Research on Bad Driving Detection Based on Behavior Recognition

Yasheng Yuan[1], Fengzhi Dai[1, 2, *], Di Yin[1], Yuxuan Zhu[1]

[1]Tianjin University of Science and Technology, China;
[2] Tianjin Tianke Intelligent and Manufacture Technology CO., LTD, China;
E-mail:[*]daifz@tust.edu.cn
www.tust.edu.cn

Abstract

Dangerous driving behavior is considered to be the direct or indirect reason of road accidents. Although artificial video surveillance is good to prevent bad driving, it wastes too much time and manpower. How to effectively identify behavior becomes the focus of the research. In recent years, deep learning showed the huge advantage in the field of computer vision. This paper adopt a number of deep learning network models, mining video integration of space and time features, introduction of analogy in human visual attention mechanism, improve the model deeply, using the LSTM to accurate and efficient video behavior analysis technology.

Keywords: deep learning, fatigue testing, convolutional neural network, action recognition

1. Introduction

With the maturity and development of internet technology, especially the development of video surveillance system, people tend to use more and more monitoring method to prevent some accidents. Video monitoring is an important part of prevention measures, the importance and actual efficiency were higher than human security. In the highly security demand of modern life, it is an essential part of a high and new technology, the action of driving of the driver is that people need to use one of monitoring. Research [1] shows that bad driving behavior of drivers is an important cause of traffic accidents. How to effectively monitor the driver's behavior is the key to safe driving.

Compared with traditional detection technology, deep learning technology shows the strong feature extraction ability. For target detection that performs well in deep learning, although it can more accurately detect the smoking [2], but for the cryptic action of fatigue, no time-level characteristics of target identification is obviously weak [3]. Especially in the video, the driver's side head, bowed their heads and normal driving operation will greatly improve the target detection model miscalculation, so network have time-level characteristic behavior identification, will have high research significance.

2. Main research method

Based on the deep learning on the driver's action recognition classification, through the analysis of the behavior recognition algorithm and the network model. Two-stream network is our choice as our main network, on the basis of the Two-stream network add amount Conv layers, improve the effect of feature extraction and increase the nonlinear. Adding BN layers and Dropout layers to prevent over fitting, and chooses appropriate activation function to further reduce the amount of calculation. The LSTM (Long Short-Term Memory) layer is added to further learn the timing information between behaviors and mine the long time dependence relationship of video sequence.

2.1 Data set processing

The training data set for this project contains 13,360 YouTube clips from the UCF101 data set and driving videos from volunteers. The videos mainly include five types of movements: human and object interaction, only body movements, human interaction, playing music equipment, and various kinds of sports. 22 kinds of videos of facial movements from UCF101 data set is selected. About 300 short videos in total, this videos have bad driving behaviors, such as smoking, talking on the phone, taking a nap and looking left and right, which were collected by ourselves from the vehicle camera during driving.

First, each video is divided into several short videos of about 120 frames (10s). After that, each short video is extracted frame by frame into RGB images, which are used as the input of spatial flow. Then, through the OpenCV and CUDA, the RGB frames extracted from the video are converted into optical flow diagrams [4].

The optical flow is the instantaneous velocity of pixels moving in space on the observed imaging plane. The optical flow method uses the change of pixels in the time domain and the correlation between adjacent frames. The method find the corresponding relationship between the previous frame and the current frame. It is a method to calculate the motion information of objects between adjacent frames. The projection of the motion for the object in the three-dimensional space on the two-dimensional imaging plane is shown in Fig.1. The result is a two-dimensional vector describing the change of position. In the case of minimal motion interval, we usually regard it as a two-dimensional vector describing the instantaneous velocity of the point u = (u, v), which is called the optical flow vector [5].

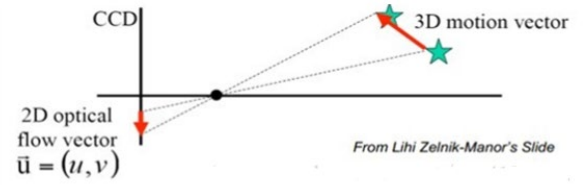

Fig.1. Optical flow method

TVL1 optical flow method in OpenCV to extraction is used in this paper. Each image obtained an optical flow diagram extracted from the X and Y directions respectively. The resulting optical flow diagram is shown in Fig.2.

