Gesture recognition based on multi-modal feature weight

Haojie Duan[1], Ying Sun[1,2*], Wentao Cheng[1,3], Du Jiang[1,3*], Juntong Yun[2,3], Ying Liu[4], Yibo Liu[4], Dalin Zhou[5]

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

[2] *Institute of Precision Manufacturing, Wuhan University of Science and Technology, Wuhan, 430081, China*

[3] *Research Center for Biomimetic Robot and Intelligent Measurement and Control, Wuhan University of Science and Technology, Wuhan, 430081, China*

[4] *Hubei Key Laboratory of Mechanical Transmission and Manufacturing Engineering, Wuhan University of Science and Technology, Wuhan, 430081, China*

[5] *School of Computing, University of Portsmouth, Portsmouth PO1 3HE, UK*

Corresponding author

E-mail address: sunying65@wust.edu.cn, jiangdu@wust.edu.cn

Abstract: With the continuous development of sensor technology, the acquisition cost of RGB-D images is getting lower and lower, and gesture recognition based on depth images and RGB images has gradually become a research direction in the field of pattern recognition. However, most of the current processing methods for RGB-D gesture images are relatively simple, ignoring the relationship and influence between its two modes, and unable to make full use of the correlation factors between different modes. In view of the above problems, this paper optimizes the effect of RGB-D information processing by considering the independent features and related features of multi-modal data to construct a weight adaptive algorithm to fuse different features. Simulation experiments show that the method proposed in this paper is better than the traditional RGB-D gesture image processing method and the gesture recognition rate is higher. Comparing the current more advanced gesture recognition methods, the method proposed in this paper also achieves higher recognition accuracy, which verifies the feasibility and robustness of this method.

Keywords: Gesture recognition; RGB-D; Multi-modal fusion; Weight adaptation

1. Introduction

In recent years, the field of computer vision and pattern recognition has been vigorously developed by the influence of human-computer interaction mode, and gesture recognition has gradually become the main human-computer interaction method of various types of devices, such as electronic equipment and industrial equipment [1, 2]. On the basis of traditional two-dimensional sensors, many scholars have developed multiple cameras to synchronize data collection in different directions to collect three-dimensional information to achieve three-dimensional human interaction. With the development of science and technology, new types of sensors have also made new breakthroughs. There are more and more types of sensors with 3D sensing, and the cost of obtaining data is cheaper and the methods are more diverse [3]. Some classic and better performing 3D sensors [4, 5], such as Kinect, Xtion and Leap Motion, have greatly reduced the complexity of 3D human-computer interaction. These three-dimensional sensors acquire both image information and depth information [6, 7]. Their appearance greatly improves the accuracy of gesture recognition, and also promotes the development of human-computer interaction [8, 9].

With the availability of depth information, multimodal data fusion has become an important research content in gesture recognition. There are three main levels of common multimodal data fusion: pixel-level fusion, feature-level fusion, and decision-level fusion. Pixel-level fusion is the direct fusion of deep color information. The advantages are small data loss and high accuracy, but the disadvantages are large data volume and poor real-time performance. Decision-level fusion combines the processing results of color images and depth images. It is fast but has poor accuracy and does not consider the correlation of multimodal data. [10, 11]. At present, most of the multi-modal data fusion in gesture recognition is based on feature-level fusion, but the fusion method does not fully consider the correlation and independence of multi-sensor data, only the weighted fusion of data [12, 13, 14]. In this paper, the multi-modal feature-level fusion method is studied by considering the independence and correlation of multi-modal data. We propose a multi-modal feature weight adaptive fusion method, which solves the problems of redundant information and missing information appearing in information fusion, and improves the superiority of fusion features and the suitability of multi-modal feature weights. Based on the idea of this method, we designed a dual-stream convolutional neural network, and carried out experiments on the self-built gesture library and ASL gesture library, and verified the effect of the method on improving the accuracy, real-time and robustness of gesture recognition [15, 16, 17].

The other parts of this paper are arranged as follows: Sect.2 introduces the latest related works in this field, and proposes the idea of multi-modal data fusion; Sect.3 details the multimodal feature fusion framework and implementation method; In Sect.4, based on the method of the Sect.3, a network framework of gesture recognition is designed. Compared with the related literature, the advantages of the proposed method are verified. In Sect.5, the work of this paper is summarized.

2. Realted work

In recent years, deep learning has also been used for RGB-D visual analysis. For example, Gupta et al.[18] proposed a method of encoding depth data into three channels: horizontal difference, ground height, point normal, and angle between inferred gravity. Then, they train the CNN on the three channels instead of RGB-D target recognition and segmentation of the sequence depth image. Huo et al. [19] extracted a feature extraction method for multimodal data for modal convolution. It uses modal convolution to describe the CNN and uses it to extract inter-modal and modal information and fuse features at the pixel level. Couprie et al.[20] proposed a multi-scale CNN for RGB-D scene markers based on hierarchical feature method. Liao et al. designed a static gesture recognition system combining depth image and color image[21], using depth image and color image acquired by Realsense, combined with generalized Hough transform to map depth image to color image. Wang et al.[22] designed a deep neural network for surface normal prediction. However, these methods ignore the relationship between data from different modalities because RGB and depth information are simply connected together. Eitel et al.[23] proposed a dual-flow CNN model combining fusion layers for RGB-D target recognition. In the RGB-D sign language recognition application, Ravi et al.[24] designed a four-stream convolutional neural network with RGB space and depth time in the mainstream RGB space and ROI stream as inputs. The multi-mode feature sharing mechanism solves the problem of the color information identifying the gesture action of the video data. Park et al.[25] proposed a multi-mode feature depth learning method for RGB-D target recognition using shared features of RGB and depth images. Elboushaki et al.[26] proposed an effective multi-dimensional feature learning method (MultiD-CNN) to solve complex background, occlusion, lighting conditions and other issues that cannot be handled by a single deep network. Hsien-I Lin et al.[27] established a skin color model based on the Gaussian mixture model, and then used the calibration output of the gesture image and the skin color model as the input image of the CNN to realize the effective recognition of the seven custom gestures. Negin et al.[28] used the CNN model to make a simple exploration of static gesture recognition. Inspired by Wang's work, Gao et al.[29] proposed a recognition multimodal feature learning method for RGB-D scene recognition.

It can be seen from these related literatures that the gesture recognition method based on multi-modal fusion does not fully consider the relationship between different modes. For example, in the research of gesture recognition using RGB-D images, most methods simply map depth images to color image learning features, or simply treat RGB-D images as undifferentiated four-channel data. [30, 31, 32]. In the actual process, the two modalities complement and repeat the information contained in each other. The complementarity of the two kinds of information can improve the robustness of the recognition, and the repeated information can be mutually verified to improve the quality of feature extraction [33, 34, 35]. Based on the multi-modal gesture recognition research, this paper considers the independence and correlation of RGB-D features, optimizes the multi-modal feature fusion method, and uses it for RGB-D image gesture recognition, which improves the effect of gesture recognition.

3. Multi-modal feature fusion framework and implementation method

3.1. Multimodal feature fusion framework design

The difference between the two different modes is fully considered before the multi-modal feature fusion. Therefore, double convolutional neural network is used to extract the features of RGB and depth mode respectively [36], and the feature extraction network extracts features at each level to generate multi-level, multi-modal abstract features. Then, a multi-modal feature weight adaptive adjustment learning structure is proposed for further explore the relationship between these two modes. The input of this structure is the feature of convolutional layer extraction of a dual-stream convolutional neural network at different levels of abstraction. The multi-modal fusion framework is shown in Figure 1.

