Detection of Distracted Driving Based on Multi-Granularity and Middle-Level Features

Min Tang, Fang Wu, Li-Li Zhao, Qi-Peng Liang, Jian-Wu Lin, Yun-Bo Zhao
College of Information Engineering, Zhejiang University of Technology, Hangzhou, China

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327078

***Abstract*—A so-called MGMN (Multiple-Granularity-Middle Network) algorithm is proposed to improve the detection accuracy of distracted driving, based on multi-granularity features and middle-level features. The algorithm is derived from the ResNet50 neural network and is the first time to use multi-granularity features and mid-level features of images in the field of distracted driving detection. The multi-granularity feature extraction module consists of three branches: the global branch to learn the global features without local information, the second branch to divide the image level into two parts and later to learn the local features of the upper and lower parts, and the third branch to divide the image level into three parts, and later to learn the local features of the upper, middle and lower parts. By extracting the features of the middle layer of the image, the feature extraction of the algorithm is enriched. While the multi-granularity features are individually input to the cross-entropy loss function, the multi-granularity features of the image and the middle-level features are combined and input into the cross-entropy loss function. The proposed algorithm has an accuracy of 99.8% on the dataset "Districted-Driver-Detection" published by State Farm Company, which is 1.5% to 3% higher than existing algorithms, and an improved accuracy of 98.7% on the dataset "AUC-Districted-Driver-Detection".**

Keywords—Distracted driving, Convolutional neural network, Multi-granularity features, Mid-level features, Image processing

I. INTRODUCTION

By the end of 2019 China has as many as 260 million automobile vehicles and 400 million drivers[1]. As the number of automobile vehicles and drivers continues to grow, more and more serious traffic safety issues arise. According to the report of World Health Organization[2], 200,000 traffic accidents happen in China every year and about 60,000 people lose their lives in traffic accidents. The Ministry of Public Security of our country said that more than 3/4 of the accidents are caused by driver's cognition and decision-making mistakes, a large proportion of these accidents are caused by the driver's distracted driving[1]. This thus raise the urgent need of detecting distracted driving for the safety purpose. This need can still be true even with the development of the automatic driving technology, as many believe that total driverless drive can be impossible and the ultimate form of drive is some form of driver-system cooperating, in which case the detection of the driver's distraction can still be important.

Distracted driving refers to the driver's inability to concentrate due to some distracted behaviors during driving[3]. Distracted driving behaviors generally include, e.g., using a mobile phone, talking with passengers, eating, drinking, tidying up appearance, using radio, etc. Distracted behaviors of the driver will lengthen the driver's reaction time and reduce the ability to respond to sudden emergencies. Distracted driving behaviors will also make the driver lack of observation of the road and easily cause traffic safety accidents.

Current distracted driving detection approaches are mainly divided into two categories. One is to indirectly measure the driver's attention state by monitoring the driving conditions of the car, and the other is to directly monitor the driver's attention state[4].

The first approach needs to consider the speed, lateral position and turning angle of the vehicle. This detection method will not interfere with the driver's operation and the vehicle driving procedure, but will be affected by the driver's experience, driving conditions and The influence of factors such as vehicle type. When considering driver fatigue detection, lane detection is very important because it is an indication of the lateral position of the vehicle[5].

The second approach consists of mainly two types of methods, either based on the driver's physiological characteristics or based on the analysis of the driver's facial features[6]. The first type requires the driver to wear some medical monitoring equipment to obtain signals such as brain electrical signals, myoelectric signals, electrocardiographic signals, eye electrical signals, and the driver's breathing and pulse for direct measurement. Despite its high accuracy, this method is usually not practically sound because of the cumbersome wearable medical monitoring equipment, high cost, and the difficulty in vehicle-mounted in real time[7]. The driver's concentration can also be detected based on, e.g., the movement of his eyelids. In order to analyze human eyelid movement, eye detection and tracking can be used as the basis for extracting other visual features, which can indicate the driver's state of concentration[8]. However, the accuracy of existing machine vision is affected by the vehicle model and the environment, and hence its practical implementation still face several restrictions [9].

