ISSN: 1813-4890

Survey of Behaviour Recognition based on Deep Learning

Tinghong Fang[a], Jianshe Dong[b]

Tianjin University of Technology and Education, Tianjin 300222, China

[a]18222363105@163.com, [b]dongjianshe@tute.edu.cn

Abstract

In the field of computer vision, human behaviour recognition is a hot problem that has received widespread attention .It has great application value in intelligent surveillance, robotics, human-computer interaction, virtual reality, smart home, intelligent security, athlete-assisted training and so on. The current mainstream human behaviour recognition algorithms based on deep learning is analyzed in depth, including 3D convolution,two-stream network models, long short term memory networks and graph convolution. The recognition effects of those algorithms were compared on UCF101, HMDB51 and Kinetics400 datasets. It is helpful for selecting suitable behaviour recognition models for different application scenarios. Finally, the future direction of the behavioural recognition field is envisaged, which provides a reference for subsequent research.

Keywords

Human Behaviour Recognition; Deep Learning; Convolutional Neural Networks.

1. Introduction

In recent years, with the development of computer vision technology, various technical achievements have emerged, and human behaviour recognition technology has gradually gained widespread attention and its application areas are becoming increasingly diverse, such as virtual reality, video surveillance and health care. The task of behaviour recognition is to recognise different actions from video clips (2D frame sequences), where the actions can be performed or not performed for the entire duration of the video. Behaviour recognition appears to be an extension of the image classification task to multiple frames and then aggregates the predictions from each frame. Despite the great results achieved in image classification, video classification and representation learning is still slow due to factors such as shooting angle, lighting, action context and human clothing.

Earlier, a method based on manual recognition was proposed in which traditional machine learning algorithms were used to extract local or overall features of an object, encode and normalise the extracted features, and finally train the extracted features to obtain the corresponding predictive classification results. For the traditional behavioural recognition methods, the local high-dimensional visual features of the video region are mainly extracted, after which the extracted features are cropped into fixed-size video segments, and finally a classifier is used for the final classification and recognition. Among them, Laptev et al [1] discovered a local representation of spatio-temporal interest point detection (STIPS), where interest points are extracted in the spatio-temporal domain as changing neighbourhood points in the spatio-temporal domain, thereby describing the local features of the human body in the video for behavioural recognition. As Laptev et al.'s method suffers from the deficiency of extracting too few useful interest points, Dollar et al [2] proposed to use Gabor filters combined with Gaussian filters to add stable interest points for behavioural recognition. Richardson et al [3] proposed Maikov Logic Networks, which Wang et al. [4] proposed a dense trajectory (DT)

ISSN: 1813-4890

algorithm that densely samples each spatial scale and then tracks each spatial scale separately to form a trajectory description to classify behavioural recognition. However, the accuracy of behavioural recognition is affected by a moving camera, so Wang et al [5] optimised the optical flow image and improved the iDT (improved Dense Trajectory) algorithm for human behavioural recognition.

Extraction based on manual features has been relatively successful, but this method only targets fixed images for feature extraction, does not guarantee the universality of the video, and is very slow to adapt to practical real-time needs. In recent years, deep learning has been developed intensively and is used in many fields. Because the principle of deep learning is to use a large number of neurons to simulate human audiovisual and thinking activities, which has the same mechanism as behavioural recognition, researchers have also tried to use deep learning to solve some problems, and have achieved a series of better results. There are currently four mainstream algorithms: two-stream neural network (Tow-stream), 3D convolutional neural network, long short term memory network (LSTM) and graph convolutional neural network (GCN), etc.

2. Deep Learning Based Behaviour Recognition Algorithms

2.1. Two-stream Neural Network Structure

The basic principle of the two-stream model is mainly to fuse spatial streams and time; spatial networks are used to capture the spatial dependencies in the video, while temporal networks capture the spatial locations of periodic motions in the video. Therefore it is very important to map the spatial features related to a specific region to the temporal feature map of the corresponding region. To achieve this goal, the network needs to fuse the same pixel locations corresponding to this in the early convolution stage. The output of the joint temporal network is performed on time frames in order to model long-term dependencies. This network model allows long-range temporal modeling with better remote loss and higher accuracy, but the correspondingly larger computational effort leads to slower speed. The basic flow of the two-stream neural network is shown in Fig. 1.

