EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Residual network based on convolution attention model and feature fusion for dance motion recognition

Dianhuai Shen[1], Xueying Jiang[2], and Lin Teng[3,*]

[1]College of Music and Dance, Huaqiao University, Xiamen 361000 Fujian, China
[2]School of Public Policy and Management, Tsinghua University, Beijing 100000 China.
[3]Software College, Shenyang Normal University, Shenyang 110034 China
Email:shenmingyudrive@163.com;jiangxue18@mails.tsinghua.edu.cn;910675024@qq.com

Abstract

Traditional posture recognition methods have the problems of low accuracy. Therefore, we propose a residual network based on convolution attention model and future fusion for dance motion recognition. Firstly, the fusion features of the relative position, angle and limb length ratio of human body are selected by combining the information of bone key points. The shallow features of the original dance image are extracted and compressed by convolution layer and pooling layer. Then it uses the stacked residual to learn deep features, the gradient dispersion and network degradation can be alleviated. The convolutional attention module is used to assign weighted values to the deep degradation features of the dance. Finally, dance motion detection in complex dance scenes can be realized. The dance movement recognition method proposed in this paper can accurately identify dance motion. Compared with other recognition algorithms, this new algorithm has the best recognition accuracy and faster recognition efficiency.

Keywords: dance motion recognition, residual network, convolution attention model, future fusion.

Received on 10 September 2021, accepted on 10 December 2021, published on 16 December 2021

doi: 10.4108/eai.16-12-2021.172434

*Corresponding author. Email: 910675024@qq.com

1. Introduction

Human posture recognition is the main way to help learn and understand human movements and behaviors. It can realize the analysis of human movements and the preservation of action information through human posture recognition [1-4]. For example, in the teaching process of dance movements, students or coaches can standardize the movements according to the recognition results of human body posture. For minority dances, human posture recognition can also obtain and preserve key information of dance movements, reducing the risk of dance disappearing in the process of inheritance [5]. At present, the main process of human posture recognition includes three steps: data acquisition and preprocessing, human feature extraction and construction, and movement recognition. Among them, the extraction structure of human body features is the key to human posture recognition. However, the current feature extraction and construction methods, including low-level feature tracking methods and semantically based methods, usually have low accuracy. For example, Hu et al [6] could effectively detect and predict abnormal events in the video by analyzing composite motion features based on the underlying feature tracking method. Based on semantic features, Kolivand et al. [7] expanded the fault-

tolerant features of under-standard sign language recognition under the condition of limited samples, which improved the accuracy of human gesture recognition to a certain extent, but there was still a problem of low recognition accuracy.

In recent years, data-driven methods based on deep learning algorithms have achieved good results in speech recognition [8-10], image processing, motion recognition and other fields due to their powerful learning and fitting abilities. Among all kinds of deep learning algorithms, convolutional neural network (CNN) has strong feature extraction ability and a certain noise reduction function. Therefore, CNN is widely used in the field of moving image processing. However, the image recognition method based on CNN has the following deficiencies. First of all, deep features extracted from images by CNN network using multiple convolutional layers contain two dimensions, channel and space, but the degradation information content of different channel features or different spatial features is different. Therefore, the contribution to image recognition is not the same, but the previous methods tend to ignore this point, which will bring adverse effects on image recognition. Secondly, the traditional CNN network needs enough depth to have good learning ability, but when the network is stacked to a certain depth, there will be gradient dispersion and network degradation, which is not conducive to the identification task. To solve the above problems, a residual network based on convolution attention model and future fusion is proposed in this paper for dance motion recognition.

2. Related works

2.1. Residual network

In order to solve the gradient dispersion and network degradation caused by deep CNN network, Xie et al. [11] introduced Residual neural network (ResNet) into CNN. ResNet adds a direct connecting edge between multiple nonlinear convolutional layers, which can alleviate network degradation and avoid gradient disappearance.

