# Multi-modal Data Fusion Method for Human Behavior Recognition Based on two IA-Net and CHMM

ZHANG Yinhuan[1,2], XIAO Qinkun[1], CHU Chaoqin[1], XING Heng[1]

[1] Xi'An Technological University, School of Mechatronic Engineering, Xi'an 710018, China

[2] Weinan Vocational & Technical College, Civil & Architectural Engineering, Weinan 714000, China

*corresponding.author

ZHANG Yinhuan, Xi'An Technological University, School of Mechatronic Engineering, Xi'an 710018, China 283811371@qq.com

Funding information

This work was supported in part by the NSFC under Grant 6207010855

## Abstract

The multi-modal data fusion method based on IA-net and CHMM technical proposed is designed to solve the problem that the incompleteness of target behavior information in complex family environment leads to the low accuracy of human behavior recognition.The two improved neural networks(STA-ResNet50、STA-GoogleNet)are combined with LSTM to form two IA-Nets respectively to extract RGB and skeleton modal behavior features in video. The two modal feature sequences are input CHMM to construct the probability fusion model of multi-modal behavior recognition.The experimental results show that the human behavior recognition model proposed in this paper has higher accuracy than the previous fusion methods on HMDB51 and UCF101 datasets. New contributions: attention mechanism is introduced to improve the efficiency of video target feature extraction and utilization. A skeleton based feature extraction framework is proposed, which can be used for human behavior recognition in complex environment. In the field of human behavior recognition, probability theory and neural network are cleverly combined and applied, which provides a new method for multi-modal information fusion.

## Key words

Attention mechanism; CHMM; LSTM; Multi-modal fusion; Human behavior recognition

## 0. Introduction

In the information-based family care for the elderly, human behavior recognition [1-3] is an important nursing value to master the situation on the spot, to judge abnormal behavior, to prevent accidents, and to ensure the safety of the elderly life. It has important nursing value. In complex environments, how to accurately recognition behaviors is a hot spot of research experts at home and abroad [4-5]. Behavior recognition using target features acquired by a single model is susceptible to environmental impact such as lighting, visual angle, background, etc. There are problems of missing and incomplete features, resulting in inaccurate recognition results [6-8]. The multi-modal fusion model can not only capture multi-modal data and solve the contradiction of data loss in the process of single sensor behavior recognition, but also improve the accuracy of behavior recognition by using the complementarity of different modal data **Error! Reference source not found.**. Multi-modal information [13-14] usually adopts an adaptive fusion method to obtain higher recognition accuracy than individual features, but there is usually no theoretical explanation for each feature weight assignment problem. A target recognition method based on fuzzy theory is proposed in [15], In order to improve the accuracy of the fusion model, an improved logsig function is introduced to express the importance of the information and then the weights are calculated using the fuzzy relationship to improve the recognition accuracy, However, the sensor weight to obtain the target characteristics has been given in advance and the recognition results are vulnerable to human factors. The article [16]presents a behavior recognition method based on Hidden Markov Model (HMM), which uses probability fusion method to provide theoretical basis for multi-modal data fusion. However, this method needs to set model parameters adaptively, introduces too many system parameters and reduces the speed of model training and calculation. At present, there is no in-depth study on the recognition of complex background and target occlusion.

Two improved attention networks (IA net) [17-19] and a pair of hidden Markov (CHMM) [20]are combined for behavior recognition in this paper. IA-net are respectively combined of the improved spatio temporal attention ResNet50 (STA-ResNet50), Google net (STA-GoogleNet) and Long Short Time Memory (LSTM) network. Model

treatment timely and ensuring their life safety of the elderly, the premise is accurate recognition of their behavior. However, due to the influence of object occlusion, light, environment and other factors in the family, the behavior recognition is inaccurate, which poses a threat to the life of the elderly. In order to improve the reliable recognition of human behavior, extracting effective behavior feature information is vitally improtant, this paper uses Microsoft Kinect equipment to obtain RGB and skeleton video from different angles, which not only overcome effectively the influence of complex background and avoid the occlusion problems, but also constructs the multi-modal fusion behavior recognition model, as shown in Fig. 1.

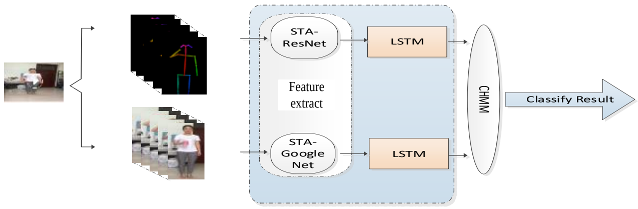


***Fig.1.** Multi-modal human behavior recognition model*

As can be seen from Fig. 1. firstly, RGB and 3D skeleton videos are input STA-GoogleNet+LSTM (expressed as LSTMr) and STA-ResNet50+LSTM network (expressed as LSTMg) to obtain two stream feature sequences. Secondly, the multi-modal features are sent to CHMM as inputs to construct the probability fusion model of human behavior recognition.

Because information redundancy and different features in neural network have distinct effects on recognition results, SE-block [18]is introduced and analyzed in this paper. It is found that the model adopts two full connections (FC) layer, which cause the network parameters to increase exponentially. the sigmoid function is lead to the problem of neuron inactivation simultaneously. The Improved SE-block(ISE-block) is proposed and embedded into ResNet50, GoogleNet respectively. The specific implementation process is shown in Fig. 2. Given a video input X, its characteristic channel number is $C'$ , after a series of convolution $F_{tr}$ transformations, the feature map with channel number C is obtained. Finally, which is introduced into the residual branch of

advantages: incomplete features will reduce the classification accuracy of the first level and the two-level fusion mechanism can be repaired at a higher level. The behavior recognition method based on HMM needs to establish an adaptive HMM classifier for each behavior in the previous research. This framework uses LSTM model to automatically extract system parameters, which can be used to improve the learning of classifier in CHMM. Finally, IA-ResNet50, IA-GoogleNet are used for behavior feature extraction to avoid the negative impact of unsupervised local features (such as HOG) strengthen important features and weaken redundant features.

## 1. Improved feature extraction network

In the family environment with complex background, Realizing the dangerous behavior recognition of the elderly, ResNet50 and inceptionand of GoogleNet added with the STA-Net.

In Fig. 2, ISE residual modal, replace FC+ReLu+FC +Sigmoid of SE-block[17] with conv_1+ReLu+ conv_2+Sigmoid to obtain ISE block. In order to avoid the problem of excessive calculation consumption caused by the increase of parameters, conv_1 convolution replaces FC layer and ReLu function connection conv_2 convolution processing to obtain 0~1 normalized weight. Finally, ISE block is introduced into different networks to form ISE-nets, so that different important features can assign different weights and improve the efficiency of feature extraction.

