# A review of action recognition based on Convolutional Neural Network

# A review of action recognition based on Convolutional Neural Network

**Yang Jiaxin[1,a], Wang Fang[1,b*], Yang Jieru[1,c]**

[1]College of Electrical and Power Engineering, Taiyuan University of Technology, Taiyuan, Shanxi Province, China

[a]email: 1315587229@qq.com, [c]email: 873636190@qq.com,

*Corresponding author: [b]email: wangfang@tyut.edu.cn

**Abstract.** At present, the development of video action recognition is very rapid in many fields, such as video understanding, intelligent monitoring, and human-computer interaction. However, there are some challenges in the development of action recognition, and researchers have tried to put forward some explorations. Convolutional neural network (CNN) is applied to action recognition, which improves the performance of action recognition. It is divided into 3 methods in this paper. In addition, C3D, Two-stream and I3D, three classic CNN algorithms, are reproduced. And their recognition rates are 72%, 78.0% and 97.6% respectively on the UCF101 dataset.

## 1. Introduction

Research on action recognition is closely related to real life needs, such as somatosensory entertainment, intelligent robots, and abnormal motion monitoring. Therefore, action recognition has broad application prospects and potential economic value. At present, the methods of action recognition can be divided into traditional artificial feature extraction methods and deep learning methods. For traditional artificial feature extraction methods, they are generally necessary to design the extractors based on spatiotemporal features. This requires designers to have strong professional qualities, so their limitations are relatively large. Deep learning methods have strong nonlinear modeling capabilities and can learn general features from raw data. Compared with the traditional artificial feature extraction methods, deep learning methods which are end-to-end methods, have strong feature expression ability, small limitations, and are more suitable for dealing with real-world action recognition problems.

Action recognition based on deep learning methods refers to automatically analyzing and recognizing the motions of a single or multiple targets in an unknown video through deep learning methods. The process of action recognition based on deep learning methods is shown in Figure 1. Deep learning methods can directly recognize the motion in the video frame, or they can be used as feature extractors combined with a classifier to recognize the motion. Convolutional neural networks (CNN) [1], as a type of deep learning, have always performed well in terms of images. Convolutional neural networks are introduced into action recognition, which improves the performance of algorithms and promotes the development of action recognition.

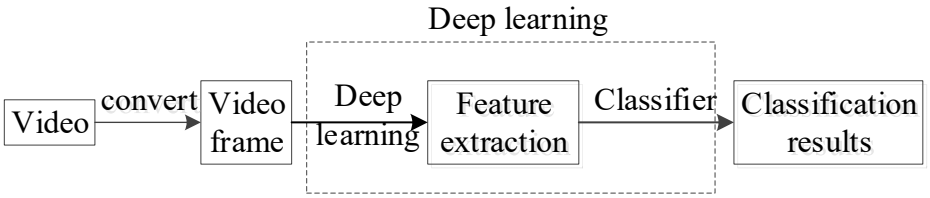


Figure 1. Action recognition based on deep learning methods.

Currently, there are many literatures that have studied in detail all aspects of action recognition. Reviews [2,3] summarized the challenges faced by action recognition and classified them into four categories. Reviews [3-5] discussed the extraction of spatiotemporal features and the performance of action recognition from the traditional artificial feature extraction methods and deep learning methods. Reviews [6-8] introduced common datasets and evaluation benchmarks of action recognition. Review [3] conducted the practical application of action recognition. Review [9] focused on CNN-based action recognition.

CNN-based action recognition is reviewed in this paper, which is organized as follows. Section 2 introduces three challenges of action recognition. In Section 3, CNN-based action recognition is divided into two categories for review. Section 4 provides some examples of action recognition codes we have run. Section 5 introduces other deep learning methods and the future direction of action recognition.

## 2. Research challenges

Action recognition has made some progress in traffic management, animal breeding, crime prevention, and human-computer interaction. However, the existence of research challenges prevents the further development of action recognition. In recent years, researchers have tried to give some solutions and provide some new ideas for action recognition.

### *2.1 Intra-class and inter-class distance*

When exhibiting the same motion, different targets may express different meanings, which leads to a large intra-class distance in action recognition. For example, some people throw their fist as a way of saying hello, while others may hit someone. When exhibiting different motions, different targets may express the same meaning, which leads to a small inter-class distance in action recognition. In the case of throwing a fist, when the speed is fast, the motion may be fighting or boxing, and when the speed is slow, it may be a way for friends to meet and greet. Both of these phenomena will cause misjudgments in action recognition and reduce the recognition accuracy.

