Computer-Aided Recognition and Analysis of Abnormal Behavior in Video

Zhongtang Zhao[1,2] and Qingtao Wu[1]

[1]School of Intelligent Engineering, Zhengzhou University of Aeronautics, Zhengzhou 450046, China, sinto0425@163.com
[2]Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing Institute of Technology, Beijing 100081, China, wuqingtao2008@163.com

Corresponding author: Zhongtang Zhao, sinto0425@163.com

Abstract. In intelligent computer-aided video abnormal behavior recognition, pedestrian behavior analysis technology can detect and handle abnormal behaviors in time, which has great practical value in ensuring social safety. We analyze a deep learning video behavior recognition network that has advantages in current research. The network first sparsely sampled the input video to obtain the video frame of each video segment, and then used a two-dimensional convolutional network to extract the characteristics of each video frame, then used a three-dimensional network to fuse them. The method realizes the recognition of long-term and short-term actions in the video at the same time. In order to overcome the shortcoming of the large amount of calculation in the 3D convolution part of the network, this paper proposes an improvement to this module in the network, and proposes a mobile 3D convolution network structure. Aiming at the problem of low utilization of long-term motion features in video sequences, this paper constructs a deep residual module by introducing long and short-term memory networks, residual connection design, etc., to fully and effectively utilize the long-term dynamic features in video sequences. Aiming at the problem of large differences in similar actions and small differences between classes in abnormal behavior videos, this paper proposes a 2CSoftmax function based on double center loss to optimize the network model, which is beneficial to maximize the distance between classes and minimize the distance between classes, so as to realize the classification and recognition of similar actions and improve the recognition accuracy.

Keywords: Abnormal behavior recognition; convolutional network; dual-stream convolutional fusion; computer-aided
DOI: https://doi.org/10.14733/cadaps.2021.S3.34-45

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

1 INTRODUCTION

Modern society is a society of information explosion. With the rapid development of information technology, people's lives have also become informatized [1]. This development trend facilitates people's production and life, and at the same time inevitably brings many hidden dangers [2]. The rapid development of computer hardware technology has made hardware equipment popularized. Aiming at security issues, computer-aided video abnormal behavior recognition has begun to be widely used in all walks of life [3]. The traditional computer-aided video abnormal behavior recognition only gathers the pictures of various monitoring points, and then relies on human eyes to monitor in real time [4]. Such a method itself has great disadvantages, such as missed detection due to fatigue of the human eyes, low efficiency of mass storage and retrieval, and so on. With the development of computer technology, computer-aided video abnormal behavior recognition has gradually developed towards intelligence. The traditional human behavior analysis system includes three stages: human detection, human tracking, and behavior recognition [5]. Human detection and tracking are in the middle and low-level part of computer vision processing, while behavior recognition belongs to the high-level part of computer vision processing. The three stages are complementary. Human detection and tracking are the basis of video-based human behavior analysis, and accurate moving human detection and tracking results are the key basis for subsequent human behavior analysis [6]. Human behavior depends on many factors, such as environment, culture, personality differences, emotions, etc. The same action of different people is different, and even the same person does the same action at different times. At the same time, due to different body types, appearances, and complex and diverse human behaviors, it is still difficult to automatically judge human behaviors at this stage. The research on the recognition of a single human abnormal behavior in a simple background has been relatively mature. Human behavior recognition libraries such as Weizmann database, KTH database and other video libraries have achieved very good recognition effects. However, due to the complexity of real scenes and the complexity of human non-rigid body movements as well as the ambiguity of behavior classification boundaries, the effects of various recognition methods for single or multiple persons in real scenes are not very satisfactory. Fernández-Ramírez et al. [7] have proposed a simplified background subtraction algorithm, which uses a non-parameter background model to extract regions of interest. This model calculates the recent historical pixel intensity probability value and shows good results in a cluttered and non-stationary background. Li et al. [8] have proposed a classification method that mainly uses 3D stereo and motion capture systems to obtain three-dimensional depth data of human activities. Anitha and Priya [9] proposed an adaptive background extraction algorithm that uses appearance model tracking technology and continuous tree classifiers to automatically detect, learn and predict abnormal behavior. Wan et al. [10] proposed a method based on integrating spatio-temporal motion images to detect human fall events. They used feature space technology to extract inherent motion features, and used the strategy of multi-level support vector machines when doing motion classification and distinguishing fall behavior. Gnouma et al. [11] comprehensively used the mixed Gaussian model foreground detection method, Hough line detection and other methods to better solve the engineering problem of pedestrian crossing detection. In order to detect the fall of the elderly and children in time, Xu et al. [12] have developed a smart home robot, and realize the recognition and detection of human fall behavior based on the side length and side width ratio of the target outer contour and the deviation of the body's centroid from the previous motion state.

