Multimedia Tools and Applications (2023) 82:20251–20268
https://doi.org/10.1007/s11042-023-14355-8

Research on human behavior recognition in video based on 3DCCA

Hong Zhao[1] · Juan Liu[1] · Weijie Wang[1]

Received: 11 January 2022 / Revised: 15 August 2022 / Accepted: 2 January 2023 /
Published online: 24 January 2023

Abstract

Human behavior is an important part of video content. Therefore, the effective recognition of human behavior in the video has attracted extensive attention. In order to solve the problem that the key features are not prominent and the accuracy rate is not high in the existing methods of human behavior recognition in video. This paper proposes a three-dimensional convolutional neural network fusing channel attention (3DCCA) model feature extraction method. Mean normalization is presented for the preprocessing of RGB video frames. The three-dimensional convolution (3DCNN) is presented for the spatiotemporal features extraction of the inputs clips. The channel attention(CA) is used to select features that are more critical for current behavior recognition from all features. Softmax classifiers to achieve in the Classification and Identification of the human behavior in video. The training results on UCF101 and HMDB51 public datasets show that the algorithm can make better use of the original information in the video, extract more effective features, correctly detect human behaviors and actions and show stronger recognition ability to the algorithm compared with other commonly used human behavior feature extraction and recognition methods.

Keywords Deep learning · 3DCNN · Spatiotemporal feature · Attention mechanism · Behavior recognition

1 Introduction

With the rapid development of the Internet and multimedia technology, a large number of videos are shared through the network all the time, and the video information grow exponentially. How to quickly and accurately identify the content of these videos has become a research hot spot currently [11, 28, 35]. Human behavior recognition in videos is an important

✉ Juan Liu
1507447025@qq.com

[1] School of Computer and Communication Technology, Lanzhou University of Technology, Lanzhou 730050, China

part of video content recognition, which is widely applied in such fields as human-computer interaction, intelligent video monitoring, medical diagnosis, abnormal behavior detection and sports analysis [7, 8, 16].

Therefore, accurate and timely recognition of human behavior in the video plays an important role in the development of artificial intelligence and computer vision [9]. The contributions of this study are as follows:

- In order to achieve efficient and accurate behavior recognition, a three-dimensional convolutional neural networks and channel attention (3DCCA) human behavior recognition model is proposed. The model is divided into two stages to extract the features of human behavior in the video.
- In the first stage, the standard human behavior recognition data set is decomposed into video frame sequence, and the temporal and spatial features of human behavior in continuous video frame sequence are extracted effectively by using three-dimensional convolutional neural networks (3DCNN).
- In the second stage, the channel attention is used to extract the key features of human behavior recognition from the features extracted in the first stage. The convolutional feature map is compressed in time and space by global pooling, and important features are extracted from the compressed feature map by the multi-layer perceptron (MLP).Finally, the convolution feature map is given different weight values.
- In these two stages, the accuracy of human behavior recognition is improved by extracting the spatiotemporal features and key features of human behavior in the video.

The rest of this paper is organized as follows. The literature review is given in Section 2. The framework and the methodology are introduced in Section 3. The experimental results are discussed in Section 4. Finally, the conclusion and future work are given in Section 5.

2 Related works

According to difference of feature extraction methods, the human behavior recognition algorithms in the video can be divided into traditional algorithms based on hand-crafted features and deep learning algorithms based on end-to-end automatic learning features [14, 29, 42].

The feature of hand-crafted is usually used to describe the local spatiotemporal changes of human motion in the video, such as space-time interest points (STIP) proposed by Laptev et al. [13], which detects the spatiotemporal range of human movement and calculates its scale-invariant space-time descriptor to realize the recognition of human walking motion in the scene with occlusion and dynamic background, Paul et al. extended scale-invariant feature transform (SIFT) to 3D-SIFT [19], and used sub-histogram to identify human behavior based on the local time and space information encoding of human movements in the video. Klaser et al. extended traditional histograms of oriented gradients (HOG) in the time domain direction to form 3D-HOG [12], and adopted the uniform distribution surface of regular polyhedra as histogram angle to avoid singularity problem and perform human behavior recognition on standard data sets such as KTH. Wang et al. proposed improved dense trajectories (IDT) [25], to densely sample local blocks in different scales from each frame of video, and then track these local blocks in the dense optical flow field for human behavior recognition.

In deep learning algorithms based on end-to-end automatic learning features, two-dimensional convolutional neural networks(2DCNN) and 3DCNN are often used. In 2DCNN algorithms, Simonyan et al. proposed the two stream CNN model for human behavior recognition in video [21]. In this model, CNN extracts features through the original single-frame RGB image and the dense optical stream image of the video frame were used to independently, and finally the output results were fused for human behavior recognition. Yue-hei Ng J et al. combined long and short term memory (LSTM) with CNN [34], extracted temporal features with LSTM and spatial features with CNN, and fused them to recognize human behavior. Wang et al. proposed temporal segment networks (TSN) [26], which combined sparse time sampling strategy and video level supervision to efficiently extract the features of the entire video, and improved the ability of long-range video modeling. Qing et al. proposed a two-stream heterogeneous network based on the basic structure of the traditional two-stream network [30], filter weight grafting is carried out for invalid filters in DenseNet network and BNIception network based on filter grafting technology, then the improved DenseNet network and BNIception network are used to form a two-stream heterogeneous network to extract spatial and temporal information of human behavior.

In 3DCNN algorithms, Wang et al. proposed a three-dimensional convolutional boltzmann machine (3DCRBM) that can extract features from the adjacent frames of the original RGB and the adjacent frames of the depth graph [24], which uses the features obtained after unsupervised data training in the deep belief network (DBN) as the input of CNN for human behavior recognition. Ji et al. proposed a 3DCNN model to realize human behavior recognition [10]. The model generates multi-channel information (gray level, horizontal gradient, vertical gradient, horizontal optical flow and vertical optical flow) among consecutive video frames, and fuses the multi-channel information for human behavior recognition. Tran et al. proposed a three-dimensional convolutional neural networks (C3D) for large-scale data sets [22]. C3D can process multiple consecutive RGB video frames at a time, and the extracted features have strong versatility. Carreira et al. proposed the inflated three-dimensional convolutional neural networks (I3D) [4], which applyed the 2D model parameters trained on Imagenet data set to 3D model, and used 64 frames receptive field for human behavior recognition. Qiu et al. proposed pseudo-3d residual Net (P3D ResNet) [18]. 2D space convolution and 1D time convolution were used to form different pseudo-3D blocks at different positions in the ResNet network, which enhanced the network's ability to identify human behavior by enhancing the diversity of model structure.