In view of the above two phenomena, one possible method is to increase the long-term dependence between the input video frames for action recognition and reduce the confusion between similar motion [10]. Another possible method is to find a more differentiated loss function or action recognition methods. Long Short-Term Memory (LSTM) [11], compressed single ordered optical flow image [12], time-domain convolution kernels of different depths [13], sparse sampling by region [14] and Temporal Excitation and Aggregation (TEA) [15] can make full use of video frame information to obtain long-term dependence between motions. The center Loss [16] and 2C-softmax of the dual center Loss [17] can maximize the inter-class distance and minimize the intra-class distance to a certain extent.

### *2.2. Complex environmental changes*

When the camera is fixed, the influence of weather such as illumination, rain and fog, and the occlusion problem between multiple targets bright some challenges to action recognition. When the camera moves, the occlusion problem can be alleviated to a certain extent. But the complex environmental changes are increasing exponentially, and even the targets are hard to be captured, which brings greater challenges to action recognition.

In view of the complex environmental changes, efficacious preprocessing is a possible solution. Some methods [18-20] used the techniques of object tracking and detection before feature extraction and other methods [21-23] jointly recognized and localized actions in videos. Factorized Action-Scene

Network [24] could encode and fuse the most relevant and informative semantic cues for action recognition.

*2.3. Insufficient training data*

Unlike the traditional artificial feature extraction methods, the deep learning methods require a large number of data to train the network. Otherwise it is easy to cause the network to overfit, and the learned features are not representative, resulting in unsatisfactory action recognition results. However, many datasets cannot provide enough data, resulting in insufficient training data.

For the problem of insufficient training data, one possible method is to expand the datasets and increase the number of data. Another possible method is to use a pre-training network through transfer learning to prevent overfitting. Data expansion can increase effectively the data. Common data expansion methods have flip, rotation, scale transformation, random picking, color dithering, and Gaussian noise. The transfer learning usually uses a predecessors trained action recognition network as a feature extractor to extract features or directly train new data with a predecessors trained action recognition method, and fine-tuning the network.

## 3. CNN-based action recognition

CNN is a type of artificial neural network based on convolution operation which is the core of extracting features. Convolution methods include 1D convolution, 2D convolution and 3D convolution. Among them, 2D convolution and 3D convolution can often be used for feature extraction in action recognition. However, 3D convolution and 2D convolution are very different when extracting features, as shown in Figure 2. When processing a L-frame video segment, the output of a 2D convolution kernel extracting features is a two-dimensional feature map, and the time domain information is compressed. The output of a 3D convolution kernel extracting features is a three-dimensional feature map, which reflects the information of multiple images individually and interconnected very well.

The essence of action recognition is to obtain the appearance information in space and the movement information in time. A single 3D convolution can capture the spatiotemporal information of video action behaviors at the same time, while a single 2D convolution requires some help in capturing time information. Therefore, according to the difference in the convolution methods used in action recognition, CNNs are divided into single-channel CNN, dual-channel CNN, and fusion of multiple methods in this paper, as shown in Figure 3.

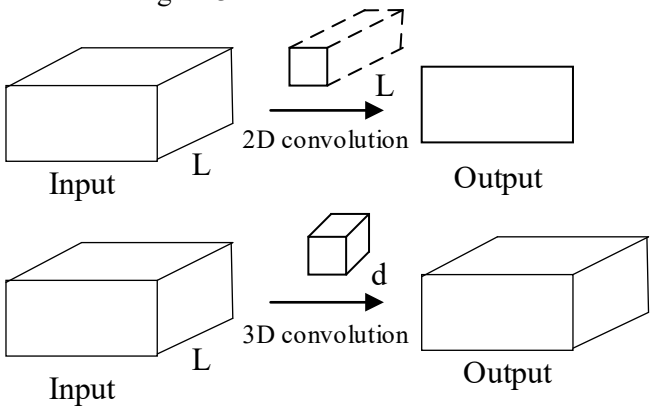


Figure 2. Comparison of feature extraction between 2D convolution and 3D convolution.