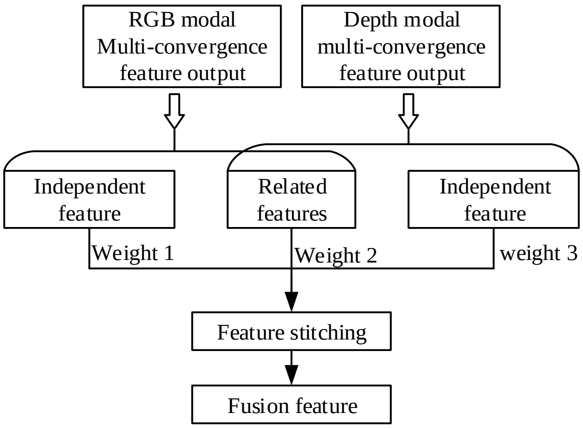

Figure 1. Multimodal feature fusion framework

As shown in Figure 1, when designing a multimodal fusion framework, the idea in literature [37] considers the consistency or sharing features between different modal features based on the features extracted from each layer of the dual-stream convolutional neural network. The characteristics of each mode should include the independent features of its own modes and the features of the two modally related parts, which are characterized by multimodal correlation and independence. The degree of influence of different modalities on the results of gesture recognition is also inconsistent, and the degree of importance between independent features and related features cannot be determined. Therefore, by determining the weights of different features, a more distinguishing and robust fusion feature is generated [38]. In order to achieve this goal, the features acquired under different modalities are forced to share a relevant partial feature. In addition, when determining the weights of different modal features, it is not necessary to know which modal data is more important, and only the loop iterative learning of the framework completes the weight adaptive control [39, 40].

The goal of the multimodal feature fusion framework is to learn a new feature representation that contains two sets of attributes: 1) related features shared by the two modalities; 2) independent features unique to the individual modalities.

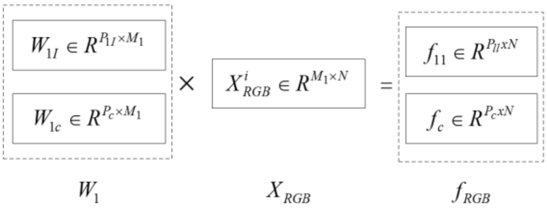
$W_{1I} \in R^{P_{1I} \times M_1}$
$W_{1c} \in R^{P_c \times M_1}$
×
$X_{RGB}^{i} \in R^{M_1 \times N}$
=
$f_{11} \in R^{P_{II} xN}$
$f_c \in R^{P_c xN}$
W_1
X_{RGB}
f_{RGB}

Figure 2. Schematic diagram of mapping operation

As shown in Figure 2, $X_1 = [x_1, x_2, ..., x_N] \in R^{M_1 \times N}$ is the M-dimensional output of a pooled layer of the RGB mode in the network layer of a batch of N images, and $W_1 = [W_{1I}; W_{1c}]$ represents the sum of the transformation matrices corresponding to the independent and related features in the RGB mode [41]. Through the change of the matrix, independent and related features in the current mode can be generated to form a distinction. Where $W_{1I} \in R^{P_{1I} \times M_1}$ denotes a transformation matrix that generates RGB modal correlation features. $W_{1c} \in R^{P_c \times M_1}$ denotes a transformation matrix that generates RGB modal independent features. $M_1' = P_{1I} + P_c$, $f_{1I} \in R^{P_{1I} \times N}$ and $f_c \in R^{P_c \times N}$ represent the independent features and related features generated in the RGB mode. At the same time, $f_{RGB} = [f_{1I}; f_c]$ and depth modes are the same as above. That is, the symbol definitions with the subscript 1 are all changed to 2 different modes [42].

Definitions $f_{RGB} \in R^{M_1' \times N}$ and $f_{DEPTH} \in R^{M_2' \times N}$ represent the characteristic representations of the two-stream convolutional neural network learned from the RGB and Depth modes, respectively. The characteristics of the two modes are mapped into two parts by the transformation matrix $W_1 = [W_{1I}; W_{1c}]$. The first part is the two modal features share the correlation feature $f_c \in R^{P_c \times N}$; the second is the two modalities respectively contain their own unique features representing $f_{1I} \in R^{P_{1I} \times N}$ and $f_{2I} \in R^{P_{2I} \times N}$. Thus the respective extracted features of the two different modalities expressed as: $f_{RGB} = [f_{1I}; f_c]$ and $f_{DEPTH} = [f_{1I}; f_c]$. Therefore, it can be seen from the above that the task is to find the transformation matrices W_1 and W_2 to obtain f_{RGB} and f_{DEPTH}, and finally splicing the two features to obtain $f = [f_c; f_{1I}; f_{2I}]$.

The relevant parts of the data features of different modalities play an important role in extracting sharable information, but it does not mean that the correlation characteristics of the two modalities can affect the results more than the characteristics of the modal itself [43, 44]. However, it is not easy to manually set the weight of each part. Therefore, different parts of the fusion feature need to be designed with adaptive weights. Finally, by designing different weights, corresponding labels and categories can be obtained. By designing the characteristics of each convolutional layer output of the dual stream, different levels of fusion features and corresponding labels can be obtained, and the sorted fusion features and labels can be input into the gesture classification model to obtain the final prediction category, and the final recognition is completed. The output is shown in Figure 3.

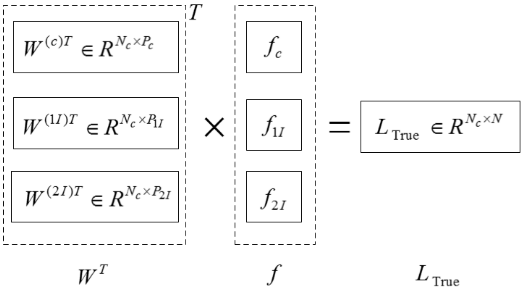
T
$W^{(c)T} \in R^{N_c \times P_c}$
$W^{(1I)T} \in R^{N_c \times P_{1I}}$
$W^{(2I)T} \in R^{N_c \times P_{2I}}$
×
f_c
f_{1I}
f_{2I}
=
$L_{\text{True}} \in R^{N_c \times N}$
W^T
f
L_{True}

Figure 3. Schematic diagram of the regression coefficient matrix operation

As shown in Figure 3, where $W^{(c)T}$, $W^{(1I)T}$ and $W^{(2I)T}$ respectively correspond to the relevant partial feature weight, the RGB mode independent feature weight and the Depth mode independent feature weight; $W^T = [W^{(c)T}; W^{(1I)T}; W^{(2I)T}]$ is a regression coefficient matrix; f is a fusion feature of the output characteristics of the same convolutional layer of the two modes; L_{True} is the real category label for the gesture sample.

3.2. Adaptive weight network objective function design and optimization

3.2.1. Objective function design

In the above framework, the original features are mapped into the associated feature space and the independent feature space by learning the transformation matrix $W_i (i = 1, 2)$ of the two model [45, 46]. To understand the characteristics of different modalities that contain shared attributes and modal-specific attributes, the objective function is designed to:

$$\min S = S_1 + S_2 + S_3 \tag{1}$$

Where, S represents the minimum cost function of the entire model; S_1 and S_2 represent the cost functions of the RGB mode and the Depth mode, respectively; S_3 is the cost function of the two modes after fusion.

By minimizing the cost function of the model to make the output of the model optimal, combined with the transformation matrix, the different modal cost functions in formula (1) can be expressed as:

$$S_1 = \mu_1(\|W_1X_1 - f_{RGB}\|_F^2 + \|W^T{}_1 f_{RGB} - X_1\|_F^2 + \alpha_1 g(f_{RGB})) \tag{2}$$

$$S_2 = \mu_2(\|W_2X_2 - f_{DEPTH}\|_F^2 + \|W^T{}_2 f_{DEPTH} - X_2\|_F^2 + \alpha_1 g(f_{DEPTH})) \tag{3}$$

$$S_3 = \lambda(\|W^T f - L_{True}\|_F^2 + \alpha_2 g(\|W\|_{2,1})) \tag{4}$$

where, $\|\bullet\|_F^2$ represents the Frobenius norm; $\|\bullet\|_{2,1}$ represents the $l_{2,1}$ norm. In formula (2), μ_1 is the weight coefficient of the RGB mode; the first Frobenius norm representation forces f_{RGB} and W_1X_1 to be more similar; the second norm represents the ability of the enhancement f_{RGB} to reconstruct the original feature X_1 in reverse by W_1^T; in the third term, g is a smooth L_1 penalty function. The formula (3) has the same meaning as the symbol in the formula (2). In formula (4), supervisory information is combined to enhance the ability to recognize learned features. λ is a very important parameter whose purpose is to balance the relationship between feature reconstruction constraints and supervisory constraints; the first norm term represents the similarity between enhanced reconstruction features and tags; α_2 is the penalty function weighting coefficient of the fusion feature; where W is the regression coefficient matrix, and the $l_{2,1}$ norm ensures that W is sparse, so it acts as an $f = [f_c; f_{1I}; f_{2I}]$ feature selector; Generally, L_{True} is selected according to the rule of thumb, which indicates the true label of the training sample. , then formula (1) can be expressed as:

$$\begin{aligned} \min_{W_1, W_2, \mu_1, \mu_2, f_{RGB}, f_{DEPTH}, W} S = {} & \mu_1(\|W_1X_1 - f_{RGB}\|_F^2 + \|W^T{}_1 f_{RGB} - X_1\|_F^2 + \alpha_1 g(f_{RGB})) \\ & + \mu_2(\|W_2X_2 - f_{DEPTH}\|_F^2 + \left\|\begin{matrix} W^T{}_2 f_{DEPTH} \\ -X_2 \end{matrix}\right\|_F^2 + \alpha_1 g(f_{DEPTH})) \\ & + \lambda(\|W^T f - L_{True}\|_F^2 + \alpha_2 g(\|W_{2,1}\|)) \quad \text{s.t. } \mu_1 + \mu_1 = 1,\ \mu_1 \geq 0,\ \mu_2 \geq 0 \end{aligned} \tag{5}$$

The definitions of S_1 and S_2 seem to indicate that the RGB and Depth modes are independently optimized, but f_{RGB} and f_{DEPTH} are not actually independent because they explicitly require sharing a common part f_c. By connecting f_c and modal specific components f_{1I} and f_{2I}, the final representation of each image is $f = [f_c; f_{1I}; f_{2I}]$.

After obtaining the matrix W, W_1 and W_2 of the training phase, the characteristics of any test image can be directly calculated as: $f_{1I} = W_{1I}X_1$, $f_{2I} = W_{2I}X_2$, $f_c = (W_{1c}X_1 + W_{2c}X_2) / 2$. Using the multimodal feature representation $f = [f_c; f_{1I}; f_{2I}]$, the final recognition result is directly calculated as $W^T f$. At this time, the features of different levels can be used to obtain the multi-modal feature representation and the corresponding recognition result by using the above transformation matrix [47]. The multi-modal features of different levels are arranged in time series to form a series of multi-modal fusion feature kernels and recognition results with different abstract levels.

3.2.2. Objective function optimization

In the optimization of the objective function, a typical alternating optimization strategy is used to obtain the local optimal solution of the formula (5). The steps of the algorithm are as follows:

```
Algorithm flow: multi-modal feature adaptive weight learning framework optimization
Input: two modal training samples $X_1$ and $X_2$, and the corresponding sample correct category label $L$
Output: feature transformation matrix $W_1$ and $W_2$; regression coefficient matrix $W$; fusion feature of an abstract level $f$
Step 1: Initialize the parameters. Including $W$, $W_1$, $W_2$, $f$, $\mu_1$, $\mu_2$, using random initialization.
Step 2: Optimize the parameters.
Loop
    2.1 fixed parameters $W$, $W_1$, $W_2$, $f$ Update $\mu_1$, $\mu_2$ according to formula (5)
    2.2 fixed parameters $W_1$, $W_2$, $f$, $\mu_1$, $\mu_2$ ,Update according to formula (4) $W$
    2.3 fixed parameters $W$, $W_1$, $W_2$, $\mu_1$, $\mu_2$ ,Update $f_{RGB}$, $f_{DEPTH}$ according to formula (1,2)
    2.4 Fixed parameters $W$, $f$, $\mu_1$, $\mu_2$ Update $W_1$, $W_2$ according to formula (3)
End loop   Loop iteration until convergence
```

First, the transformation matrices W_1, W_2 and the regression coefficient matrix W and the feature f are randomly initialized, and then the coefficients μ_1 and μ_2 are initialized to 0.5, respectively. All of these variables, including W_1, W_2, W, f, μ_1, μ_2 , will be learned and updated in the algorithm. Other parameters, such as α_1, α_2 and λ are set based on experience.

In step 2.1 of step 2 of the algorithm flow, first fix W_i, W and f to update parameter μ_i. μ_1 and μ_2 allow different modes to have different weights because they do not perform the same thing. When the parameters W_i, W, and f are fixed, the following Lagrangian functions can be constructed according to formula (5):

$$L(\alpha,\eta) = \mu_1 C_1 + \mu_2 C_2 + \lambda C - \eta(\mu_1 + \mu_2 - 1) \tag{6}$$

Where, C_1, C_2 and C are constant markers for the corresponding term when W_i, W and f are fixed. But this way is trivial for solving formula (6). For example, if C_1 is less than C_2, then the solution to minimize formula (6) would be: $\mu_1 = 1$ and $\mu_1 = 0$, which means that only one form is used in feature learning. Simulation experiments have found that such a situation may lead to locally optimal results. Therefore, in order to make full use of the information of different modalities, the cost function in formula (5) can be modified to:

$$\begin{aligned}\min_{W_1,W_2,\mu_1,\mu_2,f_{RGB},f_{DEPTH},W} S = S_1 + S_2 + S_3 = {} & \mu_1{}^q(\|W_1X_1 - f_{RGB}\|_F^2 + \|W^T{}_1 f_{RGB} - X_1\|_F^2 \\ & + \alpha_1 g(f_{RGB})) + \mu_2{}^q(\|W_2X_2 - f_{DEPTH}\|_F^2 + \|W^T{}_2 f_{DEPTH} - X_2\|_F^2 + \alpha_1 g(f_{DEPTH})) \\ & + \lambda(\|W^T f - L\|_F^2 + \alpha_2 g(\|W_{2,1}\|)) \ \text{ s.t. } \mu_1 + \mu_1 = 1,\ \mu_1 \ge 0,\ \mu_2 \ge 0\end{aligned} \tag{7}$$

As can be seen from formula (7), $q > 1$ here is an additional parameter. By adding this additional parameter q, the target of μ_i will become nonlinear, and both modes will be constrained to obtain the relevant features in feature f and the independent features of the particular mode. At the same time, most of the original information in feature f can be retained. Therefore, the Lagrangian function can be rewritten as:

$$L(\alpha,\eta) = \mu_1{}^q C_1 + \mu_2{}^q C_2 + \lambda C - \eta(\mu_1 + \mu_2 - 1) \tag{8}$$

By partial derivative of μ and η, and the following conditions are met:

$$\begin{cases} \dfrac{\partial L(\alpha,\eta)}{\partial \mu} = 0 \\ \dfrac{\partial L(\alpha,\eta)}{\partial \eta} = 0 \end{cases} \tag{9}$$

Based on formula (8), we can get the update formula for μ_i as:

$$\mu_i = \frac{(\frac{1}{C_i})^{\frac{1}{q-1}}}{\sum_{i=1}^{2}(\frac{1}{C_i})^{\frac{1}{q-1}}} \tag{10}$$

In steps 2.2-2.4 of the algorithm flow, the gradient descent algorithm is used to update other variables, using the same learning rate γ. In particular, the regression coefficient matrix W is updated in step 2.2. The derivative of the cost function relative to W is as follows:

$$\frac{\partial F}{\partial W} = 2\beta(f(W^T f - L)^T + \alpha_2 EW) \tag{11}$$

