future internet

MDPI

Article

A Data Augmentation Approach to Distracted Driving Detection

Jing Wang [1,2,3], ZhongCheng Wu [1,2], Fang Li [1,3,*] and Jun Zhang [1,3]

[1] High Magnetic Field Laboratory, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China; wj2019@mail.ustc.edu.cn (J.W.); zcwu@iim.ac.cn (Z.W.); zhang_jun@hmfl.ac.cn (J.Z.)
[2] Graduate School of Computer Applied Technology, University of Science and Technology of China, Hefei 230026, China
[3] High Magnetic Field Laboratory of Anhui Province, Hefei 230031, China
* Correspondence: lif@hmfl.ac.cn

Abstract: Distracted driving behavior has become a leading cause of vehicle crashes. This paper proposes a data augmentation method for distracted driving detection based on the driving operation area. First, the class activation mapping method is used to show the key feature areas of driving behavior analysis, and then the driving operation areas are detected by the faster R-CNN detection model for data augmentation. Finally, the convolutional neural network classification mode is implemented and evaluated to detect the original dataset and the driving operation area dataset. The classification result achieves a 96.97% accuracy using the distracted driving dataset. The results show the necessity of driving operation area extraction in the preprocessing stage, which can effectively remove the redundant information in the images to get a higher classification accuracy rate. The method of this research can be used to detect drivers in actual application scenarios to identify dangerous driving behaviors, which helps to give early warning of unsafe driving behaviors and avoid accidents.

Keywords: distracted driving; driving behavior; driving operation area; data augmentation; feature extraction

Citation: Wang, J.; Wu, Z.; Li, F.; Zhang, J. A Data Augmentation Approach to Distracted Driving Detection. *Future Internet* **2021**, *13*, 1. https://dx.doi.org/10.3390/fi13010001

Received: 6 November 2020
Accepted: 20 December 2020
Published: 22 December 2020

1. Introduction

According to the World Health Organization (WHO) global status report [1], road traffic accidents cause 1.35 million deaths each year. This is nearly 3700 people dying on the world's roads every day. The most heart-breaking statistic is that road traffic injury has become the leading cause of death among people aged 5 to 29 [2]. The investigation [3] for the cause of car collisions shows that 94% of road traffic accidents in the United States are caused by human operations and errors. Among them, distracted driving, which can reduce the driver's reaction speed, is the most dangerous behavior. In 2018 alone, 2841 people died in traffic collisions on United States roads due to driver distraction [4].

The impacts of distracted behavior of drivers are multifaceted [5], including visual behavior, operating behavior, driving pressure, and the ability to perceive danger, etc. According to the definition of the National Highway Traffic Safety Administration (NHTSA) [6], distracted driving refers to any activity that can divert attention away from driving, including (a) talking or texting on a phone, (b) eating and drinking, (c) talking to others in the vehicle, or (d) using radio, entertainment or navigation system.

Distracted driving detection can be used to give early warning of dangerous driving behavior, including using a mobile phone to call or send text messages, using navigation applications or choosing to play music, etc. [7]. Distracted driving detection methods are mainly based on the driver's facial expression, head operation, line of sight or body operation [8]. Through visual tracking, target detection, motion recognition and other technologies, the driver's driving behavior and physiological state can be detected.

In the early days, researchers mainly focused on behavior analysis based on the driver's line-of-sight direction through the eye, face, and head operation. In 2002, Liu et al. [9] proposed a method of tracking the driver's facial area and used the yaw direction angle to estimate the driver's facial operation to detect the driver's facial orientation. Eren et al. [10] successfully developed a driver's facial operation detection system based on binocular stereo vision, using hidden Markov models to predict the driver's facial operation in 2007.

Later, with the development of machine learning technology and the public of driving behavior datasets, increasing studies were added to analyze the driver's phone calling, drinking, eating and other unsafe driving behaviors. Southeast University driving posture (SUE-DP) dataset [11] was proposed in 2011. The experiment collected four types of distracted driving behaviors: "grasping the steering wheel", "operating the shift lever", "eating," and "talking on a cellular phone". A series of studies have been conducted based on the SUE-DP dataset: Zhao [12] used the multiwavelet transform method and the multilayer perceptron (MLP) classifier to recognize four predefined driving postures and obtained an accuracy of 90.61%. Zhao et al. [11] introduced a contourlet transform method for feature extraction and achieved 90.63% classification accuracy. Subsequently. Chihang [13] used support vector machines (SVM) algorithm with an intersection kernel for obtaining 94.25% accuracy. In 2013, The image pyramid histogram of oriented gradients (PHOG) features and edge features [14] were extracted comprehensively to increase the classification accuracy to 94.75% within MLP. In 2014, Yan [15] extracted the PHOG features of historical moving images containing time information and obtained 96.56% accuracy through the random forests (RF) algorithm.

