Research on Human Action Recognition Method Based on Machine Learning

Yushan Li

School of Automation, Qingdao University, Qingdao, 266071, China

Abstract. Human action recognition (HAR) is an important research direction in the field of computer vision, and human action recognition technology provides an important foundation for computers to understand and simulate human behavior. It has been widely used in many practical applications, including human-computer interaction, gesture recognition, motion analysis, motion capture, virtual reality and augmented reality. Early methods were mainly based on the characteristics of manual design and traditional machine learning methods, such as support vector machine (SVM) and Random Forest. These methods usually rely on manually extracted features, such as edge, texture and color information, but they do not perform well in complex scenes and changing environments. With the rise of deep learning, especially the successful application of Convolutional Neural Network (CNN), great breakthroughs have been made in human action recognition. Using CNN, we can learn the feature representation directly from the original image data, which avoids the trouble of manually designing features. This paper expounds the research progress and future development trend of human action recognition methods based on machine learning.

Keywords: Human action; recognition method; machine learning; research progress.

1. Introduction

Human action recognition (HAR) refers to extracting features from videos or images to represent human behavior, and analyzing the extracted features through related algorithms, so that human behavior can be recognized. Its purpose is to enable the machine to automatically analyze human activities from videos. Early research usually used manual features [1]. However, with the continuous expansion of data scale, the traditional feature extraction algorithm can no longer meet the huge demand. Nowadays, on the basis of the original traditional machine learning, deep learning has developed rapidly, and its excellent and efficient feature learning ability has made breakthrough achievements in various application fields. action recognition has become one of the research topics in the field of computer vision [2]. It has great application prospects and commercial value in various fields, such as video surveillance [3], video retrieval [4], human-computer interaction [5] and so on.

2. Machine learning method based on artificial feature extraction

The research on action recognition can be traced back to 1973. Johansson [6] found through the experiment of moving light spots that the basic motion of human body can be recognized by tracking the motion trajectories of 10-12 human skeleton joints, which has since opened the prelude to the research on action recognition. Nowadays, according to the different methods of action feature extraction, researchers mainly divide it into machine learning method based on manual feature extraction and deep learning method based on automatic feature extraction. Figure 1 shows the current mainstream action recognition methods.

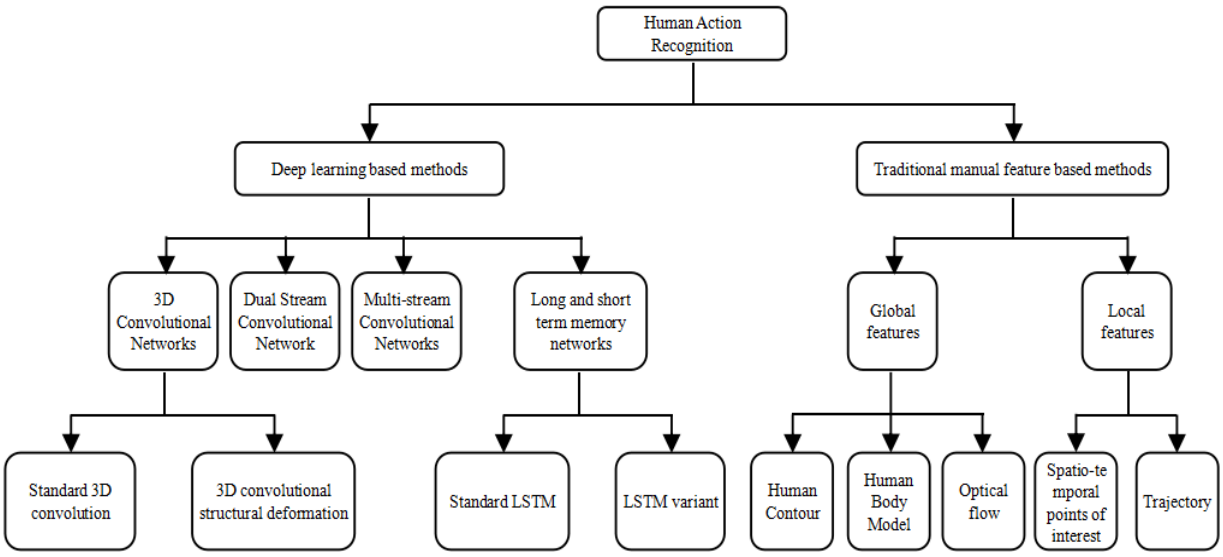

Figure 1. Research methods of human action recognition.