This paper proposes a mobile three-dimensional convolutional network, combined with the structure of the effective convolutional network, designs a mobile effective convolutional network for video behavior recognition, so as to realize the real-time performance of the algorithm. The network can recognize long-term and short-term actions in the video at the same time. We select UCF101 and HMDB51 data sets, perform experimental parameter setting and network structure training on the apparent short-term motion flow, long-term motion flow and deep residual LSTM modules in the deep network, and then transfer the trained model to CASIA computer-aided on

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

the video data set. The task of identifying abnormal human behaviors is completed by training + fine-tuning. The human body abnormal behavior recognition model constructed in this paper is tested on the UCF101 and HMDB51 data sets. The performance of the deep network model in this paper is evaluated from different input data forms, dual-stream structure fusion methods, feature fusion methods, and different loss schemes. We migrate the deep network model to the CASIA dataset, and complete the task of identifying abnormal human behaviors by training + fine-tuning. The rest of this article is organized as follows. Section 2 discusses related theories and key technologies. In Section 3, we design the video abnormal behavior recognition algorithm. Section 4 analyzes the experimental results. Section 5 summarizes the full text.

2 RELATED THEORIES AND KEY TECHNOLOGIES

2.1 Neural Network Related Theory and Technology

In a multilayer neural network, its structure roughly consists of three parts, namely the output layer, the hidden layer and the input layer. The hidden layer is usually composed of many layers. The input layer receives the incoming data, and each hidden layer obtains information from its upper layer. After processing the information, the processed information is transferred to the next layer of the network, and the output layer is the final result of the model result. As far as deep learning is concerned, most of them are solving optimization problems. In optimization problems, the gradient descent method is widely used because of its simple thinking and convenient solution. The core idea of the gradient descent method is to obtain the optimal solution by continuously seeking the differentiation of a point. The optimal solution has the possibility of local and global optimal. When designing and training, we should try our best to avoid the situation that the model has been trapped to the local optimum and cannot escape. Then you iterate continuously through this method, and finally find the global optimal solution. During the training process, all samples from 1 to m need to be calculated and updated. Its advantage is that the learning rate is always constant, so there is no need to worry about the decline of the learning rate due to model training. At the same time, with more samples, the standard deviation will gradually decrease, so that the deviation of its estimate will be small. But the shortcomings of batch gradient descent method are also more obvious, because the current graphics card memory is still limited, for the case of a large number of samples, it is obviously unrealistic to put all training samples into the video memory for training, and for traversing all samples, calculations using vectorization are also very time-consuming.

2.2 Mainstream Feature Extraction Network Analysis

In Alex Net, the total number of layers is eight layers, which are five convolutional layers plus three fully connected layers. The input scale is 227×227. After 3×3 pooling, the final output is a 27 ×27×96 feature map. The second convolutional layer is a 5×5 convolution kernel convolution and then a 3×3 pooling, and the final output feature map size is 13×13×256. The filter size in the third, fourth and fifth convolutional layers is 3×3, and the size of the feature map output through convolution is 6×6. Then there is the fully connected layer, and after three fully connected, 1000 neurons are finally softmaxed. After each convolutional layer, a nonlinear mapping is added. At the same time, because the number of parameters of the full connection is large, the optimization process of dropout is used in the full connection to reduce the amount of calculation. The final number of parameters of Alex Net is about 6 billion, and there are 65k neurons in the network. Experiments on Alex Net found that increasing and decreasing the convolutional layer will greatly reduce the final detection accuracy of the network, and its structure is also very clever. Because the size of the video memory was limited at that time, the network used two independent graphics cards for calculation, which solved the problem of insufficient hardware resources. By analyzing Alexnet, it can be concluded that the convolution kernel should not be too large. Although a large

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

convolution kernel means that the receptive field is larger, it also brings a lot of calculations, which will cause the model to overfit. At the same time, using a larger convolution kernel in the closer to the input and using a smaller convolution kernel in the high-dimensional feature map is helpful to reduce model parameters, and the use of dropout can also be well reduced.

In deep learning, the deeper the structure, the easier it is to fit the features. Under normal circumstances, it is better than the shallow network, but the deeper the network, the more parameters are involved in the calculation, and too many parameters will cause the gradient to explode, so you cannot blindly increase the network depth during network design. However, when resnet was proposed, this problem was solved. The current resnet network has a depth of more than a thousand layers, and it has achieved very dazzling results in related recognition tasks. The reason why such a good result can be achieved is due to the design of the residual model. When the reverse update is performed, the conventional network structure updates all the information, but there are many repetitions, so adding reverse update appears meaningless. A short-circuit model is designed in resnet, which only reversely updates the difference between the layers, so that the amount of calculation can be greatly reduced without affecting the effect, so a deep network and residual error can be designed. The unit of learning is shown in Figure 1.