Moreover, the feature learning of each layer network is irreversible due to the existence of nonlinear activation function. Therefore, some degenerate information will be lost more or less when spreading in the network. When the network deepens, excessive information loss may even affect the accuracy of the prediction model. However, the cross-layer jumping connection of residual module directly transfers the input information to the output, which effectively alleviates the propagation loss of the degraded information in the network and improves the utilization efficiency of the degraded information in the network.

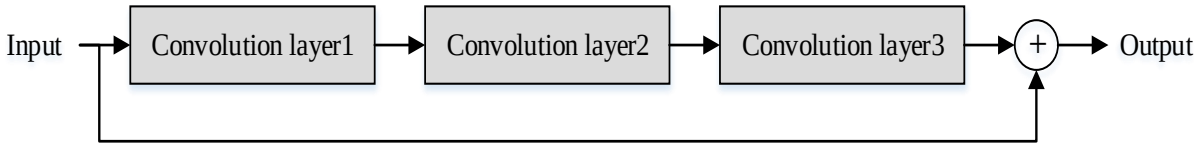

Figure 1. Residual module

Figure 1 shows the structure of the residual module used in this paper, where the number of convolutional layers 1, 2 and 3 are K/4, K/4 and K respectively. The kernel size is 1×1, S×1 and 1×1 respectively. The structure of shrinkage and expansion and the embedded 1×1 convolution kernel will greatly reduce the network parameters and improve the prediction speed of the model.

2.2. Convolution attention

The attention mechanism in deep learning borrows from the idea of human visual attention, focusing attention on the more important information to the target task and suppressing the interference of useless information, so as to improve the efficiency of neural network. At present, several neural network models have been combined with attention mechanism and achieved good results in their respective tasks. In view of this, the convolutional block attention module (CBAM) [12], which is adapted to the recognition task, is also introduced into CNN for feature re-calibration, in order to enhance network features that contribute more to the recognition task.

CBAM is mainly composed of channel attention and spatial attention, which reinforce important information and suppress invalid information in the channel dimension and spatial dimension of deep features respectively.

Channel attention

The importance of each output channel of the convolution layer to the recognition task is not consistent. Therefore, it is necessary to adopt channel attention mechanisms to assess the importance of different channels and assign

greater weight to more important channel features to enhance the impact of these channels. The principle of channel attention is shown in figure 2.

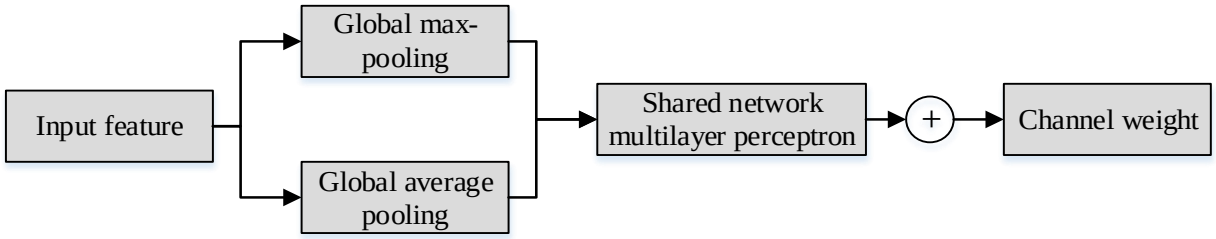

Figure 2. Channel attention

Let the dimension of input features be H×C, where H is the spatial dimension and C is the channel dimension. Channel attention first uses both global maximum pooling and global average pooling to aggregate the spatial information of input features. The spatial information in each channel is compressed into a representation, and two 1×C feature vectors are obtained. Then they are input the two feature vectors into the shared network respectively. Then, the two one-dimensional vectors 1×C of the shared network output are added and merged. Finally, sigmoid function is used to normalize and the attention weight of each channel is obtained. The shared network is a multi-layer perceptron (MLP) with two hidden layers. In order to reduce the number of parameters, the number of neurons in the first hidden layer is set to C/r, where r is the decline rate, and the number of neurons in the second hidden layer is restored to C. The specific calculation process of channel attention is as follows:

$$M_c(F) = \delta[MLP(GAP(F)) + MAP(GAP(F))] \\ = \delta[W_1(W_0(F_{GAP}^c)) + W_1(W_0(F_{GMP}^c))] \quad (1)$$