2.1. Method of human action recognition based on whole features

The whole feature representation method is to represent the whole of human body movements, human body structures and human body movements as a whole, and then analyze and identify them as a whole. The research of Aaron et al. [7] statically stores the motion energy image and the motion history image to illustrate the human motion in the image sequence. The motion energy image is used to obtain the initial position of motion, while the motion history image explains how the image moves. Then, according to Mahala Nobis distance, the storage model is matched and classified with the motion energy image and the motion history image calculated by the test video. Gorelick et al. [8] put forward the volume expansion of the motion energy image template, and expressed the three-dimensional shape as human action. Daniel et al. [9] suggested that the spatio-temporal volume should be represented as a moving historical image, and the three-dimensional feature map can improve the robustness of diverse viewpoints. Yilmaz et al. [10] proposed to recognize human actions through different attributes of the space-time volume, which is constructed by superimposing the outline of the target on the time axis, and the changes of attributes such as the shape of the space-time volume indicate the potential motion, but both of these methods need to separate the moving human body from the background, so the effect is not good in the case of complex dynamic background. Dollar et al. [11] proposed that the temporal and spatial feature points in the video should be taken as the research focus, and the movement trajectory of the feature points of the action appearance should be recognized to realize the action classification. However, the action recognition method based on the whole feature representation cannot effectively match the actions, and the recognition performance is poor.

2.2. Method of human action recognition based on local features

The local feature representation method is to use some human action areas in the scene to represent the motion features. Laptev [12] proposed a human action recognition method based on Space-Time Interest Points (STIPs). The local representation method first detects the interest points, then extracts local descriptors according to the interest points, and then aggregates the local descriptors. Laptev extends Harris corner detector to 3-3-dimensional, 3D Harris detector, and triggers 3D Harris through spatial characteristics and time information of action. 3D Harris detector can effectively identify large spatial changes and unsteady movements of human body. In order to solve a series of irrelevant spatio-temporal interest points caused by camera shake, Liu et al. [13] suggested using statistical attributes to remove irrelevant features of spatio-temporal interest points. Klaser et al. [14] in order to extract motion features more effectively, dense interest point trajectories are extracted by sampling

and tracking spatio-temporal interest points of multiple scales. The histogram of gradient and optical flow is extracted at each dense point to further improve the performance. Wang et al. [15] improved the dense motion trajectory feature, and obtained the motion feature by estimating the camera motion and Fisher vector coding each video. The proposed method can extract the motion features in fast motion and can effectively adapt to the changes of different motion speeds.

2.3. Deep learning method based on automatic feature extraction

2.3.1. Two-stream network method based on double-stream convolution

Inspired by the dual-channel theory of human vision, the dual-stream network framework divides video analysis into two streams: time stream and space stream. The dual-stream network uses two convolutional neural networks (CNN) to extract the temporal and spatial features of action videos respectively, and the periods do not interfere with each other. Finally, the extracted features are fused in a certain way and then classified.

The basic principle of Two Stream is to calculate the dense optical flow every two frames in the video sequence, and get the sequence of dense optical flow. In the traditional video recognition algorithm, the information of optical flow is always used for motion or behavior information. This network regards the optical flow as an image, and the optical flow itself is a vector. It can treat the X direction and the Y direction as two images, and then train a convolutional neural network for the optical flow image. Using two independent convolutional neural networks, CNN models are trained for video image (spatial) and dense optical flow (temporal) respectively, and the two branched networks judge the categories of actions respectively. Finally, the class score of the two networks is directly fused (including direct average and SVM) to get the final classification results.

