Neural Computing and Applications (2020) 32:6327–6342
https://doi.org/10.1007/s00521-019-04141-9

ORIGINAL ARTICLE

Decomposition algorithm for depth image of human health posture based on brain health

Bowen Luo[1] · Ying Sun[1,2] · Gongfa Li[1,3,4] · Disi Chen[5] · Zhaojie Ju[5]

Received: 22 August 2018 / Accepted: 9 March 2019 / Published online: 23 March 2019

Abstract
At this stage, brain health can be directly expressed in the human hand posture estimation. Therefore, model estimation of healthy human hand posture can also be used as a criterion for brain health. The recognition algorithm of healthy human hand gesture based on global feature extraction of depth map sequence is not enough to analyze the motion correlation of healthy human hand posture, which leads to the need to improve the accuracy of human body hand gesture description and the change of movement speed of robustness. After analyzing the characteristics of healthy human hand movement in detail, this paper proposes a hand posture decomposition algorithm based on depth map sequence. The goal is to find information that plays a key role in hand gesture recognition in the depth map sequence. The algorithm can remove redundant information and improve the robustness of the recognition algorithm.

Keywords Brain health · Decomposition algorithm · Feature extraction · Depth map sequence

1 Introduction

The rapid development of modern information technology, followed by the gradual update of computer application technology, human–computer interaction, has become an important direction of computer science research (Human–Computer Interaction, HCI). At present, the relationship between computers and users has long been adapted from users to computers, gradually moving toward user-centered, and computers are actively adapting to the new stage of users. Computers will also be designed to be more and more intelligent and more humane to meet the needs of different groups of people. The development of human–computer interaction has gone through the interface of the keyboard and mouse as the main tools. Although the keyboard and mouse are still used as the main tools, the inconvenience and unnaturalness exposed during its use have greatly limited the further development of human–computer interaction [1].

When designing the human–computer interaction system, if the human hand gesture language and movement trajectory are used as a new communication method, the communication between the human and the machine will become intelligent and convenient. Compared with the traditional keyboard, mouse and other external devices, gesture operation is intuitive and natural. However, due to the rich information represented by human gestures, especially in the gesture recognition process, it is difficult for the opponent to locate the gesture, and there is still a gap between the recognition effect and other recognition technologies. Influenced by human subjectivity and objective factors, the content of the gestures to be expressed is ambiguous, and the expression is also diversified. Therefore, gestures have the characteristics of

✉ Gongfa Li
ligongfa@wust.edu.cn

1 Key Laboratory of Metallurgical Equipment and Control Technology of Ministry of Education, Wuhan University of Science and Technology, Wuhan 430081, China

2 Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan 430081, China

3 Research Center of Biologic Manipulator and Intelligent Measurement and Control, Wuhan University of Science and Technology, Wuhan 430081, China

4 Institute of Precision Manufacturing, Wuhan University of Science and Technology, Wuhan 430081, China

5 School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK

flexibility and intuitiveness in human–computer interaction. With the development of various types of artificial intelligence devices, gesture recognition has an important significance in the field of human–computer interaction.

In general, gesture recognition can be divided into static gesture recognition and dynamic gesture recognition. Among them, the recognition of the static gesture refers to the recognition of changes in the hand shape and the orientation of the hand at a certain time, but the recognition of the dynamic gesture is the process of focusing on analyzing the movement of the hand. In other words, dynamic gesture recognition is the sequence of trajectories formed by the movement of the hand identified within a period of time. Based on different input devices, gesture recognition includes wearable device-based gesture recognition and computer vision-based gesture recognition. Gesture recognition based on wearable devices mainly utilizes data gloves and corresponding 3D tracking devices [2]. Among them, the data glove is made of a material having certain elasticity. At the same time, the data glove is also installed with a sensor corresponding to the joint of the human hand and can be used to detect the unfolding and bending of the finger. Through the data glove, the sensor's resulting information is passed to the computer to get the dynamic information of the gesture in real time. The gesture recognition using the 3D tracking device needs to be installed on the user's arm or hand, and spatial information such as the relevant position can be obtained. Although this method has a small amount of data and a high processing speed, it requires the user to wear a complicated device which greatly affects the flexibility of operation and natural comfort; furthermore, the price of the device is relatively expensive. In vision-based gesture recognition, the input data only contain the image of the hand when the input device is just a 2D camera. Therefore, it is necessary to obtain the characteristics of the gesture through a certain image processing algorithm and then the recognition can be performed [3].

In the new dynamic gesture recognition development background, dynamic gesture recognition based on depth information has the following advantages over traditional gesture recognition methods: First, equipment costs are reduced. Contact sensors have a more sophisticated structure, and they are often more expensive. On the premise of ensuring accuracy, the depth camera reduces the cost of acquiring 3D information which is an inevitable advantage for the application of gesture recognition based on depth image. Second, depth maps are not affected by lighting conditions and background complexity. Using depth maps for hand gesture recognition can easily and accurately implement segmentation of the hand region and remove the effect of the background on gestures based on the depth information, thereby further improving the robustness of the algorithm. Third, dynamic gesture recognition based on depth information has rich three-dimensional information. Compared with traditional gesture recognition algorithms based on color images, depth information-based gesture recognition introduces depth maps which allows gesture recognition algorithms to return to the analysis of gesture space shapes [4]. Furthermore, this provides more distinguishing feature information for gesture recognition. Finally, gesture recognition based on depth information is more natural and flexible. From the perspective of the development trend of human–computer interaction, the way of interaction has become increasingly natural and convenient [5]. Gesture recognition based on depth information does not require users to wear complex devices and is easy to learn and master. With the increasing demand for user experience, this way of gesture interaction has become an inevitable trend [6]. These features make the dynamic gesture recognition based on depth information closer to the requirements of practical applications. On the other hand, gesture recognition based on depth information is still in the basic research stage. The recognition efficiency of dynamic gesture recognition based on depth information in extreme environments such as strong or weak illumination is significantly better than traditional gesture recognition methods. The research on dynamic gesture recognition based on depth information can be a good complement to the traditional gesture recognition method, which will make the final recognition result more accurate. Therefore, our research should continue to deepen in this area.

This paper mainly introduces the gesture decomposition algorithm based on depth map sequence. The fourth chapter mainly introduces the direction-based depth map gesture shape feature extraction algorithm. The direction-based depth map gesture shape feature has the characteristics of low dimension and low extraction method complexity and also can well describe the edge direction information of the hand region;

The fifth chapter mainly introduces the depth map sequence gesture decomposition algorithm based on spectral clustering algorithm. This algorithm divides the depth map sequence into different sub-segments according to the similarity between frames which is to divide the gesture into different sub-processes. At the same time, we extract the corresponding key nodes from each sub-process. The sixth chapter is mainly based on the sequence decomposition of the key node extraction algorithm to extract the key point set from the decomposed subsequence, and then, the new sequence consisting of the key point set, at the same time, removes the redundant information of the time domain. In addition, the depth map sequence decomposition algorithm based on feature similarity distance is proposed to adapt to the application scenario with low

computational performance. The experimental results show that the proposed algorithm effectively implements the decomposition of depth map sequences and the extraction of key points.

2 Related works

At present, commonly used gesture recognition methods include template-matching method and state transition-based graph model method.

The basic idea of the template-matching method is to calculate the similarity between the input gesture and the known gesture template to recognize the gesture, which is the simplest gesture recognition method. Specifically, a corresponding template is established for each gesture training by using a gesture sample. When a new gesture is identified, its feature vector is first calculated and then matched with the known template one by one. The gesture type corresponding to the template with the highest similarity is the recognition result. Three functions are generally used to measure the similarity between input gestures and gesture template feature sequences: squared differences, correlation coefficients and related matches. Literature [7] uses skin history images for gesture modeling and uses *K*-means clustering algorithm to train to get gesture templates. The similarity between the two is calculated by calculating the tangent distance between the input gesture and the gesture template.

The template-matching method is simple to set up and modify, but when the dynamic gesture becomes very complicated, the difference in time and space of the gesture makes the template threshold of each gesture become larger; that is, the matching range is set to be wider. When there are many types of gestures, there may be a case where a template matches several gestures at the same time, which eventually leads to recognition errors. The recognition speed of the template-matching method is gradually reduced as the number of gestures increases. Therefore, the template-matching method cannot solve the problem of time and space difference of gestures and cannot accurately realize real-time multi-gesture recognition.

Dynamic time warping (DTW) [8] is a time-varying data sequence matching method that eliminates differences in dynamic gesture time by adjusting the sequence of gestures on the time axis. DTW adjusts the timeline to nonlinearly map the input gesture to the timeline of the template gesture, minimizing the distance between the two and then performing template matching to get the final recognition result. The advantage of the DTW algorithm is that the data training is simple and easy to implement. The disadvantage is that a large number of template-matching calculations are difficult in real time and are susceptible to noise. In addition, when gestures are more complex, such as large differences in time or large amplitude changes, or when encountering undefined interactive gestures, the recognition effect of DTW will be worse.

The state transition-based graph model method uses the nodes or states of the graph model to describe each static pose or motion state and is linked by various probabilities through corresponding graph model nodes. In the literature [9], the characteristics of the position of the dynamic gesture, the direction of the trajectory and the speed of motion are set as observation sequences. The HMM model of each gesture is trained to analyze the time series, and the timescale is not deformed under the condition of time and space changes. HMM is a widely used statistical method. Its topology is general, not only can describe the shape, position, direction and motion characteristics of the hand, but also can describe the difference in time between gestures, especially for complex dynamic gesture recognition. HMM training and recognition calculations are very large and have difficulty meeting actual requirements.