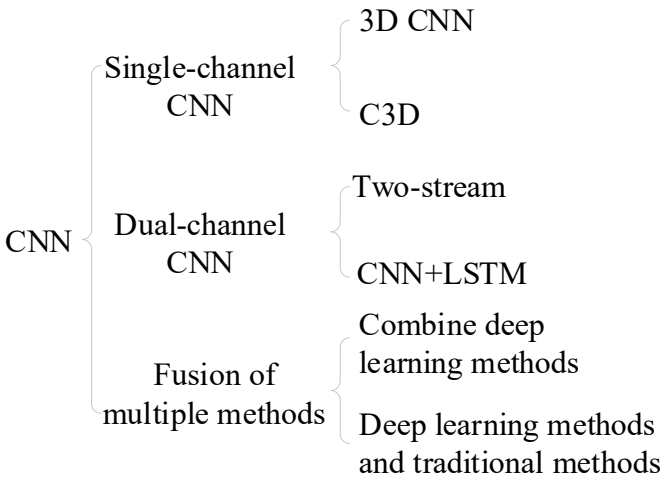


Figure 3. CNN algorithm classification

*3.1. Single-channel CNN*

The single-channel CNN can simultaneously extract the spatiotemporal features of motion behavior through a CNN. 3D CNN [25] and 3D Convolutional Network (C3D) [26] are two typical single-channel CNN. They are 3D convolutional neural networks designed by researchers directly for action recognition, and can directly extract information from multiple video frames through 3D convolution. 3D CNN proves the rationality of 3D convolution to extract spatiotemporal features, and C3D gives the optimal size of 3D convolution kernel to extract spatiotemporal features.

The single-channel CNN has many parameters and a large amount of calculation. This is mainly due to the increased dimension of 3D convolution. The 2D convolution and 1D convolution parameters and calculation amount are relatively small. Using 1x3x3 and 3x1x1 convolution combination [27-29] instead of 3x3x3 convolution could separate the time and space information and greatly reduce the amount of parameters.

The Res3D network [30] was inspired by the 2D residual network [31], and designed an 18-layer 3D residual network, weighing accuracy, computational complexity and network capacity. The Res3D network suddenly expanded the depth of the 3D network to 18 layers, greatly deepening the depth of the 3D network. Temporal 3D ConvNets (T3D) [13] designed a 3D DenseNet [32] to train a deeper 3D network and gathered time-domain convolution kernels of different depths to extract more complete motion features. T3D realizes the feature extraction of variable video frame length, which broadens the performance capabilities of 3D network models. X3D [33] drawed on the strategy of MnasNet [34] structure search, used ResNet and Fast pathway part of SlowFast [35] to train the best network, and explored the influence of frame rate, frame number, resolution, depth, width, and bottleneck width on model performance. Lightweight excellent design and dimensional expansion strategy of X3D are a new development direction for 3D networks.

*3.2. Dual-channel CNN*

When the separate 2D convolution kernel cannot solve the action recognition problem, the researchers separated appearance and motion features from the whole video. They continued to use a predecessors trained CNN to extract the appearance feature of the video, and then focused on researching a good motion feature extraction network, which is called the dual-channel CNN. Two-stream [36] and CNN+LSTM are two common types of dual-channel CNN. They used CNN in the spatial domain to extract appearance features, and optical flow [37] +CNN and LSTM in the time domain to extract motion features.

Optical flow method is a method of calculating the motion information. It relies on the correlation between the current frame and the previous frame in the image sequence to compare the motion of adjacent frames. In the two-stream network, the optical flow method extracted the motion information

of 10 frames in advance as the density optical flow images, and then CNN was used to extract the time-domain features of the density optical flow images. At the same time, CNN was used to extracted the space-domain features of single-frame RGB image.

In order to get better performance, two-stream used BN-Inception [38], VGG16 [39] and 3D ResNet50 network [14, 40, 41] instead of AlexNet network [42]. At the same time, the feature map fusion technology [43] drew the conclusion that the last spatial convolutional layer had less fusion loss, more parameter saving and higher accuracy. The traditional optical flow methods are very slow and offline extraction, which limit the running speed of the two-stream. The FlowNet network [44, 45] could use CNN to calculate optical flow, which solved the problem of slow speed of optical flow feature extraction. However, optical flow calculation is still independent of the two-stream network, which cannot achieve end-to-end training of deep learning, and occupies a certain amount of computing time and storage space on the GPU. The MotionNet model [46] was added to the time network to realize end-to-end training of the entire model. The method of using motion vector instead of optical flow [47] could not only achieve end-to-end training, but also improve the speed a lot.