Wang et al. [18] put forward an improved method on the basis of dual-stream network, using dual-stream network model to extract multi-scale convolution features. This scheme encodes advanced features into feature maps constrained by sampling trajectories, which can effectively collect information for a long time. Then, a time segment network (TSN) [19] is proposed to improve the structure of the dual-stream network model. They segment the whole video in the time dimension, then randomly extract the video frames of each segment, and obtain the context information of the long-time range of action by modeling the long-time series completely; Time-flow network uses warping optical flow diagram as input data to eliminate the influence of camera motion on model performance. At the same time, pre-training strategies and data enhancement technologies solve the problem of less training data sets, so as to further enhance the performance of the network model. Ali et al. [20] proposed to fuse temporal distinctive features with temporal coding, and then map them to low-level feature maps. The proposed method can not only extract long-time information between frames, but also encode it into dense representation. Li Qinghui et al. [21] put forward an action recognition model of ordered optical flow diagram based on double-stream convolutional neural network, which uses SVM to compress a series of continuous optical flow sequences into a single ordered optical flow diagram to realize the modeling of video context information; Then, the appearance and motion information of the action are captured by using the double-stream convolution network with RGB frames and optical flow diagrams as inputs respectively. Finally, the features are fused to recognize the action, and the network model captures the long-term motion information of the action. Double-stream convolution network takes RGB frame and optical flow diagram as the input of the model, extracts the spatial and temporal features of actions respectively, and classifies actions after fusing the features. Compared with single-stream network, this model improves the recognition ability of actions, but there is a problem that it is difficult to generate optical flow diagram. For the action with complex background, a lot of noise will be generated while generating the optical flow diagram, which will affect the identification performance of the double-flow network structure.

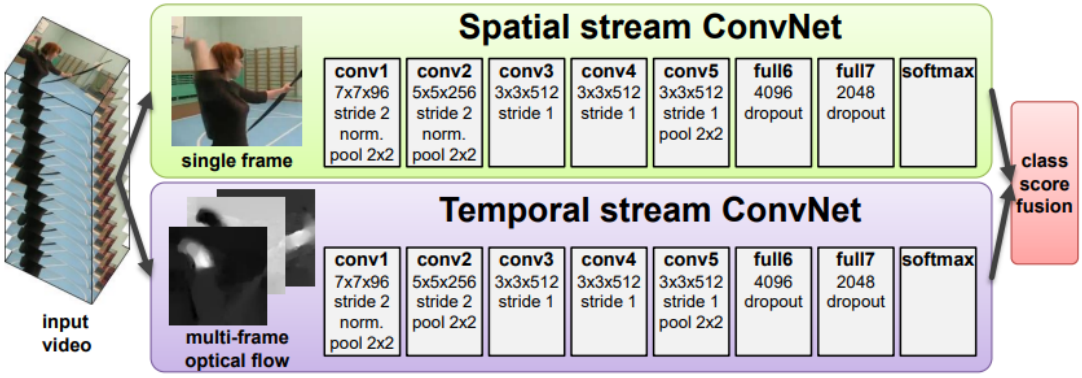

Figure 2. Two-stream architecture for video classification.

Although this method improves the performance of single-stream method by obviously capturing local time motion, there are still some shortcomings:

Because video level prediction is obtained by averaging the prediction scores of sampling clips, the long-term time information in the learned features is still lost.

Because the trained clips are evenly sampled from the video, they have the problem of wrong label assignment. The basic assumption that every clip is the same is inconsistent with the basic situation that the action may only occur in a small period of time in the whole video.

This method involves pre-calculating the amount of light flow direction and storing them separately. In addition, the training for the two streams is separated, which means that there is still a long way to go before the end-to-end training can land.