Literature [10] adopts a dynamic gesture recognition method based on feature package (BOF-based). Firstly, the local region is extracted from the sample, and the operator is described by SIFT. Then, the feature dictionary is constructed by clustering algorithm, and the feature package of each type of gesture is calculated, which is used as the basis for gesture classification. Local image blocks of the method are typically obtained by feature point detection, random sampling, and thus local features typically do not have explicit semantics.

The neural network has strong adaptive learning ability and fault tolerance. The neural network and DTW are combined. At the same time, the recognition method based on hand shape and motion trajectory is proposed in the literature [11]. Neural networks are used for classification and recognition of hand shapes, and DTW is used for motion path recognition. The neural network is not susceptible to noise and has strong fault tolerance. However, the training calculation is large and the ability to model time series is poor, which cannot solve the problem of gesture time difference.

3 Gesture decomposition core framework

At present, the process of a typical gesture recognition algorithm based on the depth map sequence is shown in Fig. 1. Firstly, we acquire the depth map sequence by the depth sensor and the hand movement is tracked, then the hand feature extraction is performed, and finally the final output gesture label is obtained by classifier classification. Feature extraction is one of the key issues, and its accuracy directly affects the accuracy of gesture recognition;

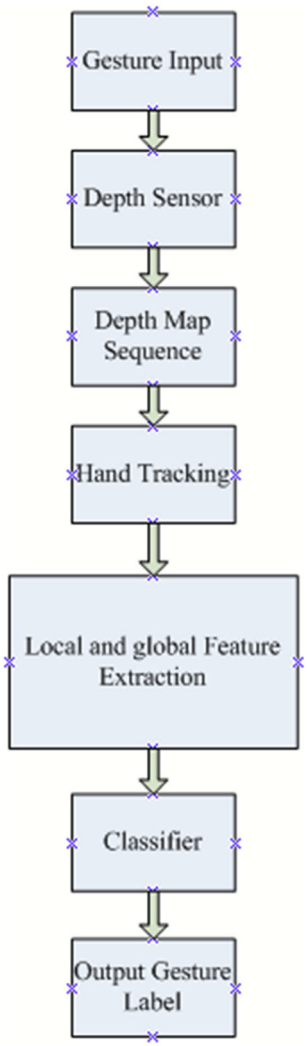

Fig. 1 Dynamic gesture recognition algorithm flow of depth map sequence

however, there are still some challenging problems in the feature extraction accuracy, robustness and computational complexity. First, for the same gesture, different people have different speeds when doing the gesture. This difference leads to the inconsistency in the time-domain motion intensity of the same gesture frame in the sequence of depth pictures and reduces the similarity between the same gestures [12]. In addition, by observing the series of depth maps acquired in real time, it is found that there is a lot of redundancy between adjacent frames. Such redundancy does not contribute to the feature extraction in gesture recognition and even reduces the distinguishability of the features. However, it takes a lot of computational cost to process this redundant information. This is the main reason for the low efficiency of the current algorithm for gesture recognition based on the global features of depth map sequences [13].

This paper proposes a method based on dynamic gesture decomposition to solve the above challenges: By analyzing the similarity between frames and frames, the gesture sequence is decomposed into multiple independent subsequences. The frames within the subsequence have relatively slow motion changes, and the key gesture action nodes in the subsequence are extracted to form the last sequence used for recognition. Identifying the sequence of these key action nodes can not only overcome the difference of the same gesture caused by the different speeds of the collected person, but also remove the redundant frames in the sequence, so as to improve the efficiency of the gesture recognition algorithm [14].

The overall block diagram of the algorithm is shown in Fig. 2. First, we extracted the image features of each frame from the depth map sequence, and then the depth map sequence decomposition strategy is adopted to divide the entire depth map sequence into several fragments. Then, a key node is extracted from each segment as a representative of this segment, so that a sequence of depth maps consisting of a set of key nodes is obtained. And the new sequence will be used as the input to the gesture recognition algorithm to further perform gesture feature extraction and gesture recognition. Among them, the key issues that need to be resolved include these points: First of all, we need to analyze the similarity between the frame and the frame and also need to design a reasonable depth map feature so that it can accurately distinguish the differences between different frames [15]. Second, we need to design a reasonable depth map sequence decomposition algorithm

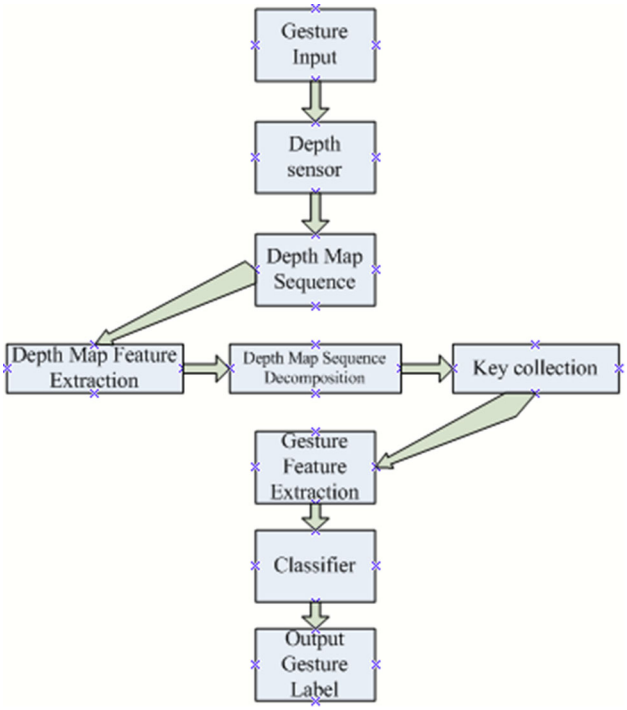

Fig. 2 Dynamic gesture recognition flowchart for gesture decomposition

which can decompose the depth map sequence into several independent subsequences [16]. The similarity between frames and frames within the subsequence is the highest, and the similarity of the images is the lowest. In the end, we need to further adopt a reasonable strategy to extract the most representative key sub-action nodes from each subsequence, so as to achieve the construction of the key point set [17].

4 Feature extraction algorithm for depth image gesture direction

Since there is no texture information in the depth map, the description of the gesture is more dependent on the shape of the hand. But the different hand types can be described by the edge information of the hand [18]. The Sobel operator can extract the direction information of the edge which is the effective expression of the shape feature of the hand. Therefore, this paper uses Sobel operator to calculate the depth map pixel gradient. The Sobel operator was proposed by Irwin Sobel in 1970 which is a one-step operator for edge detection [19]. It includes a horizontal operator and a vertical operator, and they are, respectively, expressed as:

$$S_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad S_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

Convolution of these two operators with the image can yield the transverse gradient G_x and longitudinal gradient G_y of each pixel point, as shown in Eq. 1:

$$\begin{cases} G_x = S_x * I \\ G_y = S_y * I \end{cases} \tag{1}$$

The angle between the horizontal gradient and the vertical gradient is the direction of the gradient. The direction of the gradient can describe the directionality of the edges in the image [20].

4.1 Gesture shape feature extraction

The depth map can describe the spatial geometric properties of the object well, and the difference between the hand shapes can be distinguished by the edge contour of the hand. However, depth maps often have a large area of depth-invalid regions, i.e. regions with zero pixel values. Pixel-level features exhibit sparseness and cannot describe image content well [21]. For this reason, a block-based gesture shape feature is proposed in this paper. Since pixel-level features exhibit sparseness, the image content cannot be well described. For this reason, a block-based gesture shape feature is proposed in this paper. For each depth map, the shape of the hand can be described using the edge direction of the hand. For a given depth map I, find the lateral gradient G_x and the longitudinal gradient G_y according to the Sobel operator introduced in the previous, and then use formula 2. Find the direction angle of each pixel's gradient:

$$\theta = \arctan\left(\frac{G_x}{G_y}\right) \tag{2}$$

Because the edges of the depth map are not sharp, this will cause the edge direction information obtained by the gradient operator to overlap. In order to remove the excess direction vectors, the gradient values are further processed in the form of block histograms. In order to quantify the direction of the pixel, we need to map it to a gradient pattern and divide the gradient pattern according to the block of $k*k$; finally, the gradient histogram is counted in each block. In this paper [22], the size of the block is set to 8*8 pixels, because the movement of the gesture is not only formed by the rotation of the hand, the forward and backward movement of the hand can also be used as a gesture. Therefore, we do not need to consider the hand scale problem in the decomposition of the gesture sequence. In other words, the characteristics need to be sensitive to changes in the scale of the hand, so as to decompose the gestures of the forward and backward movements [23]. Due to the spatial nature of the depth map, the gradient angle range of the gradient should be [0°, 360°]. In this paper, the quantization step length is 20°; that is, 18 equal parts are divided to ensure the integrity of the spatial data. The largest classification of the histogram of each block is taken as a feature of the block, and this component represents the main direction of the depth edge in this block. The shape feature of the gesture is to connect the main components of each block, represented as $F = \{d_1, d_2, \ldots, d_n\}$. The direction-based gesture shape feature describes the direction of the gesture edge with a low-dimensional feature vector [24].

At present, the directional characteristics of gestures in the depth map have been extracted and used to calculate the feature distance between different frames, and the feature distance can be converted into the similarity between frames. The decomposition process of the depth map sequence is actually a clustering process. Therefore, this paper adopts the clustering method to achieve the decomposition of the gestures. The sub-actions with higher similarity are grouped into one category, and the sub-actions with lower similarity are decomposed into different sub-categories [25].