In order to balance between convenience and accuracy in the distracted driving detection, we propose a distracted driving detection algorithm called MGMN (Multiple-Granularity-Middle Network) based on multi-granularity and mid-level features. Based on the ResNet50 model, the multi-granularity and mid-level features of the image are extracted to make full use of the information in the image. The algorithm is applied to the driving behavior photos in the dataset "Districted Driver detection" published by State Farm Company and the "AUC-distracted-driver" dataset published by the American University of Cairo for training. It has an accuracy rate of 99.8% on State Farm's dataset, an improvement of 1.5% to 4% compared to other algorithms. It also has the highest accuracy rate of similar algorithms on the dataset "AUC-Districted-Driver-detection".

The main contributions are:

(1) For the first time, multi-granularity features are introduced into the field of distracted driving detection, which can better extract the characteristics of driving images.

*Corresponding author. E-mail: ybzhao@ieee.org

This work was supported by the National Key Research and Development Program of China (No. 2018AAA0100801) and the National Natural Science Foundation of China under Grant 61673350.

(2) For the first time, middle-level features are introduced into the field of distracted driving detection, which enriches the feature extraction of the algorithm.

The remainder of the paper is organized as follows. Section II introduces some existing driver's distracted driving detection algorithm and then formulates the problem. MGMN algorithm design is then introduced in Section III. Section IV gives the experimental results of the algorithm and the comparative analysis with other algorithms, and Section V concludes the paper.

II. Related Work

In this section, we will introduce the existing distracted driving detection algorithms based on machine vision, and then analyze their advantages and disadvantages.

A. Methods Based on the Combination of Hand and Face Detection

Yehya Abouelnaga and Hesham M.Eraqi et al. first used detectors to detect the driver's hands and faces [10-11]. They used genetic algorithms to learn the weights of the four types of images: hands, face, overall images, and a combination of hands and face. Pictures were input to the softmax function after integration. However, it was found in gesture recognition that hand and face detector would fail, which led to a decrease in the accuracy of distracted driving detection. At the same time, during the detection process, hands have a higher proportion than face. So, the accuracy of images focusing on face and "face + hands" is lower than that of hand images. Abouelnaga et al. also proposed a system composed of genetically weighted convolutional neural network (CNN) collections, which are trained on original images, skin segmentation images, facial images, hand images, and "face + hand" images.

B. Detection of Hand Movement Based on Multi-classifiers

Ohn-ba et al. used three classifiers on different areas of the captured driver's image to detect whether the driver's hand is on the steering wheel, gearbox, or dashboard[12]. They said that these three classifiers placed in a specific area are combined into a second-stage classifier. However, they analyzed fewer hand movements, such as: (1) Both hands are on the steering wheel (2) One hand is on the dashboard and (3) The hand is in gear. At the same time, the location detection results of only these three areas are easily interfered.

C. Methods Based on Foreground Extraction

Maitree Leekhaa and Mononito Goswamib et al. proposed a simple architecture that uses foreground extraction and convolutional neural networks (CNN)[13]. Using GrabCut to merge other functions (such as the driver's posture) through foreground extraction, the performance of the model can be improved. However, the model has a large error due to the influence of light and environment. The real driving activity is a 3D perspective, and the video is 2D, and the accuracy of real-time detection is low.

D. Methods Based on Directional Gradient Histogram and Support Vector Machine

Taking the driver's distracted behavior image as the classification goal, Bu et al. proposed a behavior detection method based on Histogram of Oriented Gradient (HOG) and Support Vector Machine (SVM)[14]. By acquiring the region of interest in the image, the image is enhanced, denoised and normalized; then, the image HOG features are extracted, and then the cross-validation method is used to optimize the parameters in the SVM classifier; finally, the video image The different behaviors of the driver are classified and recognized.

III. Algorithm Design

In this section, we will design the MGMN algorithm based on the characteristics of the driver's distracted driving behavior, and introduce the structure and components of the algorithm: the multi-granularity feature extraction module, the middle-level feature extraction module and the loss function.