2.3.2. Based on 3D convolution neural network method

The 3D convolutional neural network extends the time dimension on the basis of the 2D convolutional neural network, and the 3D neural network can effectively extract the action spatio-temporal information. Ji et al. [22] proposed a 3D CNN model. Firstly, video frames were convolved in three dimensions to obtain multi-channel action information, and then multi-channel features were convolved and pooled continuously (i.e. subsampled). Finally, each channel feature was further mapped and fused to obtain action feature representation, and then the features were classified. Later, Tran et al. [23] proposed a 3D (C3D) method which is suitable for large-scale surveillance of video data sets and learning the spatio-temporal characteristics of actions. This method is a good attempt of deep learning model in the field of action recognition, and it is superior to 2D CNN method in many real scene data sets. Tran et al [24] improved the 3D CNN model in a deep residual network structure, and proposed a new 3D ResNet(Res3D) based on the Residual Networks,ResNet) model. Res3D's recognition performance for actions is much higher than that of C3D network model, and its time complexity is twice that of C3D, and its parameters are twice smaller than that of C3D. However, the capacity of 3D CNN structure is extremely limited, and the calculation cost and storage requirements are very expensive, which makes it very difficult to train very deep 3D CNN. Therefore, using 2D CNN to simulate or build 3D CNN may be an alternative.

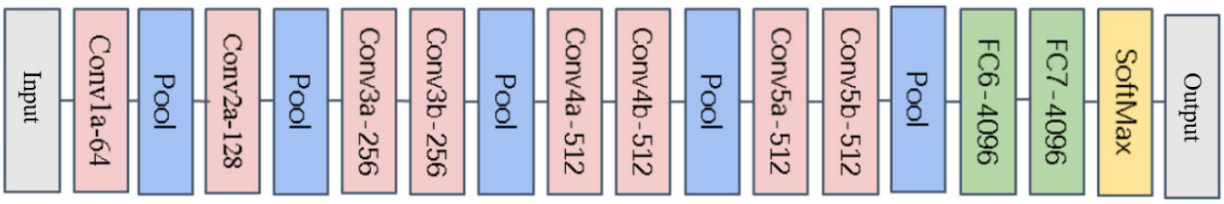

Figure 3. Architecture of C3D neural network.

2.3.3. Based on Recurrent Neural Network (RNN)

The method based on Recurrent Neural Network (RNN) usually combines recursive structures such as long short-term memory networks (LSTM) on the basis of 2D CNN to construct an action recognition model. Using the processing ability of LSTM to time series, the time information of action

can be obtained. Donahue et al. [29] studied the effectiveness of the combined structure of LSTM model and 2D CNN in modeling the action time dimension, and proposed a long-term circular convolution neural network, which abstracted the features of video frames into action text representations instead of common action feature representations. Joe et al. [30] also proposed a recursive neural network combining LSTM with the underlying CNN to identify video actions. Liu et al. [31] proposed a three-dimensional LSTM model for the classification and recognition of actions in video, and used a cyclic neural network with an increased spatial dimension to extract the temporal and spatial features of actions, further improving the performance of the network model for action recognition. Song et al. [32] take RNN as the basic framework and combine LSTM model to identify and classify actions. This method classifies actions by learning the joints of moving bones in a targeted way through attention mechanism. At the same time, a regularization model training strategy is proposed. Using this training strategy, the appearance and motion feature representation of the action are further enhanced, which fully proves the effectiveness of the proposed method. Srivastava et al. [33] proposed an LSTM action recognition model based on encoder and decoder. Firstly, the input video sequence was compiled into vectors with the same length, and then the vector was decoded by the decoder to obtain the motion category. The action recognition model based on the combination of convolutional neural network and LSTM can effectively capture the time sequence information of actions. LSTM algorithm solves the problem that RNN model cannot effectively deal with long-term action dependence due to gradient disappearance to a certain extent, and the hybrid model has inherent advantages in representing the spatial characteristics of actions.