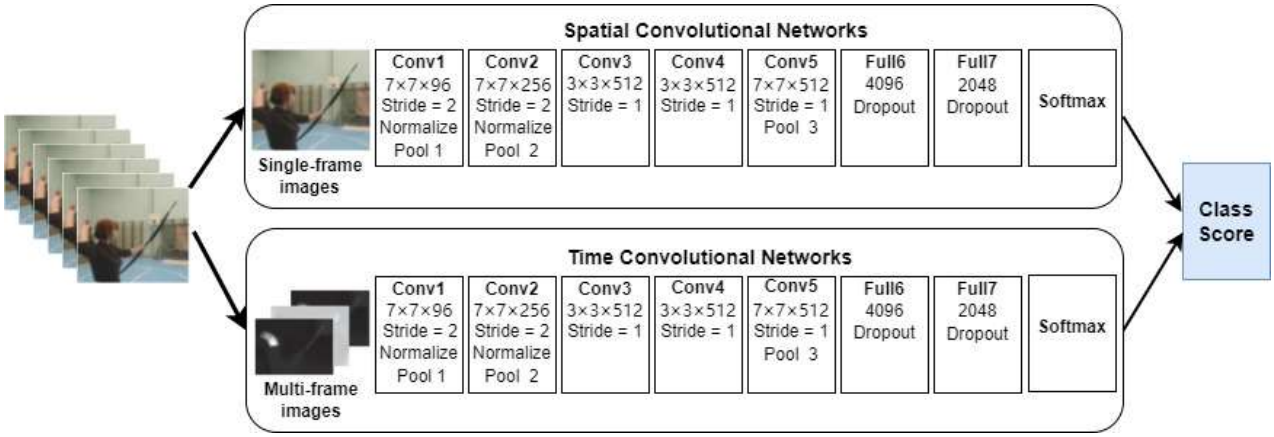


Fig.1 Structure framework of two-stream

Simonyan et al [6] proposed a two-stream model (Two-stream Netword) based on optical flow design. The model structure uses a single-frame image as input to a neural network for processing spatial dimension information, and uses a multi-frame density optical flow field as input to a convolutional neural network for processing temporal dimension information, and achieves good recognition results using a multi-task training method. A certain foundation was laid for the subsequent two-stream network.

ISSN: 1813-4890

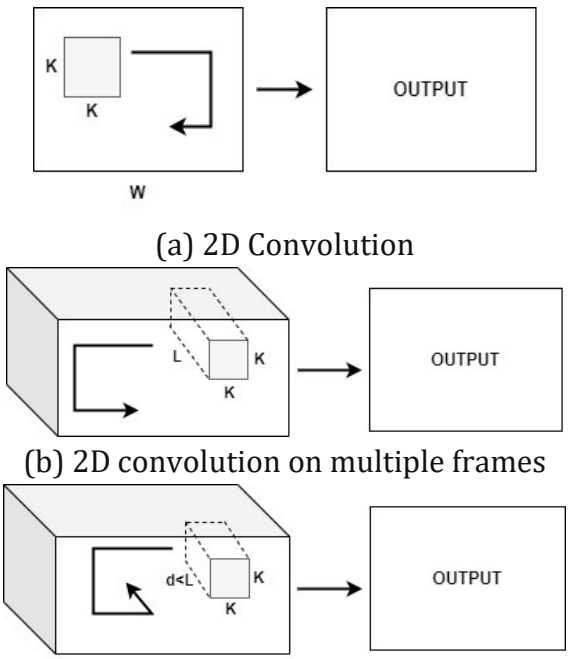
Wang et al [7] used TSN (Temporal Segments Networks) as the main network on the basis of the above dual-stream model, divided the long-time video into K segments, randomly selected a short part from each segment, used the above two-stream model algorithm for feature extraction on the final selected part, and fused the features of the extracted parts, solving the problem that the above two-stream model could not be applied to the long-time video. This solves the problem that the above two-stream model cannot model long videos.