A. Analysis of the Detection Algorithm Demand

Distracted driving behavior is the main cause of inattention of drivers. Distracted driving behaviors mainly include the following categories: (1) making and receiving phone calls with the left hand; (2) making calls with the right hand; (3) sending and receiving messages with the left hand; 4) Send and receive messages with the right hand; (5) Talk to passengers; (6) Eat and drink during driving; (7) Operate the dashboard and radio; (8) Go back and fetch things from the back seat; (9) Clean up appearance during driving. These are shown in Fig.1:

Fig.1. Distracted driving behaviors

The corresponding characteristics of different distracted driving behaviors of drivers are as follows: (1) The left hand will be placed on the head, the right hand will be placed on the steering wheel, and the head will be upright; (2) The right hand will be placed on the head; Place it on the steering wheel with the head upright; (3) The left hand will be placed between the steering wheel and the head, and the right hand will be placed on the steering wheel; (4) The right hand will be placed between the steering wheel and the head, and the left hand Will be placed on the steering wheel; (5) the head will turn to the right (left rudder car), while both hands will be placed on the steering wheel; (6) one hand will hold food or a water cup

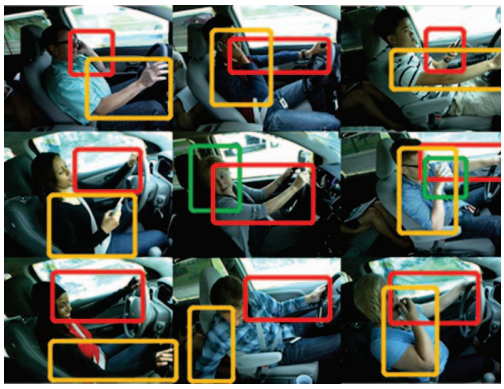

Fig.2. Distracted driving behavior characteristics

near the head's mouth, the other Only one hand is placed on the steering wheel; (7) The left hand will be placed on the steering wheel, the right hand will be stretched forward and downward, and the head will be slightly tilted forward; (8) The head will be turned backward, the right hand will be stretched back, and the left hand will be placed On the steering wheel; (9) One hand will be placed near the head and mouth, and the other hand will be placed on the steering wheel. These are shown in Fig.2:

Fig.3 shows the visualized heat map of the image features of the driver's right hand sending and receiving messages:

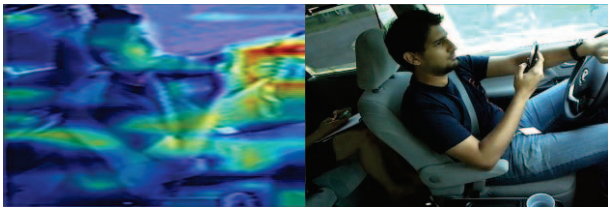

Fig.3. Feature heat map

Aiming at the above characteristics of distracted driving behavior, we design an MGMN algorithm that can simultaneously extract multi-granularity features and middle-level features in driving images. The extraction of the multi-granularity features makes every parts of the driver image being fully used to determine the driver's actions, while the extraction of middle-level features enriches the features of the image. The structure of the proposed MGMN algorithm is based on the ResNet50 network, as shown in Fig.4, with the structure of the ResNet50 network being shown in Table I. The depth of the deep learning network has a great influence on the final classification and recognition effect. The designer of the network hopes to design the network as deep as possible. However, the effect of conventional stacked networks is getting worse when the network is deep. One of the reasons is that the deeper the network, the more obvious the disappearance of the gradient, and the training effect of the network will not be very good.

Table I. ResNet50 Network

Layer Name	Output Size	50-Layer
Conv1	112×112	7×7,64,stride2
Conv2_x	56×56	3×3, max pool, stride2
		$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$
Conv3_x	28×28	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$
Conv4_x	14×14	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix}\times6$
Conv5_x	7×7	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix}\times3$
	1×1	Average pool，1000-d fc，softmax
FLOPs		3.8×10^9

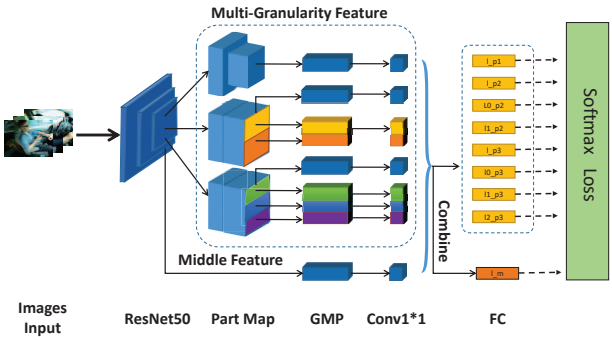

Fig.4. Structure Diagram of MGMN Algorithm

But now the shallower network can't significantly improve the recognition effect of the network, so the problem to be solved now is how to solve the problem of gradient disappearance while deepening the network. ResNet introduces the residual network structure, through the residual network, the network layer can be deepened, and the final network classification effect is also very good.