Recently, with the development of deep learning classification and detection technology, increasing researchers analyzed driving behavior through convolutional neural networks (CNNs). More researchers have also begun to build their own research datasets. A combination of pre-trained sparse filters and convolutional neural networks [16] was used to increase the classification accuracy of the SUE-DP dataset to 99.78%. Yan [17] improved the conventional regions with CNNs features (R-CNN) framework by replacing traditional skin-like region extractor algorithms and obtained a classification accuracy of 97.76% on the SUE-DP dataset. Liu [18] used an improved dual-input deep three-dimensional convolutional network structure algorithm based on a three-dimensional convolutional neural network (3DCNN), achieving 98.41% accuracy on the rail transit dataset. Jin [19] trained two independent convolutional neural networks by optimizing the size and number of convolution kernels, which can effectively identify mobile phones and hands in real time achieving 144 fps and an accuracy of 95.7% for mobile phone usage on the self-built dataset. Multiscale attention convolutional neural network [20] was proposed driver action recognition.

More recent datasets and studies include: The StateFarm dataset was published in the 2016 Kaggle distracted driving recognition competition [21] with ten types of distracted driving behaviors. Alotaibi [22] used an improved multiscale Inception model with a classification accuracy of 96.23% on the StateFarm dataset. Lu et al. [23] used the improved deformable and dilated faster RCNN (DD-RCNN) structure to obtain an accuracy of 92.2%. Valeriano [24] and Moslemi [25] used the 3DCNN algorithm to recognize driving behavior and got 96.67% and 94.4% accuracy rates, respectively. In 2018, Eraqi and others [26,27] proposed the American University in Cairo (AUC) distracted driving dataset with reference to the ten distracted postures defined in the StateFarm dataset. The ASUS ZenPhone close-range camera and DS325 Sony DepthSense camera were used to collect driving images and videos of 44 volunteers from 7 countries. The cameras were fixed to the roof handle on the top of the front passenger's seat. The resolution of the data is 1080×1920 or 640×480. A total of 17,308 frames were collected for the dataset, which was finally annotated to ten kinds of distracted driving behaviors: safe driving (3686), texting using the right hand (1974), talking on the phone using the right hand (1223), texting using left hand (1301), talking on the phone using left hand (1361), operating the radio (1220), drinking (1612), reaching behind (1159), hair and makeup (1202), and talking

to the passenger (2568). The dataset was randomly divided into 75% for training and 25% for test data. A genetically weighted ensemble of convolutional neural networks combined with the face, hand, and skin regions was proposed to obtain an accuracy of 95.98% with the AUC dataset. Bhakti and others [28] used the AUC dataset to achieve 96.31% accuracy through improved visual geometry group 16 (VGG-16) with regularization methods.

The collection of the dataset mainly used the camera to obtain images of the driver's driving process. During the collection process, it was usually recommended that the driver perform distracting subtasks to simulate distracting driving. The distracted driving methods were mainly based on the driver's facial expression, head operation, line of sight, or body operation for feature extraction. Machine-learning methods and deep learning CNN methods were used for distracting driving recognition. However, the existing datasets and related analysis methods still encounter some problems in the research: on one hand, the current distracted driving research mainly judges driving behavior by the driver's facial and head direction, hand movements, or skin segmentation information. However, the judgment of single local information is prone to classification errors. On the other hand, due to the differences in the resolution, wide-angle, installation position, and installation angle of the camera in different datasets, or the differences in the position of the seat and steering wheel, the position and angle of each driver in the dataset will be different, leading to the images in the dataset have different redundant information.

Two-stage deep architecture methods [29] were usually used in image classification of remotely collected images. A common approach in the literature was employing CNNs for feature extraction to reduce the dimensionality [30]. Data enhancement methods [31] such as flip, rotation, crop, interpolation and color convert [32] were also often used in the first stage of processing to increase robustness. In order to build a more robust driver behavior analysis model and improve the accuracy of dataset classification, this paper designs a data augmentation preprocessing model for driver behavior key areas based on faster R-CNN [33] detection algorithms to improve the accuracy of the algorithm learned from the two-stage depth architecture method. The classification results with data augmentation are verified based on AlexNet [34], InceptionV4 [35] and Xception [36], respectively. To achieve the best performance, transfer learning [37] was applied in training. The American University in Cairo distracted driver (AUC) dataset is used for the experiments.