The length of video required by different actions is not all the same, However, the classical neural network can only accept video input with a fixed length (7 frames), resulting in low behavior recognition accuracy of arbitrary length video. In order to more fully extract the features of continuous actions with different time lengths, this paper connects the LSTM model behind the improved neural network to overcome the complex human behavior representation of long video representation, after the model is connected to the full connection layer of IA-ResNet and IA-GoogleNet, the relationship between the features of continuous action sequences with various length is obtained.

## 2. Two IA-Nets+CHMM fusion calculation

In order to more accurately recognition the behavior of the elderly, the multi-modal information obtained in the second part is fused. The specific methods are as follows, In the process of home care for the elderly, for reducing the impact of background on human behavior recognition, the STA-ResNet+LSTM is used to obtain skeleton flow $I^{D}=\left(I_i^{3D}\right)_{i=1}^{T}$ , which is expressed by quaternion: $q=(v,w)$ $=(a,b,c,w)$ represents feature, and a complete 3D skeleton behavior is expressed as 4×25 =100 dimensional vector $X_{q}=(q_1,q_2,\cdots q_{25})$ . In order to avoid the background problem, this paper adopts the color video stream $I^{c}=\left(I_i^{c}\right)_{i=1}^{T}$ from another perspective, uses STA-GoogleNet+ LSTM to directly extract the RGB behavior features, then takes the

RGB and skeleton features as the CHMM model input for modeling, so as to obtain the multi-modal fused human behavior recognition model, as shown in Fig. 3. The two network models have the same structure, but have different feature vector types. They are inputed into CHMM model for fusion to produce human behavior recognition results. Where, $x_t^r$ , $x_t^g$ are RGB or skeleton related vector input signals respectively, $h_{t-1}$ is an intermediate hidden variable,

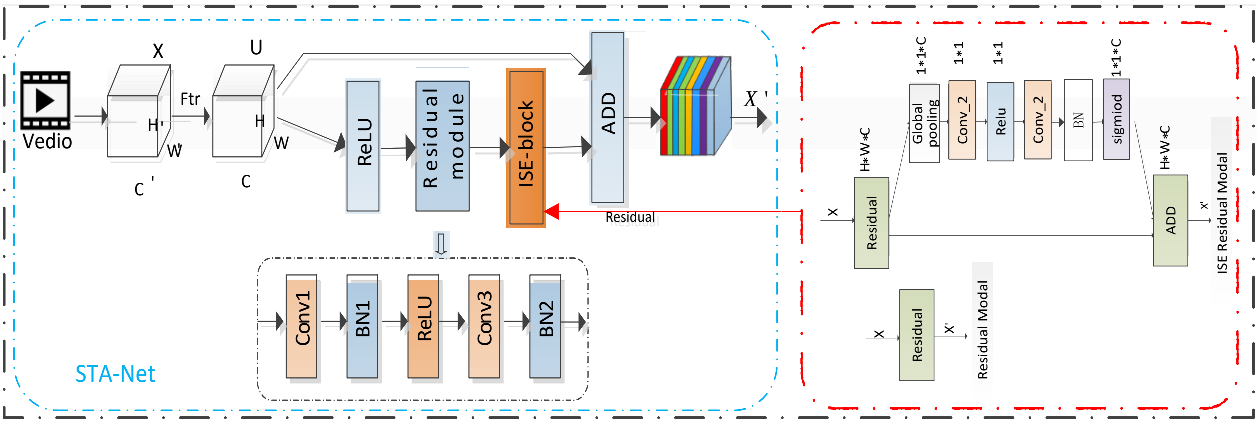


***Fig.2.*** *ISE-Net structure*

$h_0$ is the initial value, $\sigma(\cdot)$ and tanh (•) are the activation function, w is the network weight, and b is the deviation. Output $y_t = \text{soft}\max\left(w_y h_t + b_y\right)$ . In order to facilitate modeling, r and g distribution is defined to represent RGB and skeleton related information. The output of LSTM is $y_t^r$ and $y_t^g$ , which are RGB and skeleton sequences and $y_t^r$ $y_t^g$ are input to CHMM as observation signal. According to graph model theory, CHMM is divided into two basic models to simplify its calculation. It is a simple dynamic Bayesian network (DBN), which can be determined by Markov chain theory. It consist of parameters λ= (π, a, b), definition: π represents prior knowledge, A is the state transition matrix and B is the observation matrix.

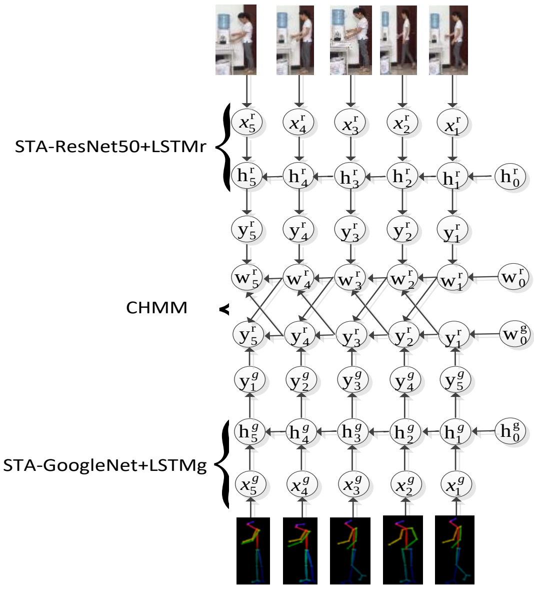

STA-ResNet50+LSTMr
$x_5^r$ $x_4^r$ $x_3^r$ $x_2^r$ $x_1^r$
$h_5^r$ $h_4^r$ $h_3^r$ $h_2^r$ $h_1^r$ $h_0^r$
$y_5^r$ $y_4^r$ $y_3^r$ $y_2^r$ $y_1^r$
CHMM
$w_5^r$ $w_4^r$ $w_3^r$ $w_2^r$ $w_1^r$ $w_0^r$
$y_5^r$ $y_4^r$ $y_3^r$ $y_2^r$ $y_1^r$ $w_0^g$
$y_1^g$ $y_2^g$ $y_3^g$ $y_4^g$ $y_5^g$
STA-GoogleNet+LSTMg
$h_5^g$ $h_4^g$ $h_3^g$ $h_2^g$ $h_1^g$ $h_0^g$
$x_5^g$ $x_4^g$ $x_3^g$ $x_2^g$ $x_1^g$

***Fig.3.*** *Two IA-Net+CHMM fusion model for human behavior recognition*

Fig. 3. can be described as $y^r = (y_1^r, y_2^r, \cdots, y_T^r)$ and $y^g = (y_1^g, y_2^g, \cdots, y_T^g)$, where $y_i^r$ $y_i^g$ represent respectively the behavior RGB and skeleton eigenvector at time t, A hybrid DBN dynamic Bayesian model is established, which use behavior observation Y and symbolic state W to represent the continuous human behavior recognition system. Assuming that RGB and skeleton have the same correlation contribution, the state sequence $w^g$ = （$w_1^g, \cdots, w_T^g$） can be updated with $w^r$ = （$w_1^r, \cdots, w_T^s$）, then the information in the two sequences can be fused through probabilistic reasoning to produce a final state with high estimation accuracy.