Zhou et al [8] focused more on temporal relationship inference based on the TSN model, extending temporal inference between two frames to more frames, performing temporal inference on the input feature map, adding three fully-connected layers to learn the weights of video frames of different lengths, which can recognise actions that cannot be discerned by key frames (single RGB image) alone, such as falling.

Feichtenhofer et al [9] proposed multiple fusion approaches based on the two-stream model, fusing temporal and spatial networks at the convolutional layer instead of at the softmax layer, resulting in substantial savings in the number of parameters without performance degradation.

2.2. 3D Convolutional Neural Network Architecture

The 3D convolutional neural network is an improvement on the 2D convolutional neural network, solving the problem that the 2D convolutional neural network cannot extract information on the temporal sequence well. In contrast to 2D convolutional neural networks, which learn on a single frame of a single channel, 3D convolutional neural networks learn on multiple frames, adding a temporal dimension. Thus 3D convolutional neural networks are the result of extending the work of 2D convolutional neural networks on extracting spatio-temporal features. 2D convolutional and 3D convolutional differences are shown in the Fig. 2.



(a) 2D Convolution

(b) 2D convolution on multiple frames

(c) 3D convolution

Fig.2 Comprarison of 2D-CNN and 3D-CNN

Ji et al [10] used 3D convolution which can extract temporal and spatial features so as to extract motion information in continuous video frames. The model generates multiple information channels from the input frames and eventually combines the information features of all channels for behaviour recognition, laying the foundation for 3D convolutional neural networks.

ISSN: 1813-4890

Tran et al [11] proposed a C3D model based on 3D convolutional neural networks to directly extract behavioural features of videos, applying 3D convolutional kernels to behavioural videos on a large scale, but the C3D network suffers from a large number of parameters, is prone to overfitting and suffers from gradient disappearance, problems that limit the depth scaling of C3D.

Diba et al [12] combined a DenseNet network model based on C3D, attempted to model 3D convolutional kernels in different length video ranges, used TTL to replace the pooling layer, and proposed a T3D model structure, which greatly reduced the number of network structure parameters, but the dense links in the network model increased the computational load.

Lv et al proposed a (2+1)D convolution approach instead of 3D convolution to deepen the network structure, and added a (2+1)D convolution layer and a 3D pooling layer to improve the drawback that the original C3D model is prone to overfitting during training, which is conducive to effective learning of videos with longer motion times and richer video pixels.

2.3. Long Short Term Memory Networks Structure

Long Short Term Memory Networks (LSTMs) are special recurrent neural networks (RNNs). LSTMs improve on the shortcomings of RNNs that cannot effectively learn features in longer time sequences and were the first improvements recognised to effectively alleviate the problem of long-term dependence. Therefore Srivastava et al [14] concluded that LSTM is an effective way to facilitate behaviour recognition models to learn long sequence relationships. the distinction of LSTM structure is shown in Fig. 3.