5 Spectral clustering algorithm and depth map sequence gesture decomposition algorithm

At present, dynamic gesture recognition algorithms based on global features often consider a dynamic gesture as a whole, but in fact a dynamic gesture often consists of multiple sub-processes. Each sub-process can be seen as a basic unit of a dynamic gesture, and the entire gesture is a combination of multiple basic units. This section will focus on the decomposition of dynamic gestures. We will analyze each sub-process and extract a set of key points from it to improve the performance of the gesture recognition algorithm [26]. Finding the segmentation points of different sub-processes is the key to dynamic gesture decomposition. At the same time, the differences between adjacent sub-processes can be measured by changes in the depth map. The decomposition process of the depth map sequence is to divide the depth maps with similar similarity between adjacent frames into the same category, and the similarity frames are divided into different classes [27]. We consider the depth map sequence as a dataset which is actually a clustering process, and each frame of depth map is a point in the dataset. Decomposition of the depth map sequence is to cluster these data points. When the sample space does not satisfy the assumption of convex optimization, the traditional clustering algorithm will fall into a local optimum and cannot obtain a global optimal solution. Spectral clustering algorithm avoids the assumption that the sample space is convex. It simplifies the optimization problem of the graph into the matrix solution problem and realizes the global optimal solution process [28]. This paper will use spectral clustering algorithm to solve the decomposition of the gesture sequence.

5.1 Spectral clustering algorithm

Spectral clustering algorithms are derived from graph optimization theory. For a given dataset, we first need to build a graph model. Assume that the graph model is $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ represents a set of vertices, each v_i represents a data point in the sample and $\{s_{11}, s_{12}, \ldots, s_{ij}, \ldots, s_{nm}\}$ denotes an adjacency matrix. Among them, $s_{ij} \geq 0$ indicates the similarity of any two sample data points v_i and v_j, and s_{ij} is greater than 0 or greater than a certain threshold, indicating that two vertices are connected; otherwise, there is no connection between the two vertices. The process of clustering is to divide the graph model and divide the graph into several subgraphs. The similarity of the vertices in the subgraph is the largest, and the vertex similarity between the subgraphs is the smallest. The optimal solution of the graph partition criterion is an NP-hard problem [29]. However, the spectral clustering algorithm considers the continuous relaxation of the problem and converts this problem into the spectral decomposition of a similarity matrix or Laplacian matrix. This solution is an approximation of the optimal solution to the graph. Finally, by clustering the selected eigenvectors, the dataset at this time satisfies the assumption of convex optimization, and the clustering results can be obtained by using a traditional clustering algorithm [30].

When we use the spectral clustering algorithm to decompose the depth map sequence, we should regard each frame image in the depth map sequence as a vertex and construct an undirected weighted graph model. This graph model connects any two frames with similarity weights [31]. However, this will bring two problems: First, the sequence of depth maps is ordered in the time domain. Constructing the above undirected graph model will break this ordering, resulting in depth maps that are far apart in the time domain which may be divided into one class. Second, the gesture needs to be decomposed into several subsequences that are unknown at the beginning, so when the feature vectors are further clustered, the number of categories cannot be initialized. In order to solve these two problems [32], this paper will further discuss the construction of similarity matrix and the clustering method without initial category parameters.

5.2 Similarity metric matrix

To achieve the decomposition of the depth map sequence, we must first construct the relationship between the frame and frame of the depth map sequence. In this paper, the similarity measure matrix is used to find the similarity between any two frames [33]. From the previous section, a gesture sequence can be represented as a set of gesture shape features. First, the Gaussian similarity function is used to find the similarity between any two frame features, as shown in Eq. 3:

$$s(F_i, F_j) = \exp\left(-\frac{||F_i - F_j^2||}{2\sigma^2}\right) \tag{3}$$

The similarity (F_i, F_j) between any two frames I_i and I_j forms a similarity metric matrix. The parameter is used to control the degree of dispersion of the similarity matrix. As shown in Fig. 3, when takes different values, the similarity metric matrix shows different degrees of dispersion [34]. The smaller the value of is, the more concentrated the matrix is. As increases, the matrix gradually becomes dispersed.

However, as described in the above section, the similarity measure matrix calculated according to the above rules ignores an important problem: The depth map

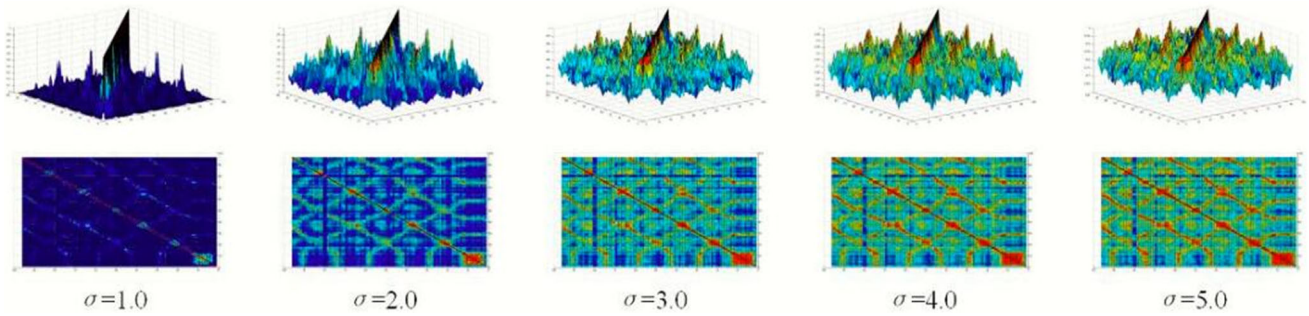

Fig. 3 Distribution map of similar matrix when takes different values

sequence is time series, and the similarity calculation between any two frames neglects the influence of the time sequence change. However, as described in the above section [35], the similarity measure matrix calculated according to the above rules ignores an important problem: The depth map sequence is time series and the similarity calculation between any two frames neglects the influence of the time sequence change. It can be clearly seen from Fig. 3 that the similarity of the depth map sequence in the time domain may also be very large [36]. There are several peaks in addition to the diagonals in Fig. 4 because the gesture movement may have a certain periodicity and the depth map of the same hand may appear multiple times in the sequence. This situation needs to be avoided by weighting in the time domain. If the image frames are only clustered based on the spatial characteristics, then the discontinuous frames will also be clustered into one category and destroy the time-domain adjacency of the subsequences. In order to avoid dividing the non-adjacent depth maps in the time domain into a class, the similarity metric matrix needs to be weighted and constrained by the time-domain conditions [37]. The weights are still obtained using a Gaussian similarity matrix, as shown in Eq. 4:

$$S_t(t_i, t_j) = \exp\left(-\frac{(t_i - t_j)^2}{2\tau_t^2}\right) \tag{4}$$

Among them, the parameter τ_t determines the degree of weight control of the time-domain window. The larger the value of τ_t, the slower the attenuation of the weight and the larger the window in the time domain. The influence of different τ_t on the similarity matrix is shown in Fig. 4, where the similarity matrix takes $\sigma = 5.0$. By weighting the time domain, the similarity measure is $w_{i,j} = s_{i,j} * s_{ti,j}$.

5.3 Laplacian matrix and feature vector selection

To further transform the graph optimization problem into a matrix solution problem, the spectral clustering algorithm converts the similarity matrix into a Laplacian matrix. There are generally two types of Laplacian matrix selections: The first type is a non-canonical Laplacian matrix: $L = D - W$ and the other is the canonical Laplacian matrix. The canonical Laplacian matrix is divided into two forms: $L = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$ and $L = D^{-1}L = I - D^{-1}W$ [38]. The Laplacian matrix first needs to calculate the degree matrix of the graph which is a diagonal matrix composed of the degrees of each vertex. The degree of each vertex is given by Eq. 5 which is the sum of the elements of each row of the adjacency matrix:

$$d_i \sum_{j=1}^{n} w_{i,j} \tag{5}$$

In this paper, we solve the actual depth map sequence and find that the degree of change of the obtained vertex is not significant. So as shown in Fig. 5, when the degree of

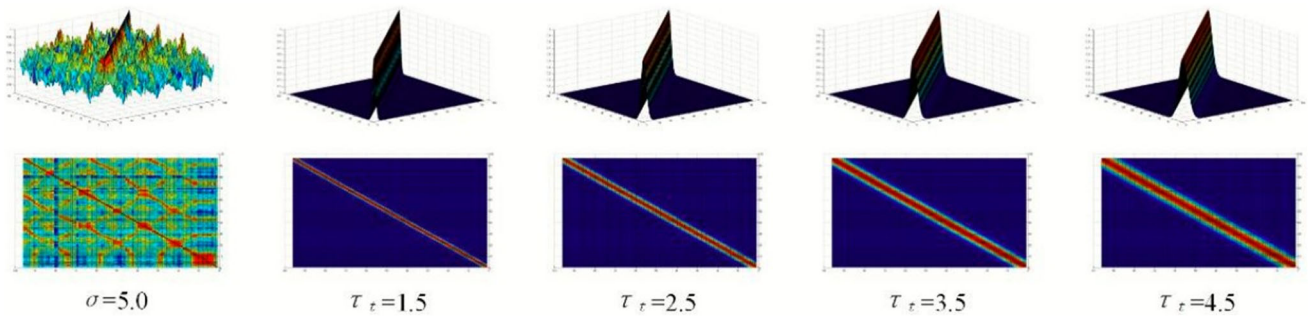

Fig. 4 Weighted similarity matrix with different values of τ_t

the vertices is close, the result of which Laplacian matrix is selected is basically the same.