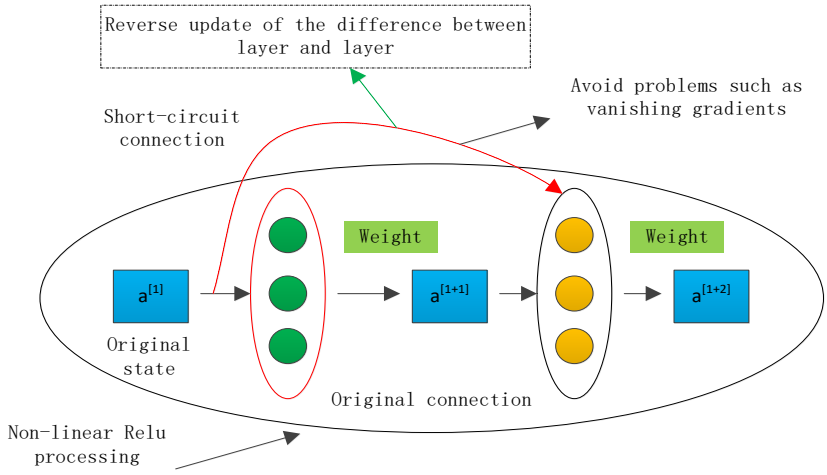

Figure 1: Schematic diagram of residuals.

2.3 Convolutional Neural Network Structure and Application Analysis

In the application of deep learning, convolutional neural network is widely used, and the main operations in its structure are convolution, pooling, full connection and classification. Compared with traditional algorithms, convolutional neural networks can handle a large number of complex problems efficiently, and its accuracy is much higher than traditional machine learning algorithms. Therefore, this section takes the C3D network, which is very important in the field of behavior recognition, as an example for analysis. This model has achieved good results in behavior recognition and has more options for processing in the field of behavior recognition.

As shown in Figure 2, the network has a small number of layers, consisting of only three convolutional layers, two pooling layers, a fully connected layer and a classification layer. In the neural network, the function of the convolutional layer is mainly to extract features through convolution calculation. The so-called convolution can be understood as the inner product operation of the input window data and a fixed-size filter matrix. For the input matrix, it is usually the input feature vector, including the number of feature layers, the length and width of the

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

image, and the channel. For three-dimensional convolution, a time dimension is added to capture the continuity of behavior in the video stream.

After the convolution operation, there is usually a pooling operation. The function of the pooling operation is mainly to reduce the feature vector output after the convolution operation, increase the receptive field of the convolution kernel, and effectively reduce the risk of overfitting. Generally, average pooling is commonly used in pooling operations. Its core idea is to perform proportional dimensionality reduction processing on the input feature values, which can avoid the distortion of the features after pooling, and the amount of calculation is also obtained.

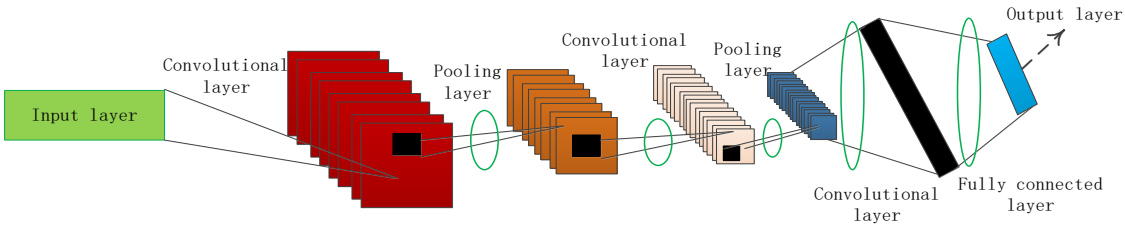

Figure 2: C3D network structure diagram.

Usually, the last part of the model is generally fully connected. The so-called fully connected means that all neurons between the last layer and the previous layer are connected, so that the fully connected layer can obtain all the feature information of the previous layer network and the obtained information is mapped to the label space of the sample, and then the loss function is used to inversely update the weight to achieve the classification ability. After the fully connected layer, the conventional network will add a classification layer. The softmax function is commonly used, which is also called the normalization function.

2.4 Model Training Related Technologies

In model training, the quality of hyperparameter settings can greatly improve the efficiency and accuracy of training. By analyzing the loss curve and accuracy during the training process and then adjusting the hyperparameters of the model, the model training can be more advanced. Therefore, correct adjustment of the model's hyperparameters is related to the key step of whether the model can quickly find the optimal solution.