The main contributions of this paper are summarized in the following three parts:

(1) First, the class activation mapping method [33] was used to analyze the driving operation key areas in the images.

(2) Then, in order to enhance the dataset, the image detection algorithm faster R-CNN was used to generate the new driving operation area (DOA) dataset. The driving operation areas were labeled on 2000 images using the AUC dataset to establish the training driving operation areas detection dataset for faster R-CNN training. Within the trained faster R-CNN model, all the AUC dataset images were tested to obtain the preprocessed AUC new DOA classification dataset, which was consistent with the original AUC dataset at the classification storage method and naming method.

(3) Next, a classification model was built to process the AUC original dataset and the DOA dataset. The experiments were tested with AlexNet, InceptionV4 and Xception separately to get the best result.

(4) Finally, the trained classification method was used to test our own dataset, which was established with a wide-angle camera different from the close-range camera in the AUC dataset.

The framework of classification model with data augmentation method is shown in Figure 1. Experiments proved that the classification accuracy of the method proposed in this paper is up to 96.97%, which can improve the accuracy of classification.

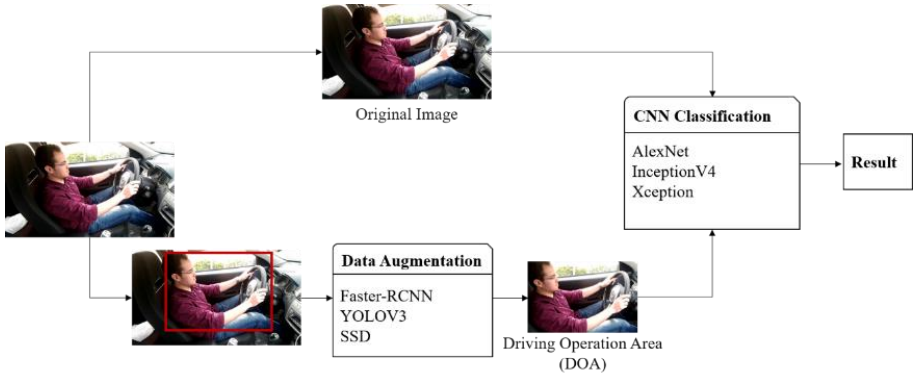

Figure 1. The framework of the classification model with the data augmentation method.

2. Materials and Methods

2.1. Driving Operation Area

In order to effectively observe which areas of the image the network focuses on, this paper used gradient-weighted class activation mapping (grad-CAM) [38] to visually display the features regions found by the Xception classification network, which displayed in the form of a class-specific saliency map or "heat map" for short. Figure 2 shows the grad-CAM result of ten different driving behaviors, which can be used to visually evaluate the key feature regions of the image. The distribution of the ribbon color from red to blue that you can see the mapping relationship between weight and color. The red area in the activation map represents the higher basis area for the model to make classification decisions.

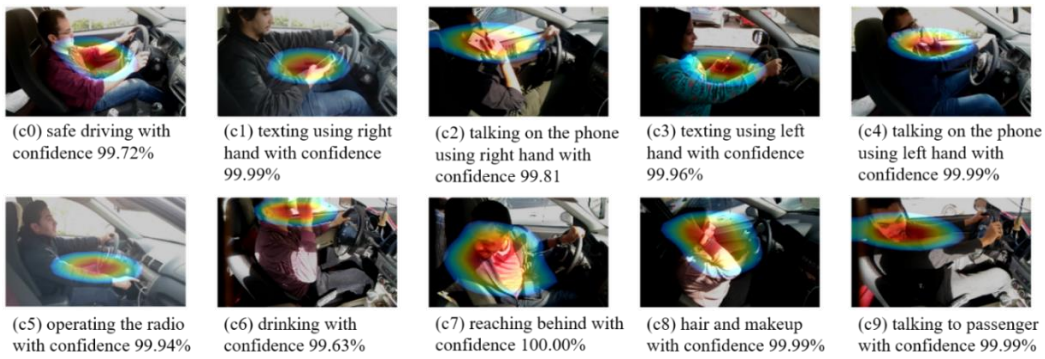

Figure 2. Grad-CAM activation maps of various distracted driving behaviors in the dataset.