Mask RCNN is a popular architecture for performing semantic and instance segmentation. The model predicts the position of the bounding box of various objects in the image and the mask of the semantic segmentation object at the same time. The basic architecture can be easily extended to human action recognition (As shown in Figure 4).

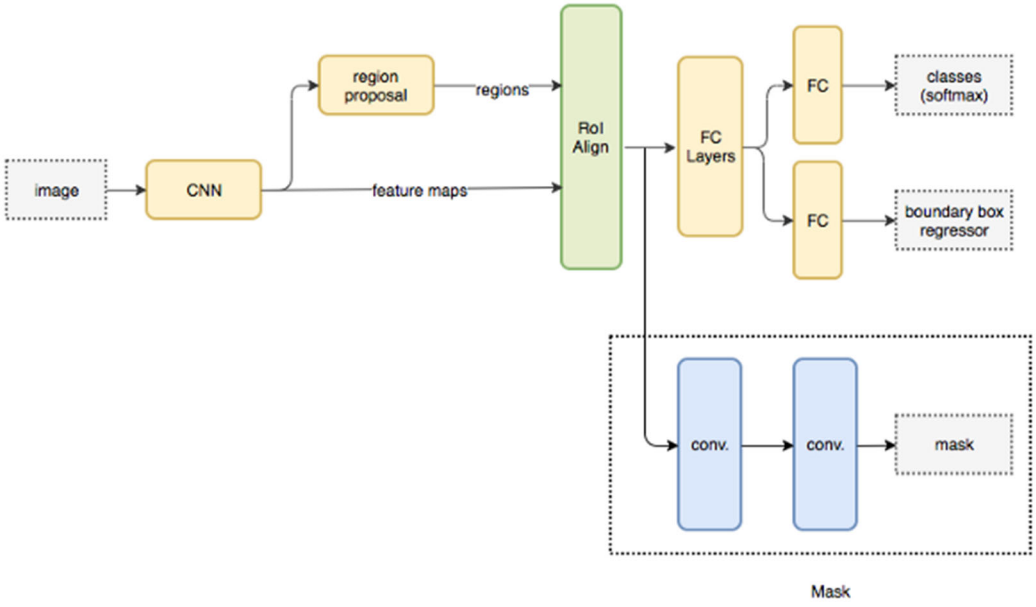

Figure 4. Mask RCNN architecture flow chart.

The basic architecture of Mask RCNN firstly uses CNN to extract feature maps from images. Region Proposal Network, RPN) uses these feature maps to obtain bounding box candidates of existing objects. The bounding box candidate selects an area (region) from the feature map extracted by CNN. Because candidate bounding boxes can have various sizes, a layer called RoIAlign is used to reduce the size of extracted features, so that they all have a uniform size. Now, this extracted feature is passed to the parallel branch of CNN for the final prediction of bounding box and segmentation mask. At the same time, the target detection algorithm can be trained to identify the position of people. By combining the position information of people and their key point sets, the human posture skeleton of each person in the image is obtained.

3. Problems and challenges in human action recognition

Although the action recognition technology combined with deep learning has been widely used in many application scenarios, and with the emergence of more and more data sets and the development of hardware technology, more researchers have joined the action recognition research, further promoting the development of this field. But so far, there are still many problems and challenges in action recognition technology, mainly from the following aspects:

3.1. Action background environmental factors

Action videos are collected not only in specific places (such as specific laboratories) but also in public places (such as shopping malls and streets). Whether the background is messy, whether the illumination is uniform, whether the camera is shaking, whether the human body is blocked, etc. will affect the integrity of the action, thus directly affecting the extraction of action features. These uncertain factors may seriously affect the accuracy of action recognition and classification.

3.2. Differences and Similarity of Actions

For the same action, different people will have different body shapes, different action ranges and different shooting angles, which will lead to differences within the action category. For different actions, the similarity of body posture usually leads to the similarity between actions. To some extent, it increases the difficulty of action recognition and classification.

