EAI Endorsed Transactions
on Scalable Information Systems

Research Article EAI.EU

Spatio-temporal weight Tai Chi motion feature extraction based on deep network cross-layer feature fusion

Naiqiu Wu[1,*] and Yang Shi[2]

[1]Department of Arts and Sports, Henan Technical College of Construction, Zhengzhou, 450064, China
[2]College of Information and Electronic Technique, Jiamusi University, Jiamusi, 154007 China

Abstract

Tai Chi is a valuable exercise for human health. The research on Tai Chi is helpful to improve people's exercise level. There is a problem with low efficiency in traditional Tai Chi motion feature extraction. Therefore, we propose a spatio-temporal weight Tai Chi motion feature extraction based on deep network cross-layer feature fusion. According to the selected motion spatio-temporal sample, the corresponding spatio-temporal motion key frame is extracted and output in the form of static image. The initial motion image is preprocessed by motion object detection and image enhancement. Traditional convolutional neural network extracts features from the shallow to the deep and builds a classifier for image classification, which is easy to ignore the shallow features. Based on the AlexNet network, a CL-AlexNet network is proposed. Batch normalization (BN) is used for data normalization. The cross-connection structure is introduced and the sensitivity analysis is performed. The Inception module is embedded for multi-scale depth feature extraction. It integrates deep features and shallow features. The spatio-temporal weight adaptive interpolation method is used to reduce the error of edge detection. From the edge features and the motion spatio-temporal features, it realizes motion features extraction, and outputs the extraction results. Compared with the state-of-the-art feature extraction algorithms, the experiment results show that the proposed algorithm can extract more effective features. The recognition rate exceeds 90%. It can be used as guidance and evidence for Tai Chi training.

Keywords: Tai Chi motion feature extraction, spatio-temporal weight, cross-layer feature fusion, deep network.

Received on 04 September 2021, accepted on 17 September 2021, published on 21 September 2021

doi: 10.4108/eai.21-9-2021.170964

*Corresponding author. Email: aqiufenga@163.com

1. Introduction

As the essence of Chinese martial arts, Tai Chi is a national intangible cultural heritage. Studies have shown that Tai Chi can not only help people reduce blood pressure [1], enhance the functional level of the immune system, relieve physical stress and improve the quality of sleep [2], but also enhance muscle strength, improve flexibility and prevent falls [3,4]. So it is attracted by more and more people.

Gesture motion recognition has always been a hot research topic in the field of computer vision, which has important academic value in many fields such as video surveillance, motion analysis, sports events and medical diagnosis [5-7]. The recognition of human posture motion will apply relevant algorithms and techniques. The motion recognition methods based on spatio-temporal feature extraction and based on motion trajectory analysis are the most frequently used motion attitude and motion recognition methods at present. In order to improve the recognition accuracy of human motion posture behavior (Tai Chi), this paper optimizes the spatio-temporal feature extraction method of motion.

The spatio-temporal weight feature extraction combines computer vision technology and image processing technology. The computer vision technology is used to extract the relevant information of human spatio-temporal posture motion, and determine whether the point of each motion image belongs to a feature of an image [8]. By dividing the points in the image into different subsets to form continuous curves or regions, the feature extraction results of human posture motion can be obtained. In practice, the results of human motion recognition can be obtained by comparing the extracted features with the information in the standard database. Traditional spatio-temporal weight attitude motion feature extraction algorithms include regular grid [9], image content analysis [10] and Mel-frequency cepstral coefficient (MFCC) [11]. Because of the rapid transformation speed of human posture movement and the diversity of behavior, the implementation of feature extraction algorithm is very difficult. In addition, light, angle and other objective factors will also affect the accuracy of the spatio-temporal weight motion feature extraction results.

In order to solve the above problems, based on the traditional motion feature extraction algorithm, the idea of cross-layer feature fusion based on deep network is introduced. The main contributions are as follows:

1) On the basis of traditional algorithm extraction steps, the improved AlexNet algorithm is introduced to collect Tai Chi motion images, and the operating process of the AlexNet algorithm is followed to detect motion feature objects, thus improving the integrity and accuracy of motion feature extraction results.
2) Processing the collected motion images, calculating the threshold values and features based on the data, and dividing the matching blocks to be used.
3) According to the weighted matching, motion feature fusion is completed, and the spatio-temporal weight Tai Chi motion feature extraction algorithm is realized, which indirectly improves the recognition accuracy of human spatio-temporal weight Tai Chi motion.
4) The weight of Tai Chi motion is calculated to obtain the fusion results of multi-scale motion feature extraction.

The structure of this paper is as follows. In section 2, we detailed introduce the proposed Tai Chi motion feature extraction. The experiments and analysis are conducted in section 3. There is a conclusion in section 4.

2. Proposed Tai Chi motion feature extraction

2.1. Extracting the spatio-temporal motion key frame

Before extracting the spatio-temporal motion key frame, the corresponding spatio-temporal sample needs to be selected first. Monitoring equipment is installed in Tai Chi sports venues, and the completed video files are the selected spatio-temporal samples. The horizontal spatio-temporal slices of the shot are extracted from the selected motion video samples [12], and the spatio-temporal slices of the video are clustered. After cluster processing, motion video files may appear time discontinuous but be clustered together. After spatio-temporal slicing and clustering processing, the motion samples are collected to form the corresponding sub-lens, so a key frame can be extracted as the object image with motion features according to the preset rules. The constraint conditions that the extracted space-time Tai Chi motion key frames need to satisfy are as follows:

$$Completeness(v) = \max_{k}^{s} \tag{1}$$

Where, s represents the range that can be selected by the motion video block v in the spatio-temporal sequence. k represents the number of objects containing video frames in the current video sequence. Constraint condition formula (1) can ensure the integrity of the key frame extraction content of spatio-temporal Tai Chi motion.

2.2. Tai Chi motion image preprocessing

The extracted spatio-temporal motion key frame is output in the form of static image. Through motion object detection, image enhancement, morphological processing, image normalization and other steps, it can achieve the Tai Chi motion image preprocessing results.

Moving object detection

The key problem of time-space weight gesture motion feature extraction is to detect high quality moving human object image. In this process, background subtraction, image extraction and other technical approaches are involved [13,14]. The specific moving object detection process is shown in figure 1.

