ISSN: 2664-9640

Research progress on action recognition based on deep learning

Qing Wen [a], Jianjun Wu, Zhihui Li *

College of Information Science and Engineering, Henan University of Technology, Key Laboratory of the Ministry of Education (Henan University of Technology), Zhengzhou, China

[a]942699502@qq.com

Abstract

With the rapid growth of video data, action recognition technology promotes the application of video data insecurity, entertainment, and other fields. Traditional action recognition algorithms are difficult to meet existing needs. And deep learning has gradually become an important method of action recognition research. In order to promote the application of action recognition, it summarizes the development of deep learning. On the basis of deep learning network structure. It mainly elaborates on the existing two-stream method, 3D convolution network and graph convolution network with good effect in the field of action recognition, as well as the research and development of algorithm improvement based on these basic networks. In addition, the development and application trend of deep learning in action recognition are prospected, which provides reference suggestions for the next research of human action recognition based on deep learning.

Keywords

Deep learning; Action recognization; Two-stream ConvNet; 3D ConvNet; GCN.

1. Introduction

The popularity of the Internet has led to a large number of video data. The recognition of human action in the video can not only pay attention to the action of people and facilitate video retrieval but also identify abnormal actions and timely alarm. Video surveillance has become an important part of our life and is widely used in transportation, urban management, grain depots, and other places [1-2]. Since the last century, the video surveillance system is gradually established [3]. And with the popularity of social networks, more and more video information can be generated predictably. Therefore, when searching or processing videos, it is necessary to accurately judge the action in videos. And in some popular VR games, gesture control, and other technologies, it also has higher requirements for action recognition technology. The rapid growth of video data is a new opportunity in the field of video information processing, and the complexity of information also brings more challenges to action recognition. The general steps of human action recognition are motion region detection, feature extraction, and training of input video clips, which are then inputted into a classifier to classify and output recognition results. The research on action recognition has been done for decades, and a lot of progress has been made. With the development of machine learning, it is an important direction to apply deep learning algorithms in the field of action recognition in recent years. At present, the commonly used action recognition data sets include HMDB-51[4], Kinetics [5], and UCF101[6], which provide the data foundation for the training of deep learning framework.

2. Overview of Deep Learning Algorithms

Deep learning imitates human neurons and extracts features layer by layer to solve problems. The concept of the neuron was first put forward in 1943, which provided heuristic influence for

ISSN: 2664-9640

the development of artificial intelligence. The concept of perceptron determines the basic neuron model of machine learning. At first, the neural network was limited to linear separable problems, but the idea of calculating problems according to the network model is of great significance to the development of artificial intelligence. The emergence of the BP neural network emphasizes the superposition of multiple hidden layers. In 2006, Hinton pointed out that the output of the upper layer in the network can be used as the input of the next layer, which makes the neural network move towards "depth". Then deep learning developed rapidly. In the ImageNet image recognition competition in 2012, the deep neural network model AlexNet reduced the error rate by half for the first time in the image recognition competition, and the error rate was lower than that of human recognition for the first time; AlphaGo defeated world Go champions Li Shishi and Ke Jie in 2016 and 2017 respectively, which is a major breakthrough in the history of artificial intelligence [7].

3. Overview of Convolutional Neural Networks

Convolution neural network is inspired by the concept of "receptive field" of the visual system, and adopts the combination of forwarding propagation and backward propagation to calculate input values and backward adjustment parameters respectively. The basic structure of CNN is composed of input layer, convolution layer, pooling layer, full connection layer and output layer. It is an end-to-end design, which avoids the manual feature extraction process. The essence of convolution layer is to extract features through convolution kernel, pool layer is to process the extracted features, so as to reduce the amount of data while retaining useful feature information, and full connection layer is to carry out regression classification on the features extracted from previous layer-by-layer transformation and mapping [8]. There is no need to activate functions in general linear problems, but in such problems as action recognition, it is necessary to change the linear outputs of neurons in neural networks into nonlinear outputs. The functions reflecting this change are called activation functions, and commonly used activation functions include sigmoid, tanh, ReLU, etc. [9].

In recent years, many improvements have been made to the convolution layer, pool layer and full connection layer of convolutional neural network, and many neural network frameworks have been put forward based on convolutional neural network for processing various information. The mainstream convolutional neural network framework is shown in Table 1.