Fig.2. Optical flow frame and image frame

2.2 Net selection

The two-stream convolutional network [6] is added by some network in this paper, which is divided into spatial convolutional network and time convolutional network.

The spatial-flow Convolutional network operates on a single video frame. Because some behaviors are strongly related to specific scenes and objects, it is effective to recognize behaviors from still images. As CNN (Convolutional Neural Networks) network is already a powerful image recognition algorithm, it can build a video recognition network based on large-scale image recognition algorithm. It can also utilize the pre-training network on the existing image classification data set. Time-flow convolutional network is different from ordinary CNN network. The input of time-flow convolutional network is formed by stacking optical flow displacement field between several consecutive frames. This input is characterized by the motion between the video frames, making the recognition process easier (no network implicit estimation of the motion is required).

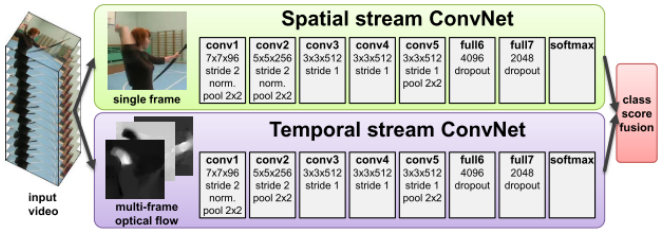

Fig.3. Two-stream Convolutional network

The two-stream convolutional network is shown in Fig.3. The algorithm is divided into two branches of convolutional neural network for feature extraction. First, a frame of RGB image in the video stream was input. After five layers of convolution and three layers of sampling, it was passed into the full connection layer of the two layers for feature extraction. The extracted spatial feature was input into the Softmax classifier for category prediction. Second, the optical stream features extracted from the video stream were input into the convolutional neural network with the same feature extraction branch of pixel space. Finally it passed into the Softmax classifier for category prediction.

2.3 Identify face areas

The network behavior identification network selected in this paper is improved based on two-stream, and the number of network layers is deepened on this basis. Before each lower sampling layer, a batchsize normalization layer, which means BN (Batch Normalization) & Relu [7], is added to prevent the over-fitting caused by too few training samples and too deep network. The network structure is shown in Table 1.

Table 1. Network structure

Layer name	Convolution kernel size	Convolution kernel number
Flow_conv_begin	7	64
BN &RELU	4	
Conv_begin_pool	3	---
Layer_1_Conv1_1	1	64
BN &RELU	4	
Layer_1_Conv1_2	3	64
BN &Relu	4	
Layer_1_Conv1_3	1	256
Layer_1_Conv1_expand	1	256
Layer_1_Conv1_sum	1	256
BN &Relu	4	
Layer_1_Conv2_1	1	64
BN &Relu	4	
Layer_1_Conv2_2	3	64
BN &Relu	4	
Layer_1_Conv2_3	1	256
Layer_1_Conv2_sum	1	256
BN &Relu	4	
Layer_2_Conv1_1	1	128
BN &Relu	4	
Layer_2_Conv1_2	3	128
BN &Relu	4	
Layer_2_Conv1_3	1	320
Layer_2_Conv1_sum	1	320
Layer_2_Conv1_expand	1	320
BN &Relu	4	
Layer_2_Conv2_1	1	128
BN &Relu	4	
Layer_2_Conv2_2	3	128
BN &Relu	4	
Layer_2_Conv2_3	1	320
Layer_2_Conv2_sum	1	320
BN &Relu	4	
Layer_2_Conv3_1	1	123
BN &Relu	4	
Layer_2_Conv3_2	3	128
BN &Relu	4	
Layer_2_Conv3_3	1	320
Layer_2_Conv3_sum	4	
Conv_final		512
BN &Relu	4	
Pool	8	---
Dropout	0.2	
Global_pool	Avg	
FC	16	
Loss		

The BN layer and Dropout mechanism is added, which is used to a certain extent and avoid network of fitting. Relu is selected as the activation function, change the sigmoid disappeared as a result of activation function gradient (sigmoid close to the saturated zone, change is too slow, derivative tends to zero) to complete the deep net training. The interdependencies between parameters are reduced, which can reduce the happen of over fitting. Through the proper superposition of 1x1 Convolutional layer and 3x3 Convolutional layer, the features are further extracted while the computation is reduced as much as possible. Finally Global pool is added to simplify the calculation.