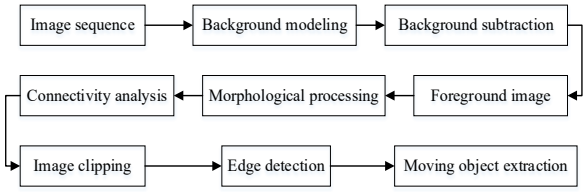

Figure 1. Detection process of moving object

According to the detection process shown in figure 1, firstly, the background in the image is modeled, and the specific formula is:

$$B(x,y) = Median_{j=1,2,\cdots,n}^{x}{}_{y} \tag{2}$$

Where $B(x,y)$ represents the pixel value of the video image at position (x,y). Then the foreground image of the object is extracted by the background subtraction technology. Assuming $f_i(x,y)$ is the extracted spatio-temporal motion image, then the calculation formula of corresponding foreground image is:

$$d_i(x,y) = | f_i(x,y) - B(x,y) | \tag{3}$$

Substituting the background model of equation (2) into equation (3), the foreground moving image of the video image can be obtained. Finally, the foreground image is cropped to get the moving object [15]. Moving object detection and processing can be used as a foundation to extract the motion features of spatio-temporal weight.

Motion image enhancement processing

Enhancement processing of moving images mainly includes the following two steps:

Step 1: Taking the foreground image in the collected image as the processing object for image denoising. This step is to avoid the noise in the image that can affect the sharpness of the moving image [16].

Step 2: Using the filter to enhance the moving image. Assuming that the noise reduction filter of the moving image is $h(x,y)$, the convolution operation of the noisy image can obtain the image after noise elimination, and the noise elimination process can be described as:

$$f(x,y) = d_i(x,y)h(x,y) \tag{4}$$

Choosing Gabor filter to sharpen and enhance image can make the filter obtain the best resolution in both spatial and frequency domain.

Image normalization

In the motion image, the location of human body area in the movement process is in a state of constantly changing, so we need to normalize the image into uniform size. And the movement area of the human body is defined in the same central position, so that the positions of the human body in all images are aligned, which is convenient for the subsequent extraction of the corresponding image features in the Tai Chi motion image. First, human edges in video sequence images are detected, denoted as $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$. Then, the calculation formula to determine the location of the center of the moving human body is:

$$\begin{cases} x = \dfrac{x_{\min} + x_{\max}}{2} \\ y = \dfrac{y_{\min} + y_{\max}}{2} \end{cases} \tag{5}$$

It cuts the image to a fixed size and ensures that the completed human movement area can be retained in the trimmed image.

2.3. Cross-layer feature fusion convolutional neural network

AlexNet network proposed by Krizhevsky et al. in 2012 triggered a boom in the field of deep neural network-based image processing. The network consists of five convolutional layers and three fully connection layers. It has successfully trained about 1.2 million images of 1000 categories with ReLu as activation function, multi-GPU parallel computing, local response normalization, overlapping pooling, and Dropout layer to coordinate network performance. The 17% top-5 error rate of ILSVRC2012 dataset is achieved with 60 million parameters. After AlexNet, various deep neural network structures have been put forward. GoogLeNet network uses global pooling and Inception moudle to cluster sparse matrices into dense sub-matrices to improve computing performance and optimize parameters. The network has 22 layers and adjusts parameters for gradient

problems caused by network depth with three Loss outputs. Another VggNet-16 also shows that network depth is the key to excellent performance of the algorithm.

In this paper, the improvement of AlexNet network is mainly studied on Tai Chi motion image recognition. The performance of GoogLeNet and VggNet-16 deep network in Tai Chi image feature classification is compared to carry out correlation analysis.

New network design

This paper proposes a new network based on the original AlexNet network. It consists of an input layer, four convolutional layers (followed by pooling layer), one Inception module, a cross-layer connection structure, two full connection layers (followed by Softmax loss function), and an output layer. It uses BN to replace Local Response Normalization (LRN), a second pooling layer is cross-linked to the full connection layer, which fuses with deep features extracted from the backbone network, and eventually plugs into the classifier.

The new AlexNet network structure is shown in figure 2. Table 1 lists the specific parameters of the new AlexNet network, including the Type, convolution kernel Size, Stride and Output Size of each network layer.

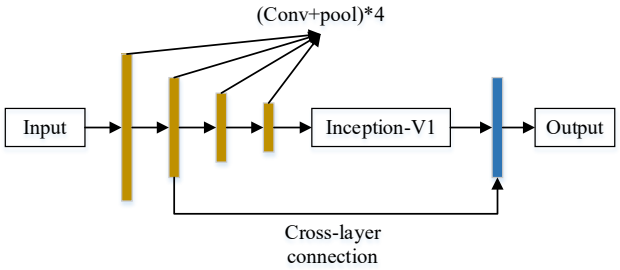

Figure 2. Proposed network structure

Table 1. Parameters in proposed network

layer	type	kernel	stride	Output size
x	Input	--	--	128×128×3
h1	convolution	5×5	2	64×64×96
h2	maxpooling	3×3	2	32×32×96
h3	convolution	5×5	1	32×32×128
h4	maxpooling	3×3	2	16×16×128
h5	convolution	3×3	1	16×16×256
h6	maxpooling	3×3	2	8×8×256
h7	convolution	3×3	1	8×8×256
h8	maxpooling	3×3	2	4×4×256
h9	Inception	--	--	4×4×256
h10	maxpooling	3×3	1	4×4×256
h11	Fc	--	--	2048
o	output	--	--	5

After the convolutional feature extraction, AlexNet network is normalized by LRN, and lateral suppression is performed on the neurons adjacent to the activated neurons to achieve local suppression and improve the model generalization ability. However, BN can effectively accelerate model convergence, prevent single samples from being frequently selected during batch training, and prevent "gradient dispersion", while abandoning the dropout layer and L2 regular term parameters [17].

$$\hat{x}^l = \frac{x^l - 1/m\sum_{l=1}^{m} x^l}{\sqrt{1/m\sum_{l=1}^{m}(x^l - 1/m\sum_{l=1}^{m} x^l)^2 + \xi}}, l = 1,2,\cdots,m \quad (6)$$

Where x^l represents the $l-th$ input sample. $\hat{x}^l$ represents the output result. ξ is the offset term, and m is the batch number of samples.