The above human behavior recognition methods have their own characteristics. References [12, 13, 19, 25] extract features through manually designed operators, which can achieve good recognition effect. However, the hand-crafted features only reflect the local information of video content, and the generalization ability is weak. Moreover, the design of operators requires a lot of professional knowledge and experience, and most of the extracted features are shallow features, therefore, these methods are only applicable to specific scenes. References [21, 26, 34] used 2DCNN algorithms for human behavior recognition, which achieved a high recognition rate, but the video content was mainly considered from the image level, lacking the extraction of temporal features in the video content. Secondly, the computation of optical flow in Two Stream network and the input of the whole video in TSN will greatly increase the computational complexity. References [4, 10, 22, 24, 30] can effectively extract the spatiotemporal features of human behavior in video, and achieve high recognition rate in human behavior recognition. However, the important difference of various features in local video region is not considered, which affects the recognition effect. Moreover, large-scale behavior recognition data set as the input of neural network makes the model training cost higher.

In summary, in view of the weak generalization ability and inconspicuous key features in existing human behavior recognition methods in videos, channel attention(CA) was introduced based on the three-dimensional convolutional neural network [27]. A 3DCCA model that uses a three-dimensional convolutional neural network to fuse channel attention for human behavior recognition.

3 Establishment of 3DCCA model

Figure 1 shows the 3DCCA model framework, which mainly includes data preprocessing, local feature extraction, key feature extraction and behavior recognition.

In Fig. 1, in data preprocessing, FFMPEG is used to decompose multiple video segments containing human actions into video frame sequence, and then normalize the video frame sequence by cropping the central picture, calculating the mean value and reducing the mean value, etc., so as to improve the accuracy of model training. In the training stage, the 3D convolution fusion attention mechanism is used to extract the features of the processed data, which is subdivided into three-dimensional convolution to extract the local features of continuous video frames, and the channel attention extracts the important features of continuous video frames. In the verification stage, through continuous optimization of the parameters obtained in the training process, the optimal 3DCCA model is selected as the model in this paper. In the test stage, the classifier Softmax is used to complete human behavior recognition. Finally, the performance of 3DCCA model is evaluated by accuracy rate.

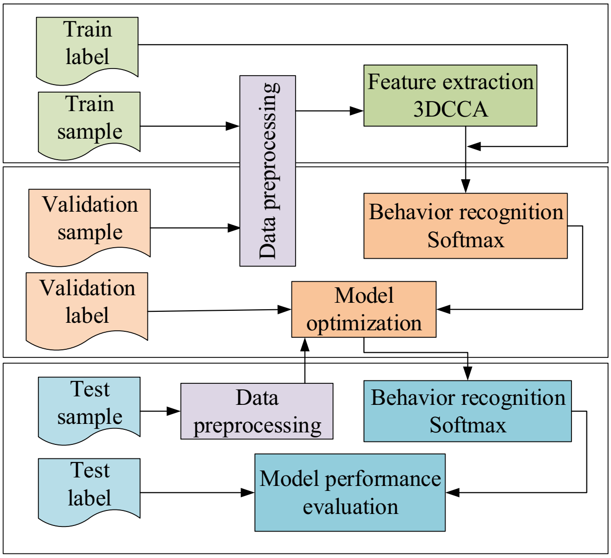

Fig. 1 3DCCA model framework

3.1 Data preprocessing

In this paper, we use UCF101 and HMDB51 standard video human behavior recognition data set. The data set is composed of multiple video segments containing human actions. Each video segment contains a specific human action and the resolution is relatively high, and it does not pay too much attention to the characteristics of individuals in the video, in order to improve the performance of the model, the following preprocessing is required for the video human behavior data set:

- Video segment to video frame sequence. The FFMPEG platform is built, and each video segment is intercepted at the frame rate of 5fps and saved as a JPG video frame sequence.
- Data set division. 75% of the video frame sequences are randomly selected as the training set, and the remaining 25% of the video frame sequence are used as the test set.
- Data set standardization. From the divided training and testing video frame sequence, i is taken as the starting frame, and consecutive $f_i(i = 1, 2, \ldots, n)$ video frames are selected to form a video clip. If i < 16, the empty list clip and the index$i$of start frame is returned. On the contrary, the starting frame with index i is selected randomly, among them, $0 < i \le f_i - 16$, 16 consecutive RGB video frames from $i$as the clip list are selected. Since the human body in the video frame is in the middle area of the frame, the size of all the video frames is 112 × 112 through the center cropping, and the average value is calculated. Finally, the video frame in the clip is subtracted from the calculated average value.

Eg: First, the resolution of the original video frame is 320 × 240, we also use jittering by using random crops with a size of 3 × 16 × 112 × 112 of the input clips during training. Then use python to calculate the average value of the three channels of input clips as [121.643, 124.514, 126.064]. Finally subtract the average value from the input clips of training set, validation set and test set.

- Data set tagging. The video frames in clip were annotated to get the corresponding label of the data set.

3.2 Feature extraction

The process of human behavior recognition is shown in Fig. 2.