The STA-ResNet50/STA-GoogleNet+LSTM+CHMM calculation process, CHMM is divided into two HMM to calculate the optimal hidden state probability. Behavior related HMM includes three parameters [16]:

$$\begin{cases} \pi^r = \left[p_i^r\right]_{1\times n} = p\left(w_o^r\right) \\ A^r = \left[a_{ij}^r\right]_{n\times n} = p\left(w_{t+1}^r \middle| w_t^r\right) \\ B^r = p\left(y_t^r \middle| w_t^r\right) = N\left(u_y^r, \sum_y^r\right)\left(y_t^r\right) \end{cases} \qquad （1）$$

Where, $\pi^r$ is the prior distribution of behavior related state $w_0^r$ . If $w_0^r$ includes n states expressing $S_1, S_2, \cdots S_n$ , where $S_i$ corresponds to $P(w_0^r) = S_i = \pi^r(i)$ at time i. The term $A^r$ is the state matrix and $a_{ij}^r$ represents the state transition probability from i time to j time, therefore $a_{ij}^r = P(w_{t+1}^r) = s_j / w_t^r = s_i$. $B^r$ is the observation matrix and $y_t^r$ is a continuous variable. The observation probability

$P(y_t^r|W_t^r = s_i)$ is a Gaussian distribution, where $u_y$ and $\sum_y$ are the mean and variance respectively. The optimal state sequence of $w$, $p(w)$ can be calculated from Bayesian theory and the estimation is as follows [7]:

$$p\left(w_1^r\right) = p\left(w_1^r \middle| w_0^r\right) p\left(w_0^r\right) \qquad (2)$$

Initial time $p\left(x_0^r\right) = p\left(w_0^r\right)$, then

$$p\left(w_1^r\right) = p\left(w_1^r \middle| w_0^r\right) p\left(x_0^r\right) \qquad (3)$$

The observation $y_1$thenyields

$$p\left(w_1^r y_1^r\right) = \frac{p\left(y_1^r \middle| w_1^r\right) p\left(x_1^r\right)}{p\left(y_1^r\right)} \qquad (4)$$

The state probability at time t can be determined from:

$$p\left(w_{t+1}^r\right) = p\left(w_{t+1}^r \middle| w_t^r\right) p\left(w_t^r\right) \qquad (5)$$

And the state can be optimized using an observation sequence:

$$p\left(w_{1+t}^r \middle| y_{1:1+t}^r\right) = p\left(w_{1+t}^r \middle| y_{1+t}^r, y_{1:t}^r\right)$$
$$= \left(1/p\left(y_{1+t}^r\right)\right) \cdot p\left(y_{1+t}^r \middle| x_{1+t}^r\right) \cdot p\left(x_{1+t}^r \middle| x_t^r\right)$$
$$p\left(x_t^r\right) \sum p\left(w_{1+t}^r \middle| w_t^r\right) p\left(w_t^r \middle| y_{1:t}^r\right) \qquad (6)$$

The fusion reasoning probability is obtained through the main network related to HMM, Similarly, the three skeleton related parameters of family care for the elderly are see(7):

$$\begin{cases} \pi^g = \left[p_i^g\right]_{1\times n} = p\left(w_o^g\right) \\ A^g = \left[a_{ijk}^g\right]_{n\times n} = p\left(w_{t+1}^g \middle| w_t^g, w_t^g\right) \\ B^g = p\left(y_t^g \middle| w_t^g\right) = N\left(u_y^g, \sum_y^g\right)\left(y_t^g\right) \end{cases} \qquad (7)$$

Where, $\pi^g$ is the prior distribution of skeleton-related state $w_0^g$ . If $w_0^g$ includes n states, then expressed $(s_1, s_2, \cdots s_n)$ , $s_i$ corresponds to skeleton state $p\left(w_o^g = s_i\right) = \pi^s(i)$ at time i, The term $A^g$ is a state transaction matrix. $a_{ijg}$ denotes the transaction probability from the i time state to the j time sign state. As a result, $a_{ijk}^g = P\left(W_{t+1}^g = s_k \middle| w_t^g = s_i\right)$. $B^g$ is an observation matrix and $y_t^g$ is a continuous variable. The observation probability $P\left(y_t^g \middle| w_t^g = s_i\right)$ is then a Gaussian distribution, where $u_y$ and $\sum_y$ are the mean and variation respectively. The initial hand-related HMM inference state is given by:

$$p\left(w_1^g\right) = p\left(w_1^g \middle| w_0^g, w_0^r\right) p\left(w_0^g, w_0^r\right)$$
$$= p\left(w_1^g \middle| w_0^g, w_0^r\right) p\left(w_0^g\right) p\left(w_0^r\right) \qquad (8)$$

From Fig. 3. $w_0^g$ can then be updated using $y_1^g$ :

$$p\left(w_1^g y_1^g\right) = \frac{p\left(y_1^g \middle| w_1^g\right) p\left(w_1^g\right)}{p\left(y_1^g\right)}$$
$$= \frac{p\left(y_1^g \middle| w_1^g\right) p\left(w_1^g \middle| w_0^g, w_0^s\right) p\left(w_0^g\right) p\left(w_0^s\right)}{p\left(y_1^g\right)} \qquad (9)$$

In general, the state at time t is:

$$p\left(w_{1+t}^g\right) = p\left(w_{1+t}^g \middle| w_t^g, w_t^r\right) p\left(w_t^g, w_t^r\right)$$
$$= p\left(w_{1+t}^g \middle| w_t^g, w_t^r\right) p\left(w_t^g\right) p\left(w_t^r\right) \qquad (10)$$