B. Design of the Multi-granularity Feature Extraction Module

After the conv4_1 module of the Resnet50 network is divided into three independent branches, the three branches are from top to bottom: the global branch, the second branch and the third branch. The three branches are the same as the original. The ResNet50 network uses the same architecture[15].

- The global branch learns global information without local features. After its conv5_1 module, it uses down-sampling with stride-2 convolutional layer, and then performs a global maximum pooling operation on the output feature map to obtain 2048-dimensional element features. After using the normalized 1×1 convolutional layer and the activation function ReLU, the pixels can be reduced to only 256-dimensional element features.
- The second branch uses a similar architecture to the global branch, but it divides the image output feature map into two pieces in the horizontal direction. These two pieces will perform the same operations as the global branch, extracting the upper and lower parts of the image respectively Information.

Fig.5. Three-layer schematic

- The structure of the three branches is exactly the same as the second branch, but it is divided into three blocks in the horizontal direction of the output feature map of the image. These three blocks will perform the

Table II. Training Results of MGMN Algorithm on State Farm's Dataset

Model	Best_acc/ epoch	Epoch									
		ep1-acc	*ep2-acc*	*ep3-acc*	*ep4-acc*	*ep5-acc*	*ep6-acc*	*ep7-acc*	*ep8-acc*	*ep9-acc*	*ep10-acc*
MGMN	99.8%/ep8	98.54	99.42	99.56	99.62	99.68	99.76	99.76	99.8	99.78	99.78
	99.72%/ep5	98.66	99.28	99.48	99.66	99.72	99.72	99.7	99.64	99.7	99.72
	99.76%/ep7	97.84	99.3	99.58	99.72	99.74	99.7	99.76	99.7	99.66	99.74
	99.8%/ep8	98.42	99.42	99.5	99.56	99.56	99.72	99.76	99.8	99.8	99.6
	99.74%/ep6	98.28	99.22	99.44	99.62	99.66	99.74	99.64	99.7	99.6	99.74
	99.74%/ep10	96.48	98.68	99.24	99.22	99.42	99.52	99.24	99.36	99.46	99.74
	99.44%/ep7	96.94	99.06	99.38	99.30	99.12	98.92	99.44	99.38	99.44	99.16
	99.72%/ep7	95.64	98.90	99.14	99.44	98.16	99.68	99.72	99.64	99.72	99.32
	99.68%/ep7	99.34	99.28	99.58	99.32	99.52	99.54	99.68	99.62	99.4	99.46

sameoperation as the global branch, respectively extracting the upper, middle and lower three in the image Part of the information, as is shown in Fig.5.

By divided into three branches, the MGMN algorithm combines the global and local features of the image, and is able to make full use of image information. In view of the difference between the overall posture and partial actions of the driver's different distracted actions, high-precision distracted driving detection is realized.

C. Mid-level Feature Extraction Module Design

In this algorithm, we have added a middle-level feature extraction module, which aims to extract more image features.

The rich image information helps to improve the accuracy of detection. After inputting the image, the neural network model will use the hidden layer in the model to transform the image into shallow, middle, and high-level features through convolution and pooling operations[16].

Shallow layer features, middle layer features and high layer features contain a lot of information about the image. Based on the ResNet50 network, this paper combines the output of the Conv5_b convolutional layer feature and the output of the multi-granularity feature extraction module after global average pooling.

D. Loss Function

The loss function (also known as the cost function) is a standard used to evaluate the output of the neural network model. The greater the difference between the output and the real label, the greater the loss function value. The smaller the value, the smaller the loss function value. In this algorithm, we use the cross-entropy loss function—Softmax. Cross entropy is the amount used in deep learning to find the gap between the target and the predicted value. The Softmax loss function is as follows:

$$\begin{cases} y_j = \dfrac{e^{z_j}}{\sum_k e^{z_k}} \\ L = -\sum_j y_j' \ln y_j \end{cases} \tag{1}$$

IV. Experimental Setup and Result Analysis

In this section, we will conduct experiments on the MGMN algorithm designed in the previous section on two public data sets and compare them with some existing deep learning algorithms. Experiments have proved that the MGMN algorithm has a significant improvement in the accuracy of detection.