As shown in Fig. 2, input is clip data. The data is extracted feature through a network structure of two module 1 and three module 2 in series, in which module 1 is one convolution, one attention layer and one pooling layer (green dashed box) successively, while module 2 is two convolution layers, one attention layer and one pooling layer (yellow dashed box) successively. The convolution in module 1 and module 2 is three-dimensional convolution, which extracts the time and space features of human behavior in video frame sequence. The channel attention layer in the two modules increases the feature sensitivity between the channels of the convolution feature map to extract the key features of human behavior recognition. The pooling in the module is three-

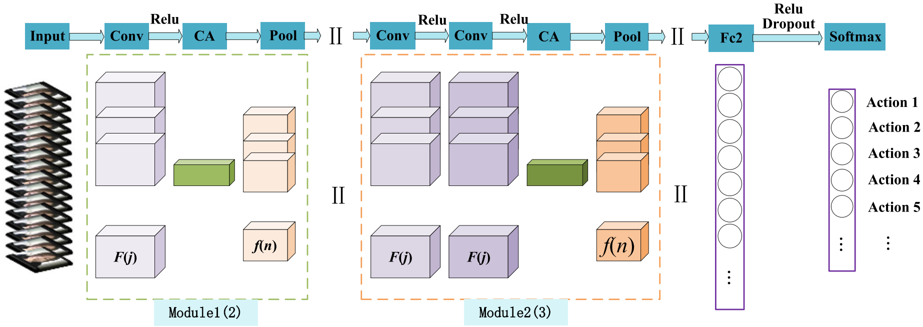

Fig. 2 Human behavior recognition process

dimensional pooling, which is used to gradually reduce the scale of the feature map of the attention layer, reduce the calculation of the parameter amount in the 3DCCA model, and reduce the difficulty of model training. Human behavior recognition uses the Softmax function.

3.2.1 Convolution layer

The video segment is composed of video frames. The recognition of human behavior in video frames is mainly realized by representing the local features of human actions and human scenes, such as edges, spots, color changes and textures. There is a certain hierarchical structure between these features. Convolution neural network can extract the local features represented by human actions and human scenes. Therefore, 3DCNN is selected for local feature extraction, the preprocessed clip is used as the input of 3DCNN. The size of clip is $c \times l \times h \times w$, in which c is the number of channels with a value of 3, $l$is the length of video frame, with a value of 16, and h and w are the height and width of frame, with the values of 112. The convolution kernel of $d \times k \times k$ is used for convolution operation, where d is the time depth, with a value of 3, and k is the space size, with a value of 3. The output feature map is connected with several adjacent video frames in the Input layer to extract the local feature information of human motion in the video. The process of 3D convolution feature extraction in module 1 and module 2 is shown in Fig. 3, and the calculation of feature value is shown in Eq. (1).

$$F_{ij}^{xyz} = \mathrm{Re}lu\left(b_{ij} + \sum_{m} \sum_{k=0}^{K_{i-1}} \sum_{q=0}^{Q_{i-1}} \sum_{d=0}^{D_{i-1}} w_{ijm}^{kkd} F_{(i-1)m}^{(x+k)(y+q)(z+d)} \right) \tag{1}$$

Among them, F_{ij}^{xyz} means that the pixel points contained in the j-th (j = 1,2,...512) feature map in the i-th (i = 1,2,...,8) convolutional layer are in (x, y, z) position feature value, Relu represents the activation function of the convolutional layer, b_{ij} represents the offset of the j-th feature map in the i-th convolutional layer, m represents the number of feature maps of the i-1-th convolutional layer, K_{i-1} and $Q_{i-1}$are the height and width of the kernel in the i-1-th convolutional layer, respectively, where D_{i-1} is the size of the 3D kernel along the temporal dimension, w_{ijm}^{kkd} represents the volume weight matrix connected to the m-th feature map in the i-1-th convolutional layer.

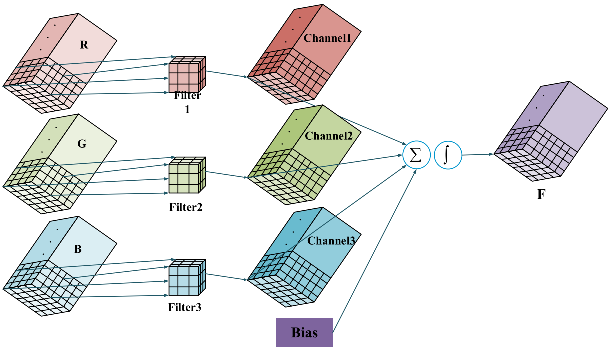

Fig. 3 Feature extraction process of 3D convolutional network

3.2.2 Attention layer

After convolution operation, each channel of the feature map F_i is regarded as a feature extractor, and the importance of different channel features in human behavior recognition is different. Channel attention layer will realize the selection of key channel features representing human behavior, and the process of calculating the attention weight of each channel is shown in Fig. 4. Among them, $F_i$represents the feature map of the *i-th* convolutional layer, F^c_{avg} represents the global average pooling feature map, $F^c_{\max}$ represents the global maximum pooling feature map, $F1$ and $F2$ represent the feature map obtained after MLP respectively, $M_{\text{c}}(F_i)$ represents the channel attention feature map. The steps to calculate the attention weight of each channel in the feature map are as follows:

- Since each convolution operation only operates in a local region and cannot provide information outside the region, global pooling is adopted to compress the spatial and temporal dimensions of the feature graph so that each feature graph is a one-dimensional array to achieve global information extraction. In this paper, the global average pooling and global maximum pooling are used to compress the spatial and temporal dimensions of the feature graph F_i into unit length, and the corresponding global average pooling feature graph F^c_{avg} and global maximum pooling feature graph $F^c_{\max}$ are obtained. The calculation of the feature map is shown in Eq. (2) and Eq. (3). Among them, F_i represents the feature map of the i-th convolutional layer, and H, W and L respectively represent the spatial and temporal dimensions of F_i.

$$F^c_{avg} = \frac{1}{HWL} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{l=1}^{L} F_i(h, w, l) \tag{2}$$

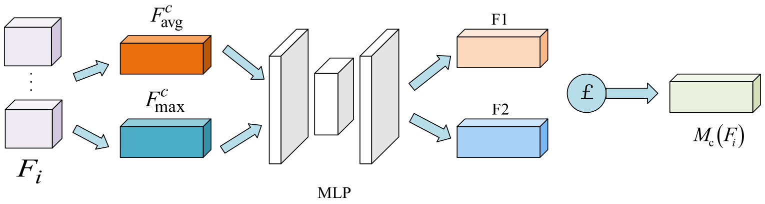

Fig. 4 Key feature extraction process

$$F_{\max}^{c}=\frac{1}{HWL}\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{l=1}^{L}F_i(h,w,l) \tag{3}$$

- The key features of F_{avg}^{c} and $F_{\max}^{c}$ are extracted through the shared network composed of MLP. After MLP, the $F1$ and $F2$ channel feature graphs are obtained. The Sigmoid function is used to normalize F1 and F2 to obtain the $M_{\text{c}}(F_i)$ weight of CA layer. The calculation is shown in Eq. (4). Among them δ represents the Sigmoid normalization function, $W_0 \in R^{C/r \times C}$ and $W_1 \in R^{C \times C/r}$ represent the weight matrix of MLP, r represents the reduction rate, C represents the number of channels of the *i-th* convolutional layer, and $F1$ and $F2$ respectively represent the feature maps obtained after MLP.