3.3. Redundancy of data

Action recognition adopts a data set in the form of video, which is a series of continuous frame images in the time dimension. Therefore, there is data redundancy in video. Redundant adjacent frame images have strong correlation among pixels, contrast and saturation, which will not only affect the performance of action recognition, but also increase the computational complexity of the model.

3.4. Limitations of the network model

The research based on two-dimensional convolutional network can only be used to extract the spatial characteristics of action, but cannot capture the temporal characteristics; Double-stream network structure and three-dimensional convolution network have too high computational complexity and too long training time; The hybrid network model does not sufficiently extract the spatio-temporal characteristics of actions; The method based on global self-attention lacks local inductive bias and needs a lot of data training to obtain effective feature representation.

4. Research trends and prospects

4.1. Multi-feature fusion

Different forms of input will get different types of features after being processed by feature extraction model, which describes the human motion pattern in the video from different aspects. The emphasis of each feature is different, and only using a single feature for subsequent recognition and judgment will easily lead to wrong classification results. Many models extract features directly based on RCB data. With the application and development of camera equipment, RCB data has the advantage of being convenient for collecting fine-grained information, and its corresponding features can directly reflect the apparent and detailed texture information of the object. However, due to camera jitter, ambient lighting and occlusion in the process of video acquisition, RCB data usually contain a lot of background noise, which causes the complexity and variability of video data in space-time dimension, resulting in a large intra-class gap between different individuals' same actions, which further affects the video representation ability of classification features. Fusion of different types of features can combine the advantages of each feature to avoid the defects of single feature classification task.

4.2. Characterization of dynamic information

Dynamic motion information is the content of multi-ton difference in video data, which is used to describe the motion history. How to design a feature extraction mechanism to accurately describe the dynamic evolution of human movements in the time dimension is of great significance to correctly distinguish human movements in video. Some researchers use the characteristics of optical flow in video to represent human dynamic information, which eliminates the influence of irrelevant background factors while compensating time information. Although it improves the accuracy, the complexity of optical flow calculation is high and the memory cost is large, which greatly reduces the effectiveness and practicability of the model. In addition, the characteristics of optical flow often need to be calculated in advance, and the generation of optical flow video needs a lot of time and cost, which cannot achieve the effect of real-time classification and prediction. Therefore, it is of great practical significance to find a simple and efficient dynamic representation to replace complex optical flow calculation and reduce memory consumption. To meet the real-time requirements, it is also an unsolved problem to integrate the dynamic feature extraction process into the action recognition network for real-time prediction and analysis.

4.3. Feature screening

Video data contains a lot of redundant information. If all features are treated equally, it will lead to a lot of unnecessary features in the feature extraction process, which will interfere with the recognition results and increase the redundant calculation. Attention mechanism can imitate the visual attention mechanism used by human beings when observing the world, focusing on the core goals in the spatial area and the action fragments in the time dimension. In recent years, researchers have designed different spatio-temporal attention mechanisms, and tend to focus on the related research of restoration-level spatio-temporal attention, so as to assist the model to automatically screen important video tilts and their corresponding prominent spatial regions. However, the action information contained in adjacent cities is almost equal, so it is difficult to distinguish their importance. Some scholars try to solve the above problems by adding complex regularization, but the calculation and complexity of the model also increase, so it is also a good idea to shift the research focus from restoration-level attention to clip-level attention and assign different importance scores to different video clips. In addition, different convolution checks correspond to different channels to extract different types of features, so the features corresponding to different channels should also be treated differently. To sum up, how to adjust the attention mechanism to assist the model to flexibly screen the beauty-bond features is the key to improve the discriminant ability of the final classification features.

References

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

[2] Kong Y, Fu Y. Human action recognition and prediction: A survey[J]. International Journal of Computer Vision, 2022, 130(5): 1366-1401.