The adjacency matrix is processed further in the form of a standard Laplacian matrix, and the matrix is solved to obtain eigenvalues and eigenvectors. The eigenvalue of the matrix is 0, and the eigenvector corresponding to the eigenvalue is a constant. The eigenvector corresponding to the second smallest eigenvalue is an approximation of the potential function which is an approximate solution to the best map partition, and this eigenvector is also named Fiedler vector [39]. The potential function referred to here is an indicator vector that indicates to which subgraph the vertex is divided. The indicator vector is a constant vector. If the vertex belongs to a subgraph, then its element is 1, otherwise it is 0. The Fiedler vector is an approximation of the vector. Figure 6 lists the feature vector curves corresponding to the first four minimum eigenvalues and the eigenvector corresponding to the eigenvalue 0. Since the eigenmatrix operation library is used to solve the matrix [40], the limitation of the operation accuracy results in the minimum eigenvalue being close to a very small number of 0, so the feature vector has slight fluctuations. Finally, feature vector 2 is the Fiedler vector.

Since similarity matrices are weighted using the information of the time-domain interval when calculating the adjacency matrix, the similarity of frames that are clustered far away in the time domain is well attenuated [41]. Therefore, the time-domain coherence of the subsequences is preserved and this limitation is also well represented in the Fiedler vector here. When there is a large difference in the independent variables in the Fiedler vector, there is no case where the dependent variables are similar. However, if we do not perform time-domain constraints on the similarity matrix, Fiedler here will have multiple extremum points which will cause the time domain to have a large number of independent variables with similar dependent variables [42]. As shown in Fig. 7, the eigenvalues and eigenvectors obtained when the Laplacian matrix is further solved for the unweighted similarity matrix are shown. Feature vector 2 has multiple peaks. After obtaining the Fiedler vector, we need to further cluster all the points. At this time, we face a new problem. We cannot determine in advance that the depth map sequence needs to be decomposed into several subsequences when clustering. That is, the clustering category number is uncertain. To solve this problem, this paper further proposes a bipartite iterative clustering algorithm [43].

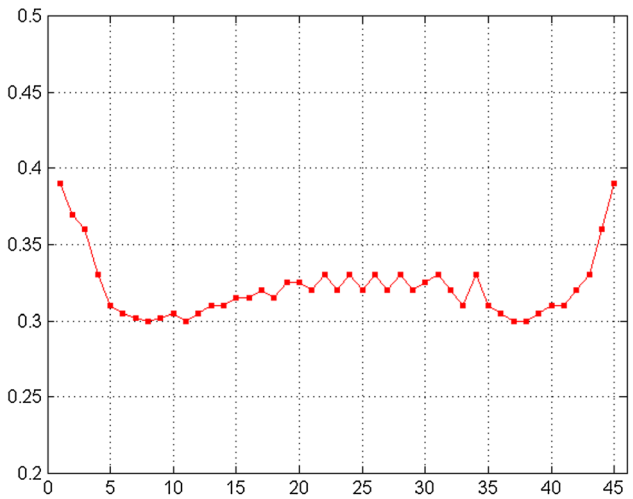

Fig. 5 Distribution of vertex degree

5.4 Iterative clustering algorithm

As mentioned earlier, a gesture depth map sequence includes several segments that cannot be predicted in advance which makes it unable to provide a category number of parameters for the clustering algorithm [3]. Because the traditional K-means clustering algorithm cannot solve this problem, this paper improves it and proposes a binary iterative clustering algorithm.

For the depth map sequence A, firstly find the similarity d_A of the similarity between the frame and the frame. The similarity mean can be obtained from formula 6 by the degree of the vertex:

$$d_A \frac{\sum_{i=1}^{n} d_i}{n^2} \tag{6}$$

In order to determine whether it is necessary to continue segmentation of sequence A, that is whether the cluster iteration stops, find the similarity mean of its two subclasses. Assuming that its two subsequences are B and C [44], then the similarity values d_B and d_c can also be obtained from the vertex degrees of B and C. If the value of $(d_B + d_c)/2$ is similar to d_A which means that the change of sequence A has been relatively smooth and it is stop the iteration. Set the iteration stop condition as shown in Eq. 7.

$$K_{\text{stop}} = \frac{d_A}{(d_B + d_c)/2} \geq \tau_{\text{stop}} \tag{7}$$

When the parameter K_{stop} gradually converges to 1, the change of the sequence A gradually becomes gentle [45]. In this time, the threshold τ_{stop} can be set to control the number of iterations and the number of categories in the sequence. The range of τ_{stop} is [0, 1].

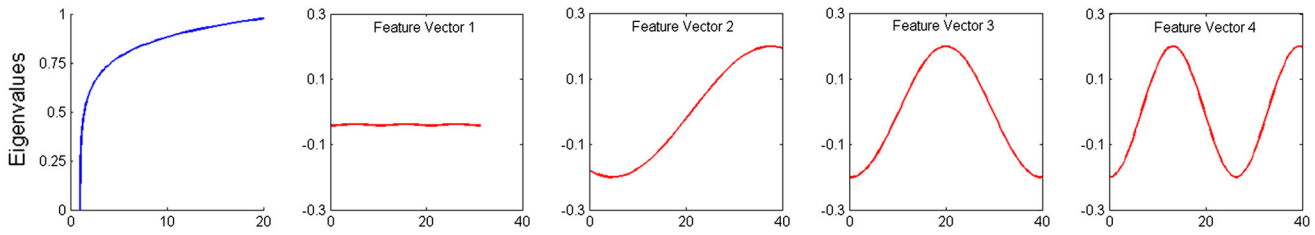

Fig. 6 Eigenvectors corresponding to the first four minimum eigenvalues

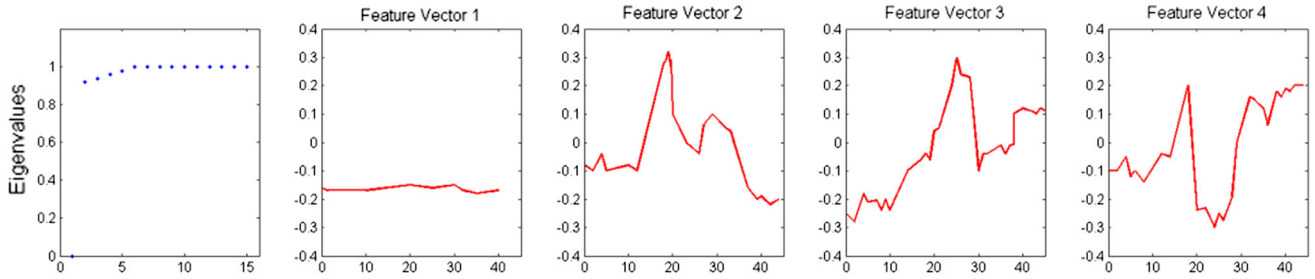

Fig. 7 Unweighted similarity matrix for solving eigenvalues and eigenvectors

6 Gesture key node extraction of sequence decomposition

The extraction of key nodes refers to finding one of the most representative frames in each depth map sequence sub-segment decomposed above and using this as a key node of the gesture. For a sub-fragment, we use the Euclidean distance between any two frames in the same segment to select the key frame, calculate the Euclidean distance between each frame in the sequence and all other frames and select the Euclidean distance between the frame and other frames [46]. Assuming that C_k is the kth segment in the m video segments, the key point selection method is given by Eq. 8:

$$kf_{i,k} = \arg\min \sum_{j \in C_k}^{i \in C_k} \mathrm{dis}\left(F_i, F_j\right) \tag{8}$$

Then, $kf_{i,k}$ represents the key point in the kth segment, and the key node set of the entire gesture is expressed as $\left\{kf_{i_1,1}, kf_{i_2,2}, \ldots, kf_{i_m,m}\right\}$. In the above manner, the most representative frame can be extracted from each subsequence as a key action node. Then, we will make a new sequence of extracted frames for gesture recognition. Key Point Set Extracted from Gesture Decomposition, the time-domain information is compressed to a certain extent which reduces a large amount of redundant information caused by high similarity between adjacent frames. In addition, the gesture decomposition according to the different degree of sequence density can decompose the depth map sequence into subsequences of different sizes. This decomposition can reduce the difference between sequences due to the speed change of the gesture [47].

7 Experimental results analysis and database testing

MSRGesture3D database The database was provided by Microsoft Research. Its depth map was acquired by the first generation of Kinect, and it is the most common depth map gesture database. The gesture database contains 12 dynamic American Sign Language (ASL) signs. Each gesture came from ten volunteers, and each volunteer repeated a gesture two or three times. The gesture set contains 336 depth map sequences, and each of which is a gesture. The gesture set has undergone a series of preprocessing, including removing the arm below the wrist. Figure 8 shows several gesture sequences in the MSRGesture3D database [48].

This paper builds a gesture dataset based on the Kinect 2 generation. In order to create a gesture dataset, four volunteers were called to collect 12 one-handed dynamic gestures. These unit dynamic gestures include a rotation-type gesture, a front-to-back variation gesture and a non-rigid body gesture. We collected eight repetitions of each gesture for each volunteer, and the speed of the same gesture varied [49]. The custom gesture set contains 384 gesture sequences, and the depth map undergoes a series of preprocessing which includes removing the arm area and uniformly setting the resolution to 160*160 pixels. Figure 9 shows several gesture sequences in a custom gestures database.