In the new AlexNet network, the Inception-V1 module in GoogLeNet network is introduced to extract deep features of Tai Chi images before full connection layer. The structure of the Inception-V1 module is shown in figure 3, and it is connected in parallel with four convolution kernel of different sizes. The first one conducts 1×1 convolution for the upper input. The second one convolves the last layer, and then it is connected with the 3×3 convolution.

The third branch convolves the upper layer with 1×1 size, and then it is connected with the 5×3 convolution. Continuous feature transformation broadens the dimension of feature expression. The fourth is the 3×3 maximum pooling to realize the compression of perceptual information. Finally, the four converged filtering layers are connected. The higher layers of Inception module has the greater efficiency [18].

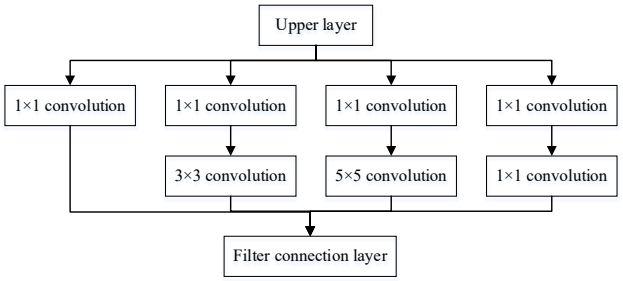

Figure 3. Inception-V1 module

The traditional convolutional neural network extracts features from shallow to deep, processes features through classifiers, and outputs probabilities under different conditions. With the deepening of network depth, this process can not effectively fuse the low-level and high-level features to form the feature classifier. In this paper, the idea of cross-layer connection proposed in DeepId [19] is introduced to connect the second pooling layer to the full connection layer for feature fusion. In general, the network firstly extracts layer features from 128×128 input images in $h_1 \sim h_4$. Two rounds of convolution operation and local neighborhood pooling alternate action are repeated $h_5 \sim h_8$. It inputs the Inception module h_9 for multi-scale and deeper feature extraction. The output features by pooling layers h_{10} and h_4 are fused and classified at full connection layer h_{11} (the node number of this layer is the sum of the nodes of h_4 and h_{10}). The output layer contains two nodes with two categories. Finally, it outputs the float probability values of the original input on each category tag. Among them, maximum pooling is selected for pooling operation, which is beneficial to preserve image texture features.

Network learning process

Assuming data set $D = \{(x^l, y^l)\}_{l=1}^{N}$, $y^l = c = \{0,1,2,3,4\}$. Where N is the total number of samples. x^l is *l-th* input sample. y^l is the category of the *l-th* sample, so the forward calculation process of the *l-th* sample x^l in the improved AlexNet network is as follows:

$$h_{1,j}^l = f(u_{1,j}^l) = f(x^l \otimes W^{1,j} + b^{1,j}) \quad (7)$$

$$h_{i,j}^{\prime l} = f(u_{i,j}^l) = f(\sum_{1}^{c_{2k}} h_{i-1,j}^l \otimes W^{i,j} + b^{i,j}), i = \{3,5,7\} \quad (8)$$

$$h_{i,j}^l = down_{\lambda,\tau}(\max(h_{i-1,j}^l)), i = \{2,4,6,8,10\} \quad (9)$$

$$h_{9,j}^l = filterCncat(h_{9\_1,j}^l \mid h_{9\_2,j}^l \mid h_{9\_3,j}^l \mid h_{9\_4,j}^l) \quad (10)$$

$$h_{11}^l = f(u_{11}^l) = f(\sum_{1}^{256} h_{10,j}^j W^{11\_10} + \sum_{1}^{128} h_{4,j}^j W^{11\_4} + b^{11}) \quad (11)$$

$$o_{12}^l = f(u_{12}^l) = f(\sum_{1}^{2} h_{11}^j W^{12} + b^{12}) \quad (12)$$

$$1 \le j \le j(i) \quad (13)$$

Where j represents the positive integer that is not greater than the output third dimension number j(i) in the i-th hidden layer, that is, $1 \le j \le j(i)$. $f(\cdot)$ is the activation function PRelu. $down_{\lambda,\tau}$ refers to the sampling process with the window size $\lambda \times \tau$ by the maximum pooling method. $\otimes$ is the inner convolution operator. $W^{i,j}$ and $b^{i,j}$ represent the j-th convolution kernel and bias in the i-th hidden layer (i=1,3,5,7). $h_{i,j}^l$ represents the j-th pooling of the i-th hidden layer (i=2,4,6,8,10). $h_{i,j}^{\prime l}$ represents the j-th convolution of the i-th hidden layer (i=1,3,5,7). The four branches of the Inception module converge together in the same dimension. $h_{9,j}^l$ is the j-th output convolution of the ninth hidden layer. h_{11} is the full connection layer output. W^{11_4} is the connection weight of the fourth hidden layer and the 11th hidden layer. W^{11_10} and b^{11} are the connection weight and bias of the 10th hidden layer and the 11th hidden layer respectively. o_{12}^l is the final output value of the output layer. The output layer uses softmax classifier. Suppose that the input of the output layer is x_o, and the classification likelihood probability is:

$$Y(c) = P(y = c \mid x_o, w) = \frac{e^{w_j \times x_0}}{\sum_{l=1}^{c} e^{w_l \times x_0}} \quad (14)$$

Loss function J(w) is:

$$J(w) = -\frac{1}{N}\sum_{l=1}^{N}[\sum_{c=1}^{c} 1\{y^l = c\} \cdot P(y = c \mid x_0, w)] \quad (15)$$

Where $P(c)$ calculates the probability when the input data classification category is c. w is the weight parameter.