2.4 Facial image processing

LSTM is an improvement based on the traditional cyclic neural network (RNN). It is very suitable for processing video sequences with time-dimension characteristics. In addition, LSTM network can avoid the gradient disappearance and gradient explosion problems in the practical application of RNN network.

In the classical LSTM structure, there are three "gates": (1) Oblivion gate, which is used to help RNN selectively forget some historical information; (2) Memory gate, which is used to strengthen the memory of some historical information; (3) Output gate, which is used responsible for comprehensive consideration of

long and short term memory information to generate output signals.

In this paper, LSTM network (Fig.4) is also used for video signal timing modeling.

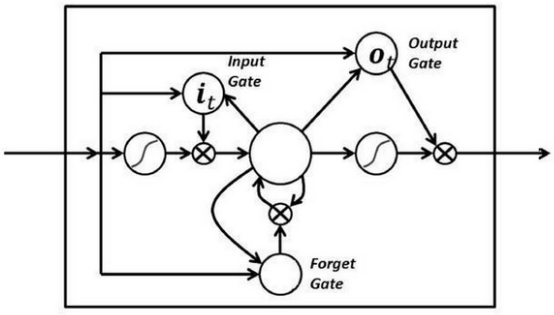

Fig.4. LSTM network structure

The key of the LSTM is cell state. The unit state is kind of like a conveyor belt. It goes all the way down the chain in a straight line, with only a few smaller linear interactions. It is very easy for information to flow unaltered.

State of cell transformation, namely value C transformation, have two operations. The two operation determines the forgotten and update of the computed tomography C_{t-1}. First of all, C_{t-1} through the multiplication of point-wise operation, this step determines whether the C_{t-1} value is forgotten. The number multiplied by C_{t-1} comes from the lower layer, the bottom of the data through a sigmoid layer, through the data values are below (0, 1). In general, when the input value is greater than 3 or less than -3, sigmoid value is close to the 1 and 0. That is to say, the value of C is generally going to be close to C_{t-1} unchanged, or 0, which means that you remember C or you forget C_{t-1}. Of course, a number between (0, 1) multiplied by C_{t-1} means how many C_{t-1} values you need to remember. Next, the value of C_{t-1} encounters an addition, which is a change to the value of C_{t-1}, such as the addition of new information, which is generally understood as something new to be remembered [8].

3. Improved PERCLOS algorithm

The network structure of this paper is shown in Fig.5. First, after the image flow and optical flow respectively pass through the convolutional network, the timing context information of the input video sequence is obtained through the LSTM layer. Two Softmax layers and two cross entropy loss layers are deployed after the feature output layer and the LSTM layer, respectively. The supervised information of the video category is used to drive the joint training of the LSTM and CONV networks. The hidden state of the LSTM layer can capture the time evolution of a specific category in the video sequence.

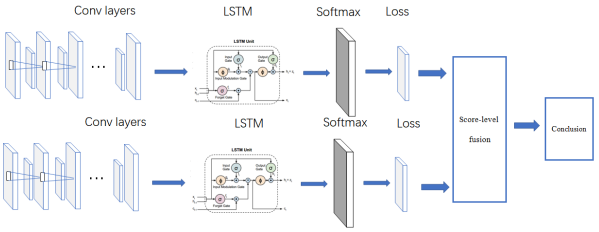

Fig.5. Two-stream-LD network structure

4. Experiment and Conclusion

All test sets are used to experiment on the model trained by the initial network, and the normalized confusion matrix of various behavior recognition rates is shown in Fig.6.