Table 1 Mainstream volume and neural network framework

Frame name	Author	Proposed time
LeNet	Yann LeCun et.al	1994
AlexNet	Alex Krizhevsky et.al	2012
VGGNet	VisualGeometry Group	2014
GoogleNet	Christian Szegedy	2014
ResNet	He Kaiming et.al	2015
DenseNet	Liu Zhuang et.al	2017
SqueezeNet	Landola	2018

4. Overview of Recurrent Neural Network

Traditional neural networks have an input layer, hidden layer and output layer, which are connected in turn. Recurrent Neural Network provides a new structure: connecting neurons in the same layer with each other. Therefore, the RNN can utilize the previously input information while utilizing the current input information. Generally speaking, RNN provides a structure suitable for the desired functions for processing context-sensitive sequence data, and has

ISSN: 2664-9640

unique advantages in the fields of text processing and speech processing. Bidirectional RNN, proposed in 1997, can learn the future input information, so as to better determine the current context. In order to solve the problems of gradient disappearance and gradient explosion in RNN training, a Long Short-Term Memory network is proposed, and the concepts of the memory unit and adjusting gate are introduced. When the adjusting door is opened, the LSTM memory unit can start to process data, otherwise, it stops learning and retains the memory. On this basis, some variants of LSTM have appeared, mainly aiming at the improvement of the regulating door [10]. In 2014, GRU (Gated Recurrent Unit) was put forward, which retained the long-term memory ability of LSTM model, and used a simpler structure to achieve similar results to LTSM [11]. According to the development trend in recent years, the research on LSTM is a hot and important issue for the direction of circulating neural network, and the improvement of LSTM can be considered from the direction of adjusting the gate. In 2018, Yang Li and others proposed to combine RNN with attention mechanism and CNN, which is also the future development direction [12].

5. Action recognition method based on deep learning

For human action recognition, we mainly detect moving images in the video. The information extracted from video includes static features of video frames and dynamic features between adjacent video frames. Traditional static feature extraction methods in video action recognition include:(1) Local Binary Pattern (LBP)[13], which has the characteristics of rotation invariance and gray invariance, and then LBP is improved to adapt to texture features of different scales [14], (2) HOG (Histogram of Oriented Gradients) features [15], (3)haar characteristics and extended Haar-like characteristics [16]. Commonly used dynamic region detection methods are as follows: (1) optical flow [17], which uses the mutual motion between the target and the background to detect the moving target due to the different motion vectors between the target and the background; (2) temporal difference [18], which makes use of the continuity of video sequence to differentiate the gray values of corresponding pixels of adjacent frame images, According to the comparison between the difference value and the threshold value, whether the target moves or not can be judged. (3) Background subtraction [19], which is similar to the idea of the inter-frame difference method, except that it differentiates with a constantly updated background image to obtain a moving image. Before deep learning was widely used in action recognition, the most commonly used method for action recognition was the dense trajectory algorithm, and then improved IDT (Improved Dense Trajectories) algorithm was proposed, which mainly focused on the description of human action by eliminating background optical flow [20]. The recognition of images or videos by machines is presented in the form of matrix data. Because the same action is different in different people or different perspectives, and there are problems such as moving objects blocking or similar actions in different backgrounds, relevant researchers turn their eyes to the depth network. At present, the mainstream video action recognition structures include 3D convolution network, two-stream network, graph convolution network ,and so on [21].

Commonly used public data sets are shown in the table 2.

5.1. Action recognition method based on two-stream network

The idea of the two-stream network is to recognize the dense optical flow between static frames and video frames, then integrate the two recognition results. Two-stream convolution network learning the method of human eyes recognizing moving objects divides video into the temporal stream and spatial stream, and integrates the results produced by different models.

In 2014, Simonyan K et al. proposed a two-stream convolutional neural network [22].In this method, the dense optical flow field of multiple frames in the video is taken as input to extract the action information in the video data. Therefore, the two-stream CNN is composed of partial

ISSN: 2664-9640

networks, which process two kinds of information respectively, and each network will output a softmax. Finally, the results are fused by the SVM method. For time-flow convolution network, in order to avoid the problem caused by too few training sets, two softmax outputs are trained by two tasks in one network, and the outputs are added together as the total error to execute BP algorithm to update the weights of the network. The structure of the two-stream convolution network is shown in Figure 1. However, this method has some limitations, such as shallow layers, limited fitting when processing time-domain information, and small training data set. Later, Wang et al. [23] proposed Very Deep Two-stream ConvNets, initialized with ImageNet training model, trained the network by clipping multi-scale, setting new Highdropout coefficient in full connection layer, and multi-GPU, which improved the experimental results on UCF101, it also solved the problem of insufficient training set by different clipping methods, and improved the network training efficiency by increasing the number of network layers and multi-GPU.