According to the grad-CAM result, the driver's upper body behaviors in the vehicle environment determine the distracted driving classification result, which means that the background and legs that are not related to the driver's operation are all redundant information in the feature extraction. We proposed the concept of the driving operation area (DOA), including the steering wheel and the driver's upper body, which include the head, torso, and arms, to describe the features related to the driver's driving behavior. The area in the red box shown in Figure 1 is what we defined DOA.

2.2. Methods for Data Augmentation

Due to the fixed nature of the distracted driving background, traditional data augmentation methods (such as flipping, rotation, trimming, and interpolation) result in unrealistic scenes, which will cause information distortion and increase irrelevant data. This paper proposed the data augmentation method based on the key area of driving behavior. The mature image detection convolutional network model is used for the data augmentation method. The AUC dataset was enhanced based on the DOA to obtain a new

dataset. Classification modules were introduced to classify the original dataset and the new dataset.

According to the requirements of the image detection model for the dataset, we randomly selected 2000 images from the AUC dataset to relabel the driving operation area using "labelImg" software tool. The labeling area included the steering wheel and the driver's upper body, including the head, torso, and arms. The annotation file was saved as an XML file in accordance with the Pascal visual object classes (PASCAL VOC) dataset format. The image detection convolutional network model was used to extract the driving operation area.

faster R-CNN [33] was chosen as the image detection model for feature extraction. faster R-CNN creatively used region proposal networks to generate proposals and shared the convolutional network with the target detection network, which can reduce the number of proposals from the original about 2000 to 300 and improve the quality of the suggested frames. The algorithm won many firsts in the ImageNet Large-scale visual recognition competition (ILSVRC) and the common objects in context (COCO) competitions that year, still used frequently by studiers now.

faster R-CNN model was trained with the labeled data; then, the trained faster R-CNN model was used to infer all the AUC dataset images to obtain the preprocessed new DOA classification dataset. The classification and naming of the DOA dataset were consistent with the original AUC dataset.

2.3. Methods for CNN Classification

This paper used the mature image detection convolutional network model for the data augmentation method. Classic models such as Alexnet [34], InceptionV4 [35], and Xception [36] had been widely used in image classification research in recent years. AlexNet successfully applied rectified linear units (ReLU), dropout and local response normalization (LRN) in CNN. The Inception network started from GoogLeNet in 2014, which had gone through several iterations of versions up to the latest InceptionV4. Xception was another improvement proposed by Google after Inception.

Transfer learning, whose initial weights of each model came from the weights obtained by pre-training on ImageNet, was used in our classification test to train the AUC dataset by optimizing the parameters to get the best result.

The AUC dataset was enhanced based on the DOA to obtain a new dataset. Classification modules were introduced to classify the original dataset and the new dataset. The classification framework with the data augmentation method is shown in Figure 1.

2.4. Wide-Angle Dataset

In order to further verify the generalization ability of our methods, A Wide-angle distracted driving dataset was collected for verification. Referring to the collection methods of the State Farm dataset and the AUC dataset, we fixed the camera to the car roof handle on top of the front passenger's seat. Fourteen volunteers sat in the car to simulate distracted driving as required in both day and night scenes. Some volunteers participated in more than one collection session at different times of day, driving roads and wearing different clothes. The 360's G600 recorder, which has a resolution of 1920 $\times$ 1080 and a wide-angle of 139 degrees, was used in the collection. In order to simulate a natural driving scene as much as possible; in some cases, there were other passengers in the car during the collection process.

The data were collected in a video format with the size of 1920 $\times$ 1080 and then cut into individual images. Our dataset finally collected 2200 pictures of ten kinds of distracted driving behaviors: safe driving (291), texting using the right hand (224), talking on the phone using the right hand (236), texting using left hand (218), talking on the phone using left hand (211), operating the radio (203), drinking (198), reaching behind (196), hair and makeup (182), and talking to the passenger (241). Part of the images of the wide-angle dataset is shown in Figure 3.

Figure 3. Part of the images in the wide-angle dataset.

3. Results

The experiments in this article were based on the PaddlePaddle framework and Python design, with the hardware environment using a Linux server with Ubuntu 16.04. A single NVIDIA GeForce GTX, 1080 Ti GPU with 12 GB RAM, was used in the experiments.

3.1. Results for Driving Operation Area Extraction

The labeled driving dataset with 2000 images was split into a training set and a validation set with a ratio of 8:2 for validating the detection model performance. Using the same training strategy as Detection, the dataset was trained with the batch size of 8, the learning rate of 0.001, and the training iterations of 50,000. The momentum 0.9 with a weight decay of 0.0001 for stochastic gradient descent (SGD) was used to converge the model. The Resnet was used for the backbone network. The Resnet weights pre-trained on ImageNet model was used for initialization.