Where, E is the diagonal matrix of the matrix $e_{kk} = \frac{1}{2}\|w_k\|_2$, and w_k is the kth row of the regression coefficient matrix W. Then, update W according to the gradient descent rule:

$$W \leftarrow W - \gamma \frac{\partial S}{\partial W} \tag{12}$$

In step 2.3 of the algorithm flow, the feature representation f is updated. Considering that f contains a related partial feature f_c and independent f_{1I} and f_{2I} of specific modalities, these three parts are updated separately. In this way, the learned features are forced to contain the relevant feature characteristics and the characteristics of the individual features of the particular modality. Therefore, the derivative of the cost function S with respect to f_c, f_{1I} and f_{2I} can be found, and the update mechanism of f_c, f_{1I} and f_{2I} can be realized as follows:

$$\begin{aligned} \frac{\partial S}{\partial f_c} = {} & 2\alpha_1^q[(f_c - W_{1c}X_1) + W_{1c}({W^T}_{1c}f_c - X_1) + \alpha_1 g'(f_c)] + 2\alpha_2^q[(f_c - W_{2c}X_2) \\ & + W_{1c}({W^T}_{1c}f_c - X_2) + \alpha_1 g'(f_c)] + 2\lambda W^{(c)}(W^{(c)T}f_c - L) \end{aligned} \tag{13}$$

$$\frac{\partial S}{\partial f_{1I}} = 2\alpha_1^q[(f_{1I} - W_{1I}X_1) + W_{1I}({W^T}_{1I}f_{1I} - X_1) + \alpha_1 g'(f_{1I})] + 2\lambda W^{(1I)}(W^{(1I)T}f_{1I} - L) \tag{14}$$

$$\frac{\partial S}{\partial f_{2I}} = 2\alpha_2^q[(f_{2I} - W_{2I}X_2) + W_{2I}({W^T}_{2I}f_{2I} - X_2) + \alpha_1 g'(f_{2I})] + 2\lambda W^{(2I)}(W^{(2I)T}f_{2I} - L) \tag{15}$$

After completing the above derivation, the relevant partial features of f and the specific modal independent partial features are updated according to following gradient descent rules:

$$\begin{cases} f_c \leftarrow f_c - \gamma \dfrac{\partial S}{\partial f_c} \\ f_{1I} \leftarrow f_{1I} - \gamma \dfrac{\partial S}{\partial f_{1I}} \\ f_{2I} \leftarrow f_{2I} - \gamma \dfrac{\partial S}{\partial f_{2I}} \end{cases} \tag{16}$$

In step 2.4 of the algorithm flow, when f, W, and μ_i are fixed, W_i is updated in a similar manner, as follows:

$$\frac{\partial S}{\partial W_1} = 2\alpha_1^q[(W_1X_1 - f_{RGB})X_1^T + f_{RGB}({W^T}_1 f_{RGB} - X_1)^T]W_1 \leftarrow W_1 - \gamma \frac{\partial S}{\partial W_1} \tag{17}$$

In the multimodal framework, X_1 and X_2 are the output characteristics of the pooled layer of each output layer, and the results of multimodal learning are backpropagated to the lower layers of the CNN network by the following formula.

$$\frac{\partial S}{\partial X_1} = 2\alpha_1^q[W^T{}_1(W_1X_1 - f_{RGB}) - (W^T{}_1 f_{RGB} - X_1)] \tag{18}$$

$$\frac{\partial S}{\partial X_2} = 2\alpha_2^q[W^T{}_2(W_2X_2 - f_{DEPTH}) - (W^T{}_2 f_{DEPTH} - X_2)] \tag{19}$$

Multimodal feature learning and backpropagation iterations are performed by formulas (18) and (19) until convergence. The designed multi-modal learning framework mainly solves the fusion of the two modes, but through the above expansion, it may also extend to the fusion of more modes, and the related parts and their respective independents are merged by integrating more modes [48]. Some of the features are connected to represent the features being learned.

4. Gesture recognition experiment results analysis

4.1. Gesture recognition experiment configuration

(1) All experiments were performed under the Ubantu 16.04 system. The graphics card is NVIDIA GTX10606G, and the sensor used for image acquisition and experimentation is Kinect1.0. The software environment being run is configured as: Python 3.5, Tensorflow-1.14-GPU, Libfreenect 1.0 and other auxiliary Python libraries.

(2) Using the experimental equipment built, construct a gesture database similar to the American Sign Language (ASL) [49, 50]. The self-built gesture database contains a total of 10 static gestures, representing Arabic numerals 1-10, which are acquired by 7 different gesture operators under different ambient lighting and background conditions. Each gesture sample contains corresponding color images and depth images. For each operator, each gesture contains about 400 color and depth maps. Therefore, the total number of images in the self-built gesture database is 28,000.

(3) For RGB-D images, the ResNet-18 network structure was used to extract the features of the two modes, and the network structure was pre-trained on ImageNet. The extracted features are output at each hidden layer, and the output of the current layer continues as the input of the next layer to perform the convolution operation extraction feature. For RGB mode and Depth mode, the size of the input image is adjusted to $224\times224\times3$. The feature extracted by the two networks is used as the input of the weight adaptive structure for feature fusion and splicing. The fusion features of each abstraction layer outputted by the weight network are sequentially input to the LSTM network for training according to the sequential method. The output of the LSTM network is connected to the Softmax layer, and the final output is the classification result of each gesture. The framework design of the established model is shown in Figure 4.

(4) We select some important influence factors and coefficients according to the related literature [51,52]. For the designed multimodal feature weight adaptive learning framework, the dimensionality M_1' and M_2' of the transformed features are set to $M_1 = 512$ and $M_2 = 512$ are the same. We ignore the fact that they may be different. Between the two modes, $P_{1I} = 256, P_{2I} = 256, P_c = 256$, it is mandatory to keep half of the M_1' and M_2' dimension features the same. In all experiments, the parameters $q_1, \lambda_1, \alpha_1, \alpha_2, \gamma$ of the RGB-D object were set to 1.5, 1500, 2, 10, 0.001, respectively, according to experience.

(5) In two different database samples, the following operations are performed respectively: 70% of the samples are randomly used as the training set, and the other 30% data samples are used as the test set. When training the model, a cross-validation method is used for training to ensure the accuracy and generalization performance of the model.

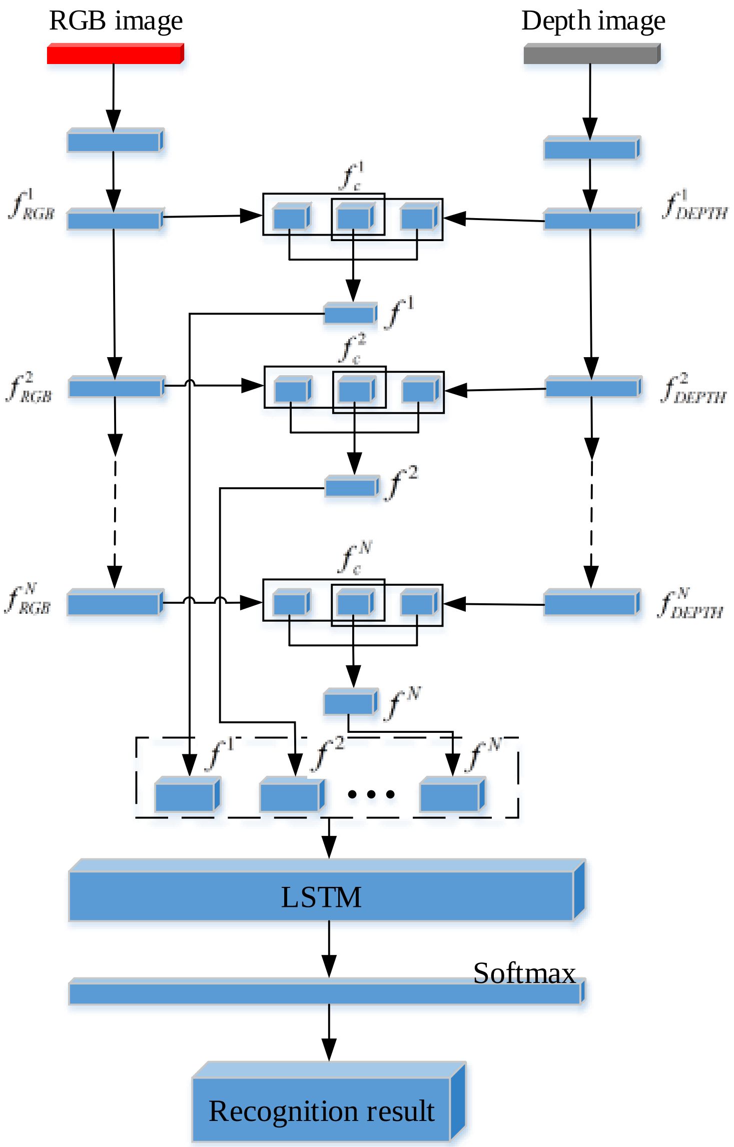

Figure 4 Multimodal fusion gesture recognition method framework

4.2. Self-built gesture database experimental results analysis

Based on the self-built gesture database, four deep learning methods for RGB-D gesture recognition are constructed by ResNet respectively, and compared with the method of this paper. The multi-modal feature weight adaptation proposed in this paper is verified. The innovative approach to gesture recognition. The four methods are designed as follows: 1) the single-mode depth image as the input ResNet network, as shown in (a) of Fig 5; 2) the single-mode RGB image as the input ResNet network, Fig 5 (a) 3) bimodal RGB-D data as a four-channel input ResNet network, shown in Fig 5 (b); 4) using RGB-D bimodal data and in the last layer of the fully connected layer The combination of the ResNet-18 network structure is shown in Figure 5(c).