A. Experimental Setup

This paper uses the "Districted-Driver-detection" data set published by State Farm, which is the same as the "AUC-Districted-Driver-detection" data set published by the American University of Cairo. Both data sets collect normal light. Under the circumstances, the different distracted driving actions of the driver in the car are classified into each type of action.

State Farm's dataset originally had 22424 images of 26 drivers between men and women from different age groups and ethnicities, demonstrating concentrated driving and 9 other distraction situations. The training set in the "Districted-Driver-detection" dataset contains 17,424 distracted driving images, and the test set contains 5000 images. The data set classification is shown in Fig.6:

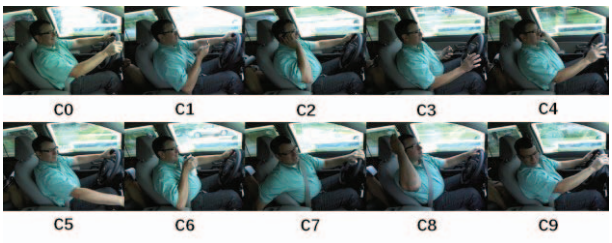

Fig.6. State Farm's Dataset

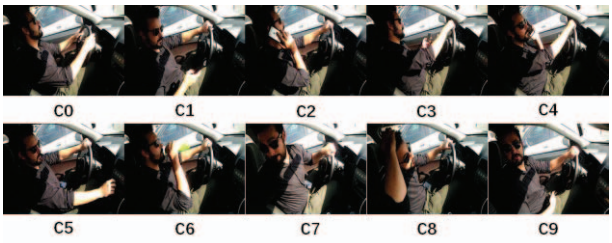

Fig.7. AUC's Dataset

Because each category in the "AUC-Districted-Driver-detection" data set is generated from the sequential screenshots of the driver's action videos, it contains a large number of wrong images. Therefore, for this dataset, we eliminated the wrong data, and then train and test the cleaned data set. The training set in the cleaned data set contains 9769 images, and the test set contains 3268 images. The data set classification is shown in Fig.8:

We use Ubuntu16.04 server to train the driver's distracted driving detection model with the graphics card being NVIDIA GeForce 1810Ti, and the video memory being 11GB. The algorithm is implemented using Pytorch, and the specific parameters are described as follows: the code learning rate is 0.0001, the batch-size is set to 64, the size of the picture is 384*128, and max epoch is set to 10.

B. *Experimental Results*

The experimental results are as follows. When the epoch s set to 10, the training accuracy of the MGMN algorithm is as high as 99.8% when there are only 8 epochs. Compared with

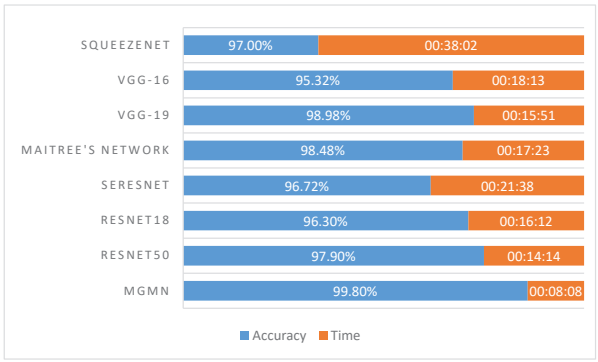

Fig.8. Comparison of experimental results on State Farm's dataset of different algorithms

Table III. Training Results of MGMN Algorithm on AUC 's Dataset

Model	Learning Rate	Best Accuracy
MGMN	0.0001	98.70%
	0.0002	98.59%
	0.0003	98.50%
	0.0004	98.20%
	0.0005	97.60%
	0.001	96.50%
	0.001	96.80%
	0.001	95.60%