Table 1 is the result of driving operation area extraction with the detection model. Faster R-CNN model was evaluated and compared with the other two models: you only look once (YOLO) [39] and single shot multibox detector (SSD) [40] models. According to the result in Table 1, the accuracy of faster R-CNN detection is 0.6271, and fps is 10.50 which can meet real-time requirements. Considering the accuracy requirements, the faster R-CNN was chosen as the detection model in our experiments. YOLOV3 and SSD models can be used as real-time detection system.

Table 1. Driving operation area detection result.

Model	fps	mAP (0.75)	mAR (0.75)
faster R-CNN	10.50	0.6271	0.6572
YOLOV3	37.21	0.5390	0.5568
SSD	49.52	0.5767	0.5812

Figure 4 shows ten different types of driving operation area detection results. Comparing Figures 2 and 4, the key regions are extracted by the image detection model.

Figure 4. Ten different types of driving operation area (DOA) detection results.

Then the trained weights of faster R-CNN were used to detect the key areas of driving behavior in the AUC dataset, and generate a dataset of driving operation area, which was recorded as the DOA dataset, which classification and naming methods were the same as the original AUC dataset.

3.2. Results for CNN Classification

In the experiment, the dataset of AUC and DOA were both 12,997 images of training set and 4331 images of test set. The image classification model AlexNet, InceptionV4 and Xception were used to train with image shape of 224 × 224 × 3, the learning rate of 0.001, batch size of 32, and epoch of 100. The top-1 accuracy was selected to evaluate the performance of the models. We performed 3 rounds of verification. Table 2 summarizes the test results for loss and accuracy of three different convolutional network models: AlexNet, InceptionV4, and Xception.

Table 2. Image classification test result with AUC and DOA dataset.

Model	Source	Loss	Top-1 Acc.
AlexNet	AUC	0.3753	0.9314
	DOA	0.3402	0.9386
InceptionV4	AUC	0.3041	0.9506
	DOA	0.2771	0.9572
Xception	AUC	0.2320	0.9531
	DOA	0.2156	0.9655

As can be seen from Table 2, the test top-1 accuracy of the AlexNet, InceptionV4 and Xception on the AUC dataset are 0.9314, 0.9506 and 0.9531, respectively, and the test results on the DOA dataset are 0.9386, 0.9572 and 0.9655, which means the DOA dataset has higher detection accuracy and lower loss value than the original AUC dataset.

Figure 5 shows the change for loss and accuracy of each method in each epoch stage. When the epoch is 10, the loss and accuracy of the DOA dataset with the Xception model begin to stabilize, and when the epoch is 14, the original AUC dataset loss and accuracy with the Xception model begin to stabilize. Moreover, The loss values in the DOA-based results are lower than original AUC dataset. It can be seen from the testing loss and accuracy curves with varying epochs, the loss of DOA dataset corresponding to the key areas of driving behavior converges faster than the original AUC dataset, and the detection accuracy rises faster too.

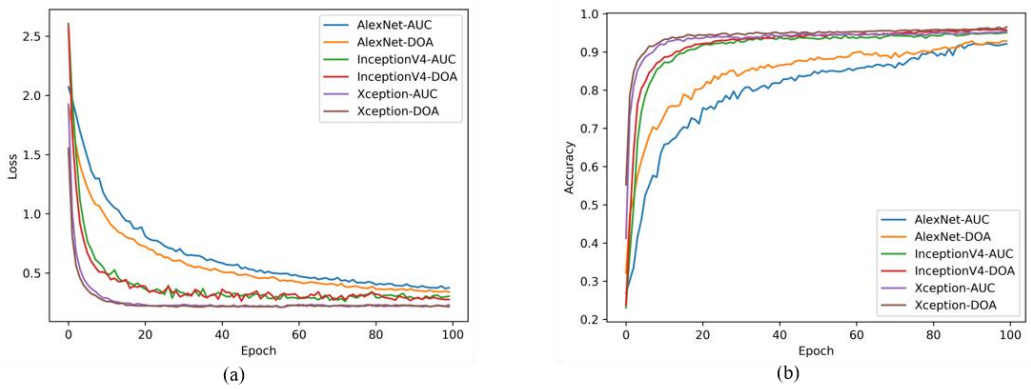

Figure 5. Accuracy and loss of image classification result. (**a**) the loss of each method in each epoch stage; (**b**) the accuracy of each method in each epoch stage.