Fig. 8 Gesture sequence of MSRGesture3D database

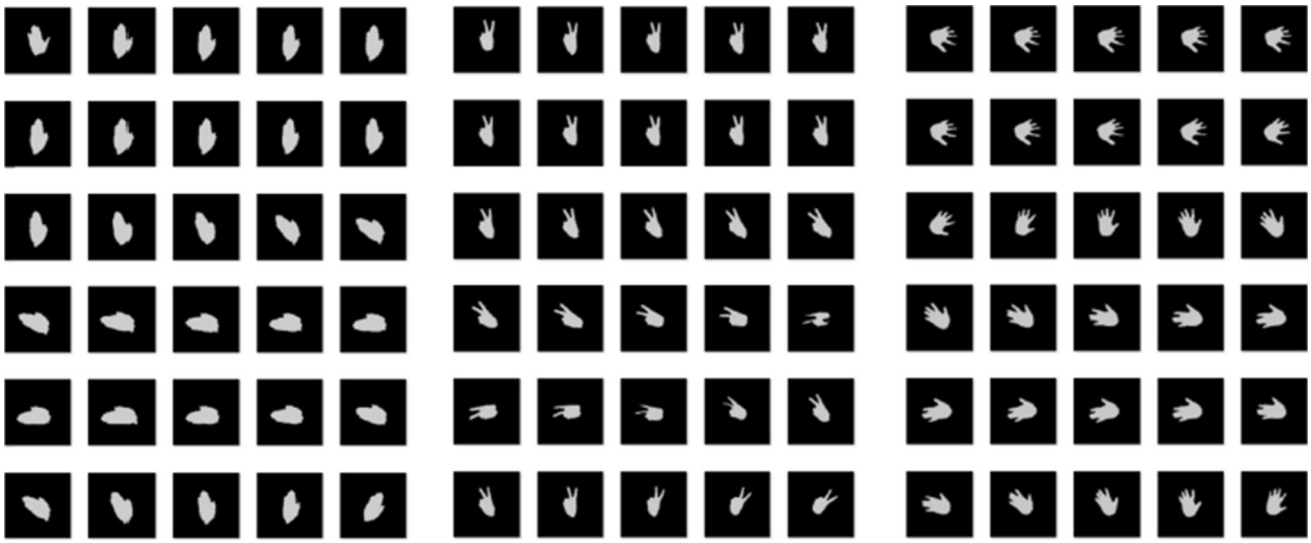

Fig. 9 Gesture sequence for custom gesture database

7.1 Analysis of depth map sequence decomposition results

Depth map sequence gesture decomposition algorithm based on spectral clustering algorithm has two important parameters that are crucial to the decomposition result of the depth map sequence, namely the parameter τ_t that limits the time-domain-weighted window size and the parameter τ_{stop} that stops the iterative clustering. The parameter τ_t controls the decay speed of the similarity matrix over the time interval. The larger the τ_t, the slower the decay of the similarity matrix, and vice versa [50]. The parameter τ_{stop} represents the threshold of the iteration stop. The closer the value is to 1, the stricter the conditions for stopping the iteration and the smaller the difference between categories; otherwise, the looser the iteration is stopped and the greater the difference between categories [51]. When the value of τ_t is large, we need to increase the depth of the iteration to increase the degree of similarity between the same classes in order to achieve a better decomposition result; in this case, it is necessary to take a large value of τ_{stop} that is to require τ_{stop} to be close to 1. When τ_t is set more strictly, that is, τ_t is smaller and smaller and the similarity between sequences has been limited in the time domain. At this time, the value of τ_{stop} does not have to be close to 1 to obtain better decomposition results [52].

Figures 10 and 11, respectively, show the changes in the number of key points extracted by the depth map sequence gesture decomposition algorithm when using the spectrum

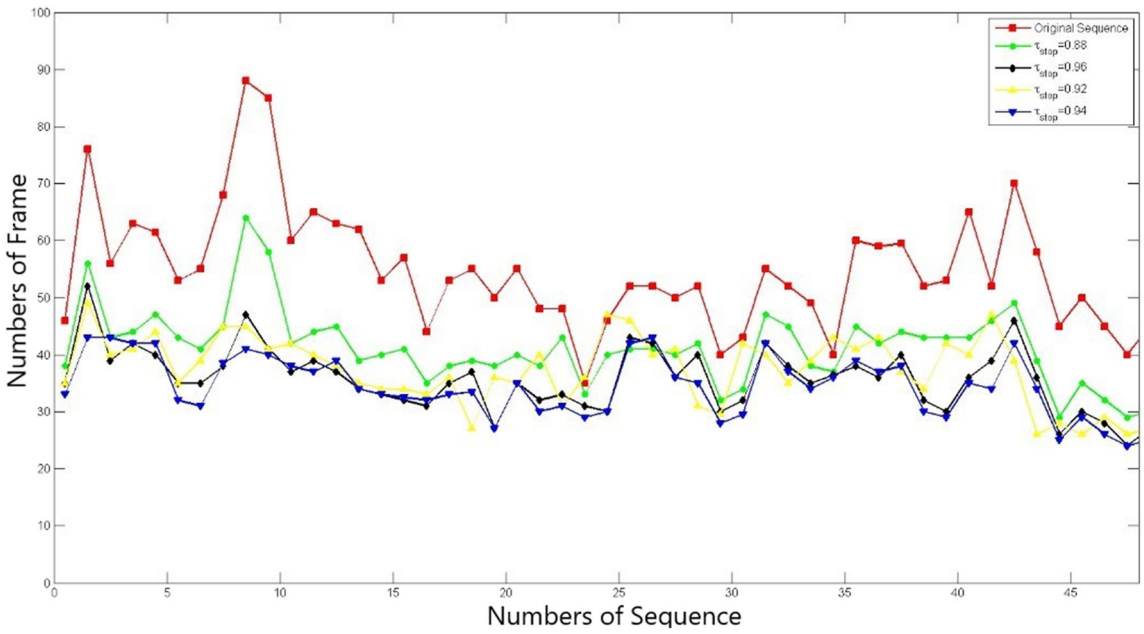

Fig. 10 Number of images in the key point set under different parameters of the MSRGesture3D gesture database at $\tau_t = 1.5$

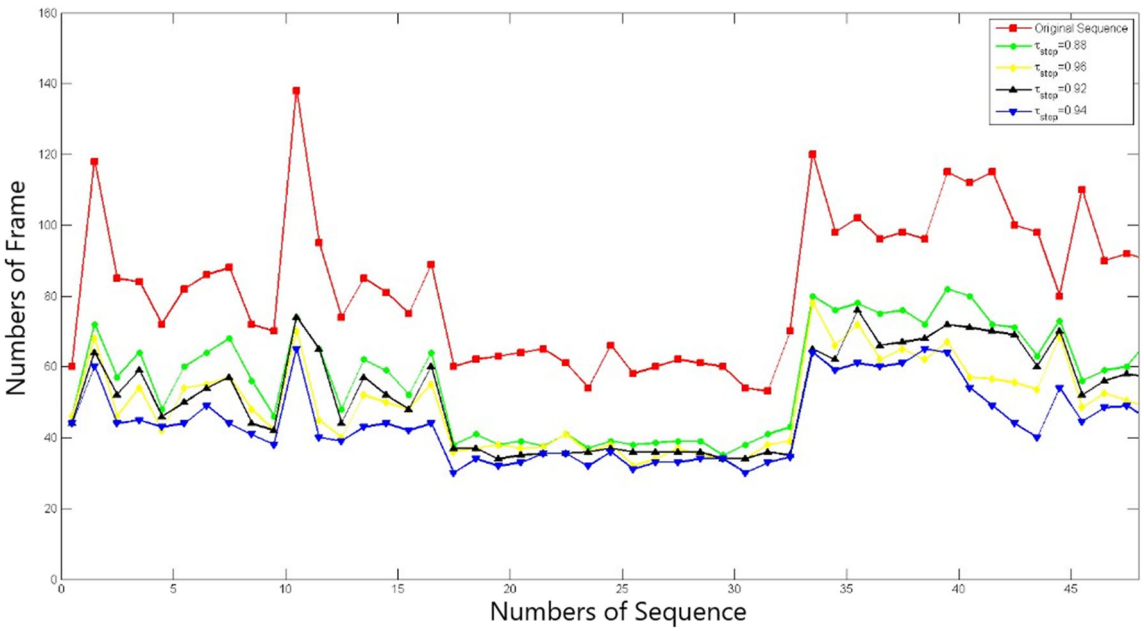

Fig. 11 Number of images in the key point set of the custom gesture database under different parameters at $\tau_t = 1.5$

clustering algorithm for the MSRGesture3D gesture database and the custom gesture database; the red curve represents the original image. The number of frames in each sequence which represents the number of images in the key points of different sequences extracted under the combination of parameters. This set of curves has shown that the gesture decomposition algorithm proposed in this paper can achieve the removal of time-domain redundant information.