Where, $p\left(w_{1+t}^g\right)$, $p\left(w_t^r\right)$ represents respectively the distribution state of g and r sequences at time t, then an optimized state can be estimated from the observed sequences:

$$p\left(w_{1+t}^g \middle| y_{1:1+t}^g\right) = p\left(w_{1+t}^g \middle| y_{1+t}^g\right) p\left(w_{1+t}^g \middle| y_{1:t}^g\right)$$
$$= \left(1/p\left(y_{1+t}^g\right)\right) \cdot p\left(y_{1+t}^g \middle| w_{1+t}^g\right) p\left(w_{1+t}^g \middle| w_t^g, w_t^r\right)$$
$$\times p\left(w_t^r\right) \cdot p\left(w_t^g\right) \times p\left(w_{1+t}^g \middle| y_{1:t}^g\right) \qquad (11)$$
$$= \left(1/p\left(y_{1+t}^g\right)\right) \cdot p\left(y_{1+t}^g \middle| w_{1+t}^g\right) p\left(w_{1+t}^g \middle| w_t^g, w_t^r\right)$$
$$\times p\left(w_t^r\right) \cdot p\left(w_t^g\right) \times p\left(w_{1+t}^g \middle| y_{1:t}^g\right) \cdot \sum_{w_t} p\left(w_{1+t}^g \middle| w_t^g\right) p\left(w_t^g \middle| y_{1:t}^g\right)$$

Bayesian theory then produces:

$$\max_{w_1, \cdots w_t} = p\left(w_{1:1+t}^g \middle| y_{1:1+t}^g\right)$$
$$= \alpha p\left(y_{1+t}^g \middle| y_t^g\right) \max_{w_t} \left(w_{1+t}^g \middle| w_t^g\right) \times \max_{w_1, \cdots w_{t-1}} \left(w_t^g \middle| y_{1:t}^g\right) \qquad (12)$$

With an optimal behavior classification prediction of:

$$\left(\hat{w}_{1:t}^g\right)^* = E\left(w_{1:t}^g \middle| y_{1:t}^g\right) = \sum_w w_{1:t}^g \cdot \left(\max_{w_1, \cdots, w_{t-1}} p\left(w_{1:t}^g \middle| y_{1:t}^g\right)\right) \qquad (13)$$

In contrast, using RGB-related HMMs as the primary probability network gives:

$$\left(\hat{w}_{1:t}^r\right)^* = E\left(w_{1:t}^r \middle| y_{1:t}^r\right) = \sum_w w_{1:t}^r \cdot \left(\max_{w_1, \cdots, w_{t-1}} p\left(w_{1:t}^r \middle| y_{1:t}^r\right)\right)$$
$$(14)$$

This leads to the final multi-modal fusion behavior classification result:

$$\left(\hat{w}^{\,fusion}_{1:t}\right)^{*} = E\left(w^{fusion}_{1:t}\middle| y^{r}_{1:t}, y^{g}_{1:t}\right)$$
$$= \sum_{w}\left\{w^{r}_{1:t}\cdot\left(\max_{w_1,\cdots,w_{t-1}} p\left(w^{r}_{1:t}\middle| y^{r}_{1:t}\right)\right)\right. \qquad (15)$$
$$\left.+\, w^{g}_{1:t}\left(\max_{w_1,\cdots,w_{t-1}} p\left(w^{g}_{1:t}\middle| y^{g}_{1:t}\right)\right)\right\}$$

## 3. Experiments and Discussion

In order to better verify the performance of the model proposed in this paper and provide a basis for the intelligent care of the elderly, experiments and analysises will be organized from five aspects: experimental environment, experimental parameter selection dataset, extracting behavior characteristics and fusion model performance evaluation.

### *3.1 Experimental environments*

The computer used in the experiment is HP Pavilion 15, Windows10 operating system, Intel quad core (TM) i5-7300 processor, 2.6GHz main frequency, 8G memory, NVIDIA GEFDRCE GTX1050 graphics card and the test running environment is MATLAB 2021a.

### *3.2 Selection of experimental parameters*

In order to train a high-precision neural network, HMDB51 dataset [21]is used for repeated training and Optimization for 30 times, as shown in Table 1. It can be obtained that if the learning rate is set too small, the convergence process becomes very slow and too large leads to failure of convergence, so the best learning rate is 0.0001. The ‘adam’ optimization algorithm with high accuracy in video classification is adopted. With the continuous increase of minbatch, the training accuracy increases, However, when it gives 32, due to hardware constraints, the network cannot train, resulting in program interruption, Therefore, the mini batch-size is 16. In order to prevent overfitting, dropout is finally selected to 0.7 through the experiment, which can obtain better verification accuracy, Finally, the processor is set as 16G graphics processing special ‘GPU’ to improve the training speed of the model.

**Table 1** Parameter Adjustment in Neural Network Training

| Learn rate | Min Batch | Algorithm | Gradient threshold | Dropout | Accuracy（%） | Loss rate（%） | Iterations PerEpoch | Max Epochs |
|---|---|---|---|---|---|---|---|---|
| 1e$^{-3}$ | 16 | adam | 1 | 0.5 | 73.03 | 1.31 | 60 | 7080 |
| 1e$^{-4}$ | 8 | SDG | 1 | 0.5 | 79.19 | 1.0 | 30 | 3540 |
| | 12 | SDG | 0.9 | 0.7 | 83.35 | 0.86 | 40 | 2360 |
| | 16 | adam | 0.7 | 0.7 | 88.94 | 0.55 | 60 | 4680 |
| | 32 | adam | 0.5 | 0.7 | -- | -- | -- | -- |
| 1e$^{-5}$ | 12 | adam | 1 | 0.6 | 78.37 | 0.79 | 60 | 4680 |
| | 16 | SDG | 0.5 | 0.7 | 83.96 | 0.75 | 40 | 2360 |
| | 32 | adam | 0.5 | 0.5 | -- | -- | -- | -- |

### *3.3 Datasets*

The two main stream datasets of UCF101 [22]and HMDB51 in behavior recognition are used to evaluate the performance of the proposed model. Among them, UCF101 contain 101 kinds of behaviors and 13320 videos, which are mainly divided into five categories, only body move-ments, human-human interaction, human-object interaction, playing music equipment and various sports. The HMDB51 dataset have 51 action categories, including 6766 video samples. In this paper, 60% of videos are used for training, 20% as validation and the rest dataset 20% for testing.

### *3.4 IA-Net extracting behavior features*

IA net is used to visualize the different actions of 'Driving' and 'brush_hair' on the two datasets, the first conv_1 and the last convolution_fire2 visually observe different frames of the action respectively and the results are shown in Fig. 4. and Fig. 5. When the first layer of convolution is obtained, the contour information of convolution action is shown in the red box of the figure. With the increase of convolution times, the finer the characteristic information of convolution is, as shown in the blue box of the figure, Therefore, in order to obtain more accurate behavior information, when improving the network, the attention mechanism is added at the end of each block convolution to improve the weight of important features and weaken redundant features.