Table 1 shows the comparison of the accuracy of different methods on a self-built gesture database. From the comparison in the table, it can be found that the accuracy of the RGB-D data as a four-channel data input is higher than that of a simple single-mode network [53, 54]. However, the gesture recognition rate on the converged network is comparable. It is better in several other methods [55, 56]. Therefore, it is proved that the multi-modal fusion method proposed in this paper is feasible for improving the accuracy of gesture recognition, and the accuracy of the proposed method is higher than that of other traditional fusion methods.

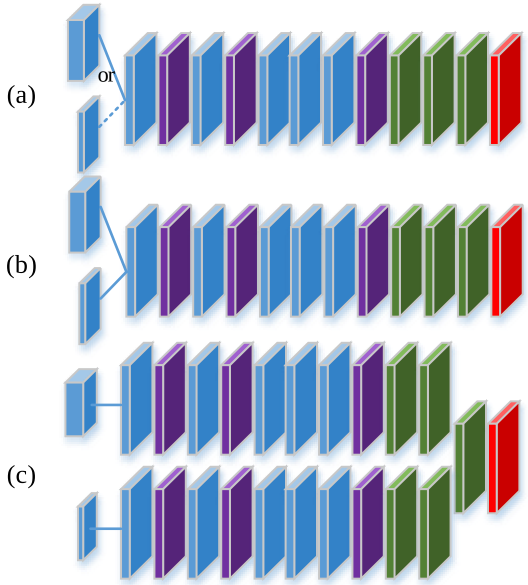

Figure 5. Different processing methods for RGB-D gesture images (The leftmost are input samples, the large cube represents the RGB image, the small represents the Depth image; the rightmost red cube represents the output of the network; in the middle of the network structure, the blue cube represents the convolutional layer, and the cyan cube represents The pooled layer, the green cube represents the fully connected layer; the dashed line indicates that the input is an RGB or Depth image, which is an optional input.)

Table 1. Experimental results of different network structure methods in the RGB-D gesture data set

Method	Maximum recognition rate(%)	Minimum recognition rate(%)	Accuracy(%)
RGB-ResNet (Single mode)	88	83.4	85.7
Depth-ResNet (Single mode)	89.2	85.4	87.3
ResNet (RGB-D as a four-channel data input)	94.8	89.8	92.3
RGB-Depth fusion(Full connection layer fusion)	96.8	94.6	95.7
The method proposed in this paper	98.8	97.8	98.3

Table 2. Experimental results of different network models in self-built gesture data sets

Method	Maximum recognition rate(%)	Minimum recognition rate(%)	Accuracy(%)
Literature[57]	89.8	85	87.4
Literature[58]	92.1	87.5	89.8
Literature[59]	93.4	87.6	90.5
Literature[20]	94.6	90.4	92.5
Literature[18]	94.9	91.9	93.4
The method proposed in this paper	98.8	97.8	98.3

In Table 2, Blum et al et al. [57] used a convolutional k-mean descriptor method; Socher et al. [58] used a recursive neural network plus CNN method; Bo et al. [59] a method of adding input channels using feature learning based on sparse coding; Coupri [20] uses CNN and feature sharing methods; Gupta [18] uses CNN and feature fusion methods. However, from the accuracy rate in the table, it can be seen that the method of this paper is better than the current cutting-edge identification method.

After analyzing the recognition accuracy of several different structures on the self-built gesture database, the feasibility and robustness of the proposed method are verified [60, 61, 62]. However, the recognition accuracy of the proposed method is unknown compared to the current state-of-the-art method [63, 64, 65]. Therefore, several current cutting-edge RGB-D image-based target recognition methods or gesture recognition methods are discussed [66, 67, 68]. The validity and accuracy of the method are shown in Table 2.

In order to understand the degree of the method proposed in this paper more intuitively and clearly, the confusion matrix of the method is calculated based on the RGB-D gesture database.

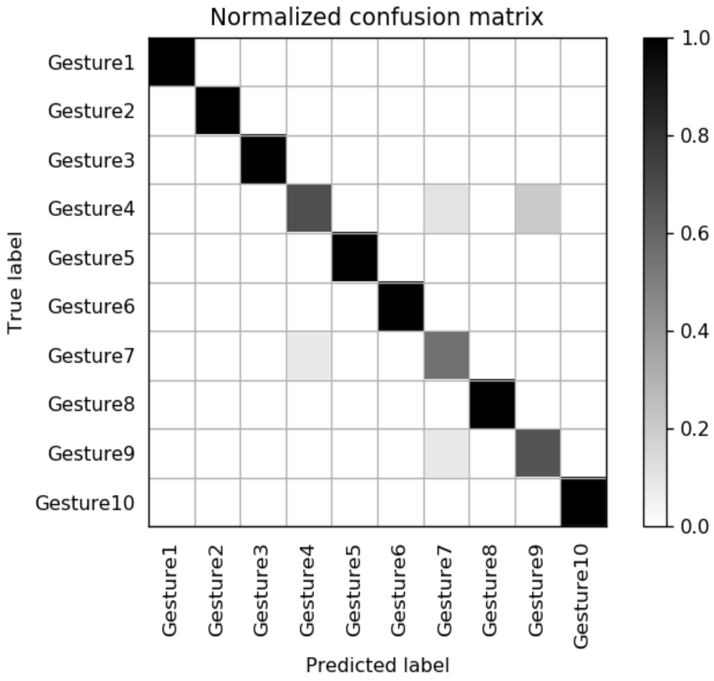

Figure 6. The confusion matrix of the proposed method in the self-built gesture database

As shown in Figure 6, the diagonal of the confusion matrix indicates the recognition accuracy of each category, with values ranging from 0 to 1, with 1 being the highest. It can be seen from the figure that there are still cases of misjudgment in the classification of some gestures, For example, the gesture numbers 4, 7, and 9 have low recognition accuracy due to their similarity, but the overall gesture recognition accuracy is relatively good.

4.3. ASL gesture database experiment results analysis

In order to repeatedly verify the advancement and effectiveness of the proposed method, experiments will be carried out on the ASL gesture database to prove that the model has better generalization performance and can display its superior performance on different data sets [69, 70]. Similarly, several other methods for superior performance on this database were analyzed on ASL, as shown in Table 3.

In Table 3, Li[71] trains a CNN framework based on a soft attention mechanism in an end-to-end manner; Wang [72] proposed the depth sequence representation method of dynamic depth image (DDI), dynamic depth normal image (DDNI) and dynamic depth motion normal image (DDMNI); Duan [73] proposed a spatial-temporal network architecture based on consensus-voting; Traver [74] uses 3D integral imaging for gesture recognition. Comparing the recognition results of the method proposed in this paper with the results in the above documents, the comparison results of the maximum recognition rate, minimum recognition rate and average recognition rate are given. It can be seen that the average recognition rate of the method in this paper is higher than other several A similar approach. Also in order to more intuitively understand the effectiveness and accuracy of this method on the ASL gesture database, the confusion matrix on the database is calculated, as shown in Figure 7.