$$\begin{aligned}M_{\text{c}}(F_i)&=\delta\left(MLP\left(F_{avg}^{c}\right)+MLP\left(F_{\max}^{c}\right)\right)\\&=\delta\left(W_1\left(W_0\left(F_{avg}^{c}\right)\right)+W_1\left(W_0\left(F_{\max}^{c}\right)\right)\right)\\&=\delta(F1+F2)\end{aligned} \tag{4}$$

- The channel attention weight $M_{\text{c}}(F_i)$ is obtained through the previous calculation, A_i is the feature map of key features extracted by CA at the *i-th* convolutional layer. The calculation of feature is shown in Eq. (5). Among them, ⊗ indicates that the attention weight of the channel is multiplied by the element in the corresponding feature map F_i.

$$A=M_{\text{c}}(F_i)\otimes F_i \tag{5}$$

3.2.3 Pooling layer

After the calculation of attention layer, the feature map A_i is reduced by three-dimensional pooling. The pooling layer of module 1 and module 2 both use maximum pooling, as shown in Fig. 5, and the calculation of the feature value of the pooling layer is shown in Eq. (6). Among them, $P_{mn}^{\partial\beta\gamma}$ indicates that the pixel points contained in the *n-th* (n = 1,2,...,512) feature map in the *m-th* (m = 1,2,...,5) pooling layer are in (∂, β, γ) position feature value, p, q and k respectively represent the sliding steps in the space and time direction of the three-dimensional pooling kernel, and X, Y and Z represent the size of the three-dimensional pooling kernel space and time dimensions, respectively.

$$P_{mn}^{\partial\beta\gamma}=\max_{0<x<X,0<y<Y,0<z<Z}\left(A_i\left(_{m-1,n-1}^{\partial\times p+x,\beta\times q+y,\gamma\times k+z}\right)\right) \tag{6}$$

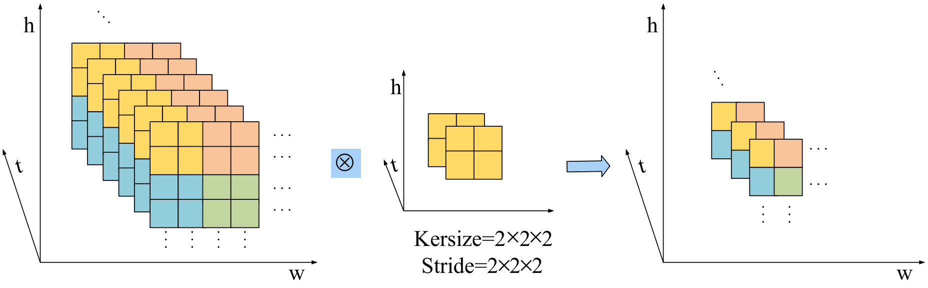

Fig. 5 3D pooling feature extraction process

3.3 Human behavior recognition

After data preprocessing, the video frame sequence is input into the 3DCCA network built in this paper, and then the key features which are conducive to human behavior recognition are extracted through module 1 and module 2 in Fig. 2. Finally, human behavior is recognized by Softmax in Fig. 2. Dropout layer is added to the full connection layer to prevent over fitting during model training. The Softmax classification function is shown in Eq. (7). Among them, w_i represents the weight matrix from the Fc2 layer to the out layer, b_i represents the corresponding bias, and d_{ij} represents the output vector of Fc2.

$$y_i = Soft\max\left(w_i d_{ij} + b_i\right) \tag{7}$$

4 Experimental design and result analysis

4.1 Experimental environment

In order to verify the effectiveness of the model proposed in this paper, using the Siteng IW4200 CPU with dual Intel®Xeon®E5-2600 V3/V4 servers, with four NVIDIA Tesla Kepler GPU, an operating system of 64-bit Windows 10, a programming language of Python, and a modeling tool of Google TensorFlow.

4.2 Experimental data

4.2.1 UCF101 data set

UCF101 data set [1] is a human behavior recognition data set with a large number of action categories and samples at present. It contains 13,320 videos from 101 types of human actions. Each type of human action is performed by 25 people, and each person does 4–7 groups. The video content has a great diversity in motion collection, which makes it one of the data sets with higher recognition difficulty. The data set consists of five categories, as shown in Fig. 6.

4.2.2 HMDB51 data set

HMDB51 data set [2] is another commonly-used data set to identify human behavior. It contains 6766 videos from 51 types of human actions. With high similarity between classes, it is currently a difficult data set for recognition. The data set mainly includes the following five categories:

- General facial movements: smiling, laughing, chewing, etc.
- Facial interaction with the object: smoking, eating, drinking, etc.
- General body movements: cartwheels, clapping, climbing stairs, etc.
- Body and object interaction: comb hair, draw sword, dribble, etc.
- Human interaction: fencing, hugging, shaking hands, etc.

The UCF101 and HMDB51 human behavior recognition data sets were preprocessed as the input of 3DCCA model as shown in Section 3.1 of 3.

4.3 Network model establishment and hyper-parameter setting

The 3DCCA model is established in Tensorflow as shown in Fig. 1. The hyper-parameter of the 3DCCA model is optimized through multiple attempts, and the parameter values are shown in Table 1.