LSTM is a special time recurrent neural network (RNN) [48], which can solve the information transmission problem of long input sequences. Using LSTM to process information in the time domain can learn and understand the features of time series very well. The Long-term Recurrent Convolutional Network (LRCN) [49] combined CNN and LSTM to extract spatial and temporal features. CNN were used to process variable-length visual input and their outputs were fed to a bunch of recurrent sequence models. Finally, they could produce variable-length predictions, and the convolution operations of different frames could be processed in parallel. So, LRCN achieved efficient end-to-end learning.
Using GoogLeNet and max pool [50] could generate relatively sparse gradient parameters, which maked the gradient reduction and learning faster. Using CNN and Deep Bidirectional LSTM (DB-LSTM) networks [51] to process video data could learn long-term sequences and process lengthy videos by analyzing the features of specific time intervals. LSTM is commonly used in fusion of multiple methods, combining 3D CNN, 2D CNN, automatic encoder [52] or human skeleton information [53] for action recognition.

### *3.3. Fusion of multiple methods*

The fusion of multiple methods combines two or more action recognition methods with good performance in extraction of spatiotemporal features, and often produce better recognition results.

The methods are roughly divided into two categories. One is to combine multiple deep learning methods to construct the network. [12, 41] used a 3D network in the time domain of the two-stream network and Two-stream Inflated 3D ConvNets (I3D) [54] separately used a 3D network in both the time and space domains of the two-stream network. They could enhance the ability of the model to discriminate data. SlowFast and Temporal Pyramid Network (TPN) [55] pay more attention to the speed of movement in action recognition and used 3D network in both the time and space domains. They can distinguish motions with similar shapes but different speeds, which improves the accuracy of action recognition. The methods of combining various CNN (including 3D CNN, 2D CNN, C3D) and LSTM [10, 12, 56] could obtain the information of long-term video and increase the recognition accuracy.

The other is to combine the deep learning methods with some traditional methods to extract features. R*CNN [57] used Fast RCNN [58] and the re-gion of interest divided into primary and secondary to extract features simultaneously. The use of Improved Dense Trajectories (IDT) [59] combined two-stream and LSTM [60] could enhance the ability to extract spatio-temporal features. Pose-based CNN descriptor (P-CNN) [61] could gather motion and appearance information along the trajectory of human body parts.

## 4. Our work

When evaluating CNN-based action recognition, it is required to evaluate with certain indicators in the same environment. The needs of the same environment have led to the creation of open datasets. UCF101 is the most commonly used dataset. It was proposed in 2012. The video comes from YouTube

and contains 101 action categories and a total of 13,320 short videos (27 hours in total). It is used to implement some classic methods in action recognition, which hopes to be helpful for beginners.

First of all, there are many codes written by original authors on the Internet. Most of them use different python environments and rely on many plug-ins. At the same time, there are many reproducible codes based on different environments by later researchers on the Internet. Their environment is relatively simple, but it is sometimes difficult to tell whether it can be implemented. Secondly, action recognition is to input multiple video frames and output one result, which is different from inputting one image and outputting one result in image recognition. So the code of the input data part is very different. Finally, due to the limited data, pre-training has a great impact on the accuracy of the final code. Of course, there are also attempts to achieve accuracy on small samples [62].

### *4.1. C3D*

C3D consists of five convolutional layers, five pooling layers, two fully connected layers and a softmax loss function layer. And the convolution kernels of all convolution layers in the C3D network are 3x3x3, and the pooling layer also uses 3D pooling.

The basic code of 3D convolution and 3D pooling in the tensorflow environment can be directly called to realize the network main. So, the implementation of C3D code is relatively easy. However, the val_accuracy and val_loss of the C3D model without pre-training is very poor. The specific results are shown in Table 1.

Table 1. C3D test result in UCF101

| Model | pre-training | val_accuracy | val_loss |
|---|---|---|---|
| C3D | no | 56% | 3..5342 |
| | yes | 72% | 1.3143 |

### *4.2. Two-stream network*

The two-stream network is divided into two-channel 2D CNN, which extracts features from an image in the spatial domain to obtain appearance information; and extracts features from 10 frames of optical flow images in the time domain to obtain motion information. The specific network structure is shown in Figure 4.