Where F is the input feature vector. $M_c(F)$ is the corresponding channel attention weight. $\delta(\cdot)$ is the sigmoid function. W_0 and W_1 are the parameters of two hidden layers of MLP. GAP and GMP are global average pooling and global maximum pooling. F_{GAP}^c and F_{GMP}^c are the feature representations obtained by pooling spatial information on each channel.

Spatial attention

Spatial attention focuses on important information in the spatial dimension of input features, and its principle is shown in figure 3.

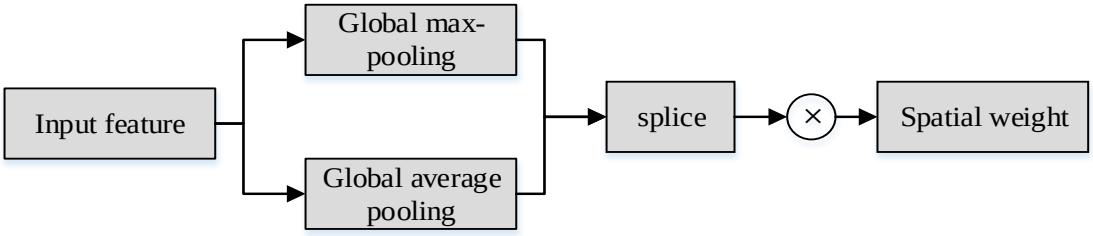

Figure 3. Spatial attention

GAP and GMP are used simultaneously on the channel axis of the input feature for spatial attention to obtain two feature vectors of dimension H×1 respectively. Then, the two feature vectors are spliced into the feature vector of H×2 along the channel direction. Then, it is mapped to the feature vector of Hx1 through the single-core convolution layer. Finally, the sigmoid function is used for normalization to obtain the spatial attention weight, and the calculation formula is as follows:

$$M_h(F) = \delta(conv([GAP(F); GMP(F)])) \\ = \delta(conv([F_{GAP}^h; F_{GMP}^h])) \quad (2)$$

Where $M_h(F)$ is the corresponding weight of spatial attention. $conv$ is single-core convolution layer. F_{GAP}^h and F_{GMP}^h are two pooled feature representations that aggregate channel information at each spatial location.

2.3. Hybrid attention

Channel attention and spatial attention focus on the more effective information in the channel dimension and spatial dimension of input features respectively. The two attention modules have complementary functions and are often used in combination [13,14]. There are three combined modes of channel attention and spatial attention. The first method is to use channel attention to weight the input feature F to get F_1, and then use spatial attention to re-weight the weighted feature F_1 to get the final output F_2. This combination method is referred to as CBAM0 in this paper, and its calculation process is as follows:

$$F_1 = M_c(F)\Theta F \tag{3}$$

$$F_2 = M_h(F_1)\Theta F_1 \tag{4}$$

Where Θ means multiplying element by element.

The second combination is CBAM1. The spatial attention module is used before the channel attention module, and the calculation process is as follows:

$$F_1 = M_h(F)\Theta F \tag{5}$$

$$F_2 = M_c(F_1)\Theta F_1 \tag{6}$$

The third combination is CBAM2. Both channel attention and spatial attention are used for input feature F to obtain channel attention weight M and spatial attention weight M respectively. Then, these two kinds of attention weights are assigned to the original input feature, and the formula is as follows:

$$F_2 = M_c(F_1)\Theta M_h(F)\Theta F \tag{7}$$

In this paper, one-dimensional convolution is used to extract the deep features of one-dimensional sequential dance images. Different from two-dimensional convolution, the "space" dimension in one-dimensional convolution contains only one dimension, which can also be regarded as the "time" dimension for one-dimensional sequential data. With the data sequence of a single sample point of the image as the input of the network, after the mapping of multiple convolutional layers, the deep features obtained still retain a certain timing, that is, the short-term timing within a single sample point.