Finally, the DOA training set obtained through data augmentation and the original AUC training set were merged to expand the dataset. The final classification accuracy is shown in Table 3. Among the three classification models, the baseline with Xception has the smallest fluctuation, the lowest loss result, and the highest accuracy, which is the most suitable for the benchmark model of this classification. For more evaluation, Figure 6 is the confusion matrix for the classification results of the ten distracting behaviors with Xception. Using the given confusion matrix, one can check that many categories can

easily be mistaken for (c0) "safe driving". Moreover, the most confusing operation is (c8) "hair and makeup". It may be due to the position of "hands on the wheel" in both classes.

Table 3. Final result after confidence comparison.

Model	Top-1 Acc.
AlexNet	0.9396
InceptionV4	0.9603
Xception	0.9697

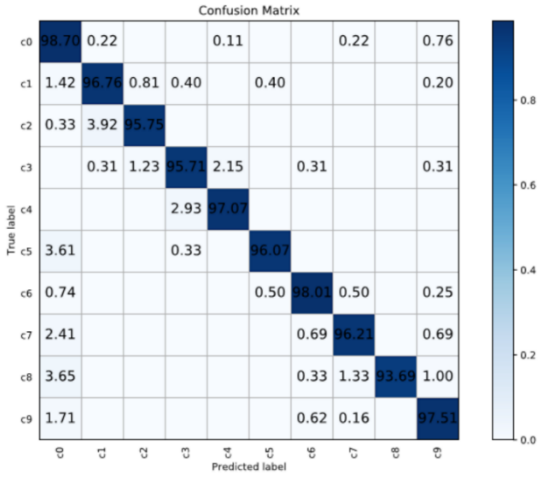

Figure 6. Confusion matrix of results Xception classification.

Our distracted driver detection result was compared with earlier methods in the literature. Compared with some early methods, our method can be applied to the preprocessing stage. We achieve the best accuracy than earlier methods as shown in Table 4. Among them, the top-1 accuracy of our module based on Xception is finally 0.9697, which is 1.66% higher than the classification accuracy of origin AUC dataset.

Table 4. Comparison with earlier methods from literature on AUC dataset.

Model	Top-1 Acc.
GA weighted ensemble of all 5 [26]	0.9598
VGG [28]	0.9444
VGG with regularization [28]	0.9631
ResNet + HRNN + modified Inception [22]	0.9236
Our method	0.9697

3.3. Tests on Wide-Angle Dataset

Due to the high correlation between the training and test data of the AUC dataset, this makes the detection of driving distraction an easier problem. Therefore, the newly collected wide-angle dataset was used to verify the generalization ability of our method. The wide-angle dataset contains 14 drivers (2200 samples). The wide-angle dataset was used to verify the feasibility of our proposed method, especially for datasets with a relatively small proportion of drivers. The trained model on the AUC dataset was used in the verification for the wide-angle dataset directly. Referring to the performance of the previous experiment with Xception-based model, this paper used the Xception-based model to verify the generalization ability.

Table 5 shows the verification result of the dataset captured by the wide-angle camera. The classification top-1 accuracy of the model is greater than 80%, which verifies a relatively good generalization ability. In addition, the classification results after extracting

the key areas of the driver operation are significantly better than the original data classification results. It proves the necessity of extracting key areas of drivers in distracted driving detection.

Table 5. Result on wide-angle dataset.

Model	Source	Top-1 Acc.
Xception	Wide-angle Dataset	0.8131
	DOA of Wide-angle Dataset	0.8394

4. Discussion

In practical applications, due to the difference in the installation position and resolution of the camera, and the difference in the position of the driver's seat and steering wheel, the driver's distribution position and angle in the image will be different. The difference in the proportion of the driver's operating area in the image will cause many pixels in the image of the collected dataset to be redundant information. This article focuses on improving the robustness and accuracy of distracted driving detection.

First, with the labeled data, faster R-CNN was used to detect the key areas of driving behavior. The extraction of DOA was a large target detection for CNN, and the general faster R-CNN has been able to achieve good accuracy. It can be seen from the experimental results that this method can extract key information and can be used in the first stage of distracted driving detection. Comparing with grad-CAM activation maps, it can be seen that our method was especially helpful for driving behavior detection in complex backgrounds.