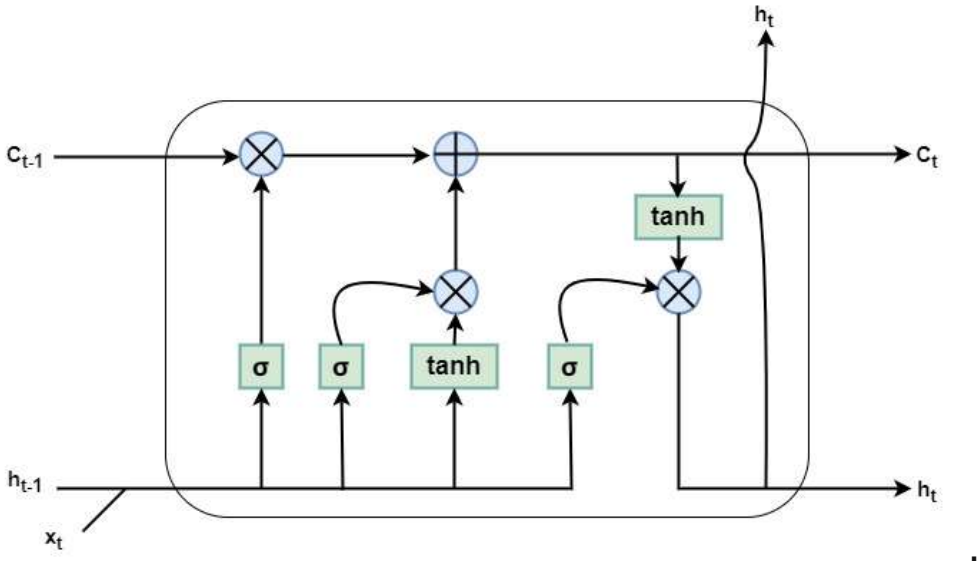


Fig.3 Internal structure of LSTM network unit

Donahue et al [15] added an LSTM network model to the convolutional network cnn, using the CNN for acquiring spatial features in the image description panel and the LSTM for acquiring temporal features, focusing on enhancing the extraction of temporal feature information of the video and achieving good results in the direction of behavioural recognition.

Ng et al [16] used global video-level convolutional networks combined with long- and short-term memory networks for spatio-temporal feature extraction in order to address the problem of inaccurate behavioural category recognition due to incomplete information extraction by convolutional networks. Global descriptions are learned using feature pools and LSTM networks. Better results were achieved by sharing parameters in time and training temporal models on optical flow maps.

ISSN: 1813-4890

Majd et al [17] improved the LSTM and proposed the C2LSTM unit. The C2LSTM unit allows the extraction of spatial and temporal features from motion data, and the learning of the spatial and motion structures of video data using convolution and correlation operators.

The LSTM somewhat enhances the long-time representation capability of CNN, but the LSTM itself is more difficult to train, plus the strict iteration of temporal sequence order affects the training efficiency.

2.4. Graph Convolutional Network Structure

Although convolutional neural networks have achieved fruitful research results in the field of behavioural recognition, the processing of regular data such as convolutional architectures and image sequences, and the fact that in complex application scenarios, the target is usually affected by lighting changes, noise, environment and other factors, with large changes in appearance, leads to degradation of algorithm performance. In contrast, graph convolutional neural networks are able to learn human skeleton features for human behaviour recognition, and are therefore robust to illumination and scene changes. The basic model of behaviour recognition based on graph convolution is shown in Fig. 4.

Yan et al [18] proposed the ST-GCN model to model dynamic skeletons by building spatio-temporal maps of skeletal sequences, and extended the graph convolutional network to learn spatio-temporal change relations by spatio-temporal graph convolutional networks, thus avoiding the drawback of designing traversal rules by hand and enabling the network to have better expressiveness and higher performance.

Shi et al [19] incorporated skeleton length-based information based on Yan et al's study, thus proposing a two-stream adaptive graph convolutional network (2S-AGCN) for skeleton-based behaviour recognition, allowing the inclusion of new connections other than natural ones to dynamically adapt the graph structure to better fit the hierarchical structure of the model. The ST-GCN model is improved by neglecting the connections between non-physically connected nodes and lacking the flexibility to further model the multi-level semantic information contained in all layers.

Li et al [20] proposed an AS-GCN model with the addition of self-supervised action and structural connections to mine potential joint connections and higher-order neighbourhood information, respectively. This model is complex in structure and not easy to compute, but can be extended to future research areas of pose prediction.

Shi et al [21] proposed the DGNN model, which turned the skeletal sequence modelling from the original undirected graph into a directed acyclic graph, not only extracting the joint points and skeletal information, but also the directional association information between them, effectively improving the problem of difficult feature propagation in disconnected subgraphs.