Using spatial (temporal) attention module, more weight can be assigned to more important time points, which is beneficial to improve image recognition performance. If the CBAM0 combination method is adopted, the timing of input features will be destroyed and the effect of spatial attention will be affected. In addition, the network itself also has a certain randomness, and there will be some deviation in the attention weight obtained by each calculation. CBAM2 simultaneously multiplies the two attention weights with the original input to amplify this deviation. Therefore, this paper finally adopts the combination method of CBAM1. Firstly, spatial attention module is used to emphasize more important spatial (temporal) features, and then channel attention is supplemented to strengthen important channels and suppress less important ones. The structure of CBAM1 is shown in figure 4.

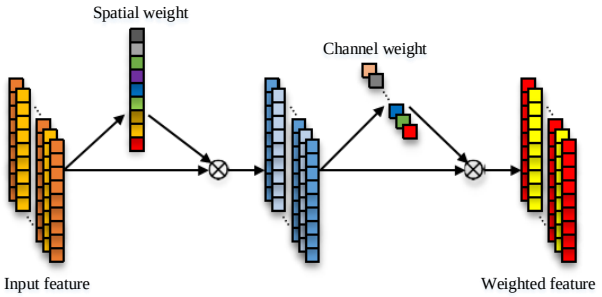

Figure 4. CBAM1 module

3. Proposed dance posture recognition method

The dance movement detection algorithm based on posture recognition is divided into three parts: posture recognition, key point feature processing and movement classification. Firstly, the input image is cut to 368*368 and the input posture recognition network is used to identify key points of human body. Then, residual network is used to detect human body regions according to the contour values of human key points. Finally, dance movement classification can be realized by integrating the feature classification of key points and image classification. The specific flow of dance movement detection based on posture recognition is shown in Figure 5, whose node classification network consists of three branches: key point feature extraction, image classification and fusion.

Aiming at the key point feature classification branch, through analyzing the dance movement characteristics, the whole connection layer is set as 6 layers. The number of neurons in the first layer is the same as the key point feature dimension of gesture recognition output, which is 18. The number of neurons in the second layer is 256. The number of neurons from layer 3 to layer 6 is 512.

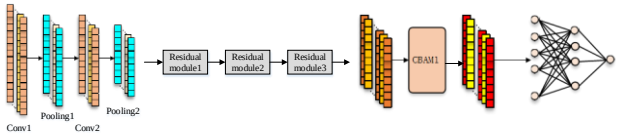

Figure 5. Proposed network

As for the residual block image classification branch, a convolutional layer and batch-normalization layer are selected to form the residual element in this study, and the residual element is stacked to form the residual network structure. In this network structure, the neural network can learn all the functions. Experiments show that the difficulty of the model can be reduced by learning the residual error directly. Therefore, residual units provided by ResNet are used to train the neural network, and a dropout layer is added between layer 3 and layer 4. Firstly, the input image was cut to 368*368, and the size of HeatMap and PAFS is set to 19*46*46. Then, according to the contour value of key points, the human body contour frame is cut out for ResNet50 training.

Two maximum pooling layers are used to compress the features to reduce the computation of subsequent network and improve the recognition speed of the model. The calculation formula of convolution layer is as follows:

$$x_{i,j}^{l+1} = \sigma(\sum_{i,j} x_i^l \otimes W_j^i + b) \quad (8)$$

Where $\otimes$ is convolution operation. W_j^i is the i-th weight on the j-th convolution kernel. b is learnable bias. $\sigma(\cdot)$ is ReLU activation function. $x_{i,j}^{l+1}$ represents the i-th feature of the $l+1$ layer, which is obtained after the convolution kernel W_j^i operation and activation function activation.