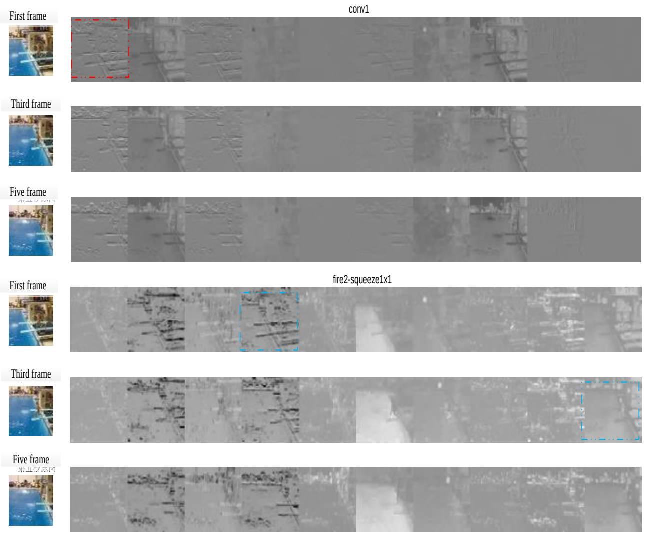


***Fig.4.*** *UCF101-Driving Behavior visualization*

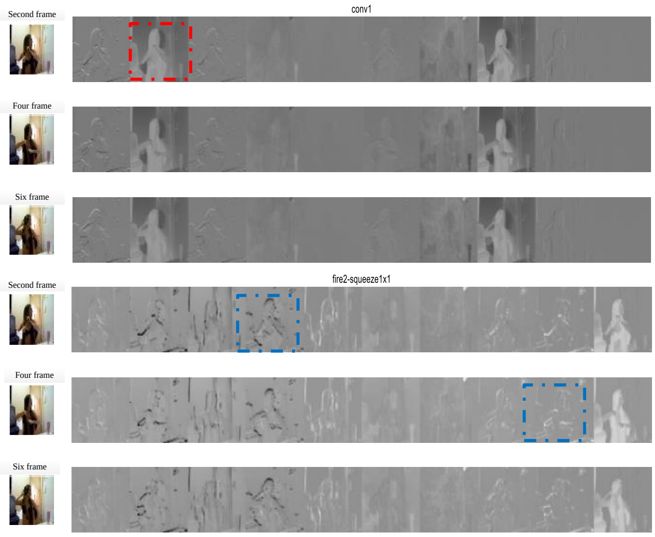


***Fig. 5.*** *HMDB51-brush_hair behavior visualization*

### *3.5 Performance evaluation of the model*

To assess the accuracy of the proposed model for identifying family behavior of the elderly in complex family environments. Two types of behavior features extracted in Section 3.4 are input to CHMM to fuse and form the final behavioral recognition model. Experiments are performed on two datasets HMDB51 and UCF101 to check the performance of the proposed model.

### *3.5.1 Evaluating performance on HMDB51 dataset:*

To evaluate the performance of the model, two improved neural networks are combined with the CHMM algorithm to form the final recognition model. Firstly, the accuracy of behavior recognition is obtained through model training. Secondly, the experimental results are analyzed to find out the specific behavior of identifying errors, so as to provide basis for model optimization. In order to further observe the identification ability of the model, finally, the dimension reduction of clustering method is used to analyze the misidentified behavior and the performance of the model is judged from aboved three aspects.

（1）Experiments process

The HMDB51 dataset contains 51 classes of actions. Ten classes are selected in turn. The training process of 1-10 classes is shown in Fig. 6. and the accuracy is 89.04%.

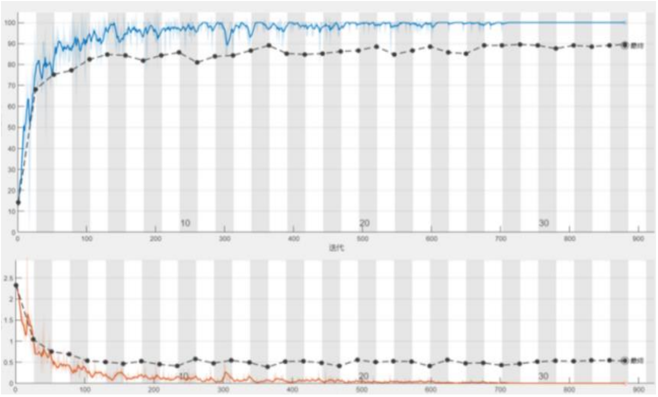


***Fig.6.*** *Training accuracy of HMDB51 dataset*

After five round experiments, the training accuracy is obtained as Table 2 and the overall training accuracy on HMDB51 dataset is 87.68%.

**Table 2** Training Accuracy on HMDB51 Dataset

| Behavior | Accuracy |
|---|---|
| 1-10 classes | 89.04% |
| 11-20 classes | 86.04% |
| 21-30 classes | 87.31% |
| 31-40 classes | 87.04% |
| 41-51 classes | 90.15% |
| Average accuracy | 87.88% |

（2）Experimental process analysis

To observe the specific recognition of each action by the model on the HMDB51 validation dataset, the confusion matrix is used. From this, there are about 20 videos for each action of the 10 types, the left vertical axis represents the real labels of the actions and the horizontal axis represents the predicted results, Fig. 7. shows that the validation accuracy is 96.37%.

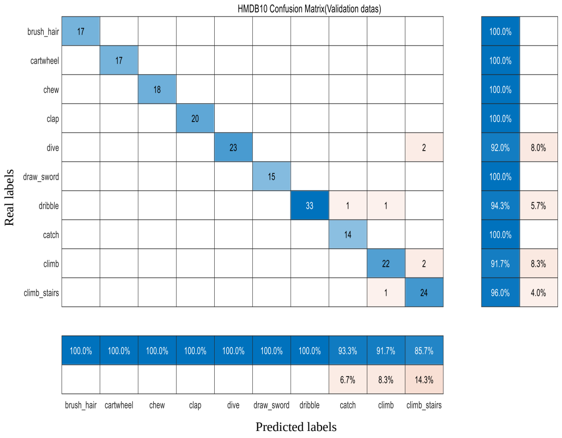
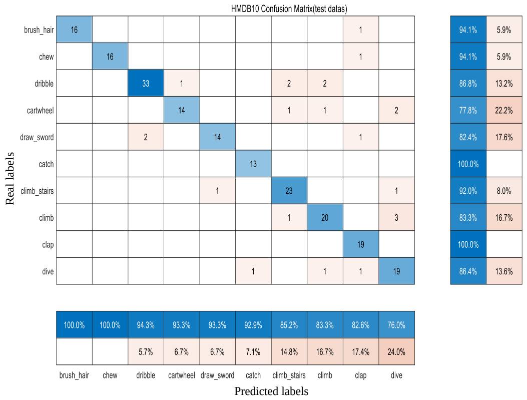