[3] Niu W, Long J, Han D, et al. Human activity detection and recognition for video surveillance[C]//2004 IEEE international conference on multimedia and expo (ICME) (IEEE Cat. No. 04TH8763). IEEE, 2004, 1: 719-722.

[4] Ramezani M, Yaghmaee F. A review on human action analysis in videos for retrieval applications[J]. Artificial Intelligence Review, 2016, 46: 485-514.

[5] Lou M, Li J, Wang G, et al. AR-C3D: Action recognition accelerator for human-computer interaction on FPGA[C]//2019 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2019: 1-4

[6] Johansson G. Visual perception of biological motion and a model for its analysis[J]. Perception & psychophysics, 1973, 14: 201-211.

[7] Aaron F, Bobick, James W, et al. The Recognition of Human Movement Using Temporal Templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001,23(3):257-267.

[8] Gorelick L, Blank M, Shechtman E, et al. Actions as Space-Time Shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007,29(12):2247-2253.

[9] Daniel W, Remi R, Edmond B. Free Viewpoint Action Recognition Using Motion History Volumes[J]. Computer Vision and Image Understanding, 2006,104(2):249-257.

[10] Yilmaz A, Shah M. Actions Sketch: A Novel Action Representation[C]. IEEE Computer Society Conference on Computer Vision & Pattern Recognition, 2005.

[11] Dollár 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.

[12] Laptev I. On Space-Time Interest Points[J]. International Journal of Computer Vision, 2005,64(2):107-123.

[13] Liu J, Luo J, Shah M. Recognizing Realistic Actions from Videos in the Wild[C]: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009.

[14] Klser A, Marszalek M, Schmid C. a Spatio-Temporal Descriptor Based on 3D-Gradients[C]: British Machine Vision Conference, 2010.

[15] Wang H, Schmid C. Action Recognition with Improved Trajectories[C]: 2013 IEEE International Conference on Computer Vision, 2014.

[16] Simonyan K, Zisserman A. Two-stream Convolutional Networks for Action Recognition in Videos[J]. Advances in Neural Information Processing Systems, 2014, 27.

[17] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal Multiplier Networks for Video Action Recognition[C]: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[18] Wang L, Qiao Y, Tang X. Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors[C], 2015.

[19] 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.

[20] Diba A, Sharma V, Van Gool L. Deep Temporal Linear Encoding Networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2329-2338.

[21] Li Qinghui, Li Aihua, Wang Tao, et al. Behavior identification based on ordered optical flow diagram and double-stream convolutional network [J]. Journal of Optics, 2018,38(06):234-240.

[22] Ji S, Yang M, Yu K. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013,35(1):221-231.

[23] Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C], 2015.

[24] Hounsfield G N. Computerized transverse axial scanning (tomography): Part 1. Description of system[J]. The British journal of radiology, 1973, 46(552): 1016-1022.

[25] Qiu Z, Yao T, Mei T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks[C]: Proceedings of the IEEE International Conference on Computer Vision, 2017.

[26] Qiu Z, Yao T, Ngo C, et al. Learning Spatio-Temporal Representation with Local and Global Diffusion[C]: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[27] Luo C, Yuille A. Grouped Spatial-Temporal Aggregation for Efficient Action Recognition[C]: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[28] Piergiovanni A, Angelova A, Toshev A, et al. Evolving Space-Time Neural Architectures for Videos[C]: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[29] Donahue J, Hendricks L A, Rohrbach M, et al. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(4):677-691.

[30] 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.

[31] Liu J, Shahroudy A, Xu D, et al. Spatio-temporal LSTM with Trust Gates for 3d Human Action Recognition[C]. European Conference on Computer Vision. Springer, Cham, 2016: 816-833.

[32] Song S, Lan C, Xing J, et al. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2018,27(7):3459-3471.

[33] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised Learning of Video Representations using LSTMs[C]. International Conference on Machine Learning. PMLR, 2015: 843-852.