Table 2 Commonly used public data sets

Data sets	Year	Number of categories	sample size	introduction
Sport-1M	2014	487	1,200,000	Sports video classification data set
UCF101	2013	101	133,320	Interaction between people and objects, simple limb movements, interaction between people, playing, sports and so on
HMDB51	2011	51	6,849	Facial action, facial operation and object operation, general body action, interactive action, human behavior action, etc.
Kinetics	2017	600	50,000	Single person behavior, double person behavior, character interaction behavior, etc.
NTU RGB+D	2017	60	56,880	RGB video, depth map sequence, 3D skeleton data and infrared video of action samples

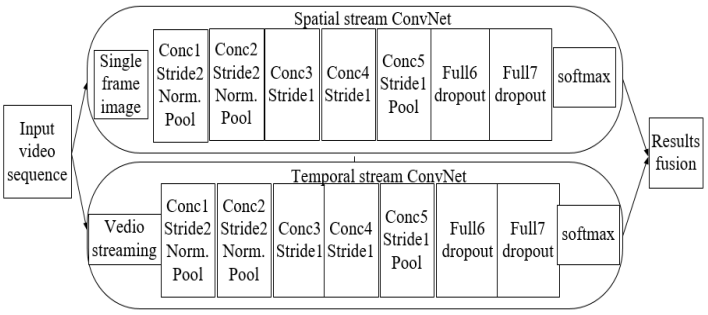

Fig. 1 Two-stream convolutional network structure

TSN(Temporal Segment Network) is proposed by Wang et al. [24], which realizes the recognition of human action by two-stream convolution network under TSN structure. The structure of TSN is shown in Figure 2. It improves the detection degree of action by increasing RGB change and distorting optical flow feature input. At the same time, TSN also added a sparse time sampling strategy to solve the problem of poor modeling of the long video in a two-stream network, randomly selected video segments and input them into the network, fused different output results, and judged the category by averaging. Karen S et al. put forward a new network architecture based on the two-stream network [22], adding a new convolution fusion layer and time fusion layer, highlighting the relationship between time and space changes. One of the problems of the two-stream network is that two streams can't influence each other. This method solves two problems: how and where to merge two streams [25]. This method learns

ISSN: 2664-9640

highly abstract features of convolutional neural networks, but it still needs to solve the problems of the small data set and much noise.

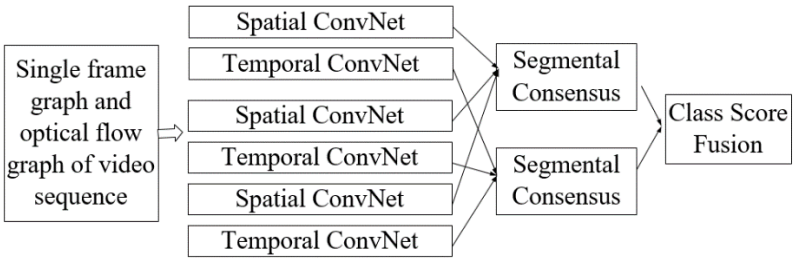

Fig. 2 TSN network structure

Fernando[26-27] proposed to use the parameters of the ranking function to encode the video frame sequence, using the parameters of the ranking machine as a new expression of action recognition, and using the Rank-Pooling method to realize end-to-end learning of model parameters and video features. Zhu Y[28] proposed Hidden Two-Stream ConvNets, adding a MotionNet before the time stream. MotionNet is a full convolution neural network, which can generate motion information inside the network and realize end-to-end action prediction, thus solving the problem that other two-stream networks [22-25] need to calculate optical flow information before input. At the same time, the standard pixel-level reconstruction error function, smoothing loss function and structural similarity loss function are used to solve the fuzzy problem of motion information and help learn the frame structure. The recognition speed is 10 times faster than that of the common two-stream neural network, and the correct rate on UCF-101 is 93.1%.

On the basis of the continuous improvement of the accuracy of the two-stream network, it is more enlightening for action recognition. Under the condition of keeping the original two-stream convolution network structure, the LSTM network is added to the time stream to meet the end-to-end training of data, and new error function combinations are tried on the traditional structure to get better optical flow information [29]. Cheron et al. [30] divide RGB and optical flow images according to human joint features as the input of the two-stream network for feature fusion and SVM recognition. In the process of action recognition, it is necessary to meet the accuracy, and it is also required to recognize human action in real-time to meet the alarm requirements in some scenarios. After trying and improving the framework, parameters, fusion algorithm and different classifiers of the two-stream network, the accuracy of action recognition can basically meet the needs. However, in the process of action recognition, the two-stream network runs through two networks and needs to calculate optical flow information, so there is still much room for improvement in real-time. At the same time, because the two-stream CovNet uses the method of averaging the decision results of continuous multiple frames to obtain video sequence information, a lot of time-domain information will be lost in practical application, which also affects the accuracy of recognition.