From Figs. 10 and 11, we can see that the number of depth maps in key points has been significantly reduced. This reduction brings two beneficial effects: First, the key point set is the result of compressing the original sequence in the time domain. It removes many redundant frames and improves the differentiation of gesture sequences. Second, gesture recognition based on key points can overcome the shortcomings of the same gesture caused by different recognition speeds, because the decomposition of the gesture sequence can overcome the influence of the change

of the gesture speed which also makes the gesture recognition based on the key point set more robust. Figures 12 and 13 show the comparison of the original data and the key point set data of the two depth map sequences in the two databases, respectively. From this, we can see that the key point set removes the redundant data between adjacent images which makes time-domain information compressed. More importantly, this kind of compression is not uniform. When the depth map changes more slowly, that is, when gesture movement is relatively slow, the extracted key points are sparse. When the depth map changes significantly, that is, the gesture movement is relatively fast, and the extracted key points are denser which improves the robustness of the recognition algorithm to the same gesture with different speed.

7.2 Feature similarity distance gesture decomposition algorithm experiment

Depth map sequence gesture decomposition algorithm based on spectral clustering algorithm can fully utilize the similarity relationship between depth map frames to achieve the decomposition of the sequence, but the clustering algorithm often requires high computational costs. In order to meet the application requirements of low-end devices, this paper also implements a fast and effective depth map sequence decomposition scheme: depth map sequence decomposition algorithm based on feature similarity distance. The algorithm realizes the time-domain compression of depth map sequences with an efficient strategy and removes the redundant information in the time domain, and then improves the robustness to the time-domain change of the same gesture which improves the accuracy of gesture recognition.

For feature extraction of each frame of the depth map sequence, a depth map sequence $\{I_1, I_2, \ldots, I_N\}$ can be represented by a feature sequence $\{F_1, F_2, \ldots, F_N\}$. Differences between frames can be represented by Euclidean distances, and the Euclidean distances of features between any frames are calculated, as shown in Eq. 9:

$$\mathrm{dis}_{i,j} = \sqrt{\sum_{k=1}^{n} \left[F_i(d_k) - F_j(d_k)\right]^2} \tag{9}$$

From the Euclidean distance between any frames, we can further obtain the correlation coefficient, as represented by Eq. 10:

$$r_{i,j} = \exp\left(-\frac{\mathrm{dis}_{i,j}}{\sigma^2}\right)^2 \tag{10}$$

The value of σ is 0.05 times the maximum distance, that is, $\sigma = 0.05 \times \max\{d_{i,j}\}$. After the correlation coefficient is obtained, the segmentation metrics between consecutive two frames of images are calculated by using the theory of normalized cut-set criteria to obtain the segmentation metrics [47]. The objective function is shown in Eq. 11 [48]:

$$\mathrm{Ncut}(A,B) = \frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(A,V)} + \frac{\mathrm{cut}(A,B)}{\mathrm{assoc}(B,V)} \tag{11}$$

where $\mathrm{cut}(A,B) = \sum_{i\in A}\sum_{i\in B} r_{i,j}$, $\mathrm{assoc}(A,V) = \sum_{i\in A}\sum_{i\in V} r_{i,j}$, $\sum_{i\in B}\sum_{i\in V} r_{i,j}$, A and B and, respectively, represent two sequence fragments and $V = A + B$. Further, the split measure between any two successive frames is obtained by Eq. 12:

$$\mathrm{Sp}_i = \ln(\mathrm{Ncut}_i) \tag{12}$$

According to the segmentation metric, we search for the segmentation points of the segmented depth map sequence.

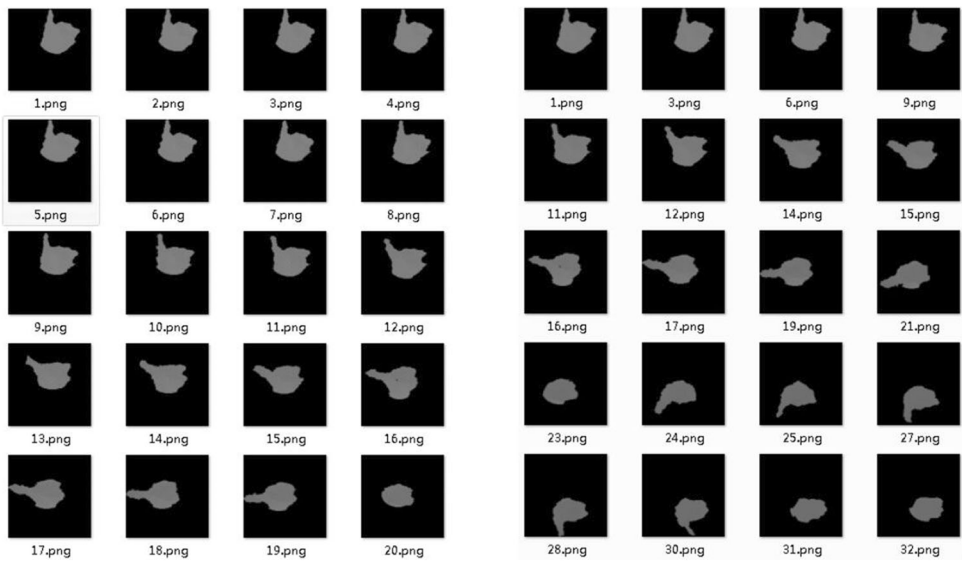

Fig. 12 MSRGesture3D gesture database original sequence and its key point set

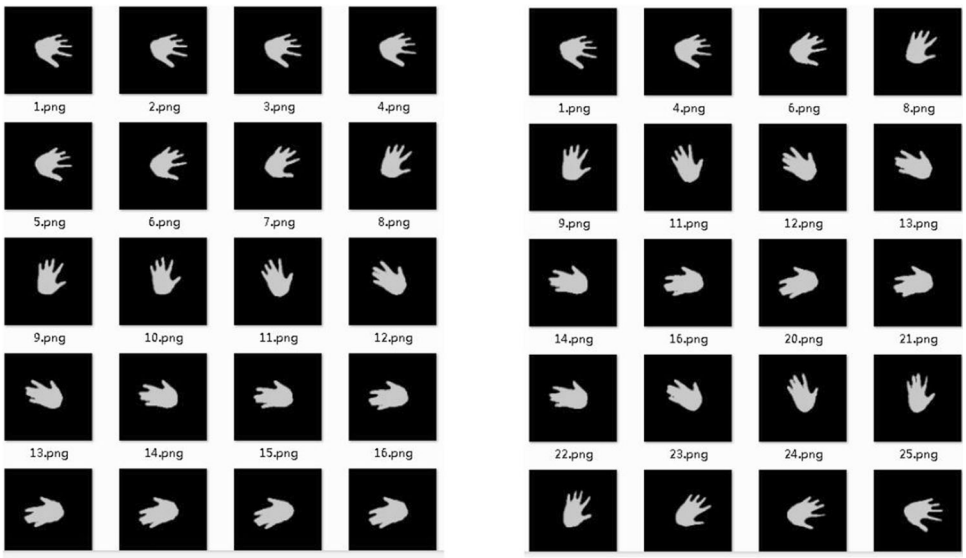

Fig. 13 Comparison of the original sequence of custom gesture database and its key point collection

When $Sp_i < \tau_s$, a segmentation point is set between the i frame and the $i + 1$ frame. Therefore, each depth map sequence will be divided into m segments, and the depth map in the segment will change smoothly while the depth maps between the segments will have large differences. Gesture Decomposition Based on Sequence of Characteristic Similar Distance Depth Map which the original sequence can also be compressed in the time domain to de-redundancy, and a more robust set of key points can be extracted. The algorithm can effectively decompose the depth map sequence and extract the key points of gesture, but the analysis of the similarity between adjacent frames is not enough which resulting in the extraction of the key point set is lower than the depth map sequence decomposition algorithm which based on spectrum clustering algorithm. In spite of this, the set of key points extracted by the depth map sequence gesture decomposition algorithm based on feature similarity distance can still improve the accuracy and robustness of the gesture recognition. Figures 14 and 15, respectively, represent the change curve of the number of frames of the key point set extracted from the two databases by the depth map sequence based on feature similarity distance.