***Fig.7.*** *Confusion matrix on HMDB51 verifies dataset*

Meanwhile, the generalization ability of the model is trained on the testset with an accuracy of 87.88% in Fig. 8. Among them, the ‘drive’ motion recognition rate is only 76.0%, because there are many video segments of sailing, aerial driving and so on in the drive that almost escape the driving object at the bottom of the sailboat.



| Real labels | brush_hair | chew | dribble | cartwheel | draw_sword | catch | climb_stairs | climb | clap | dive | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| brush_hair | 16 | | | | | | | | 1 | | 94.1% | 5.9% |
| chew | | 16 | | | | | | | 1 | | 94.1% | 5.9% |
| dribble | | | 33 | 1 | | | 2 | 2 | | | 86.8% | 13.2% |
| cartwheel | | | | 14 | | | 1 | 1 | | 2 | 77.8% | 22.2% |
| draw_sword | | | 2 | | 14 | | | | 1 | | 82.4% | 17.6% |
| catch | | | | | | 13 | | | | | 100.0% | |
| climb_stairs | | | | | 1 | | 23 | | | 1 | 92.0% | 8.0% |
| climb | | | | | | | 1 | 20 | | 3 | 83.3% | 16.7% |
| clap | | | | | | | | | 19 | | 100.0% | |
| dive | | | | | | 1 | | 1 | 1 | 19 | 86.4% | 13.6% |
| | 100.0% | 100.0% | 94.3% | 93.3% | 93.3% | 92.9% | 85.2% | 83.3% | 82.6% | 76.0% | | |
| | | | 5.7% | 6.7% | 6.7% | 7.1% | 14.8% | 16.7% | 17.4% | 24.0% | | |

***Fig.8.*** *Confusion matrix on HMDB51 test dataset*

（3）Visual Fusion Results

To further validate the model's ability to discriminate behavior, this paper visualizes the low-dimensional distribution of different actions. Selecting 10 representative action classes from HMDB51 for unsupervised clustering is shown in Fig. 9. As is showed, the overall classification results are relatively better. However, individual behaviors such as ‘brush hair’ have sparse distribution points because of the large difference in motion amplitude and angle. The ‘chew’ three actions are misidentified as ‘climb’ because three people in the dataset chew on stairs and ladder scenes causing recognition errors, which provide a basis for later model improvement.

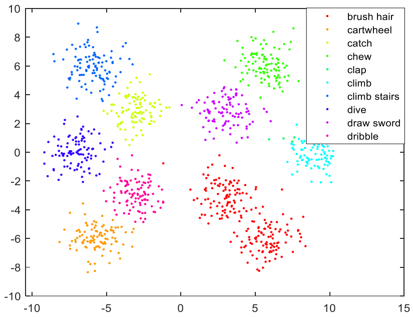


***Fig.9.*** *Clustering results on HMDB51 dataset*

### *3.5.2 Evaluating performance on UCF101 Dataset:*

（1）Experiments process

In the UCF101 experimental dataset, 10 types of actions were randomly selected, each of which contained about 107 videos, lasted 3 seconds and totaled 1241 videos. According to 6:2:2 partition randomly, 60% as training dataset, 20% as validation dataset and the rest of 20% as test dataset, the experimental validation accuracy on this dataset is up to 99% in Fig.10.

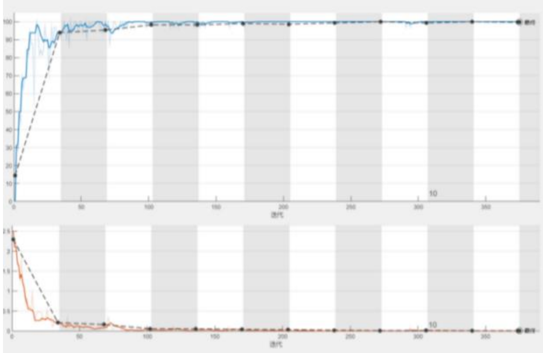
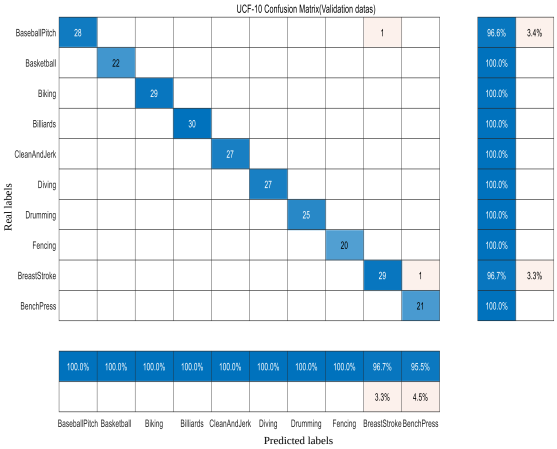


***Fig.10.*** *Training accuracy of UCF101 dataset*

（2）Experimental process analysis

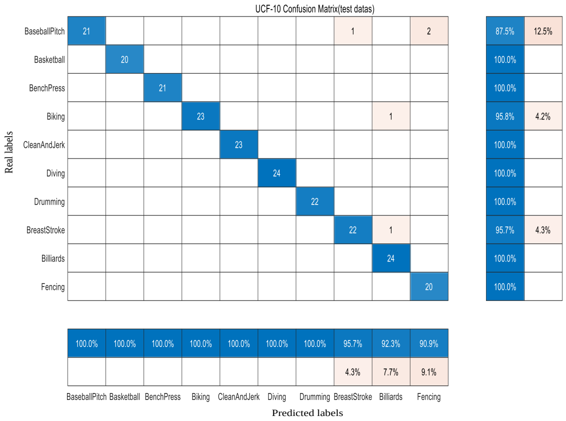
On UCF101 validation set and testset, observe the specific recognition of each action by the model and use the confusion matrix to analyze in Fig. 11. and Fig. 12.



| Real labels | BaseballPitch | Basketball | Biking | Billiards | CleanAndJerk | Diving | Drumming | Fencing | BreastStroke | BenchPress | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BaseballPitch | 28 | | | | | | | | 1 | | 96.6% | 3.4% |
| Basketball | | 22 | | | | | | | | | 100.0% | |
| Biking | | | 29 | | | | | | | | 100.0% | |
| Billiards | | | | 30 | | | | | | | 100.0% | |
| CleanAndJerk | | | | | 27 | | | | | | 100.0% | |
| Diving | | | | | | 27 | | | | | 100.0% | |
| Drumming | | | | | | | 25 | | | | 100.0% | |
| Fencing | | | | | | | | 20 | | | 100.0% | |
| BreastStroke | | | | | | | | | 29 | 1 | 96.7% | 3.3% |
| BenchPress | | | | | | | | | | 21 | 100.0% | |
| | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 96.7% | 95.5% | | |
| | | | | | | | | | 3.3% | 4.5% | | |