5.2. Action recognition method based on 3D convolution network

3D convolution is an extension of 2D convolution. Ji et al. put forward 3D convolution neural network for the first time in 2010 [31], and based on this, constructed 3D convolution neural network to convolute and pool videos in multiple channels, and enhanced features through high-level feature models. The convolution kernel of 3D convolution layer is three-dimensional, which can extract features from multiple continuous frames on a three-dimensional scale, so as to extract features in time series. See Figure 3 for the process of 3D convolution. Then, based on 3D convolution [31], 3D CNN [32] for action recognition is proposed, which includes a hard-wired layer, three convolution layers, two down-sampling layers and a fully connected layer. Each 3D convolution kernel convolution consists of seven consecutive frames. Each frame

ISSN: 2664-9640

extracts the information of five channels of grayscale, gradients in X and Y directions and optical flow, and carries out 3D convolution on each channel. After multi-layer convolution and pooling, the feature vector reflects the current motion information. This method standardizes the model by training a large number of frames, calculates dense sift descriptors based on gray images, and then combines them with moving edge historical images to construct advanced action assistant features to output the final recognition results. Combining 3D convolution networks with different network architectures, many deep networks for action recognition are derived.

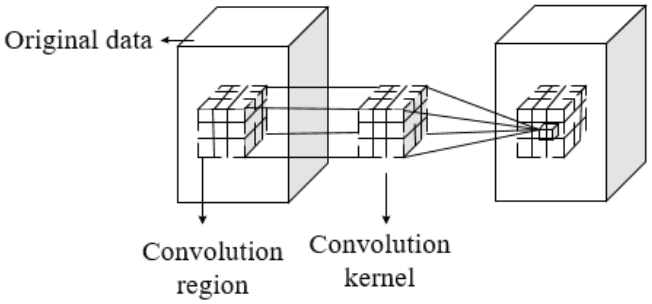

Fig.3 3D convolution process

Tran et al. [33] designed C3D network structure with convolution kernel 3*3*3 and step size 1*1*1. C3D firstly pays attention to the appearance of the previous frames, tracks the significant motion in the subsequent frames, extracts C3D features and inputs them into the multi-class linear SVM for training models, and uses PCA to reduce the dimension of C3D descriptors. Compared with ImageNet features and IDT features, C3D has better compressibility and faster retrieval speed for a large number of data. This method is more suitable for learning in time domain and space domain, and its output is better when it is input into simple classifier.

In 2017, Tran et al. [34] continued to explore a better network structure based on C3D[33] and proposed a Res3D network model based on ResNet, which uses an 18-layer network model. The final Res3D is twice as fast as C3D[27], and its model size is half of C3D's, and the test results on UCF-101 and HMDB-51 have been improved. Hara et al. proposed 3D ResNets[35] network based on 2D ResNet and Res3D[28] network, experimented with 18-layer and 34-layer networks, and opened up a variety of deep networks and effective models pre-trained on Kinetics for spatio-temporal feature extraction. Li et al. [36] put forward joint spatio-temporal feature learning operation (CoST), and constructed neural network by joint learning spatio-temporal features under the constraint of weight sharing. This method decomposes the 3D vector of video sequence into three 2D image sets, and convolves them separately, so that each frame has rich action information. Experiments show that the performance of CoST is better than C2D and C3D.Compared with C2D, this network can learn spatio-temporal features jointly. Compared with C3D[33], CoST is carried out through two-dimensional convolution. The method of learning spatio-temporal features jointly from multiple perspectives in this paper can replace C3D and C2D well, which is a new idea for future research.

On the premise of the residual network, Qiu et al. [37] put forward pseudo-3D residual network (P3D ResNet), which uses 1*3*3 convolution and 3*1*1 convolution instead of 3*3*3 convolution to extract temporal and spatial features respectively. The computation is much smaller than that of 3D convolution, and two-dimensional convolution parameters trained on image data sets can be used.P3D ResNet designed three residual network structures combining series and parallel, and finally, P3D ResNet was obtained by connecting the three structures, which showed more aggregation for the same type of features. The validity of P3D is verified by comparing it with other algorithms on several datasets such as UCF101.

