing, vol. 13, no. 12, pp. 2314-2320, 17 10 2019, doi: 10.1049/iet-ipr.2019.0650.

[12] 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," in IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 1011-1021, April 2019, doi: 10.1109/TMM.2018.2869278.

[13] Torres-V alencia C A, Garc á H F, Holgu ń G A, et al. “Dynamic hand gesture recognition using generalized time warping and deep belief networks” International Symposium on Visual Computing. Heidelberg: Springer, 2015: pp. 682-691

[14] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.

[15] D. Wu et al., "Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 8, pp. 1583-1597, 1 Aug. 2016, doi: 10.1109/TPAMI.2016.2537340.

[16] M. Abavisani, H. R. V. Joze and V. M. Patel, "Improving the Performance of Unimodal Dynamic Hand-Gesture Recognition with Multimodal Training," 2019 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 1165-1174, doi: 10.1109/CVPR.2019.00126.
[17] A. D. Deshpande et al., "Mechanisms of the Anatomically Correct Testbed Hand," in IEEE/ASME Transactions on Mechatronics, vol. 18, no. 1, pp. 238-250, Feb. 2013, doi: 10.1109/TMECH.2011.2166801.
[18] Luvizon D C, Tabia H, Picard D.: "Learning features combination for human action recognition from skeleton sequences". in Pattern Recognition Letters, vol. 99, pp. 13-20, 2017.
[19] de Smedt Q, Wannous H, V andeborre J P: "3D hand gesture recognition by analysing set-of-joints trajectories", International Workshop on Understanding Human Activities through 3D Sensors. Heidelberg: Springer, 2016: pp. 86-97
[20] Núñez J C, Cabido R, Pantrigo J J, et al.: "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition". In Pattern Recognition, vol. 76, no.43 pp. 80-94, 2018
[21] X. Chen, H. Guo, G. Wang and L. Zhang, "Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 2881-2885, doi: 10.1109/ICIP.2017.8296809.
[22] G. Evangelidis, G. Singh and R. Horaud, "Skeletal Quads: Human Action Recognition Using Joint Quadruples," 2014 22nd International Conference on Pattern Recognition, Stockholm, 2014, pp. 4513-4518, doi: 10.1109/ICPR.2014.772.
[23] Weng J W, Liu M Y, Jiang X D, et al.: "Deformable pose traversal convolution for 3D action and gesture recognition" European Conference on Computer Vision. Heidelberg: Springer, pp.142-157, 2018.
[24] O. Oreifej and Z. Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, 2013, pp. 716-723, doi: 10.1109/CVPR.2013.98.
[25] E. Ohn-Bar and M. M. Trivedi, "Joint Angles Similarities and HOG2 for Action Recognition," 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, 2013, pp. 465-470, doi: 10.1109/CVPRW.2013.76.
[26] Ma C Y, Wang A, Chen G, et al.: "Hand joints-based gesture recognition for noisy dataset using nested interval unscented Kalman filter with LSTM network" in Visual Computer, vol. 34, no.6, pp. 1053-1063, 2018.
[27] K. Lai and S. N. Yanushkevich, "CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition," 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, 2018, pp. 3451-3456, doi: 10.1109/ICPR.2018.8545718.
[28] Chen X H, Wang G J, Guo H K, et al.: "MFA-Net: motion feature augmented network for dynamic hand gesture recognition from skeletal data" in Sensors, vol. 19, no.2, Article No.23, 2019.
[29] Y. Che and Y. Qi, "Embedding Gesture Prior to Joint Shape Optimization Based Real-Time 3D Hand Tracking," in IEEE Access, vol. 8, pp. 34204-34214, 2020, doi: 10.1109/ACCESS.2020.2974551.
[30] P. K. Arachchi, N. L. Hakim, H. Hsu, S. V. Klimenko and T. K. Shih, "Real-Time Static and Dynamic Gesture Recognition Using Mixed Space Features for 3D Virtual World's Interactions," 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), Krakow, 2018, pp. 627-632, doi: 10.1109/WAINA.2018.00157.
[31] O. Köpüklü, A. Gunduz, N. Kose and G. Rigoll, "Online Dynamic Hand Gesture Recognition Including Efficiency Analysis," in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 2, no. 2, pp. 85-97, April 2020, doi: 10.1109/TBIOM.2020.2968216.
[32] H. Lei, Y. Liu and L. Yang, "Dynamic Gesture Recognition Based on DS Evidence Theory," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 6633-6638, doi: 10.23919/CCC50068.2020.9189409.
[33] O. Dospinescu, P. Brodner. Integrated Applications with Laser Technology[J]. Informatica Economica, 2013, vol. 17 no.1, pp.53-61.
[34] O. Dospinescu, I. Popa. Face Detection and Face Recognition in Android Mobile Applications. Informatica Economica Journal, 2016, vol.20, no.1, pp.20-28.
[35] X. He, S. Yan, Y. Hu, P. Niyogi and H. Zhang, "Face recognition using Laplacianfaces," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 328-340, March 2005, doi: 10.1109/TPAMI.2005.55.
[36] G. Li, H. Tang, Y. Su, et al. "Hand gesture recognition based on convolution neural network ". Cluster Computing, 2019, vol. 22, no. 2, pp.2719–2729.
[37] D. Tran, N. Ho, H. Yang, et al. "Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network," Applied Sciences, 2020, vol. 10, no. 2, pp. 722.

LI Jiashan is postgraduate studies in Harbin Medical University. She is now interested in the field of virtual simulation. And she has published several academic papers in this field in peer-reviewed journals at home and abroad.

* LI Zhonghua graduated from the School of Humanities and Social Sciences of Harbin Institute of Technology. She is now a professor at Harbin Medical University and served as a master tutor. Her research field is virtual simulation and has published several academic papers in this field in peer-reviewed journals at home and abroad.