The first is the setting of the learning rate. At present, many algorithms have been designed for adaptive learning rate. The core idea is to gradually reduce the learning rate when the epoch increases during the training process. This setting is reasonable, because in the early stage of model training, the error of the parameters is relatively large, and the amount of parameter adjustment will be relatively large during the reverse update. As the training progresses, the loss value is gradually reduced, at this time the learning rate should be reduced slowly, so as to prevent the model from skipping the global optimal solution. However, it is usually necessary to set a learning rate in the initial state. This is also called the setting of hyperparameters. If this value is set too small, the model training will be very slow. If it is set too large, the global maximum will be skipped. Excellent solution leads to the phenomenon of model overfitting. Therefore, for the setting of learning rate, it is generally reasonable to set to 0.0005. Using the optimizer can solve the problems related to learning rate to a certain extent.

Batch size and number of iterations are one of the important factors that affect the final result, because the training data is calculated by adding the processed data to the video memory. Therefore, if the video memory capacity of the graphics card is set too large, the training speed will be slow if it is too small. At the same time, because the batch is too small, the characteristics of each training data are not obvious enough, resulting in poor overall results. Therefore, when the video memory resources in the graphics card are not sufficient, the batch size is generally set to

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

[16,32,64,128]. The number of iterations generally requires observing the error between the test set and the training set. If the errors of both are very small, then the number of iterations is already appropriate. If the error value of the training set has been reduced at this time, the error value first decreases and then increases. This situation indicates that the model has over-fitting. It is necessary to stop and analyze the causes of over-fitting before re-training.

3 VIDEO ABNORMAL BEHAVIOR RECOGNITION ALGORITHM

3.1 The Structure of an Effective Convolutional Network

The effective convolutional network is to first use the two-dimensional convolutional network to process the sparsely sampled video frames to obtain the feature representation of each randomly sampled frame. Then, in order to understand how the apparent form of the action changes over time, the ECO network uses a three-dimensional convolutional network to process the feature representation of the above-mentioned randomly sampled frame. Finally, the final action classification label is generated by the three-dimensional network.

The network structure is relatively simple, and it can directly perform effective end-to-end training on large data sets. The three-dimensional convolutional network in the figure uses the 3D-Resnet18 network, an improved form of the C3D network structure. This network is a C3D network modified from the C3D network in the form of a two-dimensional convolutional network Resnet-18.

The convolutional layers conv_2x, conv_3x, conv_4x and conv_5x are the structure of the residual network, they just change the two-dimensional convolution in the two-dimensional residual network to the form of three-dimensional convolution. In order to avoid repetitive explanation, only the framework of conv_2x is given, as shown in Figure 3.

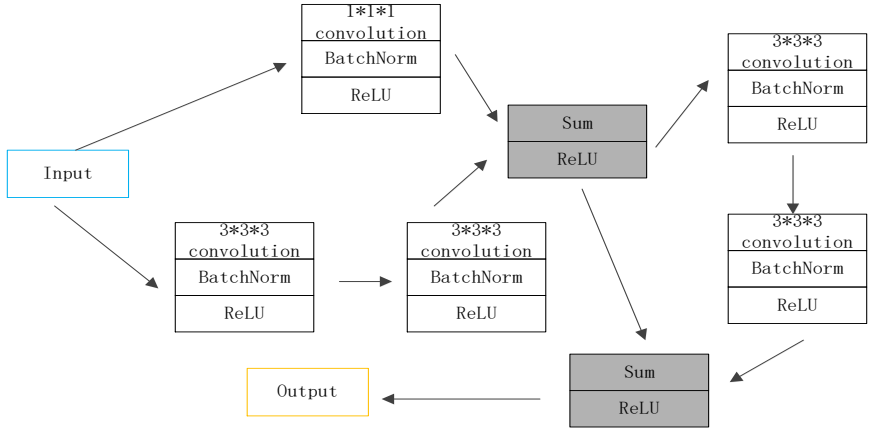

Figure 3: Network structure diagram of the three-dimensional residual convolution layer conv_2x.

The reason for adopting the form of residual network is that residual network has achieved good results in the field of two-dimensional images. After his experiment, he found that under the 3D-Resnet18 network, the performance that can be achieved is higher than that of the C3D network, and the model size of the network is smaller, and the amount of calculation is also smaller. This is because the amount of calculations and parameters in the C3D network are mainly concentrated in the fully connected layer, while the fully connected layer of the 3D-Resnet18 network has only one

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

layer, and the number of input nodes is 512, which is the same as the fully connected layer of the C3D network. In comparison, a lot of parameters are reduced.