Diba et al. proposed T3D(Temporal 3D ConvNets) model, which introduced TTL (Temporary Transition Layer) on the basis of DenseNet network architecture [38] to process video information with different lengths. The TTL layer structure is shown in Figure 4. The network

ISSN: 2664-9640

outputs features through DenseNet pre-trained in ImageNet and 3D convolution network with random initialization weights, connects the two features into the full connection layer, judges whether the input image sequences of the two networks are consistent, and realizes 3D network migration learning. The best experimental results were obtained on UCF101 and HMDB51 data sets.

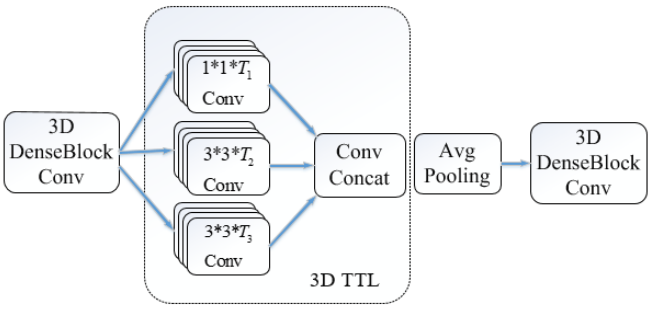

Fig.4 TTL layer structure

Carreira et al. put forward a new video data set Kinematics and I3D networks [39], which has a better ability to obtain information from the original video. The network model trained by Kinematics training set is migrated to UCF101 and HMDB51, and the recognition effect is very good.I3D expands 2D convolution kernel into 3D convolution kernel, optimizes I3D on the basis of 2D convolution kernel parameters pre-trained by ImageNet, and adds optical flow input into the network to obtain the fusion output result of RGB and optical flow. This method solves the problem of re-training parameters in 3D network, and further improves the performance of the algorithm by combining with the two-stream idea. On this basis, R(2+1)D[40] trained by Tran et al. decomposes the 3D convolution factor into 2D+1D, and processes the temporal and spatial information of information. As a result, it can achieve the effect similar to 3D convolution network, but the parameters are much reduced and easy to optimize. This kind of network structure increases the nonlinear mapping of the network, and the model with higher complexity can solve the fitting problem better.

Zhou et al. [41] considered the disadvantage that 3D CNN has lower accuracy when processing pictures than traditional convolution networks, and proposed a model combining 2D and 3D CNN for action recognition, namely mixed convolutional tube(MiCT), which combines 3D and 2D convolution networks to generate feature maps with more information and reduce the training complexity of spatio-temporal fusion. MiCT models are stacked together to form a new end-to-end MiCT-Net, which is used to explore the spatio-temporal information of human body movements. This method greatly reduces the complexity of the model and does not need to learn redundant data. Combining 3D and 2D CNN, Luvizon et al. [42] highlighted that using multi-task network to do 2D and 3D attitude estimation and 2D and 3D action recognition at the same time, at the same time, using the results of attitude estimation to improve the accuracy of action recognition, and inputting static RGB images into the whole frame to perform attitude estimation and action recognition at the same time, and combining the results to generate action tags. This method improves the accuracy of recognition by mutual promotion of two tasks, and provides a new way of thinking for action recognition.

3D convolution processes the video sequence through stereo convolution kernel, and the feature information of time and space can be obtained at the same time by one convolution, which is more convenient and faster than the two-stream network with 2D convolution; However, 3D convolution network also requires higher storage capacity and computing capacity, and the computing power of computer in general scenes is difficult to meet the demand, and the recognition accuracy of 3D convolution is relatively lower than that of two-stream network. Because of these drawbacks, the action recognition method based on 3D

ISSN: 2664-9640

convolution should be combined with 2D convolution or two-stream networks in the future to explore the action recognition method with less computation and higher accuracy.

5.3. Other methods of action recognition

Yan et al. [43] proposed a spatio-temporal graph convolution network model (ST-GCN), which is used to solve the problem of human action recognition based on key points of the human skeleton. At the same time, considering the adjacent joints of the skeleton in space and time, the concept of neighborhood is extended to the time axis. Si et al. [44] proposed that attention-enhanced graph convolution LSTM network (AGC-LSTM) combined with human skeleton information for action recognition, and explored different graph processing methods, which can not only capture the discriminant features in space and time, but also explore the relationship between time domain and space domain. The network uses the attention network to enhance the information of key nodes in each AGC layer, and achieves the best recognition accuracy at present. Skeleton graph is a special topological structure. Compared with general topological graph, human skeleton topological graph has better stability and invariance. It is a new method in the field of action recognition to combine the skeleton information in human action with graph convolution. The bending and falling actions of human body contain complex inter-frame time information and inter-frame space information. By comprehensively exploring the temporal and spatial characteristics of motion through graph convolution network, we can dig the characteristics and laws of actions more deeply.