Let the actual output of network for the input data x^l be $O = \{o^l\}|_1^N$. According to the error transfer process of back propagation, the parameters are updated by gradient descent algorithm. In the process of transmission error feedback, the fourth hidden layer h_4 and the 10th hidden layer h_{10} are connected to the 11th hidden layer h_{11}. Therefore, the feedback error of h_4 layer is decomposed into two paths. According to the calculation method of feedback transmission error of each layer in traditional CNN, the transmission error of each layer in

the backpropagation of the new AlexNet network is as follows:

$$\delta_{12}^{l} = (o^{l} - y^{l}) \circ f'(h_{12}^{l}) \quad (16)$$

$$\delta_{11}^{l} = [(W^{12})^{T} \delta_{12}^{l}] \circ f'(h_{12}^{l}) \quad (17)$$

$$\delta_{10,j}^{l} = [(W^{11\_10})^{T} \delta_{11}^{l}] \circ down'_{\lambda,\tau} f'(h_{q,j}^{l}) \quad (18)$$

$$\delta_{4\_all,j}^{l} = (\delta_{4,j}^{l} + \delta_{4\_2}^{j}) \quad (19)$$

$$\delta_{4\_2,j}^{l} = [(W^{11\_4})^{T} \delta_{11}^{l}] \circ down'_{\lambda,\tau} f'(h_{3,j}) \quad (20)$$

$$\delta_{i,j}^{l} = [\delta_{i+1,j}^{l} \oplus W^{i+1,j}] \circ down'_{\lambda,\tau} f'(h_{i-1,j}^{l}), i = 2,4,6,8 \quad (21)$$

$$\delta_{i,j}'^{l} = \sigma'(h_{i,j}^{l}) \circ up(\delta_{i+1,j}^{l}), i = 1,3,5,7,9 \quad (22)$$

$$1 \le j \le j(i) \quad (23)$$

Where δ_{12}^{l} and δ_{11}^{l} are the feedback errors of the output layer and the full connection layer respectively. " $\circ$ " represents the Hadamard product. up(·) is the up-sampling process. $\oplus$ represents the outer convolution operation. W^{12} is the weight between the output layer and the full connection layer. W^{11_4} and W^{11_10} are the weight parameters between the fourth hidden layer and the 10th hidden layer and the full connection layer respectively. $W^{i+1,j}(i = 2,4,6,8)$ is the weight parameter from the i-th hidden layer to the (i+1)-th hidden layer. $\sigma_{4\_all}^{l}$ is the total transfer feedback error received by the fourth hidden layer. $\sigma_{4\_2,j}^{l}$ is the feedback error passed from the 11th hidden layer to the fourth layer. $\sigma_{i,j}^{l}$ is the feedback transmission error of the j-th pooling layer of the i-th (i=2,4,6,8) hidden layer. Similarly, $\sigma_{i,j}'^{l}$ is the feedback transfer error of the j-th convolution of the i-th (i=1,3,5,9) hidden layer. Through the feedback transmission error of each layer, the weights and partial derivatives of the bias of each hidden layer are calculated:

$$\frac{\partial J}{\partial W^{12}} = \sum_{l=1}^{N} \delta_{12}^{l} (h_{11}^{l})^{T}, \frac{\partial J}{\partial b^{12}} = \sum_{l=1}^{N} \delta_{12}^{l} \quad (24)$$

$$\frac{\partial J}{\partial W^{11\_10}} = \sum_{l=1}^{N} \delta_{11}^{l} (down_{\lambda,\tau}(h_{q,j}^{l}))^{T}, \frac{\partial J}{\partial b^{11}} = \sum_{l=1}^{N} \delta_{11}^{l} \quad (25)$$

$$\frac{\partial J}{\partial W_{i,j}^{l}} = \sum_{l=1}^{N} (h_{i-1,j}^{l} \otimes \delta_{i,j}^{l}), \frac{\partial J}{\partial b^{i,j}} = \sum_{l=1}^{N} \delta_{i}^{l} (i = 3,5,7) \quad (26)$$

$$\frac{\partial J}{\partial W_{1,j}^{l}} = \sum_{l=1}^{N} (x^{l} \otimes \delta_{i,j}^{l}), \frac{\partial J}{\partial b^{1,j}} = \sum_{l=1}^{N} \delta_{1,j}^{l} \quad (27)$$

$$\frac{\partial J}{\partial W^{11\_4}} = \sum_{l=1}^{N} \delta_{11}^{l} (h_{1,j}^{l})^{T} \quad (28)$$

$$1 \le j \le j(i) \quad (29)$$

Where $\frac{\partial J}{\partial W}$ and $\frac{\partial J}{\partial b}$ represent the partial derivative of the loss function with respect to the weight and bias of the corresponding layer respectively.

The new AlexNet network in this paper uses the gradient descent algorithm [20] to update the weight and bias. It is known that the training set D, momentum is M and learning rate is lr. Te specific algorithm process is as shown in figure 4.

2.4. Adaptive interpolation of spatio-temporal weights

Adaptive interpolation of spatio-temporal weight can effectively solve the problem of interpolation errors caused by motion estimation errors and inaccurate edge detection. It has the advantage of automatic fusion of edge adaptive field interpolation [21,22]. Firstly, the absolute difference of the pixels before and after the moving image element should be calculated by using the spatio-temporal weight, and the corresponding weight coefficient should be calculated. Then the weighted average of adjacent pixels is carried out to obtain the estimated value of the points to be interpolated. The final expression of adaptive pixel P of spatio-temporal weight is:

$$P = Median(X(i,j)-1,t),(X(i,j)+1,t,P') \quad (30)$$

Where P' is the pixel value after interpolation. $(X(i,j)-1,t)$ and $(X(i,j)+1,t)$ are the two pixels of the current field respectively.