Since only static gesture samples were selected when selecting data samples, the category is 24 categories, and the diagonal line of the confusion matrix represents the recognition accuracy of each category. Its value ranges from 0 to 1, with 1 being the highest. It can be more It can be seen that the recognition effect of the method proposed in this article is better [75]. But there are also some gestures that are very similar, such as gestures A, E, M, N, S, T, etc. Examples of related gestures are shown in Figure 8. These gestures are all developed from the fist posture, similar to the four-finger clenching, the only difference is the position of the thumb, so in the recognition process, the recognition accuracy is average, which is also one of the tasks we will study in the future.

Table 3. Experimental results of different network models in the ASL gesture dataset

Method	Maximum recognition rate(%)	Minimum recognition rate(%)	Accuracy(%)
Literature[67]	98	92	95.6
Literature[68]	99.9	85.6	95.4
Literature[69]	97.3	89.7	94.4
Literature[70]	98.2	91	94.6
The method proposed in this paper	99	95.4	97.2

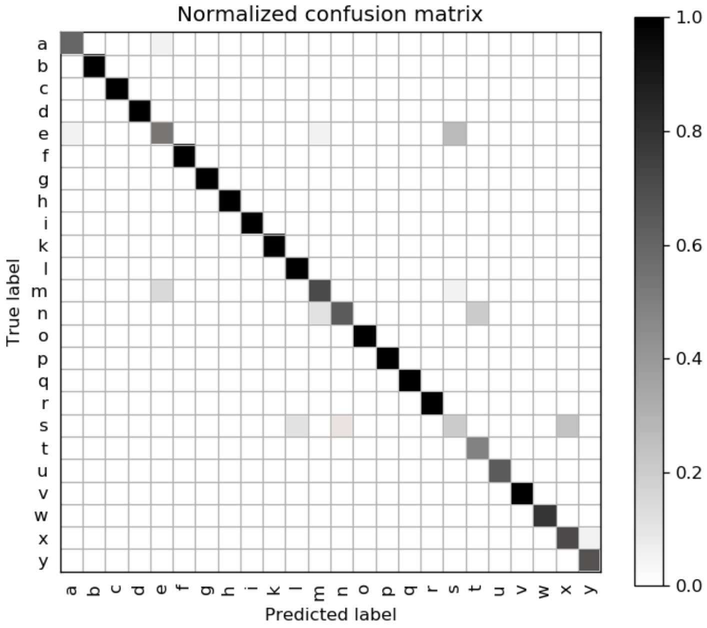

Figure 7. Confusion matrix of the proposed method on the ASL gesture database

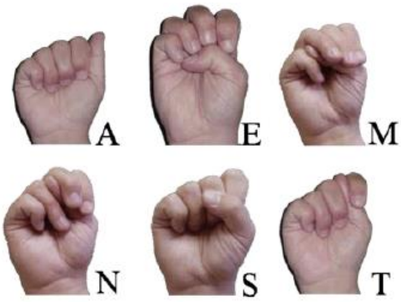

Figure 8. Some similar gestures in the ASL database

5. Conclusions

Gesture recognition technology based on computer vision is widely used in human-computer interaction systems due to its convenience and speed. Compared with the gesture recognition of RGB images, the gesture recognition based on RGB-D images contains the characteristics of spatial information, which can avoid the influence of factors such as complex background and lighting changes on the accuracy of the gesture recognition algorithm. At present, the research on gesture recognition of RGB-D images is not efficient enough in the use of RGB-D image information. To solve this problem, this paper designs a multi-modal feature weight adaptive fusion method, fully considering different modalities and different abstractions Horizontal relationship to realize multi-modal fusion gesture recognition. According to the multi-modal feature weight adaptive fusion method designed in this paper, simulation experiments of different modal fusion methods are designed, which proves that the method proposed in this paper is better than the traditional RGB-D image processing method, and the accuracy of gesture recognition is higher. Then by comparing with several current more advanced methods, it is found that the method proposed in this paper is feasible and the result of gesture recognition is more accurate and robust. The method in this paper is expanded on the basis of double convolutional neural network, but it does not analyze the influence of different network structures on the method. In the future, this article will discuss the influence of different network structures on this method.

Acknowledgements: This work was supported by grants of National Natural Science Foundation of China (Grant Nos.51575407, 51505349, 61733011, 41906177); the Grants of Hubei Provincial Department of Education (D20191105); the Grants of National Defense Pre-Research Foundation of Wuhan University of Science and Technology (GF201705) and Open Fund of the Key Laboratory for Metallurgical Equipment and Control of Ministry of Education in Wuhan University of Science and Technology (2018B07,2019B13).

References

1. Jiabin Hu, Ying Sun, Gongfa Li, Guozhang Jiang and Bo Tao, Probability Analysis for Grasp Planning Facing the Field of Medical Robotics. Measurement, 2019, 141， 227-234.
2. Ruyi Ma, Leilei Zhang, Gongfa Li, Du Jiang, Shuang Xu, Disi Chen, Grasping force prediction based on sEMG signals. Alexandria Engineering Journal, 2020, 59(3), 1135-1147.
3. Du Jiang, Gongfa Li, Ying Sun, Jianyi Kong, Bo Tao, Disi Chen, Grip strength forecast and rehabilitative guidance based on adaptive neural fuzzy inference system using sEMG. Personal and Ubiquitous Computing, 2019, DOI: 10.1007/s00779-019-01268-3.
4. Mingchao Yu, Gongfa Li, Du Jiang, Guozhang Jiang, Bo Tao, Disi Chen, Hand medical monitoring system based on machine

learning and optimal EMG feature set. Personal and Ubiquitous Computing, 2019, DOI: 10.1007/s00779-019-01285-2.