RNN has good performance in time series modeling. LSTM solves the problems of gradient disappearance and gradient explosion in the RNN training process to a certain extent, and has a little restriction on the length of the input video.In the process of action recognition by LSTM method, usually, the spatial features of video are extracted by convolution neural network, and then the extracted features are input into LSTM network in time order to characterize the temporal information of video, and then the action is recognized by spatio-temporal information [45]. Donahue et al. [46] put forward Long-term Time Convolution Recursive Network (LRCN), which combines CNN and LSTM to extract features from video data. For the experiment of activity recognition, two structures of LRCN are studied: adding LSTM after the first fully connected layer (LRCN-fc6) of CNN and adding LSTM after the second fully connected layer (LRCN-fc7) of CNN. Experiments show that the recognition accuracy is improved by 1.6% when LSTM has 1024 hidden units. This model supports longer input and output video sequences and provides more methods for action recognition in video. Qin et al. [47] combined 3D convolution network and LSTM to get a fusion model. The processed features containing time information are sent to LSTM for processing, which has obvious video learning effect on human action, and its generalization performance and robustness are higher than those of pure 3D convolution and LSTM network. The biggest difference between the cyclic neural network and other networks is that it can realize some kind of "memory function", which can better deal with timing problems such as action recognition in video.

Deep Belief Nets (DBNs) is a Boltzmann machine formed by every two hidden layers, and the Boltzmann machines obtained by all hidden layers are connected in series to obtain a network model of deep learning. The Boltzmann machines combined with the constraint that they cannot be connected by themselves constitute a restricted Boltzmann machine, and a group of RBM obtained by multiple RBMs through greedy learning constitutes DBNs[48].On the basis of the deep belief network, the whole deep learning process is divided into pre-training, encoding and decoding, and fine-tuning. The main idea of using greedy unsupervised learning algorithm is to learn unsupervised on each layer in DBNs, and finally supervise and fine-tune the whole network [49]. Due to the problem of over-fitting in action recognition, random Dropout DBN is designed on the basis of traditional DBN, and the probability parameters in Dropouts algorithm are randomly changed. After experimental verification, the accuracy of deep belief network

ISSN: 2664-9640

with random Dropout is 4.5% higher than that of traditional DBN [50]. Deep belief network can not only be used for unsupervised learning, but also acts like a self-coding machine; It can also be used as a classifier in supervised learning. Using DBNs for action recognition, we can use unlabeled data for unsupervised learning.

6. Summary and prospect

Deep learning is experiencing a rapid development era, and it is a challenging direction to apply deep learning to action recognition. Whether it is two-stream method, 3D convolution or graph convolution network, it is to establish dynamic information in video frame information, so as to realize the recognition of actions, and then distinguish actions by classifiers.

At present, in the field of action recognition, the main problems to be solved are:(1) On the existing database, the accuracy of action recognition has achieved good results. However, with the generation of more and more video data, it is still facing great challenges to realize efficient and accurate online action recognition. Using a large amount of data to train deep learning networks requires higher hardware. How to use a small amount of data to complete network training and improve recognition efficiency is a problem worth exploring. (2) The video information has the characteristics of complex background and unstable illumination intensity. In the process of action recognition, the video is preprocessed by combining foreground extraction and other methods to provide convenience for action recognition; (3) At present, most action recognition methods of deep learning use optical flow information as dynamic information to process video sequences, but the extraction of optical flow information is more computationally intensive than RGB, LBP and other features, and exploring more dynamic features is the next direction to be worked hard; (4) Abnormal action alarm, real-time action detection and other fields have different development needs for action recognition, (5) How to further identify detailed actions and distinguish similar actions; (6) How to reduce network parameters and computation; (7) What kind of network structure to design, etc., is still worth studying in the following work.

Acknowledgments

The authors acknowledge the National key research and development project (No: 2018YFD0401404). Doctor Fund of Henan University of Technology (2017BS034). Project of Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education.

References

[1] Liu Ying, Hu Nan, Yang Jingwei. Detection and recognition of staff in power grid monitoring video based on deep learning [J] . Journal of Shenyang University of Technology, 2019,41 (05): 544-548.