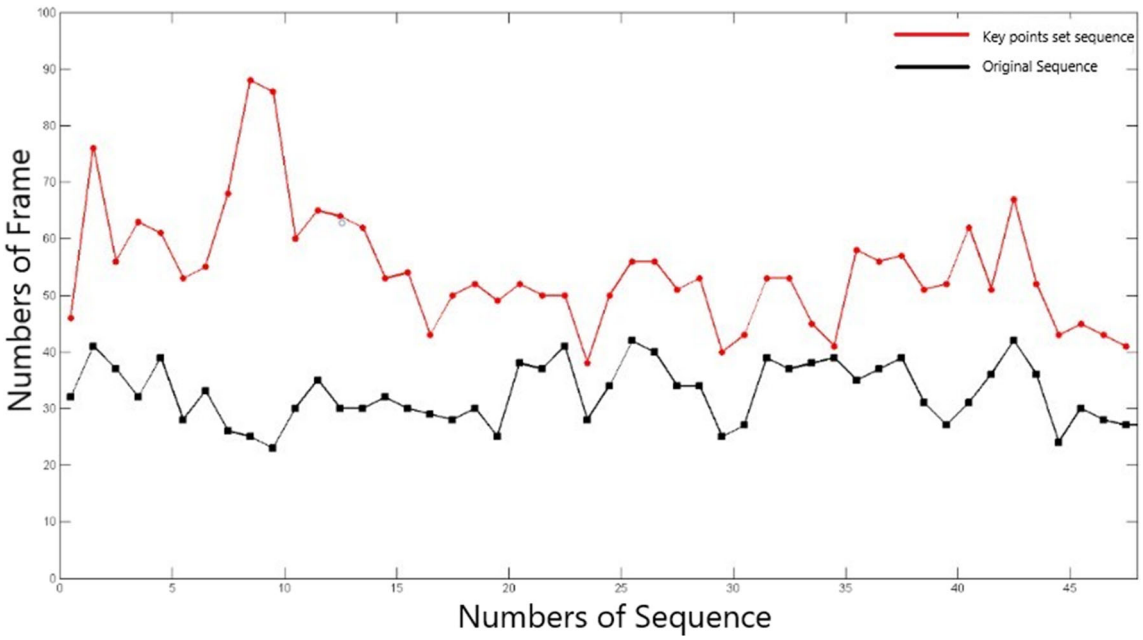

Fig. 14 Number of images in key point set of MSRGesture3D gesture database

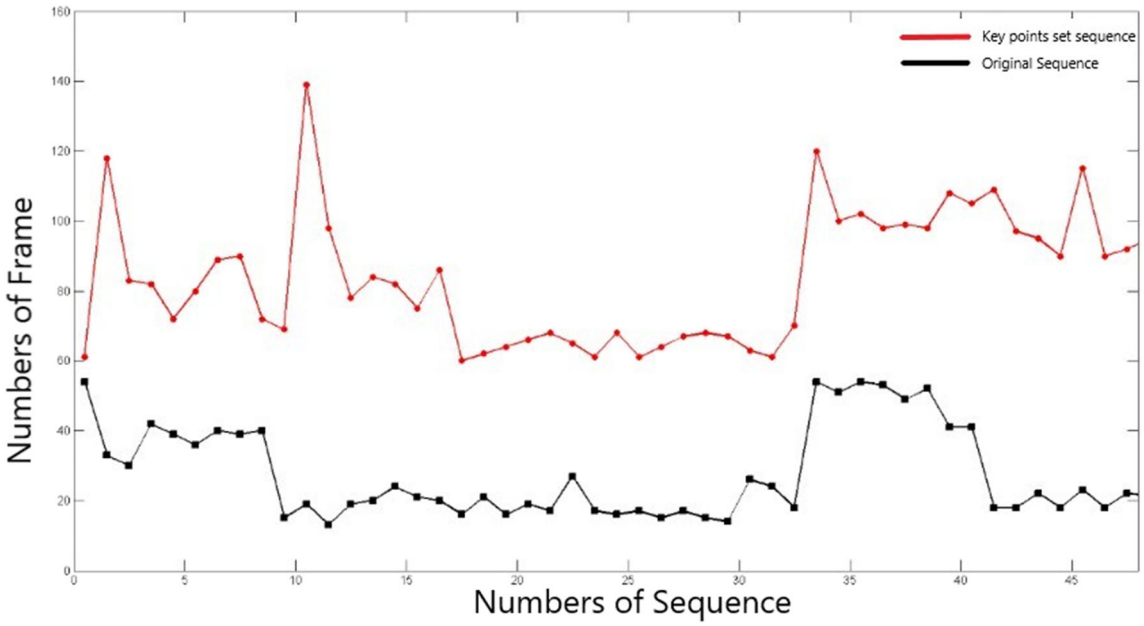

Fig. 15 Number of images in the custom key database key point set

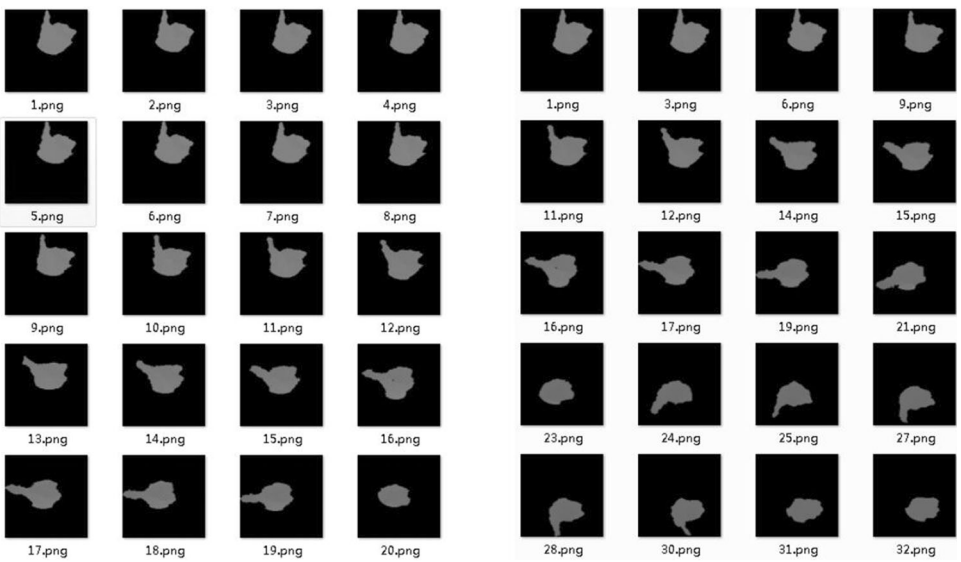

Fig. 16 MSRGesture3D gesture database original sequence and its key point set

Figures 16 and 17 show the comparison of the two sets of original data and key set data of two kinds of depth maps based on feature similarity distance-based depth map sequence gesture decomposition algorithm, respectively. From this, we can see that the key point set removes the redundant data between adjacent images and has similar rules to the extraction of key points and the depth map sequence gesture decomposition algorithm based on spectrum clustering algorithm. This rule can also overcome the effects of gesture speed changes on recognition algorithms to some extent.

8 Conclusion

This paper introduces a depth-sequence-based gesture decomposition algorithm. Direction-based depth maps have low-dimensional shape features and low extraction complexity and can be well used to describe the edge direction information of the hand area. This paper further introduces the depth map sequence gesture decomposition algorithm based on spectral clustering algorithm. The algorithm realizes that the depth map sequence is divided into different sub-fragments according to the similarity

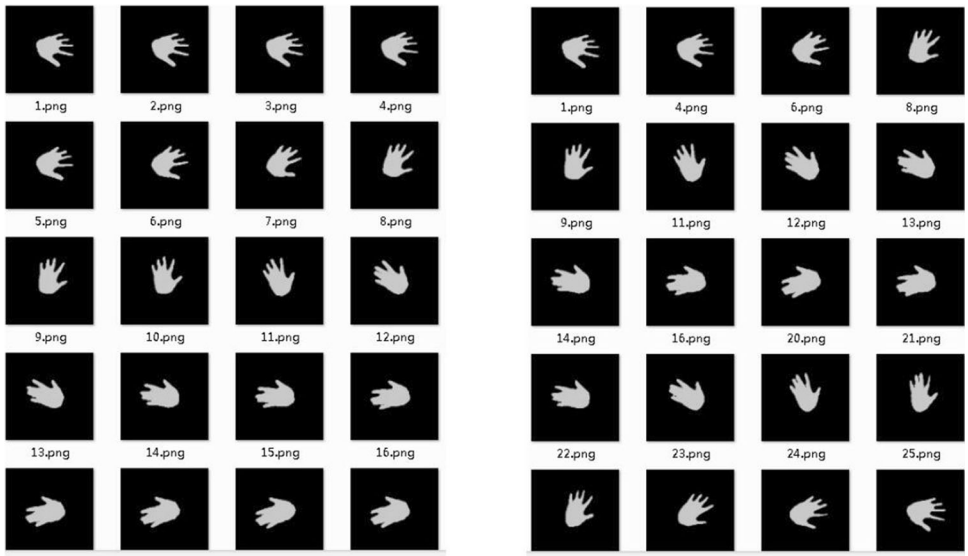

Fig. 17 Comparison of the original sequence of custom gesture database and its key point collection

between frames. This decomposition is to divide gestures into different subsections process and extract the corresponding key nodes from each sub-process. The gesture key node extraction algorithm based on sequence decomposition extracts the key point set from the decomposed subsequences. The new sequence composed of the key point set removes the time-domain redundant information and improves the robustness of change of the gesture speed. In addition, depth map sequence decomposition algorithm based on feature similarity distance is proposed to adapt to low computation performance application scenarios. Experimental results show that the algorithm proposed in this paper effectively implements the decomposition of the depth map sequence and the extraction of the key point set. In future research, we can consider the complementarity of gesture recognition based on depth information and gesture recognition based on color information, which makes the gesture recognition in different environments less affected by the environment and the recognition result more accurate.