Then residual network is used for deep feature extraction. The network consists of three residual modules superimposed. The jump connection in the residual module can effectively alleviate the propagation loss of degraded information in the network. Then the deep features learned from the residual network are input into the CBAM1 module of the attention network. In this module, spatial attention and channel attention are used successively to strengthen the relatively important spatial information and channel information in deep features.

Finally, the deep features weighted by the attention network are leveled and input into the prediction network for recognition. The identification network is composed of three full connection layers, and the calculation formula is as follows:

$$x_i^{l+1} = \sigma(\sum_{j=1}^{D} W_{i,j} x_j^l + b) \quad (9)$$

Where x_i^{l+1} is the output of the i-th neuron in layer $l+1$. D is the total number of neurons at layer l. $W_{i,j}$ is the weight parameter between the i-th neuron in layer $l+1$ and the j-th neuron in layer l.

The network in this paper includes 4 residual blocks, and the pre-processed image size is 224*224. The specific training process is as follows:

Step 1: Input the pre-processed image into the convolution layer with the size of 7*7*64 and step size of 2 for convolution to obtain the feature map of 112*112;
Step 2: The feature map is successively passed through a 3*3 pooled window with a step of 2 and three blocks, each of which consists of 3 layers. The convolution kernel at the first layer is 1*1*64, the convolution kernel at the second layer is 3*3*64, and the convolution kernel at the third layer is 1*1*256.
Step 3: After the residual unit passes through the block, it passes through an average pooling layer;
Step 4: Finally connect two full connection layers with 2048 and 512 layers successively.

For fusion action classification, it is designed as a six-layer full-connection layer network structure, and the number of neurons in the first layer is determined by the fusion of key point feature classification and residual block image classification, which is 1024. The number of neurons at the second and third layers is 512, the number of neurons at the fourth and fifth layers is 256, and the number of neurons at the sixth layer is 6.

4. Experiments and analysis

The proposed algorithm is verified by experiments on pyTorch open source neural network. The network framework contains some commonly used data sets, which are convenient for transfer learning. The torch.

4.1. Data sets

The experimental data set includes 3485 image frames extracted from concert videos and dance videos. The data set includes singers and dancers at home and abroad, etc. The stage includes all kinds of large, medium and small performance platforms and daytime scenes and night scenes. Firstly, the key points of the image are obtained by gesture recognition, and then the working features and image features are calculated to obtain the single frame image and 18-dimensional data set of the character actions. Among them, there are 6 kinds of action data of characters, as shown in figure 6. 3200 images in the data set are randomly selected as the training set, and the remaining 350 images are used as the test set. The specific division of training set and test set is shown in table 1.

Figure 6. Dance samples

Table 1. Training and testing data

Number	akimbo	hold high	Single arm open	wave	walk
Training	795	470	470	250	647
Testing	88	51	52	27	71

4.2. Parameter setting

The selection of parameters has a great influence on the detection and recognition of dance movements. Therefore, in order to improve the performance of the algorithm, it is necessary to determine the optimal network parameters of the algorithm through multiple experiments, including the number of neurons and the size of batch processing. Firstly, to determine the number of neurons, the transform.compose is used to enhance the input image data, and the image is trimmed into 256*256 size. Then, the image with the center size of 224*224 is trimmed through random rotation and horizontal reversal successively. Then, normalization operation is carried out, and the fused human skeleton is moved. Finally, the recognized information is input into Resnet and Posenet, and the 18-dimensional features are transmitted to the full connection layer. Therefore, the number of neurons in the first layer of the full connection layer is 18. After each linear function, it adds a torch. Relu() function and randomly disable the neurons at a ratio of 0.5.

If the batch size is set too large, the algorithm will converge too fast. If its setting is too small, local optimal solution is easy to appear. In this experiment, the batch size is set to 64 according to the hardware environment and data set size. Considering that the convolution layer of the algorithm network is RelU function, the dropout layer is used to reduce the parameters of the full connection layer, and the random gradient descent SGD optimizer is used to optimize the network. Epoch is set to 30, and the learning rate is set to 0.001. After every 10 epochs, the learning rate became 1/10 of the original.