5. Jinrong Tian, Wentao Cheng, Ying Sun, Gongfa Li, Du Jiang, Guozhang Jiang, Bo Tao, Haoyi Zhao, Disi Chen, Gesture recognition based on multilevel multimodal feature fusion. Journal of Intelligent & Fuzzy Systems, 2020, 38(3), 2539-2550.
6. Ying Sun, Chao Xu, Gongfa Li, Wanfen Xu, Jianyi Kong, Du Jiang, Bo Tao, Disi Chen, Intelligent human computer interaction based on non redundant EMG signal. Alexandria Engineering Journal, 2020, 59(3), 1149-1157.
7. N. Rahim, J. Ahmad, K. Muhammad, A.K. Sangaiah, S. W. Baik, Privacy-preserving image retrieval for mobile devices with deep features on the cloud. Computer Communications, 2019, 127, 75-85.
8. Gongfa Li, Jiahan Li, Zhaojie Ju, Ying Sun, Jianyi Kong, A novel feature extraction method for machine learning based on surface electromyography from healthy brain. Neural Computing and Applications, 2019, 31(12), 9013-9022.
9. Li Huang, Qiaobo Fu, Gongfa Li, Bowen Luo, Disi Chen, Hui Yu, Improvement of maximum variance weight partitioning particle filter in urban computing and intelligence. IEEE Access, 2019, 7, 106527-106535.
10. Yangwei Cheng, Gongfa Li, Jiahan Li, Ying Sun, Guozhang Jiang, Fei Zeng, Haoyi Zhao, Disi Chen, Visualization of activated muscle area based on semg. Journal of Intelligent & Fuzzy Systems, 2020, 38(3), 2623-2634.
11. Wentao Cheng, Ying Sun, Gongfa Li, Guozhang Jiang, Honghai Liu, Jointly network: a network based on CNN and RBM for gesture recognition. Neural Computing and Applications, 2019, 31(Supplement 1), 309-323, DOI:10.1007/s00521-018-3775-8.
12. Ying Sun,Yaoqing Weng,Bowen Luo,Gongfa Li,Bo Tao,Disi Chen, Du Jiang, Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. IET Image Processing, 2020, DOI:10.1049/iet-ipr.2020.0148 .
13. Ying Sun, Cuiqiao Li, Gongfa Li, Guozhang Jiang, Du Jiang, Honghai Liu, Zhigao Zheng, Wanneng Shu, Gesture Recognition Based on Kinect and sEMG Signal Fusion. Mobile Networks and Applications, 2018, 23(4), 797-805.
14. Mingchao Yu, Gongfa Li, Du Jiang, Guozhang Jiang, Fei Zeng, Haoyi Zhao, Disi Chen, Application of pso-rbf neural network in gesture recognition of continuous surface emg signals. Journal of Intelligent & Fuzzy Systems, 2020, 38(3), 2469-2480.
15. Shangchun Liao, Gongfa Li, Jiahan Li, Du Jiang, Guozhang Jiang, Ying Sun, Bo Tao, Haoyi Zhao, Disi Chen, Multi-object intergroup gesture recognition combined with fusion feature and KNN algorithm. Journal of Intelligent & Fuzzy Systems, 2020, 38(3), 2725-2735.
16. Li Huang, Meiling He, Chong Tan, Du Jiang, Gongfa Li ,Hui Yu, Jointly Network Image Processing: Multi-task Image Semantic Segmentation of Indoor Scene Based on CNN. IET Image Processing, 2020, DOI: 10.1049/iet-ipr.2020.0088.
17. Q. Gao, J. Liu, Z. Ju, Robust real-time hand detection and localization for space human robot interaction based on deep learning. Neurocomputing, 2019, DOI: 10.1016/j.neucom.2019.02.066.
18. S. Gupta, R. Girshick, P. Arbeláez, et al. Learning rich features from RGB-D images for object detection and segmentation. European Conference on Computer Vision. Springer, Cham, 2014, 345-360.
19. D. Huo, Y. F. Chen, F. X. Li, Z. C. Lei, Modality-convolutions: Multi-modal Gesture Recognition based on Convolutional Neural Network. ICCSE 2017, pp:349-353.
20. L. Duval, M. Moreaud, C. Couprie, D. Jeulin, H. Talbot, J. Angulo, Image processing for materials characterization: Issues, challenges and opportunities. ICIP 2014, pp:4862-4866.
21. R. Huang , Y. Xing and Z. Z. Wang, RGB-D salient object detection by a CNN with multiple layers fusion. IEEE Signal Processing Letters, 2019, 1-1.
22. P. Wang, W. Li, P. Ogunbona, et al. RGB-D-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding, 2018, 171, 118-139.
23. M. Loiimitz, A. Eitel, A. Vasquez and W. Burgard, Deep 3D perception of people and their mobility aids. Robotics & Autonomous Systems, 2019, 114, 29-40.
24. S. Ravi, M. Suman, P. V. V. Kishore, K. Kumar E, T. K. Kumar M, A. Kumar D, Multi Modal Spatio Temporal co-trained CNNs with Single Modal testing on RGB–D based Sign Language Gesture Recognition. Journal of Computer Languages, 2019, 52, 88-102.
25. S. Park, M. Ji and J. Chun, 2D Human Pose Estimation based on Object Detection using RGB-D information. Transactions on Internet and Information Systems, 2018, 12, 800-816.
26. A. Elboushaki, R. Hannane, K. Afdel, L. Koutti, MultiD-CNN: A Multi-dimensional Feature Learning approach based on deep

Convolutional Networks for Gesture Recognition in RGB-D Image Sequences. Expert Systems with Applications, 2020, DOI: 10.1016/j.eswa.2019.112829.

27. H. Lin and Y. Chiang, Understanding Human Hand Gestures for Learning Robot Pick-and-Place Tasks. International Journal of Advanced Robotic Systems, 2015, 12.
28. F. Negin , P. Rodriguez , M. Koperski , et al. PRAXIS: Towards Automatic Cognitive Assessment Using Gesture Recognition. Expert Systems with Applications, 2018, 106, 21-35.
29. M. Gao , J. Jiang, G. Zou , et al. RGB-D-Based Object Recognition Using Multimodal Convolutional Neural Networks: A Survey. IEEE Access, 2019, 7, 43110-43136.
30. L. Pigou, A. Van Den Oord, S. Dieleman, et al. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. International Journal of Computer Vision, 2018, 126, 430-439.
31. Gongfa Li, Du Jiang, Ying Sun, Guozhang Jiang, Bo Tao, Life prediction mechanism of ladle composite structure body based on simulation technology. Archives of Metallurgy and Materials, 2019, 64(4), 1555-1562.
32. Chong Tan, Ying Sun, Gongfa Li, Guozhang Jiang, Disi Chen, Honghai Liu, Research on Gesture Recognition of Smart Data Fusion Features in the IoT. Neural Computing and Applications, 2019, DOI：10.1007/s00521-019-04023-0.
33. Zhihua Cui, Xianghua Xu, Fei Xue, Xingjuan Cai, Yang Cao, Wensheng Zhang, Jinjun Chen, Personalized Recommendation System based on Collaborative Filtering for IoT Scenarios. IEEE Transactions on Services Computing, 2020, DOI:10.1109/TSC.2020.2964552
34. Du Jiang, Gongfa Li, Ying Sun, Jianyi Kong, Bo Tao, Gesture recognition based on skeletonization algorithm and CNN with ASL database. Multimedia Tools and Applications, 2019, 78(21), 29953-29970.
35. Tinggui Chen, Shiwen Wu, Jianjun Yang, Guodong Cong, Risk Propagation Model and Its Simulation of Emergency Logistics Network Based on Material Reliability. International Journal of Environmental Research and Public Health, 2019, 16(23), 4677.
36. Gongfa Li, Leilei Zhang, Ying Sun, Jianyi Kong, Towards the sEMG hand: internet of things sensors and haptic feedback application. Multimedia Tools and Applications, 2019, 78(21), 29765-29782.
37. N. Hesse , S. Pujades , M. J. Black , et al. Learning and Tracking the 3D Body Shape of Freely Moving Infants from RGB-D sequences. IEEE transactions on pattern analysis and machine intelligence, 2018, DOI:10.1109/TPAMI.2019.2917908.
38. Du Jiang, Zujia Zheng, Gongfa Li, Ying Sun, Jianyi Kong, Guozhang Jiang, Hegen Xiong, Bo Tao, Shuang Xu, Honghai Liu, Zhaojie Ju. Gesture recognition based on binocular vision. Cluster Computing, 2019, 22(Supplement 6), 13261-13271, DOI:10.1007/s10586-018-1844-5.
39. Zhihua Cui, Fei Xue, Shiqiang Zhang, Xingjuan Cai, Yang Cao, Wensheng Zhang, Jinjun Chen, A Hybrid BlockChain-Based Identity Authentication Scheme for Multi-WSN. IEEE Transactions on Services Computing, 2020, 13(2), 241-251,
40. Gongfa Li, Heng Tang, Ying Sun, Jianyi Kong, Guozhang Jiang, Du Jiang, Bo Tao, Shuang Xu, Honghai Liu, Hand gesture recognition based on convolution neural network. Cluster Computing, 2019, 22(Supplement 2), 2719-2729, DOI:10.1007/s10586-017-1435-x.
41. Yang He, Gongfa Li, Yajie Liao, Ying Sun, Jianyi Kong, Guozhang Jiang, Du Jiang, Honghai Liu, Gesture recognition based on an improved local sparse representation classification algorithm. Cluster Computing, 2019, 22(Supplement 5), 10935-10946, DOI:10.1007/s10586-017-1237-1.
42. Tinggui Chen, Qianqian Li, Jianjun Yang, Guodong Cong, Gongfa Li, Modeling of the Public Opinion Polarization Process with the Considerations of Individual Heterogeneity and Dynamic Conformity. Mathematics, 2019, 7(10), 917.
43. Bei Li, Ying Sun, Gongfa Li, Jianyi Kong, Guozhang Jiang, Du Jiang, Bo Tao, Shuang Xu, Honghai Liu, Gesture recognition based on modified adaptive orthogonal matching pursuit algorithm. Cluster Computing, 2019, 22 (Supplement 1), 503-512, DOI:10.1007/s10586-017-1231-7.
44. Disi Chen, Gongfa Li, Ying Sun, Jianyi Kong, Guozhang Jiang, Heng Tang, Zhaojie Ju, Hui Yu, Honghai Liu, An interactive image segmentation method in hand gesture recognition. Sensors 2017, 17(2), 253.
45. Yajie Liao, Ying Sun, Gongfa Li, Jianyi Kong, Guozhang Jiang, Du Jiang, Haibin Cai, Zhaojie Ju, Hui Yu, Honghai Liu, Simultaneous calibration: a joint optimization approach for multiple kinect and external cameras. Sensors, 2019, 17(7), 1491.
46. Zhihua Cui, Jiangjiang Zhang, Di Wu, Xingjuan Cai, Hui Wang, Wensheng Zhang, Jinjun Chen, Hybrid Many-Objective Particle Swarm Optimization Algorithm for Green Coal Production Problem. Information Sciences, 2020, 518, 256-271.
47. Tinggui Chen, Shiwen Wu, Jianjun Yang, Guodong Cong, Gongfa Li, Modeling of Emergency Supply Scheduling Problem Based