***Fig.11.*** *Confusion matrix on UCF101 validation datase*

Fig. 11. shows that the model is trained on UCF101 validation set, with about 25 videos for each type of action, 258 videos input and 257 videos recognized correctly with a recognition rate of 99.6%.



| Real labels | BaseballPitch | Basketball | BenchPress | Biking | CleanAndJerk | Diving | Drumming | BreastStroke | Billiards | Fencing | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BaseballPitch | 21 | | | | | | | 1 | | 2 | 87.5% | 12.5% |
| Basketball | | 20 | | | | | | | | | 100.0% | |
| BenchPress | | | 21 | | | | | | | | 100.0% | |
| Biking | | | | 23 | | | | | 1 | | 95.8% | 4.2% |
| CleanAndJerk | | | | | 23 | | | | | | 100.0% | |
| Diving | | | | | | 24 | | | | | 100.0% | |
| Drumming | | | | | | | 22 | | | | 100.0% | |
| BreastStroke | | | | | | | | 22 | 1 | | 95.7% | 4.3% |
| Billiards | | | | | | | | | 24 | | 100.0% | |
| Fencing | | | | | | | | | | 20 | 100.0% | |
| | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% | 95.7% | 92.3% | 90.9% | | |
| | | | | | | | | 4.3% | 7.7% | 9.1% | | |

***Fig.12.*** *Confusion matrix on UCF101 test dataset*

doi:10.20944/preprints202112.0244.v1

When the testset is used to predict, the training accuracy is high 97.92% in Fig.12. The action 'Fencing' and 'Breastroke' are recognized as 'Baseball pitch' due to the background and light, resulting in a low recognition rate.

（3）Visual Fusion Results

To further validate the model's ability to discriminate behavior, 10 types of actions from UCF101 dataset are selected to cluster in Fig. 13. From this, the results of 'billards' and 'Basketball' clustering are crossed, because they belong to spherical motion, the main features are human, sphere and there are many similar features, leading to partial confusion in the recognition results.

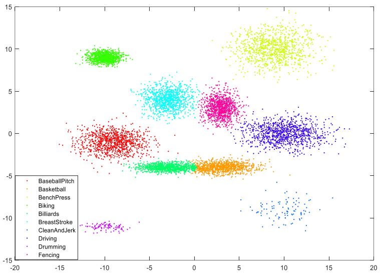


***Fig.13.*** *Clustering results on UCF101 dataset*

## 4. Comparison and analysis of experimental results

On UCF101 and HMDB51 datasets, this model is compared with the most advanced methods of behavior recognition. The results are shown in Table 3 and table 4. which are compared with different methods, including single-modal and multi-modal fusion models.

**Table 3** Comparison of accuracy between ours model and other models on HMDB51 dataset

| Method | Modality | Accuracy（%） |
|---|---|---|
| HDL[30][26] | VEDIO | 61.65 |
| Mandal[23] | RGB | 70.40 |
| MF+MVF[24] | OPTICAL | 71.28 |
| Two Stream[25] | RGB | 78.8 |
| Shou[27] | OPTICAL | 69.83 |
| Feichtenhofer[28] | VEDIO | 70.65 |
| CNN Two Stream+iDT[25] | RGB+BONE | 81.5 |
| TLE: Bilinear[29] | RGB+OPTICAL | 71.4 |
| I3D[26] | VEDIO | 77.8 |
| ST-ResNet[31] | RGB | 84.8 |
| Ours（IA-Net+CHMM） | RGB+BONE | 87.88 |

It can be seen from table 3 that for HMDB51 dataset, the behavior recognition rate of traditional methods HDL and mandal algorithms using single-modal recognition is only 61.65% and 70.40%, its performance is far lower than that of multi-modal fusion method, In addition, some of the latest models have significantly improved the recognition accuracy by optimizing the network structure. Compared with other methods, the accuracy of the two IA Net+CHMM fusion model in this paper is as high as 87.88%, which has achieved obviously remarkable results. This is because it makes full use of the spatio-temporal attention model to give different weights to different features and adopts the method of probabilistic reasoning for fusion, which is 3.08 percentage points higher than the most advanced ST-ResNet model in Table 3.

**Table 4** Comparison of accuracy between ours model and other models on UCF101 dataset

| Method | Modality | Accuracy（%） |
|---|---|---|
| Two Stream[25] | RGB | 88 |
| 3D CNN[32] | RGB | 82.3 |
| HDL[30] | VEDIO | 89.03 |
| STIAM **Error! Reference source not found.** | RGB+BONE | 94.9 |
| TLE: Bilinear[29] | RGB+OPTICAL | 95.6 |
| Mandal[23] | RGB | 95.7 |
| Shou[27] | VEDIO+OPTICAL | 95.97 |
| I3D[26] | VEDIO | 96.5 |
| Feichtenhofer[28] | VEDIO | 96.82 |
| MF+MVF[25][24] | OPTICAL | 97.34 |
| Ours（IA-Net+CHMM） | RGB+BONE | 97.92 |

On the UCF101 dataset, ours method is compared with some two flow methods and some latest methods, as shown in Table 4. Behavior recognition is more based on RGB, from 3D CNN to mandal Model recognition accuracy increased from 82.3% to 95.7%, indicating that RGB has strong advantages in feature extraction, so it provides an important model information for multi-modal data fusion. In addition to MF+MVF, this method is much better than most of the latest methods. Specifically, the result of this method is 3.02% higher than that of STIAM fusion method and 1.95% higher than that of Shou fusion method. Finally, by comparing the test results based on RGB and RGB+bone models, the highest accuracy are 95.7% and 94.9% respectively, which shows that the recognition accuracy based on multi-modal fusion model is higher than that based on single-modal. The accuracy of ours model is as high as 97.92%, indicating its advanced.

Ours model has achieved excellent results on both UCF 101 and HMDB51 dataset in this paper. The main reason is that the model introduces the ISE-Block into ResNet and GoogleNet networks to form two IA-Nets, which extract behavior features with different weights based on RGB and bone flow respectively and strengthens the key feature recognition ratio. After that, splicing LSTM network is conducive to the classification of long-time video behavior. Finally, CHMM probability method is used for data fusion, so as to improve the accuracy of human behavior recognition.

## 5. Conclusion

A fusion method combining two IA-Nets and CHMM is proposed in this article. Based on multi-modal data fusion, probabilistic reasoning and deep learning analysis, improved IA-ResNet50 and IA-GoogleNet networks are designed respectively, then CHMM is used to fuse the feature information of the two models. The complementary advantages of different modal features are used to improve the behavior recognition rate. However, the problem of interactive behavior recognition has not been deeply discussed and relevant research will be carried out in the future to improve the application scene and scope of ours model.