Inception network is an improvement of the traditional two-dimensional convolution kernel. The purpose of the Inception network is to increase the width and depth of the network as much as possible while the amount of calculation is unchanged or even reduced, so there is no large-size convolution kernel used here. It can be seen from the figure that it only has three convolution kernels of different sizes, 1*1, 3*3, and 5*5. Convolution kernels of different sizes mean that the size of the receptive field is different during convolution, so that features of different scales can be obtained by convolution. Among them, the role of the 1*1 convolution kernel is very important. While increasing the depth of the network, it can reduce the number of channels during convolution, thereby reducing the amount of calculation. However, when increasing the depth of the network, it will bring a serious impact, that is, the gradient is easy to disappear and optimization is difficult. Therefore, when the Inception network is further improved, batch normalization (Batch Norm) is introduced, which can effectively solve the problem of network optimization. Batch Norm refers to the whitening operation on the input data. The so-called whitening operation is to convert the input data into a normal distribution with zero mean and one variance. When the convolutional neural network is trained, the input is often not one piece of data, but a batch composed of several pieces of data. Performing a whitening operation on this batch is batch normalization. The proposal of Batch Norm effectively speeds up the convergence speed of the network, and also successfully prevents the disappearance of the gradient.

3.2 Mobile 3D Convolutional Network

There is also a three-dimensional network part in the ECO network, and it uses the structure of the 3D-Resnet18 network, in which the calculation amount of the three-dimensional convolution part is much larger than that of the two-dimensional convolution, so the above problems will also occur. So, this section will improve the three-dimensional network, reduce the amount of calculation of the three-dimensional network and its model size, so that the ECO network is more conducive to the realization of mobile terminal transplantation.

The difference between deep convolution and the original traditional convolution is that the convolution kernel of traditional convolution acts on each input channel, and then adds these results to get the final value, the different convolution kernels of the input channels act on each other, and then get their respective results. The point-by-point convolution is the traditional convolution, and its convolution kernel size is 1*1, which is mainly used to change the number of channels in the network layer.

Although Mobile Net can effectively reduce network parameters and calculations, its structure is too simple, resulting in a decrease in network effectiveness. The success of the residual network (Res Net) in recent years shows the importance of skip connection (Skip Connection). In addition, the unit of the deep convolution part is more likely to fail in training. This is because the convolution kernel dimension of the deep convolution is much smaller than the traditional convolution kernel. The gradient of the Re LU function when a negative number is input is 0, which causes the network to be unable to recover once it enters the 0 output state.

For the convolution stride of 1 or 2 in the convolution process, Mobile Net V2 proposes two structures. The difference between the two is that when stride =1, there is a jump connection like the residual network. But the backbone branches of both of them are the same. They first use 1*1 point-by-point convolution to increase the input dimension, and then use 3*3 deep convolution operation, and finally use 1*1 point-by-point convolution. The reason for this separation is that if stride = 2, the output size of the convolutional layer will be twice as small as the input size, so skip connections cannot be used directly. It should be noted that the Re LU activation function is not used in the last point-wise convolution, but used after the initial point-wise convolution. This is to avoid the low-dimensional convolution result of the Re LU function analyzed above. In addition, the reason why the dimension upgrade is used here instead of the dimensionality reduction

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

operation like in Res Net is that although the dimension upgrade will increase the amount of calculation, due to the use of deep convolution, the increased amount of calculation is not very large. And after the dimension upgrade operation, the input information will be richer, and then the Re LU operation and then dimension reduction can ensure that the necessary information is retained.

The Inception series network proposes a new convolution kernel improvement based on BN-Inception. It also uses the idea of decomposing convolution kernels. The difference from the decomposition idea of the Mobile Net series is that it decomposes the original 3*3 convolution kernel into two convolution kernels of 1*3 and 3*1. form. The advantage of this is that the network depth can be further increased, while improving performance, it can also reduce network parameters and calculations.

4 ANALYSIS OF EXPERIMENTAL RESULTS

4.1 Comparison of Different Input Data Forms

In this paper, the long-term motion flow in the dual-stream convolutional fusion network structure takes an ordered optical flow graph as input. The essence of the ordered optical flow graph is to arrange and compress multiple optical flow frames in an orderly manner, which can be extracted more effectively. In order to verify the effectiveness of ordered optical flow diagrams, this paper compares different input data of long-term motion flow: Static Image RGB (SI-RGB), Stacked Optical Flow (SOF), Dynamic Image RGB (DI-RGB). The optical flow diagram SOFI is shown in Figure 4 and Figure 5.

From Figure 4 and Figure 5, it can be found that the SOFI data used in this paper has achieved the best recognition effect. On the UCF101 and HMDB51 data sets, the highest recognition results were achieved at 9 and 69 frames respectively. Therefore, for different data sets, different numbers of optical flow frames need to be used to make ordered optical flow graphs. The apparent short-term motion flow in this paper is SI-RGB as the input. If the long-term motion flow also uses SI-RGB as the input, the recognition result is not good. The experimental results show that the ordered optical flow diagram has a significant effect on improving the accuracy of human behavior recognition, and can make better use of the long-term motion information in the video sequence.