on Reliability and Its Solution Algorithm under Variable Road Network after Sudden-Onset Disasters. Complexity, 2020, Volume 2020, Article ID 7501891, 15 pages, DOI: https://doi.org/10.1155/2020/7501891.

48. Xingjuan Cai, Penghong Wang, Lei Du, Zhihua Cui, Wensheng Zhang and Jinjun Chen, Multi-Objective Three-Dimensional DV-Hop Localization Algorithm With NSGA-II. IEEE Sensors Journal, 2019,19(21), 10003-10015.
49. Jinxian Qi, Guozhang Jiang, Gongfa Li, Ying Sun, Bo Tao, Surface EMG hand gesture recognition system based on PCA and GRNN. Neural Computing and Applications, 2020, 32(10), 6343-6351.
50. M. Hassan, M. Rehmani, J. Chen, Differential Privacy Techniques for Cyber Physical Systems: A Survey. IEEE Communications Surveys and Tutorials, 2020, 22(1), 746-789.
51. Penghong Wang, Jianrou Huang, Zhihua Cui, Liping Xie and Jinjun Chen, A Gaussian Error Correction Multi-Objective Positioning Model with NSGA-II. Concurrency and Computation Practice and Experience, 2020, 32(5), e5464.
52. Chengcheng Li, Gongfa Li, Guozhang Jiang, Disi Chen, Honghai Liu, Surface EMG data aggregation processing for intelligent prosthetic action recognition. Neural Computing and Applications, 2018, DOI：10.1007/s00521-018-3909-z.
53. Tinggui Chen, Qianqian Li, Peihua Fu, Jianjun Yang, Chonghuan Xu, Guodong Cong, Gongfa Li, Public Opinion Polarization by Individual Revenue from the Social Preference Theory. International Journal of Environmental Research and Public Health, 2020, 17(3), 946.
54. Xingjuan Cai, Yun Niu, Shaojin Geng, Jiangjiang Zhang, Zhihua Cui, Jianwei Li and Jinjun Chen, An under-sampled software defect prediction method based on hybrid multi-objective cuckoo search. Concurrency and Computation: Practice and Experience, 2020, 32(5), e5478.
55. Gongfa Li, Hao Wu, Guozhang Jiang, Shuang Xu and Honghai Liu, Dynamic Gesture Recognition in the Internet of Things. IEEE Access, 2019, 7, 23713-23724.
56. Bowen Luo, Ying Sun, Gongfa Li, Disi Chen, Zhaojie Ju, Decomposition Algorithm for Depth Image of Human Health Posture Based on Brain Health. Neural Computing and Applications, 2020, 32(10), 6327-6342.
57. M. Blum, J. T. Springenberg, J. W ülfing, et al. A learned feature descriptor for object recognition in rgb-d data. IEEE, 2012, 1298-1303.
58. R. Socher, B. Huval, B. Bath, et al. Convolutional-recursive deep learning for 3d object classification. Advances in neural information processing systems, 2012, 656-664.
59. L. Bo, X. Ren, D. Fox, Unsupervised feature learning for RGB-D based object recognition. Experimental Robotics, Springer, Heidelberg, 2013, 387-402.
60. J. Feng, L.T. Yang, R. Zhang, W. Qiang and J. Chen, Privacy Preserving High-Order Bi-Lanczos in Cloud-Fog Computing for Industrial Applications. IEEE Transactions on Industrial Informatics, 2020, DOI: 10.1109/TII.2020.2998086.
61. Zhihua Cui, Lei Du, Penghong Wang, Xingjuan Cai and Wensheng Zhang, Malicious code detection based on CNNs and multi-objective algorithm. Journal of Parallel and Distributed Computing, 2019, 129, 50-58.
62. Gongfa Li, Du Jiang, Yanling Zhou, Guozhang Jiang, Jianyi Kong, Gunasekaran Manogaran, Human Lesion Detection Method Based on Image Information and Brain Signal. IEEE Access, 2019, 7, 11533-11542.
63. S. Li, B. Yu, W. Wu, et al. Feature learning based on SAE–PCA network for human gesture recognition in RGBD images. Neurocomputing, 2015, 151, 565-573.
64. Ying Sun, Jinrong Tian, Du Jiang,Bo Tao, Ying Liu, Juntong Yun, Disi Chen, Numerical simulation of thermal insulation and longevity performance in new lightweight ladle. Concurrency and Computation: Practice and Experience, 2020, DOI:10.1002/CPE.5830.
65. Jinxian Qi, Guozhang Jiang, Gongfa Li, Ying Sun, Bo Tao. Intelligent human-computer interaction based on surface EMG gesture recognition. IEEE Access, 2019, 7: 61378-61387, DOI:10.1109/ACCESS.2019.2914728.
66. Y. Wang, P. Wang, J. Zhang, Z. Cui, X. Cai, W. Zhang and J. Chen, A Novel Bat Algorithm with Multiple Strategies Coupling for Numerical Optimization, Mathematics, 7(2), Article Number: 135, 2019.

67. Zhihua Cui, Feixiang Li, Wensheng Zhang, Bat algorithm with principal component analysis, International Journal of Machine Learning and Cybernetics, 2019, 10(3): 603-622

68. Yaoqing Weng, Ying Sun, Du Jiang, Bo Tao, Ying Liu, Jutong Yun, Dalin Zhou, Enhancement of Real-Time Grasp Dectection by Cascaded Deep Convolutional Neural Networks, Concurrency and Computation:Practice and Expericence, 2020, DOI: 10.1002/cpe.5976.

69. Hong Zhu, Peng Yao, Xizhao Wang, Weight learning from cost matrix in weighted least squares model based on genetic algorithm. International Journal of Bio - Inspired Computation, 2019, 13, 4, 269-276.

70. Ying Sun, Jiabing Hu, Gongfa Li, Guozhang Jiang, Hegen Xiong, Bo Tao, Zujia Zheng, Du Jiang, Gear reducer optimal design based on computer multimedia simulation. The Journal of Supercomputing, 2020,76(6), 4132–4148.

71. Y. Li, X. Wang, W. Liu, et al. Deep attention network for joint hand gesture localization and recognition using static RGB-D images. Information Sciences, 2018, 441, 66-78.

72. P. Wang, W. Li, Z. Gao, et al. Depth pooling based large-scale 3-d action recognition with convolutional neural networks. IEEE Transactions on Multimedia, 2018, 20, 1051-1061.

73. J. Duan, J. Wan, S. Zhou, et al. A unified framework for multi-modal isolated gesture recognition. ACM Transactions on Multimedia Computing Communications and Applications, 2018, 14(1s), 21.

74. V. J. Traver, P. Latorrecarmona, E. Salvadorbalaguer, et al. Three-Dimensional Integral Imaging for Gesture Recognition Under Occlusions. IEEE Signal Processing Letters, 2017, 24(2), 171-175.

75. D. Puthal, S. Nepal, R. Ranjan, J. Chen, A Dynamic Key Length based Approach for Real-time Security Verification of Big Sensing Data Stream, International Conference on Web Information Systems Engineering (WISE 2015), pp. 93-108, 2015.