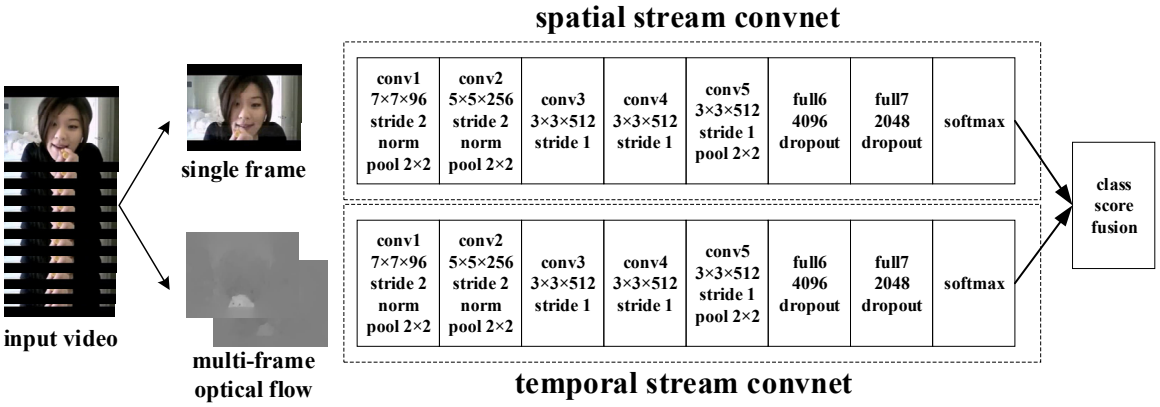


Figure 4. Two-stream network structure.

Optical flow adopts the TV-L1 method and use the denseflow tool to extract, which takes a long time. Then vgg_16 and res_v1_50 are used as pre-training models to train and test rgb, flow and two-stream respectively. Finally, the results on the verification set and test set are obtained, as shown in Table 2.

Table 2. two-stream test result in UCF101

| Category | rgb | flow | two-stream |
|---|---|---|---|
| Verification set | 0.743 | 0.646 | 0.813 |
| Test set | 0.710 | 0.541 | 0.780 |

*4.3. I3D*

I3D inputs 16 frames of images in the spatial domain instead of one frame to extract appearance information, and also inputs 16 frames in the time domain to extract motion features. It uses the mature 2D CNN which is Inception-v1, and changes the 2D in the network to 3D. The parameters corresponding to H and W are directly obtained from Inception.

Its most prominent is the use of large-scale ImageNet and Kinetics data sets to preprocess the model, and a high accuracy rate was obtained on UCF101, as shown in Table 3.

Table 3. I3D test result in UCF101

| Model | pre-training | rgb | flow | rgb+flow |
|---|---|---|---|---|
| I3D | Kinetics | 0.8660 | 0.918 | 0.953 |
| | Kinetics+ImageNet | 0.9472 | 0.9568 | 0.976 |

From the results of the code we ran and the researchers ran, it can be seen that action recognition has higher requirements on hardware devices, and the better the device, the better the results of the operation. In the case of ordinary equipment conditions and data, optical flow has a good auxiliary effect on motion recognition, making the accuracy of two-stream series networks relatively high.

## 5. Conclusion

The demands of video understanding, intelligent monitoring, human-computer interaction, etc. in real life attract researchers' attention to video action recognition. Firstly, we introduce three challenges in action recognition, and have tried to give some researchers' suggestions in this paper. Secondly, we divide CNN-based action recognition into single-channel CNN, dual-channel CNN, and fusion of multiple methods for a detailed review. Finally, we give some examples of action recognition codes and their results. In addition to CNN in deep learning methods, Restricted Boltzmann Machines (RBM) [63], Automatic Encoders [64], Deep Belief Networks (DBN) [65], and Independent Subspace Analysis (ISA) [66] can all be used as behavior recognition networks to extract spatiotemporal features. In the future, action recognition will develop towards integration, portability, and more practicality. And action recognition will be closer to humans, with voice and text for smarter recognition.

## Acknowledgments

This study was funded by the Shanxi Province Youth Fund Project (201901D211079).