**Conflict of interest statement:**None declared.

## References

[1] Liu M Y, Yuan J S. 'Recognizing human actions as the evolution of pose estimation maps'. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA.2018: 1159–1168.

[2] Ying X, Chen J, Wang Y C, et al. 'Action recognition for depth video using multi-view dynamic images'. Informatics and Computer Science Intelligent Systems Applications of 2019 Information Sciences 480 2019: 287–304. journal homepage: www.elsevier.com.

[3] Moon G, Kwon H, Lee K M, et al. 'Integral Action: Pose-driven Feature Integration for Robust Human Action Recognition in Videos'[J]. 2020.

[4] Wang Y J, Zhang W, Liu Y Y. ‘Multi-scale feature fusion network for person re-identification’[J]. IET Image Processing,2020,14(17).

[5] Meng M, Hu J H, Gao Y Y, Ma yu liang. 'EEG Emotion Recognition Based on Normalized Mutual Information Channel Selection and Hybrid Deep Neural Network'[J]. Chinese Journal of sensors and Actuators, 2021,34(08):1089-1095.

[6] Hsueh Y L, Lie W N, Guo G Y. 'Human Behavior Recognition from Multi-view Videos'[J]. Information Sciences, 2020, 517: 275-296.

[7] Chaaraoui A A, Jr P, Florea F. 'Fusion of Skeletal and Silhouette-Based Features for Human Action Recognition with RGB-D Devices'[C]//3rd Workshop on Consumer Depth Cameras for Computer Vision (CDC4CV), ICCV 2013 Workshop. IEEE Computer Society, 2013.

[8] Lee I, Kim D, Kang S, et al. 'Ensemble deep learning for skeleton based action recognition using temporal sliding LSTM networks'[C]. 2017 IEEE International Conference on Computer Vision (ICCV).

[9] Lu L P, Zhang X Q. Data fusion method of multi-sensor target recognition in complex environment[J].Journal of Xidian University,2020,47(04):31-38.

[10] Hao M, Liu G Y,Xie D S. 'Hyperspectral face recognition with a spatial information fusion for local dynamic texture patterns and collaborative representation classifier'[J]. IET Image Processing, 2021,15(8).

[11] Wu H Q. Human behavior recognition based on attention mechanism and multimodal feature fusion[D]. Anhui University, 2019.5.42.

[12] Zhang P, Zhang J X. 'Research of Motion Intention Recognition Method Based on CNN Improved Serial Hybrid Network and Multi-Sensor Data Fusion'[J]. Chinese Journal of sensors and Actuators, 2021,34 (07):932-938.

[13] Feichtenhofer C, Pinz A, Zisserman A. 'Convolutional Two-Stream Network Fusion for Video Action Recognition'[C]. Computer Vision & Pattern Recognition. IEEE, 2016: 1933-1941.

[14] Donahue J, Hendricks L A, Guadarrama S, et al. 'Long-term recurrent convolutional networks for visual recognition and description'[C].2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[15] Hao H, Wang M, Tang Y, et al. 'Research on data fusion of multi-sensors based on fuzzy preference relations[J]. Neural Computing and Applications', 2019, 31(1): 337-346.

[16] Xiao Q K, Qin M Y, Guo P, et al. 'Multi modal fusion based on LSTM and a Couple conditional hidden markov model for chinese sign language recognition'[C]. Proceedings of 2019.4.IEEE access on Information system.

[17] Jie H, Li S, Gang S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99).

[18] Li X, Wang W, Hu X, et al. 'Selective Kernel Networks'[J]. CVPR, 2019.

[19] Zhang H, Wu C, Zhang Z, et al. 'ResNeSt: Split-Attention Networks'[J]. CVPR, 2020.

[20] Ding Ran, LIN Chunli, XIA Yu. 'Design of behavior recognition system based on CHMM'[J]. Journal of Liaoning University of science and technology. 2010, 33(05).

[21] Kuehne H, Jhuang H, Stiefelhagen R, et al. 'HMDB51: A Large Video Database for Human Motion Recognition'[J]. Springer Berlin Heidelberg, 2013.

[22] Soomro K, Zamir A R, Shah M. 'UCF101: A dataset of 101 human actions classes from videos in the wild'. arXiv: 1212.0402, 2012.19.

[23] Feichtenhofer C, Fan H, Malik J, et al. 'Slowfast networks for video recognition'[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.

[24] Xie S, Girshick R , Tu Z. 'Aggregated residual transformations for deep neural networks'. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[25] Tran D, Bourdevl L, Fergus R, et al. 'Learning spatio temporal features with 3D convolutional networks'. Proceedings of 2015 IEEE International Conference on Computer Vision(ICCV). Santiage, Chile.2015: 4489-4497.

[26] Shou Z, Lin X, Kalantidis Y, et al. 'Dmc-net: Generating discriminative motion cues for fast compressed video action recognition'[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1268-1277.

[27] Meng L L, Zhao B, Chang B, et al. 'Interpretable spatio temporal attention for video action recognition'. arXiv: 1810.04511,2018.

[28] Wang P C, Li Z Y, Hou Y H, et al. 'Action recognition based on joint trajectory maps using convolutional neural networks'. Proceedings of the 24th ACM International Conference on Multimedia Conference. Amsterdam, the Netherlands. 2016:102–106.

[29] Diba A, Sharma V, Gool L. 'Deep temporal linear encoding networks'[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2017: 2329 -2338.

[30] Jaouedi N, Boujnah N, Bouhlel M S. 'A new hybrid deep learning model for human action recognition'[J]. Journal of King Saud University-Computer and Information Sciences, 2020, 32(4): 447-453.

[31] Karpathy A, Toderici G, Shetty S, et al. 'Large-scale video classification with convolutional neural networks'[C]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725–1732.

[32] Bazzani L, Larochelle H, Torresanil L. 'Recurrent mixture density network for spatio temporal visual attention'. arXiv: 1603.08199,2016.

**ZHANG Yinhuan** (1984- ), female, the Doctor of Philosophy, is the lecturer and her research fields are mainly image and video processing and recognition, deep learning, education and teaching, etc.

**XIAO Qinkun** (1974- ), male, the Post Dr., is the professor and his research fields are artificial intelligence, pattern recognition, image and video processing and recognition, etc.

**CHU Chaoqin** (1990- ), male, the Doctor of Philosophy, are mainly engaged in the research of gesture recognition, pose estimation, etc.

**XING Heng** (1996- ), male, the Master of Philosophy, are mainly engaged in the research of behavior recognition, data fusion and pose estimation.