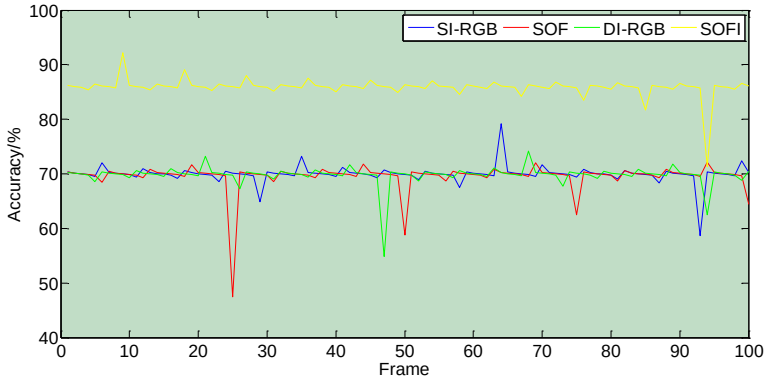

Figure 4: Comparison of different input data forms on UCF101 (%).

4.2 Comparison of Feature Fusion Methods of Different Network Layers

The features extracted by the dual-stream convolutional fusion network are still weak in characterizing complex behaviors. Therefore, this paper uses the deep fusion features obtained by feature fusion of different network layers as the input of the deep residual LSTM module to

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

continue to learn the long-term characteristics of the video. By comparing the recognition accuracy of the fusion of different network levels, it can be found that the feature fusion of the fc6 layer of the apparent short-term motion flow and the fc7 layer of the long-term motion flow will have better experimental results, as shown in Figure 6.

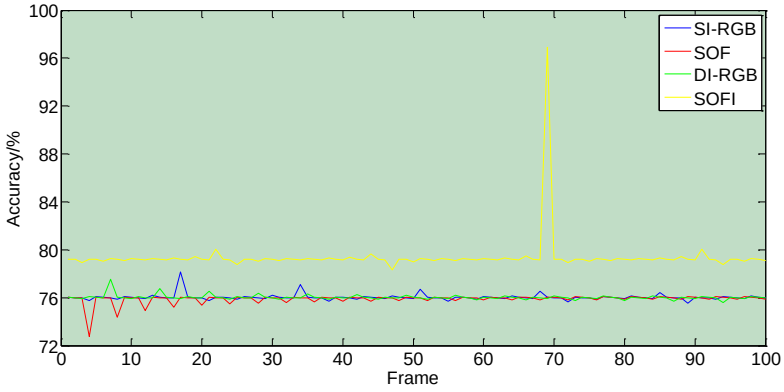

Figure 5: Comparison of different input data formats on HMDB51.

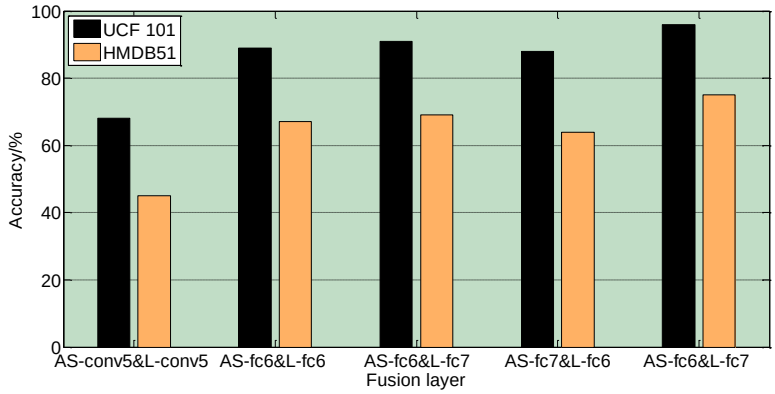

Figure 6: Comparison of the accuracy of fusion feature recognition at different network layers.

It can be seen from Figure 6 that feature fusion in the convolutional layer will have poor recognition accuracy, which is significantly lower than that of the fully connected layer. This is because the fully connected layer will perceive the model input data as a whole and has higher semantic information; when feature fusion is performed at the fc6 layer of the two streams, the recognition accuracy will be worse than when the features are fused at the fc7 layer. It shows that deep-level fusion can more effectively use the depth information in the video sequence; when the fc7 layer of the apparent short-term motion stream and the fc6 layer of the long-term motion stream are feature fused, the recognition accuracy will decrease, but When the fc6 layer of the apparent short-term motion flow and the fc7 layer of the long-term motion flow are feature fused, the recognition accuracy reaches the optimal effect, indicating that the long-term dynamic information-led model training will indeed have a better recognition accuracy. If appearance information dominates model training, it will produce poor results.