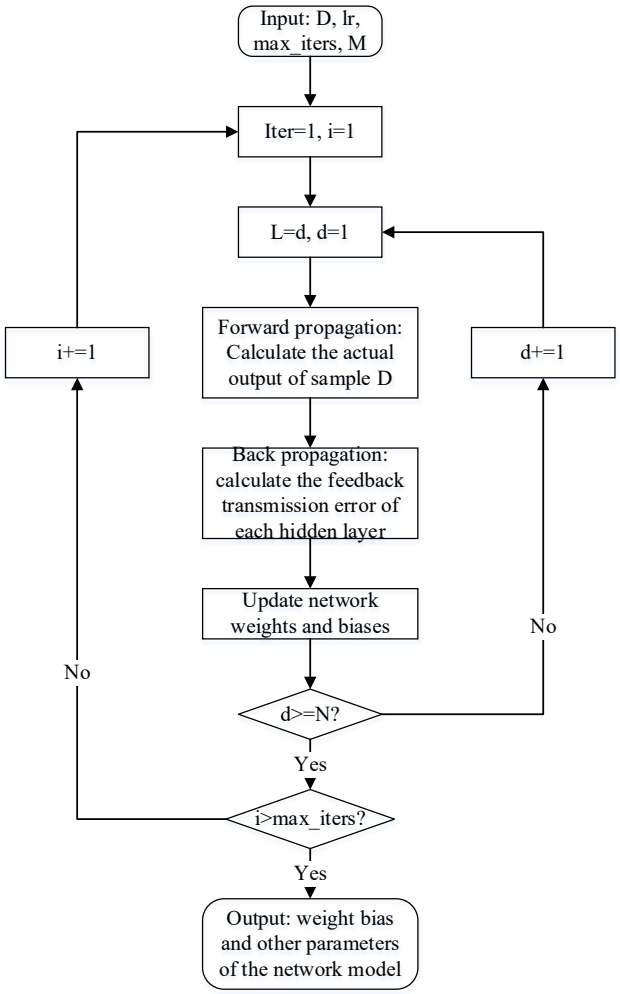

Figure 4. Flow chart of new AlexNet network parameter update

2.5. Motion feature extraction fusion

Posture edge feature

The structure of human body is fixed in a short time, so the edge contour features of posture motion can represent the motion features of human body. According to the sequence of video image collection and the training results of CNN algorithm, contour features of Tai Chi motion can be expressed in the form of coordinates, then the extracted feature vector of edge contour with spatio-temporal weight is:

$$\phi = [\phi_1, \phi_2, \cdots, \phi_z] \tag{31}$$

Where z is the change of invariant matrix of human body in a period.

Temporal and spatial features of motion

The temporal and spatial features of Tai Chi posture movement include skeletal features of human body and joint angles of limbs [23]. Under the condition of constant topological structure, the outer pixels of the gait image are stripped layer by layer by iteration. The skeleton with single pixel width is obtained, which is the feature result of extracted motion limb joint. The extraction results of temporal and spatial features of the moving skeleton are shown in figure 5. The joint angle of human body is expressed in the form of coordinate, and the rotation angle of human limb joint at different time is calculated respectively. It arranges the calculation results of rotation angles in chronological order. The temporal and spatial variation of human movement is analyzed under the corresponding skeleton model. It can be seen from figure 5 that in the actual movement process, the displacement of human limb joint is small, so the spatio-temporal features of motion can be directly represented by the joint angle features [24].

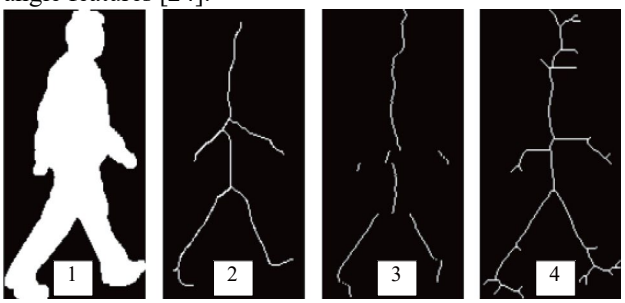

Figure 5. Spatio-temporal feature extraction results of motion skeleton

Motion features fusion

The motion feature fusion is realized by using the calculated weight of spatio-temporal. The specific fusion process is shown in figure 6.

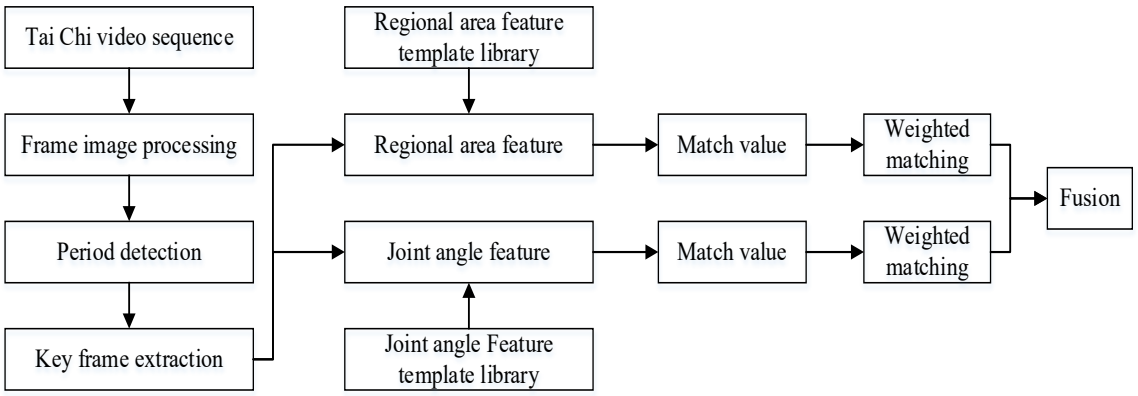

Figure 6. Process of motion feature fusion

Figure 6 shows that the reliability of different feature matching quantized values is different in the process of feature fusion. Therefore, according to the distribution of spatial and temporal weights, the fusion of motion features is realized from feature layer, data layer and decision layer. The data of each video sequence is analyzed step by step, and the data threshold is obtained according to the key frame to obtain the extraction results of the area and joint angle. The motion feature fusion is completed according to the weighted matching.

Therefore, we can summarize our proposed algorithm as shown in figure 7.

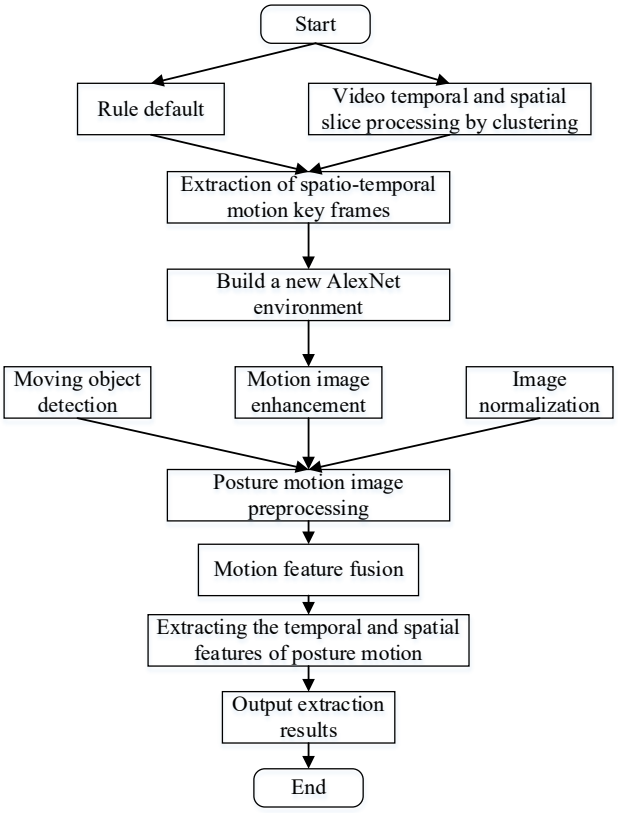

Figure 7. Proposed method in this paper.