Second, the convolutional neural network classification model was used to test the loss and accuracy of the AUC dataset and the DOA dataset. It can be seen from the result that the DOA dataset has higher detection accuracy and lower loss value than the original AUC dataset. Testing with the combined dataset of AUC and DOA, the experiment got a 96.97% top-1 accuracy. Compared with some early methods in the literature, our method can extract the overall characteristics of key areas of driving behavior. The loss of InceptionV4 and Xception dropped to a better result when the epoch was 4, and reached relatively stable when the epoch reached 40. The results showed the effectiveness of transfer learning for CNN models.

Third, The wide-angle dataset collected by actual scene was used to verify our method. Our results demonstrated that detect the key areas of driving behavior has a great significance for driving behavior analysis of wide-angle camera shooting and long-range shooting.

It can find that if the extracted features come from the entire image, which means all the information in the image (regardless of whether it is related to driving behavior) are used as a training input, the result will lead to more redundant information and larger calculation. Considering the diversity of the driver's position and the complexity of the cab environment, our method is suitable for practical application fields.

5. Conclusions

Distracted driving detection has become a major research in transportation safety due to the increasing use of infotainment components in vehicles. This paper proposed a data augmentation method for driving position area with the faster R-CNN module. The convolutional neural network classification model was used to identify ten distracting behaviors in the AUC dataset, reaching the top-1 accuracy of 96.97%. Extensive results carried out show that our method improves the accuracy of the classification and has strong generalization ability. The experimental results also showed that the proposed method was able to extract key information. This provided a path for the preprocessing stage of driving behavior analysis.

In the future, the following aspects can be continued for further research:

First, more distracted driving datasets with multi-angle and night scenarios should be collected and published for more comprehensive research. We need to verify our model on more practical large-scale datasets.

Second, the current classification algorithm divides dangerous driving behaviors into multiple categories, but in actual driving behaviors, multiple dangerous behaviors may co-exist, such as watching around when making a call. We can use detection modes such as YOLO (or any other object detector) to detect the face, hand, and other information on the basis of the work of DOA for more driving behavior identification.

Author Contributions: Conceptualization, J.W. and J.Z.; methodology, J.W.; software, J.W.; validation, J.W.; resources, J.W.; writing—original draft preparation, J.W.; writing—review and editing, F.L.; funding acquisition, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This project is supported by the Internet of Vehicles Shared Data Center and Operation Management Cloud Service Platform of Anhui Province.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. *Global Status Report on Road Safety 2018: Summary*; World Health Organization: Geneva, Switzerland, 2018.
2. Peden, M. Global collaboration on road traffic injury prevention. *Int. J. Inj. Control Saf. Promot.* **2005**, *12*, 85–91. [CrossRef]
3. Singh, S. *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*; National Highway Traffic Safety Administration: Washington, DC, USA, 2015.
4. Vasilash, G.S. Distraction and Risk. *Automot. Des. Prod.* **2018**, *130*, 6.
5. Kaber, D.B.; Liang, Y.; Zhang, Y.; Rogers, M.L.; Gangakhedkar, S. Driver performance effects of simultaneous visual and cognitive distraction and adaptation behavior. *Transp. Res. Part F-Traffic Psychol. Behav.* **2012**, *15*, 491–501. [CrossRef]
6. Strickland, D. How Autonomous Vehicles Will Shape the Future of Surface Transportation. 2013. Available online: https://www.govinfo.gov/content/pkg/CHRG-113hhrg85609/pdf/CHRG-113hhrg85609.pdf (accessed on 21 December 2020).
7. Liu, D. Driver status monitoring and early warning system based on multi-sensor fusion. In Proceedings of the 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Vientiane, Laos, 11–12 January 2020. [CrossRef]
8. Yanfei, L.; Yu, Z.; Junsong, L.; Jing, S.; Feng, F.; Jiangsheng, G. Towards Early Status Warning for Driver's Fatigue Based on Cognitive Behavior Models. In Proceedings of the Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management: 4th International Conference, DHM 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, 21–26 July 2013. [CrossRef]
9. Liu, X.; Zhu, Y.D.; Fujimura, K. Real-time pose classification for driver monitoring. In Proceedings of the IEEE 5th International Conference on Intelligent Transportation Systems, Singapore, 3–6 September 2002; pp. 174–178.
10. Eren, H.; Celik, U.; Poyraz, M. Stereo vision and statistical based behaviour prediction of driver. In Proceedings of the 2007 IEEE Intelligent Vehicles Symposium, Istanbul, Turkey, 13–15 June 2007; pp. 657–662.
11. Zhao, C.H.; Zhang, B.L.; He, J.; Lian, J. Recognition of driving postures by contourlet transform and random forests. *IET Intell. Transp. Syst.* **2012**, *6*, 161–168. [CrossRef]
12. Zhao, C.; Gao, Y.; He, J.; Lian, J. Recognition of driving postures by multiwavelet transform and multilayer perceptron classifier. *Eng. Appl. Artif. Intell.* **2012**, *25*, 1677–1686. [CrossRef]
13. Chihang, Z.; Bailing, Z.; Jie, L.; Jie, H.; Tao, L.; Xiaoxiao, Z. Classification of Driving Postures by Support Vector Machines. In Proceedings of the 2011 Sixth International Conference on Image and Graphics, Hefei, China, 12–15 August 2011; pp. 926–930. [CrossRef]
14. Zhao, C.H.; Zhang, B.L.; Zhang, X.Z.; Zhao, S.Q.; Li, H.X. Recognition of driving postures by combined features and random subspace ensemble of multilayer perceptron classifiers. *Neural Comput. Appl.* **2013**, *22*, S175–S184. [CrossRef]
15. Yan, C.; Coenen, F.; Zhang, B.L. Driving Posture Recognition by Joint Application of Motion History Image and Pyramid histogram of Oriented Gradients. *Int. J. Veh. Technol.* **2014**, 846–847. [CrossRef]
16. Yan, C.; Zhang, B.; Coenen, F. Driving Posture Recognition by Convolutional Neural Networks. In Proceedings of the 2015 11th International Conference on Natural Computation (Icnc), Zhangjiajie, China, 15–17 August 2015; pp. 680–685.
17. Yan, S.; Teng, Y.; Smith, J.S.; Zhang, B. Driver Behavior Recognition Based on Deep Convolutional Neural Networks. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (Icnc-Fskd), Changsha, China, 13–15 August 2016; pp. 636–641.
18. Liu, Y.Q.; Zhang, T.; Li, Z. 3DCNN-Based Real-Time Driver Fatigue Behavior Detection in Urban Rail Transit. *IEEE Access* **2019**, *7*, 144648–144662. [CrossRef]