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

4.3 Comparison of Different Behavior Recognition Methods

In order to verify the effectiveness and advancement of the algorithm in this paper, on the UCF101 and HMDB51 data sets, we compare the algorithm in this paper with the current mainstream human behavior recognition algorithms, as shown in Figure 7.

From the experimental results, it can be found that the recognition effect of iDT in the traditional method still has a high accuracy rate, and the algorithm tested on the UCF101 data set has a method based on the combination of deep learning behavior recognition method. The algorithm in this paper is compared with the current mainstream human behavior recognition methods, and the recognition effect is better on the UCF101 and HMDB51 data sets. By comparing the Two-stream Fusion method with the original dual-stream network model, it can be found that the fusion strategy is used to interact between the two streams, which is conducive to the extraction of spatiotemporal features in the video sequence, and is more conducive to the improvement of behavior recognition. On the UCF101 data set, comparing the original dual-stream network with Two-stream+LSTM, it can be found that the long- and short-term memory network is helpful for the extraction of time dimension features, and the accuracy of behavior recognition has been improved. Further comparing the experiments done in this article, comparing the network model without the deep residual LSTM module with the method in this article, we can find that the method in this article adds the deep residual LSTM module to improve the accuracy of behavior recognition on the two data sets. Compared with 2.3% and 6.8%, it shows that the deep residual LSTM module can make full use of the long-term dynamic information of human movements in the video sequence, which helps to improve the recognition effect. This article also verifies the dual-center loss function. When the 2C-softmax scheme of the dual-center loss function is added, the behavior recognition accuracy is improved on the UCF101 and HMDB51 data sets.

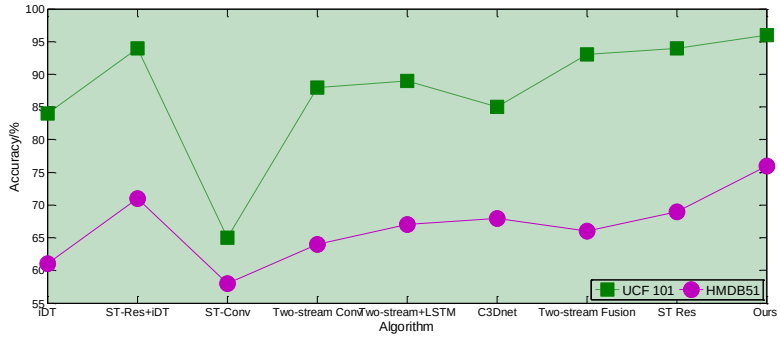

Figure 7: Comparison of accuracy of different behavior recognition methods (%).

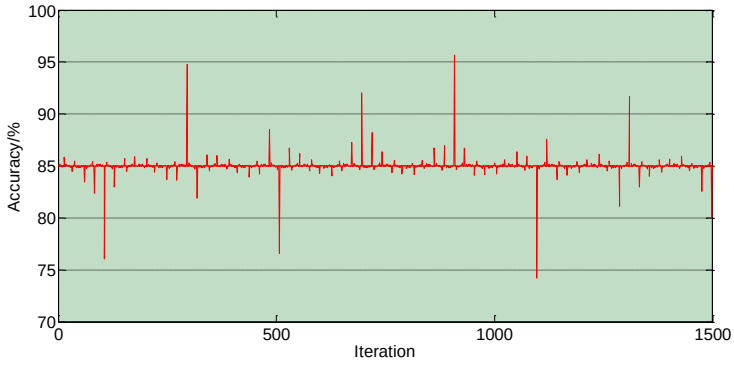

Figure 8: Abnormal behavior test results on the CASIA dataset.

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

4.4 Model Test Results on CASIA

Because the CASIA data set of the computer-aided video data set is small, it is difficult to meet the requirements of deep network training. Therefore, this article uses the large-volume UCF101 and HMDN51 data sets for model training, and then migrates the abnormal human behavior recognition model to the CASIA data set, the task of identifying abnormal human behaviors is accomplished by training + fine-tuning. The test results on the CASIA data set are shown in Figure 8.