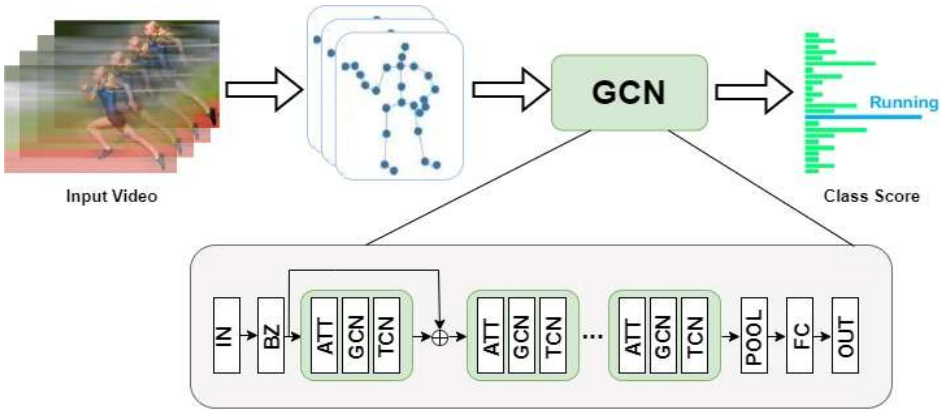


Fig.4 Basic model of behavior recognition based on graph convolution

ISSN: 1813-4890

3. Human Behaviour Recognition Datasets and Accuracy Evaluation

3.1. Commonly Used Public Datasets

The main commonly used behaviour recognition datasets are Hollywood2, HMDB51, UCF101, Sports-1M, NTU-RGB+D and Kinetics, this is shown in Tab. 1.

The Hollywood series [22] is derived from action scenes from Hollywood films. the Hollywood dataset is derived from 32 films and is divided into 8 categories, with different actors performing the same action in different scenes. hollywood2 is an extension of the Hollywood dataset, with 3669 videos cut from 69 films, divided into 12 behavioural categories and 10 scene categories, the dataset contains both a behavioural sub-dataset and a scene sub-dataset.

The HMDB51 dataset [23] is derived from digitised film and public repositories with 6,849 videos in 51 categories. Many factors such as uniqueness of dataset source, variation in filming perspective, background clutter, and obscured appearance make dataset identification difficult.

The UCF series dataset [24] was mainly intercepted from sports radio and television channels and the video website YouTube, and is rich in scenes and variety. UCF101 is an expansion of UCF50, with the number of action categories increased to 101, for a total of 13,320 videos, and the actions in each group can be further divided into five categories.

The Sports-1M dataset [25] is a large dataset of 487 sports videos with 1,000,000 videos, some with multiple tags and small differences between categories at the leaf level, obtained by Google from a number of video sequences on the video site YouTube.

The NTU-RGB+D dataset [26] contains 60 categories of actions, of which 40 are everyday behavioural actions, 9 are health-related actions and 11 are pairwise mutual actions. These movements were performed by 40 individuals ranging in age from 10 to 35 years. The dataset was acquired by the Microsoft Kinectv2 sensor and used three different camera angles to capture data in the form of depth information, 3D skeletal information, RGB frames and infrared sequences.

The Kinetics series [27] was mainly obtained by capturing high quality videos from the video site YouTube. 2017's Kinetics 400 contained 400 action categories with about 400 videos each, 2018's Kinetics 600 was produced, containing 600 action categories with at least 600 video sequences per category, each lasting about 10 s. In 2019, the Kinetics dataset was expanded again, with a total of about 700 categories, a considerable amount of data.