[2] Yang Zhenhe, Su Zhenhua, Wang Dong. Study on risk analysis and prevention of injury accidents of personnel in and out of grain depot [J]. Grain Technology and Economy, 2017, 42 (04): 10-15.

[3] Zheng Dingchao, Chen Caiwei. Intelligent video surveillance system design [J] . Automation and applications, 2020,039(003) : 91-93.

[4] Kuehne H,Jhuang H,Stiefelhagen R .HMDB-51:a large cideo database for human motion recognition[C]//IEEE International Conference on Computer Vision.2011:2556-2563.

[5] Kay W,Carreira J,Simonyan K. The Kinetics human action video dataset[EB/OL].2017-03-19[2019-11-27]. https://arxiv.org/pdf/1705.06950.pdf.

[6] Soomro K, Zamir A R, Shah M. UCF101:a dataset of 101 human actions classes from videos in the wild[EB/OL].2012-12-01[2019-11-27].http://crcv.ucf.edu/data/UCF101.php.

[7] Tan Xiaofeng. An overview of deep learning development [C]//national security geophysics Committee of China Geophysics Society, military geophysics Committee of Shaanxi Geophysics Society. Gansu: China Geophysics Society, 2019:252-260.

[8] Su Xuewei. Research on Human Abnormal Behavior in Video Surveillance Based on Deep Learning [D] .Xi'an: Xi'an University of Science and Technology, 2019.

[9] Yang Bin, Zhong Jinying. Review on the research progress of convolution neural network[J]. Journal of Nanhua University (Natural Science Edition), 2016 (3): 66-72.

[10] Wu Xiaoying, Li Rui, Wu Shengxi. Behavior Recognition Algorithm based on CNN and Bidirectional LSTM. Computer Engineering and design, 2020,41(02) : 361-366.

[11] Su Tongtong, sun Huazhi, Ma Chunmei. Human behavior recognition based on cyclic neural network [J].Journal of Tianjin Normal University (Natural Science Edition), 2018, 38 (06): 58-62 + 76.

[12] Yang Li, Wu Yuxi, Wang Junli. A review of cyclic neural network research [J]. computer applications, 2018,38 (S2): 1-6 + 26.

[13] Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions[J]. Pattern Recognition,1996,29(1):51-59.

[14] Zhang Qian,Li Hai Gang,Li Ming,Ding Lei. Feature extraction of face image based on LBP and 2-D Gabor wavelet transform.[J]. Mathematical biosciences and engineering : MBE,2019,17(2).

[15] Li penglin, Zou Jiacheng, Li Wei. Face detection and tracking based on hog and feature descriptors [J]. Journal of Zhejiang University of technology, 2020,48 (02): 133-140.

[16] Liu Ji. Research on face detection algorithm based on self-learning feature fusion [D]. Qingdao: Ocean University of China, 2015.

[17] Yu Shuchun,Wang Qi,Ru Changhai,Pang Ming. Location detection of key areas in medical images based on Haar-like fusion contour feature learning[J]. Technology and Health Care,2020,28(S1).

[18] Xiong Wei, Wang Chuansheng, Li Lirong, Liu Min, Zeng Chunyan. [J] video stabilization algorithm combining optical flow and Kalman filter. Computer Engineering and science, 2020,42(03) : 493-499.

[19] Liu Zhongmin, He Shengjiao, Hu Wenjin. Moving object detection of video sequence based on background subtraction [J]. Computer Application, 2017,37 (06): 1777-1781.

[20] Guian Zhang,Zhiyong Yuan,Qianqian Tong,Mianlun Zheng,Jianhui Zhao. A novel framework for background subtraction and foreground detection[J]. Pattern Recognition,2018,84.

[21] Jun Wang,Limin Xia. Abnormal behavior detection in videos using deep learning[J]. Cluster Computing: The Journal of Networks, Software Tools and Applications,2019,22(11).

[22] Karen S, Andrew Z. Two-stream convolutional networks for action recognition in videos[EB/OL].2014-11-12. https://arxiv.org/pdf/1406.2199v2.pdf.

[23] Wang L, Xiong Y, Wang Z. Towards good practices for very deep two-stream convnets[EB/OL]. 2015-07-08. https://arxiv.org/pdf/1507.02159.pdf.

[24] Wang L , Xiong Y , Wang Z . Temporal Segment Networks: Towards Good Practices for Deep Action Recognition[J]. Computer Science, 2016:20-36.