5 CONCLUSION

This paper analyzes the ECO network in the video behavior recognition algorithm, improves the three-dimensional convolution part of the ECO network, and proposes an M3D network, and uses the M3D network to replace the three-dimensional convolution part of the traditional ECO network. Compared with the existing three-dimensional convolutional network, M3D network not only speeds up the operation speed, reduces the amount of calculation and the amount of parameters, but also maintains or exceeds the classification accuracy of these networks. The M-ECO network uses sparse sampling of the input video, then uses a two-dimensional convolutional network for feature extraction, and finally uses the feature fusion of the three-dimensional convolutional network to improve the performance of the network. Aiming at the characteristics of abnormal human behavior in videos, this paper improves the original dual-stream convolutional network and its derivative models. We construct a dual-stream convolution structure based on C3Dnet, and use multiplicative cross-stream residual unidirectional connection for internal network integration. A scheme based on the dual-center loss function is proposed to optimize the network model. This paper conducts model training on the UCF101 and HMDB51 data sets, then transfers the model to the computer-aided video data set CASIA for abnormal behavior identification, and analyzes the experimental results to verify the effectiveness of the model algorithm.

ACKNOWLEDGEMENTS

This work is supported in part by the National Natural Science Foundation of China (Grant No.U1504609), by the Key Scientific and Technological Project of the Higher Education Institutions of He'nan Province, China (Grant No.15A520003), by the Scientific and Technological Planning Project of He'nan Province, China (Grant No.172102210525), and by the research and practice general project on the reform in higher education of He'nan Province, China (Grant No.2017SJGLX400).

Zhongtang Zhao, https://orcid.org/0000-0003-3601-1943
Qingtao Wu, https://orcid.org/0000-0002-2782-446X

REFERENCES

[1] Yang, Y.; Li, L.; Liu, Z.; Liu, G.: Abnormal behavior recognition based on spatio-temporal context, Journal of Information Processing Systems, 16(3), 2020, 612-628. https://doi.org/10.3745/JIPS.02.0134

[2] Xie, S.; Zhang, X.; Cai, J.: Video crowd detection and abnormal behavior model detection based on machine learning method, Neural Computing and Applications, 31(1), 2019, 175-184. https://doi.org/10.1007/s00521-018-3692-x

[3] Wang, J.; Xia, L.: Abnormal behavior detection in videos using deep learning, Cluster Computing, 22(4), 2019, 9229-9239. https://doi.org/10.1007/s10586-018-2114-2

© 2021 CAD Solutions, LLC, http://www.cad-journal.net

[4] Zhang, J.; Wu, C.; Wang, Y.; Wang, P.: Detection of abnormal behavior in narrow scene with perspective distortion, Machine Vision and Applications, 30(5), 2019, 987-998. https://doi.org/10.1007/s00138-018-0970-7
[5] Geng, Y.; Du, J.; Liang, M.: Abnormal event detection in tourism video based on salient spatio-temporal features and sparse combination learning, World Wide Web, 22(2), 2019, 689-715. https://doi.org/10.1007/s11280-018-0603-0
[6] Li, Y.; Zhai, Q.; Ding, S.; Yang, F.; Li, G.; Zheng, Y.-F.: Efficient health-related abnormal behavior detection with visual and inertial sensor integration, Pattern Analysis and Applications, 22(2), 2019, 601-614. https://doi.org/10.1007/s10044-017-0660-5
[7] Fernández-Ramírez, J.; Álvarez-Meza, A.; Pereira, E.-M.; Orozco-Gutiérrez, A.; Castellanos-Dominguez, G.: Video-based social behavior recognition based on kernel relevance analysis, The Visual Computer, 36(8), 2020, 1535-1547. https://doi.org/10.1007/s00371-019-01754-y
[8] Li, Y.; Miao, Q.; Qi, X.; Ma, Z.; Ouyang, W.: A spatiotemporal attention-based ResC3D model for large-scale gesture recognition, Machine Vision and Applications, 30(5), 2019, 875-888. https://doi.org/10.1007/s00138-018-0996-x
[9] Anitha, G.; Priya, S.-B.: Posture based health monitoring and unusual behavior recognition system for elderly using dynamic Bayesian network, Cluster Computing, 22(6), 2019, 13583-13590. https://doi.org/10.1007/s10586-018-2010-9
[10] Wan, S.; Qi, L.; Xu, X.; Tong, C.; Gu, Z.: Deep learning models for real-time human activity recognition with smartphones, Mobile Networks and Applications, 25(2), 2020, 743-755. https://doi.org/10.1007/s11036-019-01445-x
[11] Gnouma, M.; Ladjailia, A.; Ejbali, R.; Zaied, M.: Stacked sparse autoencoder and history of binary motion image for human activity recognition, Multimedia Tools and Applications, 78(2), 2019, 2157-2179. https://doi.org/10.1007/s11042-018-6273-1
[12] Xu, Y.; Lu, L.; Xu, Z.; He, J.; Zhou, J.; Zhang, C.: Dual-channel CNN for efficient abnormal behavior identification through crowd feature engineering, Machine Vision and Applications, 30(5), 2019, 945-958. https://doi.org/10.1007/s00138-018-0971-6

© 2021 CAD Solutions, LLC, http://www.cad-journal.net