## References

[1] Lecun Y , Bottou L . Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
[2] Poppe R . A survey on vision-based human action recognition[J]. Image & Vision Computing, 2010, 28(6):976-990.
[3] Kong Y, and Fu Y, Human Action Recognition and Prediction: A Survey[J]. 2018.
[4] Herath S , Harandi M , Porikli F . Going Deeper into Action Recognition: A Survey[J]. Image and Vision Computing, 2017, 60:4-21.
[5] Luo H, Wang C, and Lu F, Survey of video behavior recognition[J]. Journal on Communications, 2018, pp. 173-184.
[6] Kuehne H, Jhuang H, Stiefelhagen R, and Serre T, HMDB51: A Large Video Database for Human Motion Recognition[J]. 2013.
[7] Soomro K, Zamir AR, and Shah M, UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild[J]. Computer Science, 2012.
[8] Kay W , Carreira J , Simonyan K , et al. The Kinetics Human Action Video Dataset[J]. 2017.
[9] Yao G , Lei T , Zhong J . A Review of Convolutional-Neural-Network-Based Action Recognition[J]. Pattern Recognition Letters, 2018, 118(FEB.):14-22.
[10] Meng, Bo, Liu, et al. Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos[J]. Multimedia Tools & Applications, 2018.

[11] Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.
[12] Li Q, Li A, Wang T, and Cui Z, Double-Stream Convolutional Networks with Sequential Optical Flow Image for Action Recognition[J]. Acta Optica Sinica, 2018, pp. 234-240.
[13] Diba A, Fayyaz M, and Sharma V, Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification[J]. 2017.
[14] Wang L , Xiong Y , Wang Z , et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition[J]. 2016.
[15] Li Y , Ji B , Shi X , et al. TEA: Temporal Excitation and Aggregation for Action Recognition[J]. 2020.
[16] Wen Y , Zhang K , Li Z , et al. A Discriminative Feature Learning Approach for Deep Face Recognition[C]// European Conference on Computer Vision. Springer, Cham, 2016.
[17] Zhi-Qiang M , Cui-Hong M A , Jin-Long C , et al. Research on action recognition based on two-stream convolution and double center loss[J]. Microelectronics & Computer, 2019.
[18] Wang H , Schmid C . Action Recognition with Improved Trajectories[C]// 2013 IEEE International Conference on Computer Vision. IEEE, 2014.
[19] Raptis M , Kokkinos I , Soatto S . Discovering discriminative action parts from mid-level video representations[J]. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012.
[20] Ma Y，Tan L，Dong X，and Yu C, Action recognition for intelligent monitoring[J]. Journal of Image and Graphics, 2019, pp. 128-136.
[21] Ma S , Zhang J , Ikizler-Cinbis N , et al. Action Recognition and Localization by Hierarchical Space-Time Segments[C]// IEEE International Conference on Computer Vision. IEEE, 2013.
[22] Liu C , Wu X , Jia Y . Transfer Latent SVM for Joint Recognition and Localization of Actions in Videos[J]. IEEE Transactions on Cybernetics, 2016, 46(11):2596-2608.
[23] Zhou Z , Shi F , Wu W . Learning Spatial and Temporal Extents of Human Actions for Action Detection[J]. IEEE Transactions on Multimedia, 2015, 17(4):512-525.
[24] Hou J , Wu X , Sun Y , et al. Content-Attention Representation by Factorized Action-Scene Network for Action Recognition[J]. IEEE Transactions on Multimedia, 2017:1-1.
[25] Ji S , Xu W , Yang M , et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1):221-231.
[26] Tran D , Bourdev L , Fergus R , et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// IEEE International Conference on Computer Vision. IEEE, 2015.
[27] Qiu Z, Yao T, and Mei T, Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks[J]. 2017.
[28] Tran D, Wang H, Torresani L, et al, A Closer Look at Spatiotemporal Convolutions for Action Recognition[J]. 2017.
[29] Sun L , Jia K , Yeung D Y , et al. Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks[C]// IEEE International Conference on Computer Vision. IEEE, 2015:4597-4605.
[30] Tran D, Ray J, Shou Z, and Chang SF, ConvNet Architecture Search for Spatiotemporal Feature Learning[J]. 2017.
[31] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.
[32] Landola F, Moskewicz M, Karayev S, et al, DenseNet: Implementing Efficient ConvNet Descriptor Pyramids[J] .Eprint Arxiv, 2014.
[33] Feichtenhofer C . X3D: Expanding Architectures for Efficient Video Recognition[J]. 2020.