3. Experimental analysis and results

The experimental platform of this paper is Ubuntu16.04 operating system, Deep learning Caffe framework, Python2.7 interface language, GPU GTX2080Ti, processor Intel Core I7-7820x CPU@3. 60GHz× 16, and memory 64G. Setting the initial learning rate as 0.001 and using "step" attenuation. Multi-classification cross entropy loss function is used. Comparison results with SVM and traditional AlexNet is shown in figure 8.

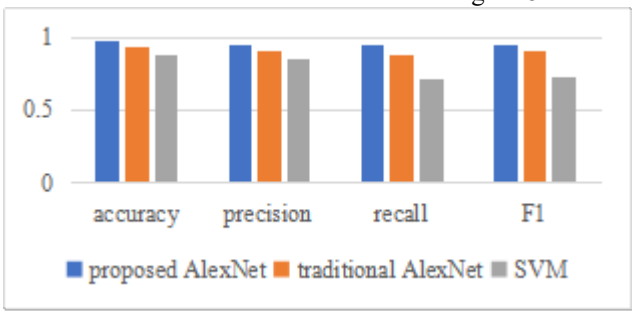

Figure 8. Classification results

By comparing the results, it can be seen that the new AlexNet network constructed in this study has significantly improved the average classification accuracy, average accuracy, average recall rate and average comprehensive index F1 value compared with SVM [25] and traditional AlexNet network [26]. In particular, the difference in recall rate indicates that the new AlexNet network has high classification accuracy, outstanding classification effect for Tai Chi images, and strong model expression power.

3.1. Cross-layer connection analysis

It is necessary to discuss the influence of the difference of cross-connection modes on network performance. The new AlexNet network by cross-connection structure is introduced in this paper to make the deep features and shallow features fusion. Therefore, it is determined that the cross-connection terminal is the unchanged full

connection layer. Now, the reliability analysis is carried out for the front segment of the network and the cross-connection initial end is changed to h2 and h6 respectively. The training process is performed under the same conditions and compared with the test results of new AlexNet network, as shown in Table 2:

Table 2. Comparison with different connection ways/%

Cross-connection	Average A	Average P	Average R	Average F1
No	92.96	89.16	82.24	82.44
h2-h11	94.62	87.34	86.38	86.44
h8-h11	95.52	90.12	88.64	88.89
Proposed	98.24	95.64	95.44	95.46

According to the test results, h4 hidden layer has certain advantages as the initial end of cross-connection. Compared with h2 and h6 as the initial end of cross-connection, the average classification accuracy is improved by 3.62% and 2.72% respectively. Other evaluation indicators also have obvious advantages. The feature graph output by h2 layer retains obvious edge information and has high overlap with the original input image information. The output features of h3 layer are more abstract than the previous layer, but it still retains some specific contour edges. h4 layer output features are already extremely abstract. Therefore, it can be concluded that, compared with new AlexNet without cross-connection structure, the classification accuracy of cross-connection network is significantly improved. However, when the shallow features of h2 layer output are fused with the depth features across the connecting end, the specific features are overemphasized and the contribution to the classification results is not high enough. The shallow feature output of h6 layer is too abstract, which has the disadvantage of parameter redundancy in feature fusion, and has less effect on the improvement of classification accuracy than h6 layer. Therefore, this paper chooses h4 layer as the initial end of cross-connection, which is the optimal solution.

Through the comparison of experimental results and visualization of training process, it is verified that the new AlexNet has higher classification accuracy than conventional SVM and AlexNet method in image classification, and there is no need to manually extract image features. The sensitivity analysis and visualization of the intermediate process verify the reliability of the cross-connection structure.

In order to verify the effectiveness of AlexNet network in the detection and classification of Tai Chi motion features, this paper conducts experiments on AlexNet network, GoogLeNet network [27] with Inception module, and VGGNet-16 network [28] on the same motion data set. In the process of model training, the method of control variable is adopted, the initial learning rate is set to 0.001, and the attenuation mode of "step" is used. The same loss function, optimization function, maximum iteration (10000 times) and parameter update method (Momentum+SGD) are used in different networks. The experimental results are shown in table 3.

Table 3. Comparison of different networks

network	A	P	R	F1	Training time	Testing speed/fps	Model size
AlexNet	90.56%	78.49%	76.11%	0.7624	32min	34	228.5MB
GoogLeNet	93.50%	85.37%	83.58%	0.8350	24min	44	24.1MB
VggNet-16	95.52%	90.78%	88.64%	0.8875	51min	12	538.2MB
Proposed	98.24%	95.64%	95.44%	0.9546	17min	50	12.8MB

According to the test results of Tai Chi motion with different methods listed in table 3, new AlexNet not only has a great advantage in average recognition accuracy, but also has the optimal average recognition accuracy, recall rate and comprehensive evaluation index F1 value. In the classical convolutional network method, VGGNet-16 highly inherits AlexNet framework and expands in depth. The classification effect of deeper GoogLeNet shows that the increase in network depth does not maintain performance growth, but the sparse connection and Inception module in GoogLeNet can reduce parameter redundancy and simplify the model. In addition, the CNN method is superior to the traditional SVM method in each evaluation index on the solder joint data set,

demonstrating the effectiveness of CNN in independent feature extraction.

3.2. Comparison with other methods

In this subsection, we select matching degree of feature extraction (MD), weighted matching elasticity (WME), multi-scale motion features fusion degree (MFD).