[25] Feichtenhofer C , Pinz A , Zisserman A . Convolutional Two-Stream Network Fusion for Video Action Recognition[J]. 2016 IEEE Conference on Computer Vision and Pattern Recognition,2016:1933-1941.

[26] Fernando B,Gould S.Learning end-to-end video classification with rank-pooling[C]//Proceesings of the 33rd International conference on international conference on machian learning. 2016:1187-1196.

[27] Fernando B , Gavves E , Oramas J . Rank Pooling for Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(4):773-787.

[28] Zhu Y，Lan Z Z，Newsam S．Hidden two-stream convolutional networks for action recognition[J] // Computer Vision-ACCV 2018 Cham:Springer International Publishing，2019: 363-378.

[29] Zeng Mingru, Zheng Zisheng, Luo Shun. Dual flow convolution human behavior recognition combined with LSTM [J]. Modern Electronic Technology, 2019,42 (19): 37-40.

[30] Cheron G,Laptev I,Schmid C.P-CNN:Pose-based CNN features for action recognition[J].Proceeding of the IEEE International Conference on Computer Vision.2015:3218-3226.

[31] Zhang Ying, Yuan Hejin. Human behavior recognition based on 3D convolutional neural network [J]. Software Guide, 2017,16 (11): 9-11.

[32] Ji S W， Xu W, Yang M. 3D convolutional neural networks for human action recognition[J]. Proceedings of the International Conference on Machine Learning. 2010:495-502.

[33] Tran D , Bourdev L , Fergus R . Learning spatiotemporal features with 3D convolutional networks[J]. 2015 IEEE International Conference on Computer Vision,2015:4489-4497.

[34] Tran D , Ray J , Zheng S . ConvNet architecture search for spatiotemporal feature learning[EB/OL].2017-08-16. https://arxiv.org/pdf/1708.05038.pdf.

[35] Hara K , Kataoka H , Satoh Y . Learning spatio-temporal features with 3D residual networks for action recognition[J].2017 IEEE International Conference on Computer Vision Workshops, 2017:3154-3160.

[36] Li C , Zhong Q , Xie D . Collaborative spatio-temporal feature learning for video action recognition[EB/OL]. 2019-03-04 [2020-1-25]. https://arxiv.org/pdf/1903.01197.pdf.

[37] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks [C]//The IEEE International Conference on Computer Vision (ICCV), 2017: 5533-5541.

[38] Diba A , Fayyaz M , Sharma V . Temporal 3D ConvNets: new architecture and transfer learning for video classification[J]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)， 2017：1063-6919.

[39] Carreira J , Zisserman A . Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)， 2017：1063-6919.

[40] Tran D, Wang H, Torresani L. A closer look at spatiotemporal convolutions for action recognition[C]//IEEE Computer Vision and Pattern Recognition(CVPR),2018:6450-6459.

[41] Zhou Y , Sun X , Zha Z J . MiCT: Mixed 3D/2D convolutional tube for human action recognition[C]// 2018 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018：449-458.

[42] Luvizon D C , Picard D , Tabia H . 2D/3D pose estimation and action recognition using multitask deep learning[J].2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018:5137-5146.

[43] Yan S , Xiong Y , Lin D . Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[EB/OL].2018-01-25[2020-2-20]. https://arxiv.org/pdf/1801.07455.pdf.

[44] Si C, Chen W, Wang W.An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition[EB/OL].2019-02-25[2020-2-20]. https://arxiv.org/pdf/1902.09130.pdf.

[45] He Qingqiang. Research on video content recognition technology based on deep learning [D]. Chengdu: University of Electronic Science and Technology, 2017.

[46] Donahue J, Hendricks L A, Rohrbach M. Long-term recurrent convolutional networks for visual recognition and description[J].IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014,39(4):677-691.

[47] Qin Yang, Mo Lingfei, Guo Wenke, Li v. combination and application of 3D CNNs and lstms in behavior recognition [J]. Measurement and Control Technology, 2017,36 (02): 28-32.

[48] Luo Xiaohuan. Polarimetric SAR image classification based on depth confidence network [D]. Xi'an: Xi'an University of Electronic Science and Technology.

[49] Fan Heng, Xu Jun, Deng Yong. Human behavior recognition based on deep learning [J]. Journal of Wuhan University (Information Science Edition), 2016,41 (04): 492-497.

[50] Wang Zhongmin, Wang Xi, Song Hui. Mobile user behavior recognition method based on random dropout deep belief network [J]. Sichuan: computer application research, 2017,34 (12): 3797-3800.