[34] Tan M , Chen B , Pang R , et al. MnasNet: Platform-Aware Neural Architecture Search for Mobile[J]. 2018.
[35] Feichtenhofer C , Fan H , Malik J , et al. SlowFast Networks for Video Recognition[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019.
[36] Simonyan K , Zisserman A . Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Advances in neural information processing systems, 2014, 1.
[37] Barron J L , Fleet D J , Beauchemin S S . Performance Of Optical Flow Techniques[J]. International Journal of Computer Vision, 1994, 12(1):43-77.
[38] Ioffe S, and Szegedy C, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015.
[39] Szegedy C, Liu W, Jia Y, et al, Going Deeper with Convolutions[J]. 2014.
[40] Wang L , Xiong Y , Wang Z , et al. Towards Good Practices for Very Deep Two-Stream ConvNets[J]. Computer ence, 2015.
[41] Feichtenhofer C , Pinz A , Wildes R P . Spatiotemporal Residual Networks for Video Action Recognition[J]. 2017.
[42] Simonyan K, Zisserman A, Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer ence, 2014.
[43] Feichtenhofer C , Pinz A , Zisserman A . Convolutional Two-Stream Network Fusion for Video Action Recognition[C]// Computer Vision & Pattern Recognition. IEEE, 2016.
[44] Fischer P , Dosovitskiy A , Ilg E , et al. FlowNet: Learning Optical Flow with Convolutional Networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2016.
[45] Ilg E , Mayer N , Saikia T , et al. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[46] Zhu Y , Lan Z , Newsam S , et al. Hidden Two-Stream Convolutional Networks for Action Recognition[J]. 2017.
[47] Zhang B , Wang L , Wang Z , et al. Real-Time Action Recognition with Enhanced Motion Vector CNNs[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.
[48] Hinton G E . Learning distributed representations of concepts.[C]// Eighth Conference of the Cognitive Science Society. 1989.
[49] Donahue J , Hendricks L A , Rohrbach M , et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017:677-691.
[50] Ng Y H , Hausknecht M , Vijayanarasimhan S , et al. Beyond short snippets: Deep networks for video classification[J]. 2015.
[51] Ullah A , Ahmad J , Muhammad K , et al. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features[J]. IEEE Access, 2018, 6(99):1155-1166.
[52] Rumelhart D E , Hinton G E , Williams R J . Learning Representations by Back Propagating Errors[J]. Nature, 1986, 323(6088):533-536.
[53] Jiang Z , Lin Z , Davis L . Recognizing Human Actions by Learning and Matching Shape-Motion Prototype Trees[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(3):533-547.
[54] Carreira J, Zisserman A, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[J]. 2017.
[55] Yang C, Xu Y, Shi J, et al, Temporal Pyramid Network for Action Recognition[J]. CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[56] Zhu G , Zhang L , Shen P , et al. Continuous Gesture Segmentation and Recognition Using 3DCNN and Convolutional LSTM[J]. IEEE Transactions on Multimedia, 2018, 21:1011-1021.

[57] Gkioxari G , Girshick R , Malik J . Contextual Action Recognition with R*CNN[J]. International Journal of Cancer Journal International Du Cancer, 2015, 40(1):1080-1088.
[58] Girshick R . Fast R-CNN[J]. Computer Science, 2015.
[59] Wang H , Schmid C . Action Recognition with Improved Trajectories[C]// 2013 IEEE International Conference on Computer Vision. IEEE, 2014.
[60] Yu S , Cheng Y , Xie L , et al. A novel recurrent hybrid network for feature fusion in action recognition[J]. Journal of visual communication & image representation, 2017, 49(nov.):192-203.
[61] Chéron, Guilhem, Laptev I , Schmid C . P-CNN: Pose-based CNN Features for Action Recognition[C]// IEEE International Conference on Computer Vision. IEEE, 2015.
[62] Ji J , Krishna R , Fei-Fei L , et al. Action Genome: Actions As Compositions of Spatio-Temporal Scene Graphs[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[63] Mcclelland J . Information Processing in Dynamical Systems: Foundations of Harmony Theory[C]// MIT Press, 1986.
[64] Baccouche M, Mamalet F, Wolf C, Garcia C, and Baskurt A, Spatio-Temporal Convolutional Sparse AutoEncoder for Sequence Classification[C]. Proc. of British Machine Vision Conf. (BMVC'12). 2012.
[65] Hinton G E , Osindero S , Teh Y W . A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2014, 18(7):1527-1554.
[66] Aapo Hyvärinen, Hurri J , Hoyer P O . Natural Image Statistics[M]// Natural Image Statistics: a probabilistic approach to early computational vision. Springer London, 2009.