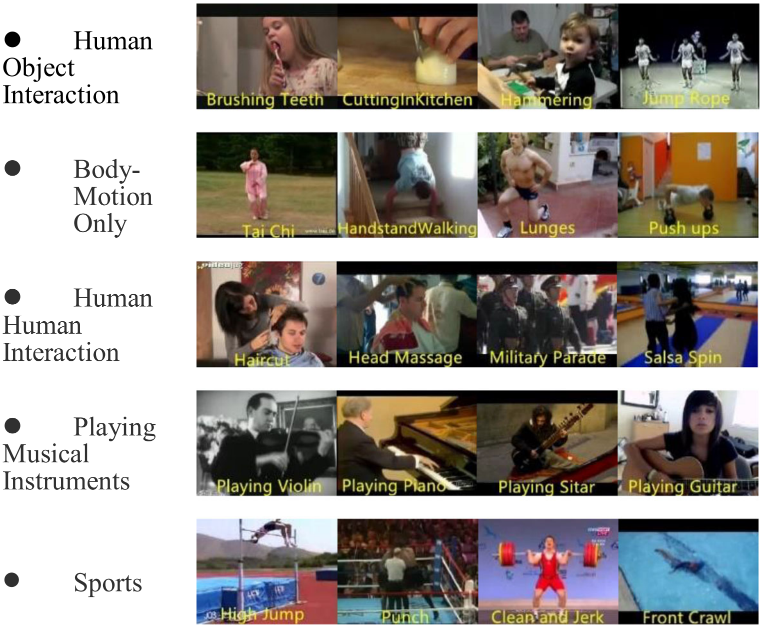

Fig. 6 Examples of 5 types of actions in the UCF101 data set

Table 1 Model hyper-parameter setting

Hyper-parameter	value
Video frames	16
Loss	Cross_Entropy
Learning_rate	0.0001
Batch_size	20
Dropout	0.5
Optimizer	Adam
Reduction_ratio	16

4.4 Ablation experiment

In the same experimental environment, UCF101 and HMDB51 data sets were used as experimental data. First, the video segments were decomposed into video frame sequence, and then the processed video frame sequence were typed into the 3DCNN, 3DCNN-CA, Pre-3DCNN and 3DCCA human behavior recognition training networks, among them, 3DCNN and Pre-3DCNN represent different training situations in the same three-dimensional convolutional neural network model, and 3DCNN represents training the three-dimensional convolutional network model from scratch. Pre-3DCNN indicates that the network model trained on Sports_1m data set is transferred to UCF101 and HMDB51 data sets for training. 3DCNN-CA and 3DCCA respectively represent different training situations after fusing channel attention on the basis of 3DCNN and Pre-3DCNN, 3DCNN-CA represents trained from scratch for the three-dimensional convolutional network fusion channel attention. 3DCCA represents model the network trained by fusing channel attention in the transferring network. Finally, the experimental comparison is performed. The experimental results are shown in Table 2.

It can be seen from Table 2 that the accuracy rate of 3DCNN-CA in human behavior recognition is increased by 2.9% and 3.9% compared with 3DCNN, and the accuracy rate of 3DCCA in human behavior recognition is increased by 0.6% and 7.1% compared with pre-3DCNN, respectively. The results show that CA has a role in improving the model recognition rate. HMDB51 data set is used to test the calculation performance of 3DCCA model, and the experimental results are shown in Table 3.

It can be seen from Table 3 that compared with 3DCNN, the accuracy rate of behavior recognition of 3DCNN-CA is increased by 3.9%, but the calculation time for each picture is increased by 0.6 ms, and the calculation time for each epoch is increased by 34.2 s. similarly, the recognition accuracy rate of 3DCCA is increased by 7.1% compared with pre-3DCNN, the calculation time of each picture is increased by 0.5 ms, and the calculation time of each epoch is increased by 31.3 s, indicating that the time overhead has increased after the introduction of CA, but the accuracy rate has been greatly improved

Table 2 Model accuracy in UCF101 and HMDB51 data sets

method	UCF101(%)	HMDB51(%)
3DCNN	77.5	36.2
3DCNN-CA	80.4	40.1
Pre-3DCNN	90.3	52.4
3DCCA	90.9	59.5

Table 3 Computational reasoning time of 3DCCA model

method	HMDB51(%)	Inference Time per epoch/image (s/ms)
3DCNN	36.2	277.6/4.3
3DCNN-CA	40.1	311.8/4.9
Pre-3DCNN	52.4	280.0/4.4
3DCCA	59.5	311.3/4.9

4.5 Analysis of experimental results

4.5.1 Analysis of experimental results of UCF101 data

In this paper, by training the human behavior recognition model on the UCF101 training set, the accuracy and the loss rate of human behavior recognition in the training process can be obtained as the curve of the change of the model iteration number epoch as shown in Fig. 7. It shows that in the 26th epoch, compared with the 3DCNN and pre-3DCNN without CA, the accuracy rate of the 3DCNN-CA and the 3DCCA for human behavior recognition has a certain degree of improvement, the loss rate has dropped to a certain extent, which proves that CA can improve the accuracy rate of human behavior recognition

Secondly, the trained model is tested on the test set of the data set, and the confusion matrix as shown in Fig. 8 is obtained, among them, the abscissa of the confusion matrix represents the predicted label of the test sample, and the ordinate represents the real label of the test sample.