Table 1. Deep skeleton sequence datasets

Year	Dataset	*Nu*mber of action types	Number of samples	Soure
2009	Hollywood2	12	3669	Action in film
2011	HMDB51	51	6849	Physical interaction
2012	UCF101	101	13,320	Interactive actions
2014	Sports-1M	487	1,000,000	Sports videos
2016	NTU-RGB+D	60	56,880	Daily behavioural actions
2020	Kinetics	700	650,000	YouTube videos

3.2. Performance Comparison of Different Algorithms

This paper mainly compares the recognition accuracy of currently popular algorithms on different datasets, without considering the number of parameters, computation, training times, data pre-processing, hardware and software configurations of each algorithm, using the accuracy (Accutacy) evaluation function on the UCF101, HMDB51 and Kinetics datasets. This is shown in Tab. 2.

ISSN: 1813-4890

Table 2. Common algorithms used in action recognition research

Methods		UCF101/%	HMDB51/%	Kinetics/%	Improvements	Disadvantages
Based on a Two-Stream model	Two-Stream	88.0	59.4	91.3	Highly accurate and scalable.	Poor time scale and real-time
	TSN	94.2	69.4	-	Recognition of long videos	Difficulty in recognising single frame images
Based on 3D Convolution	C3D	82.3	51.6	-	Simple construction	Large number of participants
	C3D+DenseNet	93.2	63.5	-	Reduced amount of parameters	Increased calculated load
Based on LSTM	CNN-LSTM	88.6	54.0-	-	Reduced noise impact	Complex parameters
	C^2LSTM	92.8	61.3	-	Capturing spatial and temporal features	General effect for RGB as input
Based on Graph Convolution	ST-GCN	-	88.3	52.8	Not susceptible to cosmetic factors	Difficult to learn relationships between joints without physical connections
	DGNN	-	-	36.9	High recognition accuracy	Feel the Wild Little

From Tab. 2, it can be seen that there is a small difference between the four deep learning models on the UCF101 dataset, where the TSN model based on the dual stream model achieves the highest accuracy of 94.2%, relatively speaking, the C3D model based on 3D convolution has a relatively low accuracy of only 82.3%; on the HMDB51 dataset there is a large difference in performance, where the ST The accuracy of the GCN model reached a maximum of 88.3%, compared to the C3D model based on 3D convolution, which was only able to reach 51.6%; on the Kinetics dataset, the graph convolution-based model was used more often in this dataset and was able to reach a maximum accuracy of 52.8%, compared to the two-stream model, which had a relatively low accuracy.

Tab. 3 collates some recent research results on deep skeleton sequence datasets, which are mostly used in graph convolution-based behaviour recognition models. By analysing the data in the table, it can be found that the four models based on graph convolution are less different, with the DGNN model having the best recognition results, being able to achieve an accuracy of 59.6% and 96.1% on the Kinetics and NTU-RGB+D datasets respectively.

Table 3. Performances of different algorithms on depth skeleton dateset

Methods	Kinetics/%	NTU-RGB+D/%	Improvements
ST-GCN	52.8	88.3	Greater expressive power and stronger generalization capability
2S-AGCN	58.7	95.1	Ability to model both first- and second-order information
AS-GCN	56.5	94.2	Capture action-specific potential dependencies directly from the action
DGNN	59.6	96.1	Representation of skeletal data as a directed acyclic graph (DAG)

ISSN: 1813-4890

4. Conclusion

This paper provides a detailed introduction and comparative analysis of the four mainstream methods under deep learning, and points out the advantages and disadvantages of each type of method, as well as introducing some mainstream human behaviour recognition datasets. Deep learning methods also have problems that need to be solved, as they require a lot of time and large datasets to train the models, while in practical applications such as video surveillance and home monitoring, human behaviour recognition is often needed in real time, and how to apply it quickly and effectively is a major challenge for future research work.