Considering that a single feature cannot fully and accurately express dance movements. The three selected features are fused. First, the key point position is transformed into the relative position of the neck position of the human body. And then it computes the limb vector. Finally, the relative position is normalized, that is, the fusion of features is realized.

4.3. Results analysis

The accuracy of the algorithm in the training set and test set is shown in Figure 7. The recognition accuracy on specific test sets is shown in Table 2.

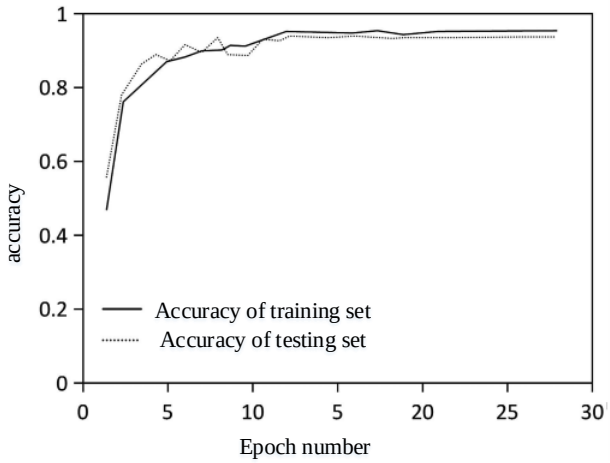

Figure 7. Accuracy of training set and testing set

Table 2. Recognition result

Action type	Accuracy/%
akimbo	98.9
hold high	84.3
Single arm open	94.2
wave	84.3
walk	96.6

As can be seen from Table 2, the average identification accuracy of this research method on the test set is more than 92%, and the overall identification accuracy is high. But arm raised and one-handed waving are less than 85%. The reason is that the arm amplitude of arm raising and one-hand waving is larger than that of the other four movements, and the state of the other hand is uncertain when waving with one hand, thus reducing the accuracy of algorithm recognition. In addition, arm raising and one-hand waving will have certain deviation due to the different distance and shooting Angle between the human body and the camera, resulting in recognition error. Therefore, the recognition accuracy of these two movements is low. In addition, the difference of data sets

also leads to the low recognition rate of arm raising and one-hand waving in this algorithm. In dance movements, one-arm outstretched and arm raised movements are less than other movements, so the recognition accuracy of the algorithm is affected to some extent.

In order to solve the problem of low recognition accuracy caused by data set differences, the confusion matrix of the above five dance movements is constructed during data set processing, as shown in table 3 to ensure that the number of data sets of each movement is basically the same. Then, the algorithm is used for recognition. According to the recognition results, the classification accuracy of the five dance movements reaches more than 90%, indicating that increasing the number of data sets of arm raising and one-hand waving by confusion matrix can effectively reduce the influence of human differences on the recognition results and improve the recognition accuracy.

Table 3. Confusion matrix of 5 dance motions

Type	Action type	akimbo	hold high	wave	walk
Action type	**98.87**	0	0	0	1.02
akimbo	0	**84.4**	2.96	0	3.17
hold high	0	0	**94.3**	0	1.84
wave	0	0.79	1.25	**96.5**	1.77
walk	0	0	2.69	1.24	**95.9**

In order to verify the superiority of the algorithm, the new algorithm, STR [15] and QES [16] are used to conduct recognition tests on the test set, and the results are shown in Table 4. As can be seen from Table 4, compared with the comparison algorithm, the accuracy of this algorithm is higher, reaching more than 92%, indicating that the algorithm has an ideal recognition effect on dance movements with high accuracy. In addition, the experimental time shows that the running rate of the algorithm on TeslaP4 graphics card is 0.75 frame/s, and the multi-person action in a single image can be recognized.