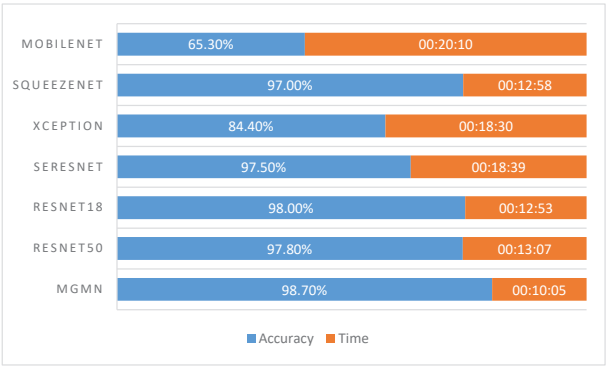

Fig.9. Comparison of experimental results on AUC's dataset of

other algorithms such as ResNet50, Squeezenet, VGG-16, VGG-19, the results have significant improvements. The results of State Farm's Dataset are shown in Table II and Fig. 8. The results of AUC's dataset are shown in Table III and Fig. 9.

C. *Result analysis*

From the above experimental results, we can see that the MGMN algorithm has a significant improvement in accuracy compared to other algorithms. In the experiment, the learning rate is 0.0001 and the highest accuracy is obtained. This is because the lower the learning rate, the shorter the step size. , The higher the accuracy, the more precise the optimization process. In order to better see the algorithm's ability to extract image features, we have drawn the distracted driving image "p3.1.0.conv1" layer of the State Farm dataset and the AUC dataset. This paper uses the Grad-CAM method to visualize the extracted features of the model in this paper. Grad-CAM is widely used to explain the mechanism of image classification by convolutional network models, and to show the discriminative features that the model focuses on in the form of heat maps[17]. In order to prevent the algorithm from

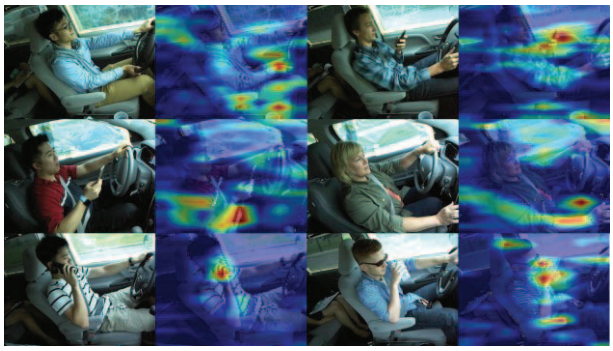

Fig.10. Heat Map of State Farm's Dataset

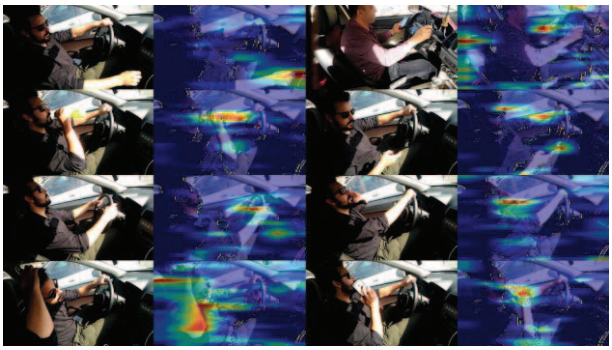

Fig.11. Heat Map of AUC's Dataset

overfitting, we selected different distracted driving actions of the same driver in the AUC dataset to prove that the algorithm has excellent feature extraction capabilities. These are shown in Fig.10 and Fig.11:

From the above experimental results, we can see that the MGMN algorithm has a significant improvement in accuracy compared to other algorithms. In the experiment, the learning rate is 0.0001 and the highest accuracy is obtained. This is because the lower the learning rate, the shorter the step size. , The higher the accuracy, the more precise the optimization process. In order to better see the algorithm's ability to extract image features, we have drawn the distracted driving image "p3.1.0.conv1" layer of the State Farm dataset and the AUC dataset. This paper uses the Grad-CAM method to visualize the extracted features of the model in this paper. Grad-CAM is widely used to explain the mechanism of image classification by convolutional network models, and to show the discriminative features that the model focuses on in the form of heat maps[17]. In order to prevent the algorithm from overfitting, we selected different distracted driving actions of the same driver in the AUC dataset to prove that the algorithm has excellent feature extraction capabilities. These are shown in Fig.11 and Fig.12:

V. CONCLUTION

Aiming at the accuracy problem of distracted driving detection, this paper proposes a distracted driving detection method based on multi-granularity and mid-level features. Based on the ResNet50 model, the multi-granularity and mid-level features of the image are extracted to make full use of the image. For the first time, it is proposed to introduce multi-granularity features into the field of distracted driving detection, which can better extract the characteristics of driving images; for the first time, it is proposed to introduce middle-level features into the field of distracted driving detection, which enriches the feature extraction of the algorithm. The MGMN algorithm achieves better results than existing distracted driving detection algorithms.