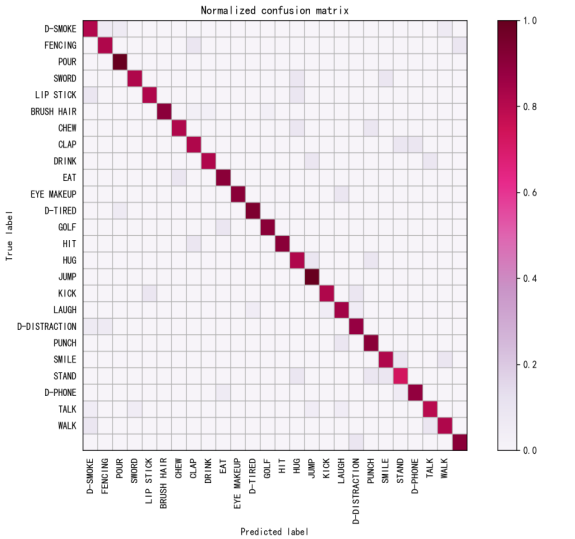

Fig.6. The normalized confusion matrix of various behavior recognition rates

In the matrix, the vertical axis represents the real label of a certain video behavior, and the horizontal axis

represents the predictive label of a certain video behavior.

The darker the grid matrix color is, the higher the recognition rate is. And the lighter the grid matrix color is, the lower the recognition rate is. In the darker diagonals, the data contained represent the correct recognition rate for a certain type of behavior, while the data in the other light colored squares represent the wrong recognition rate when a certain type of behavior is predicted to be other behaviors.

Altogether 25 types of human behaviors were identified in this experiment. Among them, there are similar frames among the behaviors that need to use hands in front of the face, such as smoking, making phone calls and making up, so the error recognition rate among these people may be relatively high.

Based on the data in the confounding matrix, the dual flow convolutional neural network designed in this paper can effectively complete the task of recognizing and classifying bad driving behaviors.

Finally, the results are fused at the decision level. For the classification results obtained by multiple classifiers in the classification problem, if each classifier has different influences on the final prediction results, then the fusion can be considered in the way of weighted sum. We will find that, compared with speech and subtitle, the visual features contained in the image sequence are more helpful to determine the accurate behavior contained in the video. Therefore, we choose to give a larger weight to the classifier results of the video images, and give a smaller weight to the classification results of other classifiers.

The final result is shown in Table 2, from which can be seen that the Two-stream did well in UCF101 data set, this means that the Two-stream network in extracting time has good extraction effect when the flow characteristics, but in RGB space flow identification process is not ideal, but in our network of Two-stream-LD, both the space and time flows recognition effect has a significant improvement.

In the detection of bad driving behaviors, the Two-stream-LD network can be used to better identify various actions of the driver in the video. After comprehensive comparison, the method proposed in this paper is of certain research significance.

Table 2. Final result

Network	UCF101		
	RGB	Flow	RGB+Flow
Two-stream	83%	85.6%	89%
TSN	81.4%	79.6%	87.2%
IDT	78.2%	81%	86.4%
C3D	85.2%	--	--
LSTM	81.3%	--	--
3D Fused	82%	83.8%	86%
Two-stream-LD	88%	86%	90.3%

References

1. State of the States Report on Drowsy Driving, *National Sleep Foundation*, Nov. 2007, U.S.A.
2. Yasheng, Y, Fengzhi, D, Yunzhong, S, et al., On Fatigue Driving Detection System Based on Deep Learning. *Proceedings of 2020 Chinese Intelligent Systems Conference (CISC). Lecture Notes in Electrical Engineering*, 2020, pp.734-741.
3. Yasheng, Y., Fengzhi, D., Lingran, A, et al., Research on fatigue detection method based on deep learning. *Proceedings of 2020 International Conference on Artificial Life and Robotics*, Oita, Japan, 2020: pp. 640–643.
4. Zach C, Pock T, Bischof H. A Duality Based Approach for Realtime TV-L1 Optical Flow. *Proceedings of the 29th DAGM conference on Pattern recognition,* 2007: pp. 214–223.
5. Wang L, Xiong Y, Wang Z, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, *European Conference on Computer Vision. Springer,* Cham, 2016: pp.20-36.
6. Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal Residual Networks for Video Action Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017: pp.3476-3484.
7. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society*, 2016: pp.770-778.
8. Donahue J, Hendricks L A, Rohrbach M, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, 39(4): pp.677-691.