Table 4. Comparison of recognition accuracy of different algorithms

Method	Average accuracy/%
STR	78.9
QES	89.3
Proposed	92.6

5. Conclusion

To sum up, this study designed a dance movement detection method based on posture recognition. Through the combination of bone key point information and residual network, dance movements in complex scenes can be automatically detected, and the recognition accuracy can reach more than 92%. Compared with traditional dance movement recognition methods, this algorithm has the highest recognition accuracy, and the recognition efficiency of the algorithm is almost not affected by the number of people in the image, which can meet the actual dance movement detection needs, and has a certain reference significance.

Acknowledgements.

The author greatly appreciates the reviewers' anonymous comments.

References

[1] Wan Q, Zhao H, Li J, et al. Hip Positioning and Sitting Posture Recognition Based on Human Sitting Pressure Image[J]. Sensors, 2021, 21(2):426.

[2] Liu J, Wang Y, Liu Y, et al. 3D PostureNet: A unified framework for skeleton-based posture recognition[J]. Pattern Recognition Letters, 2020, 140(8):143-149.

[3] Shoulin Yin, Hang Li, Asif Ali Laghari, et al. A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection[J]. EAI Endorsed Transactions on Scalable Information Systems. 21(33), e8, 2021. http://dx.doi.org/10.4108/eai.6-10-2021.171247

[4] Dongling Wang, Xiaowei Wang, and Shoulin Yin. A New Recursive Neural Network and Center Loss for Expression Recognition [J]. International Journal of Electronics and Information Engineering. Vol. 13, No. 3, pp. 97-104, 2021.

[5] Liu Q. Aerobics posture recognition based on neural network and sensors[J]. Neural Computing and Applications, 2021:1-12.

[6] Hu Q, Tang X, Tang W. A Real-Time Patient-Specific Sleeping Posture Recognition System Using Pressure Sensitive Conductive Sheet and Transfer Learning[J]. IEEE Sensors Journal, 2020, PP(99):1-1.

[7] Kolivand H, Joudaki S, Sunar M S, et al. A new framework for sign language alphabet hand posture recognition using geometrical features through artificial neural network (part 1)[J]. Neural Computing and Applications, 2020:1-19.

[8] Desheng Liu, Linna Shan, Lei Wang, Shoulin Yin, et al. P3OI-MELSH: Privacy Protection Point of Interest Recommendation Algorithm Based on Multi-exploring Locality Sensitive Hashing[J]. Frontiers in Neurorobotics, 2021. doi: 10.3389/fnbot.2021.660304.

[9] Ting-Ting Gao, Hang Li, and Shou-Lin Yin. Adaptive Convolutional Neural Network-based Information Fusion for Facial Expression Recognition [J]. International Journal of Electronics and Information Engineering. Vol. 13, No. 1, pp. 17-23, 2021.

[10] Jisi A and Shoulin Yin. A New Feature Fusion Network for Student Behavior Recognition in Education [J]. Journal of Applied Science and Engineering. vol. 24, no. 2, pp.133-140, 2021.

[11] Xie S, Girshick R, P Dollár, et al. Aggregated Residual Transformations for Deep Neural Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[12] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[J]. European Conference on Computer Vision, 2018.

[13] Shoulin Yin, Hang Li, Desheng Liu and Shahid Karim. Active Contour Modal Based on Density-oriented BIRCH Clustering Method for Medical Image Segmentation [J]. Multimedia Tools and Applications. Vol. 79, pp. 31049-31068, 2020.

[14] S. Yin and H. Li. Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5862-5871, 2020, doi: 10.1109/JSTARS.2020.3025582.

[15] Pham H H, Salmane H, Khoudour L, et al. Spatio–Temporal Image Representation of 3D Skeletal Movements for View-Invariant Action Recognition with Deep Convolutional Neural Networks[J]. Sensors, 2019, 19(8).

[16] Wu Z, Zhang J, Chen K, et al. Yoga Posture Recognition and Quantitative Evaluation with Wearable Sensors Based on Two-Stage Classifier and Prior Bayesian Network[J]. Sensors (Basel, Switzerland), 2019, 19(23).