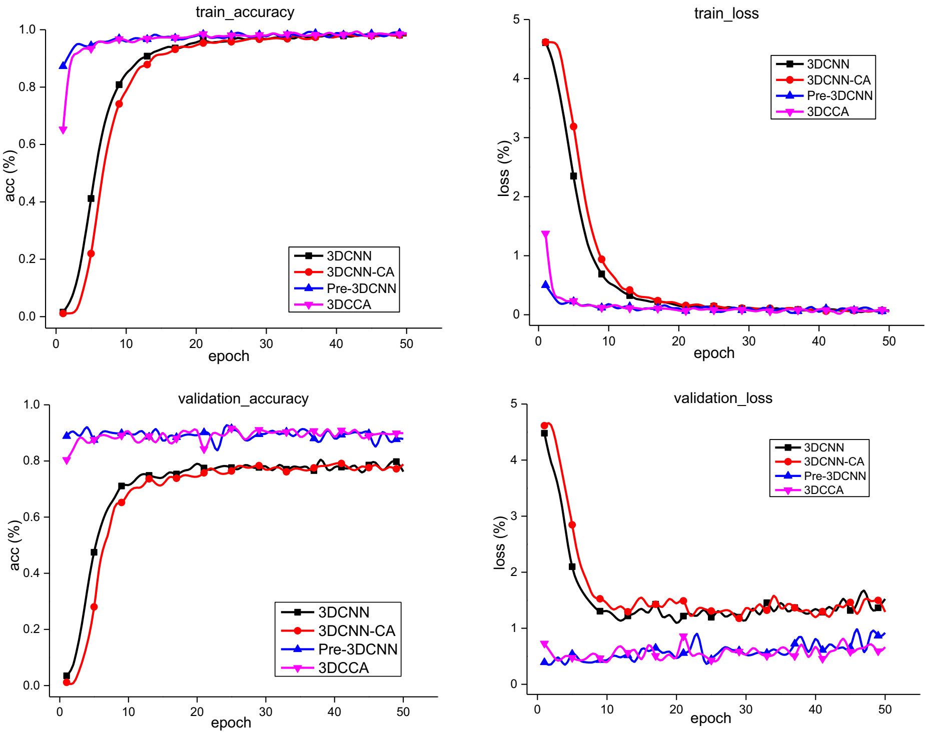

Fig. 7 Accuracy and loss rate on UCF101 data set

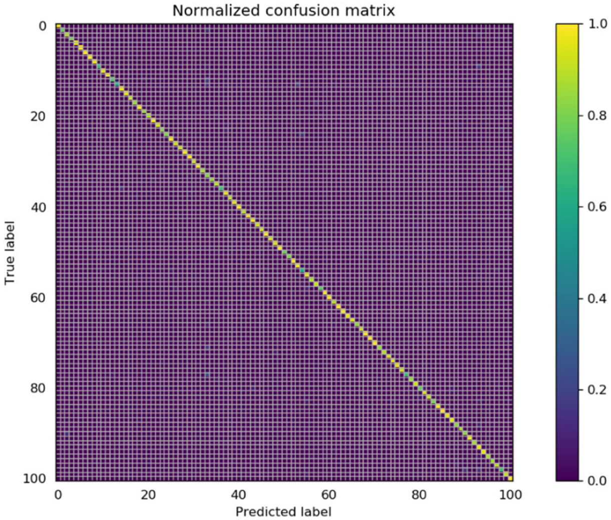

Fig. 8 Confusion matrix on UCF101 data set

The label is from 0 to 100, and each number represents a kind of behavior. The scale on the right side of the picture represents the lighter the color (the more the color is), the higher the accuracy rate of behavior classification. The recognition rate of 3DCCA in UCF101 data set is higher as a whole. Among them, the error rate of weightlifting, candle blowing, floor cleaning and other categories is higher, these categories of videos are relatively fuzzy and the background is chaotic, which makes the recognition rate lower.

4.5.2 Analysis of experimental results of HMDB51 data

Similarly, by training the human behavior recognition model on the HMDB51 training set, the accuracy and loss rate of human behavior recognition in the training process can be obtained as the curve of the change of the model iteration number epoch as shown in Fig. 9. According to Fig. 9, when the 36th epoch, compared with 3DCNN and pre-3DCNN without CA, 3DCNN-CA and 3DCCA for human behavior recognition, both of which showed a certain degree of improvement in accuracy rate and a certain degree of decrease in loss rate, which further proved that CA could improve the accuracy rate of human behavior recognition

Secondly, the trained model is tested on the test set of the data set, and the confusion matrix as shown in Fig. 10 is obtained, among them, the abscissa of the confusion matrix represents the predicted label of the test sample, and the ordinate represents the real label of the test sample. The labels are from 0 to 100, and each number represents a kind of behavior. The scale on the right side of the picture represents the lighter the color is (loser to 1.0), the higher the accuracy rate of behavior classification. Compared with 3DCCA on UCF101 data set, the recognition accuracy rate on HMDB51 data set is lower, which is mainly due to the high similarity between classes. Among them, golf, push-up, sit ups and other categories have higher recognition rate, and these types of actions have a larger range, and have lower similarity with other actions, which makes the recognition accuracy higher.