However, the MGMN algorithm also has some shortcomings and needs further improvement. The structure of the MGMN algorithm is more complex, and the volume is larger than other existing algorithms. On the other hand, we here extract the multi-granularity features and mid-level features of the driver's distracted driving detection image. Later, we may try to extract features such as high-level features.

ACKNOWLEDGMENT

We thank Professor Qing-Qian Yan from Zhejiang University of Technology for his encouragement and useful discussions, and Mr. Yehya Abouelnaga from Technical University of Munich for his help on the dataset.

REFERENCES

[1] Ministry of Public Security of the People's Republic of China,(2020, Jan.) The number of private cars in the country exceeded 200 million for the first time, and the number of cars in 66 cities exceeded one million.[Online].Available:https://www.mps.gov.cn/n2254314/n6409334/c6852472/content.html.

[2] Chao Sun, Research and Implementation of Fatigue Driving Detection Technology Based on Human Eye State, Dalian University of Technology, 2014.

[3] Zaeem Ahmad Varaich, and Sidra Khalid, "Recognizing Actions of Distracted Drivers using Inception v3 and Xception Convolutional Neural Networks" in 2019 2nd International Conference on Advancements in Computational Sciences (ICACS), 2019.

[4] Meng Chai. Identification and Early Warning of Long-distance Bus Driver Fatigue State, Jilin University, 2019.

[5] Min Hu, Research on Driver Attention Detection and Discrimination Based on Vision, Hunan University, 2018.

[6] Yu-Feng Zhang, Driver State Detection and Its Application in Human-machine Co-driving, Chongqing University, 2018.

[7] Li Li, Research on Driver Fatigue and Distraction Recognition Based on CNNs and LSTM, Hunan University, 2018.

[8] Ying-Yu Ji, Research on fatigue detection algorithm based on facial feature analysis and multi-index fusion, Jilin University, 2019.

[9] Zhi-Xiao Zheng, Research on the Influence of Visual Distraction on Driver Behavior[, Tsinghua University, 2017.

[10] Yehya Abouelnaga, Hesham M. Eraqi, and Mohamed N. Moustafa, "Real-time Distracted Driver Posture Classification" in 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada, 2018.

[11] Hesham M. Eraqi,Yehya Abouelnaga, Mohamed H. Saad, and Mohamed N. Moustafa, Driver Distraction Identification with an Ensemble of Convolutional Neural Networks, Journal of Advanced Transportation, 2019.

[12] E. Ohn-Bar, S. Martin, and M. M. Trivedi, Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies challenges , algorithms , and experimental studies. Journal of Electronic Imaging, 22(4), 2013.

[13] Maitree Leekha, Rajiv Ratn Shah, and Yi-Fang Yin, "Are You Paying Attention?Detecting Distracted Driving in Real-time" in 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM),2019.

[14] Qing-Zhi Bu, Jun Qiu, and Hu Chao Hu. Research on Driver Attention Behavior Detection Method Based on HOG Feature Extraction and SVM, Integrated Technology, 2019, 8(04), PP: 69-75.

[15] Guan-Shuo Wang, Yu-Feng Yuan, Xiong Chen, Ji-Wei Li, andXi Zhou,Receptive Multi-Granularity Representation for Person Re-Identification, IEEE Transactions on Image Processing, 2020, PP: 6096-6109.

[16] Qian Yu, Xiao-Bin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales, The Devil is in the Middle: Exploiting Mid-level Representations for Cross-Domain Instance Matching, 2017.

[17] Hao Li, Min Tang, Jian-Wu Lin,and Yun-Bo Zhao, "Cross-modal pedestrian re-recognition framework based on improved difficult triple loss,"unpublished.