After AlexNet environment is formed, the algorithm in this paper undergoes object detection, moving image enhancement processing and normalization processing. Then it obtains the extraction environment that best matches the motion features. Therefore, the MD is set as one of the experimental indicators, and its calculation formula is:

$$MD = \varphi[x^3(t)] \tag{32}$$

Where x(t) represents the normalized processing result.

In the process of motion feature fusion, it is necessary to calculate key frames to obtain data threshold, extract area and joint angle. Then it achieves WME by dividing matching blocks. The WME reflects the elasticity of weighted matching, that is, affects the recognition accuracy.

In order to realize the fusion operation of extracting motion features from the feature layer, data layer and decision layer, the multi-scale motion feature fusion degree (MFD) of the three layers is compared. The calculation formula of MFD is:

$$MFD = (MP / L) \times 100\% \tag{33}$$

Where L is the general feature to be fused. M is the scale optimization degree.

In here, we select TSF [29], CORR-OMP [30], SEMG [31] to make comparison. The experimental results of MD are shown in table 4. By analyzing the data in table 3, it is found that the number of effective features with the traditional motion feature extraction algorithm accounts for 96.6%, 93.2% and 94.3%, while the feature extraction number of the proposed algorithm accounts for 98.6%, the number of effective features accounts for 98.1%. In contrast, the number of effective features extracted by the new method is increased by 1.5%, 4.9% and 3.8%.

Table 4. Comparison of motion feature extraction

Group	TSF			CORR-OMP			SEMG			Proposed		
	EN	MN	MMN	EN	MN	MMN	EN	MN	MMN	EN	MN	MMN
1	9035	8884	55	8887	8834	66	8885	8822	64	8885	8865	21
2	8534	8217	102	8163	8113	68	8161	8112	50	8152	8111	42
3	9816	9665	48	9634	9531	104	9633	9529	105	9629	9587	43
4	9440	9142	96	9174	9023	76	9173	9022	152	9172	9121	52
5	7796	7574	83	7554	7363	72	7553	7362	192	7415	7362	54
6	8214	8018	80	8013	7973	39	7012	6971	41	8011	7991	21

Note: EN: number of extraction. MN: number of matching. MMN: number of mismatching.

The WME results are shown in figure 9. It can be seen from figure 9 that under the limit of 25 iterations, the elastic curve of the proposed method fluctuates greatly, but it is superior to other methods in matching block distribution. It shows that the proposed method can complete motion feature fusion based on weighted matching method, and realize the feature extraction of spatio-temporal weight posture motion, which provides a basis for the recognition of Tai Chi spatio-temporal weight posture motion.

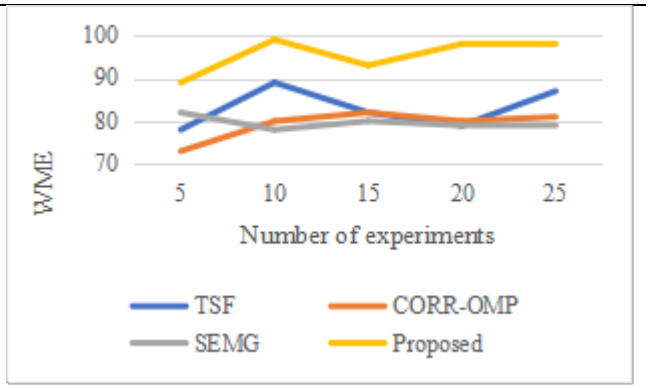

Figure 9. Test results of WME by different methods

The test results of MFD are shown in figure 10. It can be seen from figure 10 that the fusion results of multi-scale motion features with the proposed method are stronger than those of the TSF, CORR-OMP, SEMG methods at

the feature layer, data layer and decision layer, respectively. Therefore, the proposed method can not only achieve higher efficiency than the traditional methods, but also obtain more accurate extraction results of spatio-temporal weight attitude motion features.

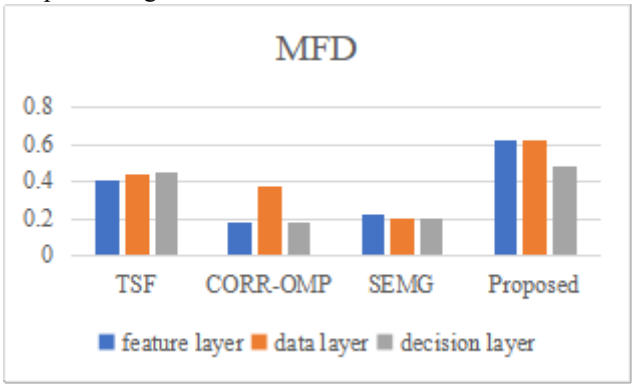

Figure 10. Test results of MFD by different methods

4. Conclusion

This paper combines cross-connection architecture and Inception to propose a new AlexNet network to realize the optimal weight design for the traditional algorithm. At the same time, BN is used for data normalization, and the gradient descent algorithm is used for optimization to accelerate the convergence of the network and avoid the gradient problem. By analyzing the time and space weights of moving objects, the problem of low extraction efficiency in traditional algorithms is solved. The motion feature fusion is completed according to weighted matching, which solves the problems of poor extraction efficiency and low recognition accuracy of traditional feature extraction algorithm in Tai Chi motion, and provides reference for related research in this field. However, the sample collection environment selected in the experiment is relatively simple, and the sample contains only one motion object. The actual identification work environment is complicated and there are many interference factors. Therefore, precise object positioning will be an important research direction in the future. In the future, we will apply this project to practical engineering applications.

Acknowledgements.

The authors appreciate the anonymous reviewers for their meaningful comments.

References

[1] Barffour, M, A., Guy-Marino, H., Ryan, W, K., et al. (2020) Effects of therapeutic zinc supplementation for diarrhea and two preventive zinc supplementation regimens on the incidence and duration of diarrhea and acute respiratory tract infections in rural Laotian children: A randomized controlled trial. Journal of global health 10(1): 010424.

[2] Sudrajat, A., Yetti, K., Waluyo, A. (2021) The effect of the range of motion exercises combined with tai chi intradialysis on the adequacy of hemodialysis in patients at lebak district hospital. Enfermería Clínica 31: S113-S116.