The key to human behaviour recognition is the extraction of robust behavioural features, both spatial and temporal. This paper provides the following outlook on future research directions in this area.

a) Behaviour recognition based on two-stream neural networks.Two-stream convolutional neural networks can effectively extract temporal and spatial features from videos, resulting in good recognition results. However, the more networks are designed, the more complex the model will be, further leading to an increase in computational effort. Therefore, how to design multi-stream networks to extract effective features is one of the important difficulties for future research.

b) Behaviour recognition based on 3D convolutional neural networks. the proposed C3D network frees researchers' thinking from 2D convolution, the 3D convolution generalises well, can be applied in combination with many networks and improves the performance of the original network. the major difference between C3D and dual-stream networks and LSTM networks is that C3D reduces the network parameters and speeds up the training of the network, but various The behaviour recognition algorithms are not yet very effective in recognising small detailed actions, which is still a challenging problem.

c) Behavioural recognition based on long and short term memory networks. the LSTM network can handle both the temporal information of the video and the problem of gradient extinction, so the LSTM enhances the long term characterization ability of the CNN to some extent, but the LSTM itself is more difficult to train, plus the strict iteration of the temporal order sequence affects the training efficiency.

d) Behaviour recognition based on graph convolutional networks. Deep skeleton data is more robust to complex scenes. The human 3D skeleton data itself can be seen as a natural topological graph data structure, with vertices representing joints and edges representing limb segments connecting joints, and GCN is able to perform deep learning on graph data. In addition, with the application of depth sensors such as Kinect, generating large deep skeleton sequence datasets such as NTU-RGB + D, GCN-based methods will be a very popular research direction in the future.

References

[1] LAPTEV I. On space-time interest points[J]. International journal of computer vision, 2005, 64(2): 107-123.

[2] DOLLAR P, RABAUD V, COTTRELL G, et al. Behavior recognition via sparse spatio-temporal features[C]//2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance. IEEE, 2005: 65-72.

[3] RICHARDSON M, DOMINGOS P. Markov logic networks[J]. Machine learning, 2006, 62(1): 107-136.

[4] WANG H, KLÄSER A, SCHMID C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.

[5] WANG H, SCHMID C. Action recognition with improved trajectories[C]//Proceedings of the IEEE international conference on computer vision. 2013: 3551-3558.

[6] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27.

[7] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, Cham, 2016: 20-36.

[8] Zhou B, Andonian A, Oliva A, et al. Temporal relational reasoning in videos[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 803-818.

[9] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1933-1941.

[10] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.

[11] TRAND B L. Learning Spatio temporal Features with 3D Convolutional Networks[C]//Proceedings oftheIEEE International Conference on Computer Vision. 2015.

[12] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3d convnets: New architecture and transfer learning for video classification[J]. arXiv preprint arXiv:1711.08200, 2017.

[13] Lv Shuping,Huang Yi,Wang Yingying. Improvement of human action recognition method based on C3D convolutional neural network[J]. Experimental Technology and Management, 2021,38(10): 168-171+176.DOI:10.16791/j.cnki.sjg.2021.10.031.

[14] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms [C]// Proc of the 32th International Conference on Machine Learning. Cambridge MA: JMLR, 2015: 843- 852.

[15] DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.

[16] YUE-HEI Ng J, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.

[17] Mahshid Majd,Reza Safabakhsh. Correlational Convolutional LSTM for human action recognition[J]. Neurocomputing,2020,396(prepublish).

[18] YAN S J, XIONG Y J, LIN D H. Spatial temporal graph convolutional networks for skeleton- based action recognition[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018.

[19] SHI L, ZHANG Y F, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16- 20, 2019. Washington: IEEE Computer Society, 2019: 12026-12035.

[20] Li M S, Chen S H, Xu C, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE CS, 2019: 3590-3598.

[21] Shi L, Zhang Y F, Cheng J, et al. Skeleton-based action recognition with directed graph neural networks[C]. Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE CS, 2019: 7904-7913.

[22] Information on http://www.di.ens.fr/~laptev/actions/hollywood2/.

[23] Information on http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-da-tabase/#dataset.

[24] Information on http://crcv.ucf.edu/data/UCF101.php.

[25] Information on https://cs.stanford.edu/people/karpathy/deepvideo/.

[26] Information on https://rose1.ntu.edu.sg/Datasets/actionRecognition.asp.

[27] Information on https://www.deepmind.com/research/open-source/open-source-da-tasets/ kinetics.