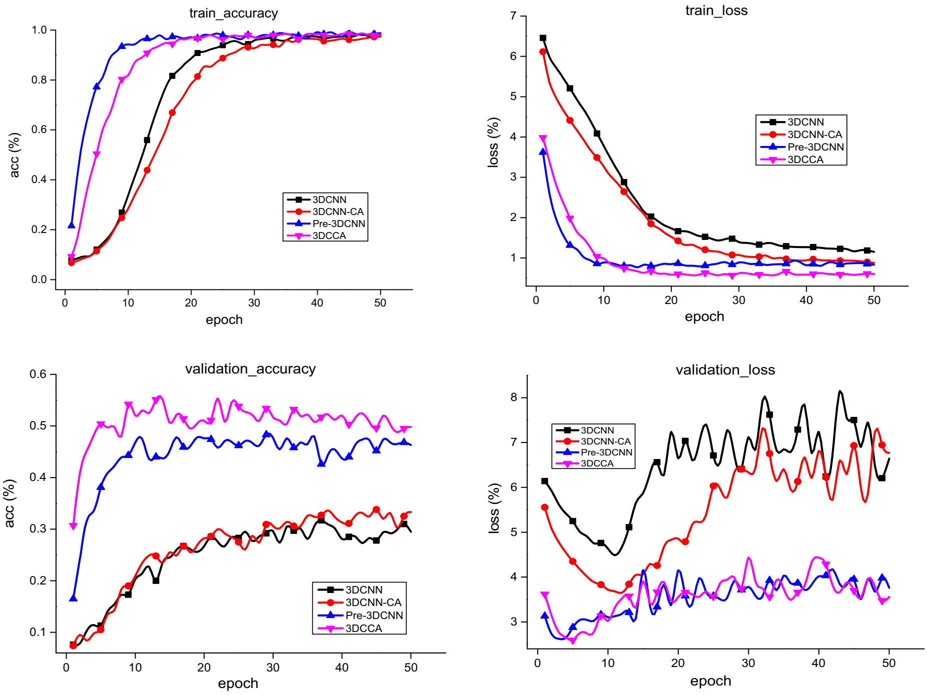

Fig. 9 Accuracy and loss rate on HMDB51 data set

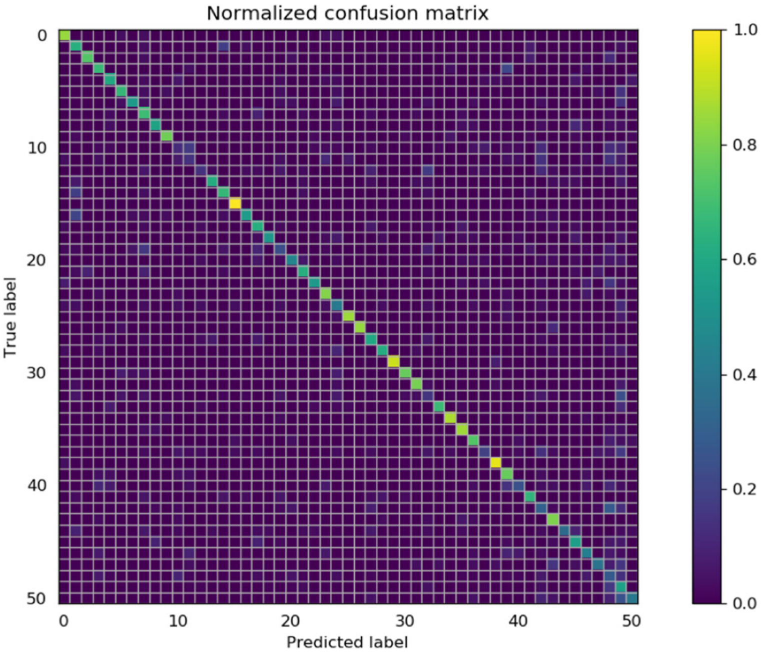

Fig. 10 Confusion matrix on HMDB51 data set

4.6 Model performance comparison

The proposed method is compared with existing state-of-theart methods, including improved dense trajectories (IDT) [3], 3D convolution [4, 22], multilayers features [41], CNN/LSTM models [15–30], attention-based models [5, 15, 17, 32, 37]. Tables 4 display the performance comparisons on UCF11, HMDB51 and UCF101.

According to the comparisons between deep-learning methods and traditional approaches, deep-learning models introduced in recent years have exceeded traditional approaches (performance gaps are~ 8% on HMDB51 and ~ 5% on UCF101) because deep-learning models can capture ample semantic information.

In Table 4, benefit from the use of spatial attention in Soft Attention [20] and VideoLSTM [15], attention models are capable of searching salient region in each frame adaptively. However, they just employ single spatial attention to select the key spatial information, thus exhibiting weak temporal stress. To this end, the spatial-temporal dual-attention network (STDAN) [40], which is mainly composed of feature extraction, attention and fusion modules, is designed. Although STDAN achieves good performance, it has high computational complexity due to the LSTM structure. Our model can obtain better action recognition results by inputting complete and continuous video frames at one time. Although the work [41] obtains an improvement by the utilization of proper layer combinations, the features are only from convolutional layers thus lack of key feature description. Therefore, our 3DCCA still performs better due to the key features from CA layer. In addition, our model outperforms some methods [31, 39], which performed a residual learning through stacked residual spatial-temporal attention blocks or introduced attention mechanism to strengthen the focus on discriminative foreground targets, but both of them omit complementary discriminative multi-level features. Compared with the two-stream methods [21], which jointly employ RGB and optical flow as input, with less computing resources, our 3DCCA still exhibits a

Table 4 Recognition accuracy rate of different models on UCF101 and HMDB51 data sets

Method	UCF101 (%)	HMDB51 (%)
DT+MVSV**Error! Reference source not found.**	83.5	55.9
Two stream CNNs [21]	88.0	59.4
C3D [22]	82.5	44.0
I3D [4]	49.8	84.5
Zhang et al. [36]	87.5	55.3
ML-HDP+iDT+TDD [23]	89.3	N.A.
Multi-layers features [41]	86.5	53.2
Spatial–temporal relations [37]	75.8	45.4
VideoLSTM [15]	79.6	43.3
Residual STAB [31]	86.0	54.4
JSTA [32]	88.6	59.8
RSTAN [5]	80.2	53.4
MFA [39]	87.6	55.1
TCLSTA (Frame + STA) [17]	85.9	54.8
Recurrent Attention [38]	75.8	45.4
IP-LSTM [33]	88.5	58.6
STDAN [40]	87.7	57.0
T-VLAD [6]	89.0	N.A
two-stream heterogeneous graft network [30]	89.3	N.A
3DCCA	90.9	59.5

distinct improvement and performs better than the classical dual-stream models and attentive 3DCNN-LSTM model [32] which leads a higher precomputation burden. References [4, 22] can effectively extract the spatiotemporal features of human behavior in video, and achieve high recognition rate in human behavior recognition. However, the important difference of various features in local video region is not considered, which affects the recognition effect. Yu et al. proposed the IP-LSTM model [33], which enhances the representation ability of actions by enhancing feature propagation, and experiments show that the LSTM architecture is more suitable for human action recognition in videos.However, for the action category with fast action features, IDT features can not be well extracted to effectively describe the action. T-VLAD [6] encoding approach lacks local temporal information due to which it did not perform better than 3DCCA, in addition, UCF101 contain diverse action classes and end-to-end training instead of formation of codebook is better Qing et al. proposed a two-stream heterogeneous network based on the basic structure of the traditional two-stream network [30], which is relatively simple, and the complexity of the algorithm needs to be optimized. In a word, our 3DCCA performs the better among the models listed in Table 4.

5 Conclusion

In order to solve the problems of weak generalization ability and unprominent key features of existing human behavior recognition methods in video, this paper constructs a 3DCCA model of three-dimensional convolution neural network fusion channel attention for video human behavior recognition research. The model mainly includes two parts: 1) extracting the spatio-temporal features of human behavior in video using the original RGB video frame; 2) the key features of the extracted spatiotemporal features are extracted, which can better represent the information of human behavior in the video, so as to improve the accuracy rate of human behavior recognition. The effectiveness of the proposed algorithm is verified on the standard human behavior recognition data sets UCF101 and HMDB51.

In this paper, the standard human behavior recognition data sets UCF101 and HMDB51 are used to train and verify the performance of 3DCCA model. In the future research work, we will use the collected real scene data set for human behavior recognition, so that the model can be effectively applied in the real environment. Besides, the traditional short-term video timing information is used as the input of the model, that is, a continuous 16-frame clip is input each time. In the next work, we must consider the video long-term timing information to further improve the accuracy rate of behavior recognition. Finally, the combination of manual feature construction algorithm and deep learning algorithm is also an effective way to improve the model recognition rate.

Acknowledgments This research work was supported in part by the National Science Foundation of China under Grant 51668043, and Grant 61262016, in part by the CERNET Innovation Project under Grant NGII20160311, and Grant NGII20160112, and in part by the Gansu Science Foundation of China under Grant 18JR3RA156.