[3] Yin, S., Li, H., Teng, L. (2018) Semantics automatic annotation in medical image based on deep learning. Basic & Clinical Pharmacology & Toxicology.

[4] Yang, F, C., Desai, A, B., Esfahani, P, et al. (2021) Effectiveness of Tai Chi for Health Promotion of Older Adults: A Scoping Review of Meta-Analyses. American Journal of Lifestyle Medicine 2021:155982762110012.

[5] Karim, S., Zhang, Y., Yin, S. et al. (2019) Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery. Multimedia Tools and Applications 78: 32565-32583.

[6] Teng, L., Li, H., Yin, S., Karim, S., Sun, Y. (2020) An active contour model based on hybrid energy and fisher criterion for image segmentation. International Journal of Image and Data Fusion 11(1): 97-112.

[7] Kushagra, N., Kirti, G., Deepshi, S., et al. (2021) An Improved Approach for Stress Detection Using Physiological Signals. EAI Endorsed Transactions on Scalable Information Systems. http://dx.doi.org/10.4108/eai.14-5-2021.169919

[8] Hou, P., Zhang, Y. (2020) Dynamic Image Sampling and Swimming Motion Image Recognition in Immersive Virtual Reality. Microprocessors and Microsystems 82(3):103760.

[9] Horn, D,. Dror, G., and Quenet, B. (2004) Dynamic proximity of spatio-temporal sequences. IEEE Transactions on Neural Networks 15(5): 1002-1008.

[10] Tu, Z., Li, H., Zhang, D., et al. (2019) Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition. IEEE Transactions on Image Processing 28(6): 2799-2812.

[11] Gao, T., Li, Hang., and Yin, S. (2021) Adaptive Convolutional Neural Network-based Information Fusion for Facial Expression Recognition. International Journal of Electronics and Information Engineering. Vol. 13, No. 1, pp. 17-23, 2021.

[12] Lu, X., D, Xu., Mao, X., et al. (2017) Feature Extraction and Fusion Using Deep Convolutional Neural Networks

for Face Detection[J]. Mathematical Problems in Engineering 2017:1-9.

[13] Sudimanto. (2020) Rekayasa Perangkat Lunak Penghitung Jumlah Tempat Parkir Tersedia Menggunakan Kamera dengan Metode Background Subtraction. Media Informatika 19(1):1-5.

[14] Li, L., Qin, S., Lu, Z., et al. (2021) Real-time one-shot learning gesture recognition based on lightweight 3D Inception-ResNet with separable convolutions. Pattern Analysis and Applications, 2021:1-20.

[15] Fan, S., Jia, Y., Liu, J. (2019) Feature selection of human activity recognition based on tri-axial accelerometer. Journal of Applied Sciences 37(3): 427-436. (in Chinese)

[16] Qiu, L., Sang, D., Fengy, H., et al. (2018) Analysis of particle motion characteristics in a funnel-shape moving bed. Journal of Engineering Thermophysics 39(12): 2708-2713.

[17] Yin, S., Li, H. & Teng, L. (2020) Airport Detection Based on Improved Faster RCNN in Large Scale Remote Sensing Images. Sensing and Imaging 21. https://doi.org/10.1007/s11220-020-00314-2

[18] Liu, Z., Abdukeyim, N., and Yan, C. (2019) Image classification of hepatic echinococcosis based on convolutional neural network," 2019 6th International Conference on Systems and Informatics (ICSAI), 1280-1284, doi: 10.1109/ICSAI48974.2019.9010184.

[19] W. Ouyang et al. (2017) DeepID-Net: Object Detection with Deformable Part Based Convolutional Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(7): 1320-1334.

[20] Wang, X., Yin, S., Sun, K. (2020) GKFC-CNN: Modified Gaussian Kernel Fuzzy C-means and Convolutional Neural Network for Apple Segmentation and Recognition. Journal of Applied Science and Engineering 23(3): 555-561.

[21] Wu, J., Song, Z., Jeon, G. (2017) GPU-Parallel Implementation of the Edge-Directed Adaptive Intra-Field Deinterlacing Method. Journal of Display Technology 10(9):746-753.

[22] Zhang, T., Chen, Y., Lei, Z. (2018) Hybrid De-Interlacing Algorithm with Adaptive Interpolation Based on Temporal-Spatial Characteristics. Journal of Tianjin University Science and Technology 51(1):73-78.

[23] Fabrizio, D., Nicolas P.Smith, et al. (2015) Non-invasive pressure difference estimation from PC-MRI using the work-energy equation. Medical Image Analysis 26(1): 159-172.

[24] Wang, J., Wang, L., Miran, S., Xi, X., and Xue, A. (2019) Surface Electromyography Based Estimation of Knee Joint Angle by Using Correlation Dimension of Wavelet Coefficient. IEEE Access 7: 60522-60531.

[25] Yin, S., L, J., Teng, Lin. (2018) A new krill herd algorithm based on SVM method for road feature extraction[J]. Journal of Information Hiding and Multimedia Signal Processing 9(4): 997-1005.

[26] Tong, J., Li, H., and Yin, S. (2020) Research on Face Recognition Method Based on Deep Neural Network. International Journal of Electronics and Information Engineering 12(4): 182-188.

[27] Kawakura, S., Shibasaki, R. (2020) Suggestions of a Deep Learning Based Automatic Text Annotation System for Agricultural Sites Using GoogLeNet Inception and MS-COCO. Journal of Image and Graphics 8(4):120-125.

[28] Ak, A., Sks, A., Ss, B., et al. (2020) Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer. Information Sciences 508:405-421.

[29] Y, Gu., Liu, H., Wang, T. , Li, S. , & Gao, G. (2020) Deep feature extraction and motion representation for satellite video scene classification. Science China. Information Sciences 63(4), 140307.

[30] Cai, W., Xia, S., Sun, R., et al. (2021) A Micro-Motion Feature Extraction Method Based on CORR-OMP," 2021 IEEE 4th International Conference on Electronics Technology (ICET), 2021, pp. 411-416, doi: 10.1109/ICET51757.2021.9451048.

[31] Shi, X., Qin, P., Zhu, J., et al. (2020) Feature Extraction and Classification of Lower Limb Motion Based on sEMG Signals. IEEE Access 8: 132882-132892.