Acknowledgements This work was supported by Grants of National Natural Science Foundation of China (Grant Nos. 51575407, 51505349, 51575338, 51575412, 61733011) and the Grants of National Defense Pre-Research Foundation of Wuhan University of Science and Technology (GF201705). This paper is funded by Wuhan University of Science and Technology graduate students' short-term study abroad special funds.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Li C, Li G, Jiang G, Chen D, Liu H (2018) Surface EMG data aggregation processing for intelligent prosthetic action recognition. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3909-z
2. Nešetřil J, Ossona de Mendez P (2016) A distributed low tree-depth decomposition algorithm for bounded expansion classes. Distrib Comput 29(1):39–49
3. Ju Z, Ji X, Li J, Liu H (2015) An integrative framework of human hand gesture segmentation for human-robot interaction. IEEE Syst J 99:1–11
4. Feng B, He F, Wang X et al (2017) Depth-projection-map-based bag of contour fragments for robust hand gesture recognition. IEEE Trans Hum Mach Syst 47(4):511–523
5. Ma X, Peng J (2018) Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information. J Sens 2018:1–9
6. Jadooki S, Mohamad D, Saba T et al (2017) Fused features mining for depth-based hand gesture recognition to classify blind human communication. Neural Comput Appl 28(11):3285–3294
7. Devanne M, Berretti S, Pala P et al (2017) Motion segment decomposition of RGB-D sequences for human behavior understanding. Pattern Recogn 61:222–233
8. Licciardi A, Agostinetti NP (2016) A semi-automated method for the detection of seismic anisotropy at depth via receiver function analysis. Geophys J Int 205(3):ggw091
9. Nguyen BP, Tay WL, Chui CK (2017) Robust biometric recognition from palm depth images for gloved hands. IEEE Trans Hum Mach Syst 45(6):799–804
10. Jung C, Joo S, Su MJ (2015) Depth map upsampling with image decomposition. Electron Lett 51(22):1782–1784
11. Gonzalez-Sanchez T, Puig D (2011) Real-time body gesture recognition using depth camera. Electron Lett 47(12):697–698
12. Wang C, Dubnov S (2015) Variable Markov oracle: a Markov-constrained online clustering algorithm with applications to human gesture recognition and following. IEEE Multimed 22(4):1
13. Despinoy F, Bouget D, Forestier G et al (2016) Unsupervised trajectory segmentation for surgical gesture recognition in robotic training. IEEE Trans Biomed Eng 63(6):1280–1291

14. Li G, Peixin Q, Kong J, Jiang G, Xie L, Gao P, Zehao W, He Y (2013) Coke oven intelligent integrated control system. Appl Math Inf Sci 7(3):1043–1050
15. Arunraj M, Srinivasan A, Juliet AV (2018) Online action recognition from RGB-D cameras based on reduced basis decomposition. J Real Time Image Process. https://doi.org/10.1007/s11554-018-0778-8
16. Kim K, Kim J, Choi J et al (2015) Depth camera-based 3D hand gesture controls with immersive tactile feedback for natural mid-air gesture interactions. Sensors 15(1):1022
17. Pang S, Chuang L et al (2011) A workflow decomposition algorithm based on invariants. Chin J Electron 20(1):1–5
18. Li G, Jiang D, Zhou Y, Jiang G, Kong J, Gunasekaran M (2019) Human lesion detection method based on image information and brain signal. IEEE 7:11533–11542
19. Liu K, Chen C, Jafari R et al (2014) Fusion of inertial and depth sensor data for robust hand gesture recognition. IEEE Sens J 14(6):1898–1903
20. Ju Z, Ji X, Li J et al (2017) An integrative framework of human hand gesture segmentation for human–robot interaction. IEEE Syst J 11(3):1326–1336
21. He Y, Li G, Liao Y, Sun Y, Kong J, Jiang G, Jiang D, Liu H (2017) Gesture recognition based on an improved local sparse representation classification algorithm. Clust Comput. https://doi.org/10.1007/s10586-017-1237-1
22. Li B, Sun Y, Li G, Kong J, Jiang G, Jiang D, Liu H (2017) Gesture recognition based on modified adaptive orthogonal matching pursuit algorithm. Clust Comput. https://doi.org/10.1007/s10586-017-1231-7
23. Chen D, Li G, Sun Y, Kong J, Jiang G, Tang H, Ju Z, Hui Y, Liu H (2017) An interactive image segmentation method in hand gesture recognition. Sensors 17(2):253. https://doi.org/10.3390/s17020253
24. Chen D, Li G, Sun Y, Jiang G, Kong J, Li J, Liu H (2017) Fusion hand gesture segmentation and extraction based on CMOS sensor and 3D sensor. Int J Wirel Mob Comput 12(3):305–312
25. Miao W, Li G, Sun Y, Jiang G, Kong J, Liu H (2016) Gesture recognition based on sparse representation. Int J Wirel Mob Comput 11(4):348–356
26. Li G, Tang H, Sun Y, Kong J, Jiang G, Jiang D, Tao B, Xu S, Liu H (2017) Hand gesture recognition based on convolution neural network. Clust Comput. https://doi.org/10.1007/s10586-017-1435-x
27. Sun Y, Li C, Li G, Jiang G, Jiang D, Liu H, Zheng Z, Shu W (2018) Gesture recognition based on kinect and sEMG signal fusion. Mob Netw Appl 23(4):797–805
28. Sun Y, Hu J, Li G, Jiang G, Xiong H, Tao B, Zheng Z, Jiang D (2018) Gear reducer optimal design based on computer multimedia simulation. J Supercomput 00:00. https://doi.org/10.1007/s11227-018-2255-3
29. Li G, Zhang L, Sun Y, Kong J (2018) Internet of things sensors and haptic feedback for sEMG based hands. Multimed Tools Appl. https://doi.org/10.1007/s11042-018-6293-x
30. Liao Y, Sun Y, Li G, Kong J, Jiang G, Jiang D, Cai H, Ju Z, Hui Y, Liu H (2017) Simultaneous calibration: a joint optimization approach for multiple kinect and external cameras. Sensors 17(7):1491
31. Fang Y, Liu H, Li G, Zhu X (2015) A multichannel surface EMG system for hand motion recognition. Int J Humanoid Rob 12(2):1550011. https://doi.org/10.1142/S0219843615500115
32. Yin Q, Li G, Zhang J (2015) Research on the method of step feature extraction for EOD robot based on 2d laser radar. Discrete Contin Dyn Syst Ser S 8(6):1415–1421
33. Li Z, Li G, Sun Y, Jiang G, Kong J, Liu H (2017) Development of articulated robot trajectory planning. Int J Comput Sci Math 8(1):52–60
34. Ding W, Li G, Sun Y, Jiang G, Kong J, Liu H (2017) D-S evidential theory on sEMG signal recognition. Int J Comput Sci Math 8(2):138–145
35. Chen D, Li G, Jiang G, Fang Y, Ju Z, Liu H (2015) Intelligent computational control of multi-fingered dexterous robotic hand. J Comput Theor Nanosci 12(12):6126–6132
36. Ding W, Li G, Jiang G, Fang Y, Ju Z, Liu H (2015) Intelligent computation in grasping control of dexterous robot hand. J Comput Theor Nanosci 12(12):6096–6099
37. Li Z, Li G, Jiang G, Fang Y, Ju Z, Liu H (2015) Intelligent computation of grasping and manipulation for multi-fingered robotic hands. J Comput Theor Nanosci 12(12):6192–6197
38. Li G, Liu J, Jiang G, Liu H (2015) Numerical simulation of temperature field and thermal stress field in the new type of ladle with the nanometer adiabatic material. Adv Mech Eng 7(4):1–13. https://doi.org/10.1177/1687814015575988
39. Du J, Zujia Z, Gongfa L, Ying S, Jianyi K, Guozhang J, Hegen X, Bo T, Shuang X, Honghai L, Ju Z (2018) Gesture recognition based on binocular vision. Clust Comput. https://doi.org/10.1007/s10586-018-1844-5
40. Li G, Liu Z, Jiang G, Xiong H, Liu H (2015) Numerical simulation of the influence factors for rotary kiln in temperature field and stress field and the structure optimization. Adv Mech Eng 7(6):1–15. https://doi.org/10.1177/1687814015589667
41. He Y, Li G, Zhao Y, Sun Y, Jiang G (2018) Numerical simulation-based optimization of contact stress distribution and lubrication conditions in the straight worm drive. Strength Mater 50(1):157–165
42. Li G, Miao W, Jiang G, Fang Y, Ju Z, Liu H (2015) Intelligent control model and its simulation of flue temperature in coke oven. Discrete Contin Dyn Syst Ser S (DCDS-S) 8(6):1223–1237
43. Li G, Gu Y, Kong J, Jiang G, Xie L, Wu Z, Li Z, He Y, Gao P (2013) Intelligent control of air compressor production process. Appl Math Inf Sci 7(3):1051–1058
44. Du F, Sun Y, Li G, Li Z, Kong J, Jiang G, Du J (2017) Adaptive fuzzy sliding mode control for 2-DOF articulated robot. J Wuhan Univ Sci Technol 40(6):446–450
45. Li G, Wu H, Jiang G, Xu S, Liu H (2019) Dynamic gesture recognition in the internet of things. IEEE 7(1):23713–23724
46. Miao W, Li G, Jiang G, Fang Y, Ju Z, Liu H (2015) Optimal grasp planning of multi-fingered robotic hands: a review. Appl Comput Math 14(3):238–247
47. Bu XY, Dong HL, Han F, Li GF (2018) Event-triggered distributed filtering over sensor networks with deception attacks and partial measurements. Int J Gen Syst 47(5):395–407
48. Li G, Qu P, Kong J, Jiang G, Xie L, Wu Z, Gao P, He Y (2013) Influence of working lining parameters on temperature and stress field of ladle. Appl Math Inf Sci 7(2):439–448
49. Chang W, Li G, Kong J, Sun Y, Jiang G, Liu H (2018) Thermal mechanical stress analysis of ladle lining with integral brick joint. Arch Metall Mater 63(2):659–666
50. Li G, Kong J, Jiang G, Xie L, Jiang Z, Zhao G (2012) Air-fuel ratio intelligent control in coke oven combustion process. Inf Int Interdiscip J 15(11):4487–4494
51. Tan C, Sun Y, Li G, Jiang G, Chen D, Liu H (2019) Research on gesture recognition of smart data fusion features in the IoT. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04023-0
52. Li J, Li X, Tao D (2008) KPCA for semantic object extraction in images. Pattern Recogn 41(10):3244–3250