Declarations

Conflict of interest The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Action recognition. UCF101: a large human motion database (n.d.). https://www.crcv.ucf.edu/data/UCF101.php
2. Action recognition. HMDB: a large human motion database (n.d.). http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/
3. Cai Z, Wang L, Peng X, et al, "Multi-view super vector for action recognition," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 596–603, Columbus, OH, USA, June 2014.
4. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308, Honolulu, HI, USA
5. Du W, Wang Y, Qia Y (2018) Recurrent spatial-temporal attention network for action recognition in videos. IEEE Trans Image Process 27:1347–1360
6. Hbn A, Fmb C, Mhya C et al (2021) T-VLAD: temporal vector of locally aggregated descriptor for Multiview human action recognition. Pattern Recognition Letters
7. Hsueh YL, Lie WN, Guo GY (2020) Human Behavior Recognition from Multiview Videos. Inf Sci 517: 275–296
8. Hu H, Cheng K, Li Z, Chen J, Hu H (2020) Workflow recognition with structured two-stream convolutional networks. Pattern Recogn Lett 130:267–274
9. Huang J, Lin S, Wang N, Dai G, Xie Y, Zhou J (2020) TSE-CNN: a two-stage end-to-end CNN for human activity recognition. IEEE J Biomed Health Inform 24(1):292–299
10. Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
11. Kim JH, Cho YI (2020) A new residual attention network based on attention models for human action recognition in video. J Korea Soc Comp Inform 25(1):55–61
12. Klaser A, Marszalek M, Schmid C (2008) A Spatio-Temporal Descriptor Based on 3D-Gradients. In: Proceedings of the 19th British Machine Vision Conference, pp. 1–10, Leeds, United Kingdom
13. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(3):107–123
14. Li R, Wang L, Wang K (2014) A review of research on human movement and behavior recognition. Pattern Recogn Artif Intell 27(1):35–48
15. Li Z, Gavrilyuk K, Gavves E, Jain M, Snoek CGM (2018) VideoLSTM convolves, attends and flows for action recognition, Comput. Vis Image Underst 166:41–50
16. Liciotti D, Bernardini M, Romeo L, Frontoni E (2020) A sequential deep learning application for recognising human activities in smart homes. Neurocomputing 396:501–513
17. Peng Y, Zhao Y, Zhang J (2019) Two-stream collaborative learning with spatialtemporal attention for video classification. IEEE Trans Circuits Syst Video Technol 29(3):773–786
18. Qiu Z, Yao T, Mei T (2017) Learning spatio-temporal representation with pseudo-3d residual networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5533–5541, Venice, Italy
19. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia, pp. 357–360, Augsburg, Germany
20. Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. CoRR
21. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 28th Neural Information Processing Systems, pp. 568–576, Montreal, Canada
22. Tran D, Bourdev L, Fergus R, et al (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497, Santiago, Chile
23. Tu NA, Huynh-The T, Khan KU, Lee YK (2019) ML-HDP: a hierarchical Bayesian nonparametric model for recognizing human actionsin video. IEEE Trans Circuits Syst for Video Technol 29(3):800–814
24. Wang L (2018) Three-dimensional convolutional restricted Boltzmann machine for human behavior recognition from RGB-D video. EURASIP J Image Video Process 120:1–11
25. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558, Sydney, Australia
26. Wang L, Xiong Y, Wang Z, et al (2016) Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the European Conference on Computer Vision, pp. 20–36, Springer, Cham
27. Woo S, Park J, Lee J Y, et al (2018) CBAM: Convolutional Block Attention Module. In: Proceedings of the European Conference on Computer Vision, pp. 3–19, Springer, Cham
28. Yao F (2020) Deep learning analysis of human behaviour recognition based on convolutional neural network analysis. Behav Inform Technol 40:1–9

29. Yao G, Lei T, Zhong J (2019) A review of Convolutional-neural-network-based action recognition. Pattern Recogn Lett 118:14–22
30. Ye Q, Liang Z, Zhong H, et al (2022) "Human behavior recognition based on time correlation sampling two-stream heterogeneous grafting network," in Optik - International Journal for Light and Electron Optics, vol. 251, Elsevier,168402
31. Yeung S, Russakovsky O, Jin N et al (2015) Every moment counts: dense detailed labeling of actions in complex videos. Int J Comput Vis 126(2–4):375–389
32. T. Yu, C. Guo, L. Wang, et al, "Joint spatial-temporal attention for action recognition," Pattern Recognit, Lett, vol. 112, pp. 226–233, 2018.
33. Yu S, Xie L, Liu L, Xia D (2020) Learning long-term temporal features with deep neural networks for human action recognition. IEEE Access 8:1840–1850
34. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al (2015) Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4694–4702, Boston, MA, USA
35. Zhang J, Hu H (2019) Domain learning joint with semantic adaptation for human action recognition. Pattern Recogn 90:196–209
36. Zhang B, Wang L, Wang Z, Qiao Y, Wang H (May 2018) Real-time action recognition with deeply transferred motion vector CNNS. IEEE TransImage Process 27(5):2326–2339
37. Zhang M, Yang Y, Ji Y, Xie N, Shen F (2018) Recurrent attention network using spatial-temporal relations for action recognition. Proceed Signal Process 145:137–145
38. Zhang M, Yang Y, Ji Y, Xie N, Shen F (2018) Recurrent attention network using spatial-temporal relations for action recognition. Signal Process 145:137–145
39. Zhang J, Hu H, Lu X (2019) Moving foreground-aware visual attention and key volume mining for human action recognition. ACM Trans Multimedia Comput Commun Appl 15(3):1–16
40. Zufan Zhang, Zongming Lv, Chenquan Gan et al, "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions," in Proceedings of the Neurocomputing, vol. 410, pp. 304–316, 2020.
41. Zhao S, Liu Y, Han Y, Hong R, Hu Q, Tian Q (2018) Pooling the convolutional layers in deep convNets for video action recognition. Proceed IEEE Trans, Circuits Syst, VideoTechnol 28(8):1839–1849
42. Zhu F, Shao L, Xie J, Fang Y (2016) From handcrafted to learned representations for human action recognition: a survey. Image Vis Comput 55:42–52

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.