19. Jin, C.C.; Zhu, Z.J.; Bai, Y.Q.; Jiang, G.Y.; He, A.Q. A Deep-Learning-Based Scheme for Detecting Driver Cell-Phone Use. *IEEE Access* **2020**, *8*, 18580–18589. [CrossRef]
20. Hu, Y.C.; Lu, M.Q.; Lu, X.B. Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network. *Signal Process. Image Commun.* **2020**, 81. [CrossRef]
21. Kaggle. State Farm Distracted Driver Detection. Available online: https://www.kaggle.com/c/state-farm-distracted-driver-detection/data (accessed on 21 December 2020).
22. Alotaibi, M.; Alotaibi, B. Distracted driver classification using deep learning. *Signal Image Video Process.* **2019**. [CrossRef]
23. Lu, M.Q.; Hu, Y.C.; Lu, X.B. Driver action recognition using deformable and dilated faster R-CNN with optimized region proposals. *Appl. Intell.* **2020**, *50*, 1100–1111. [CrossRef]
24. Valeriano, L.C.; Napoletano, P.; Schettini, R. Recognition of driver distractions using deep learning. In Proceedings of the 2018 IEEE 8th International Conference on Consumer Electronics, Berlin, Germany, 2–5 September 2018.
25. Moslemi, N.; Azmi, R.; Soryani, M. Driver Distraction Recognition using 3D Convolutional Neural Networks. In Proceedings of the 2019 4th International Conference on Pattern Recognition and Image Analysis, Tehran, Iran, 6–7 March 2019; pp. 145–151. [CrossRef]
26. Eraqi, H.M.; Abouelnaga, Y.; Saad, M.H.; Moustafa, M.N. Driver Distraction Identification with an Ensemble of Convolutional Neural Networks. *J. Adv. Transp.* **2019**. [CrossRef]
27. Abouelnaga, Y.; Eraqi, H.M.; Moustafa, M.N. Real-time Distracted Driver Posture Classification. *arXiv* **2017**, arXiv:abs/1706.09498.
28. Baheti, B.; Gajre, S.; Talbar, S.; IEEE. Detection of Distracted Driver using Convolutional Neural Network. In Proceedings of the 2018 IEEE/Cvf Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, Utah, USA, 18–22 June 2018; pp. 1145–1151. [CrossRef]
29. Petrovska, B.; Zdravevski, E.; Lameski, P.; Corizzo, R.; Štajduhar, I.; Lerga, J. Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification. *Sensors* **2020**, *20*, 3906. [CrossRef]
30. Petrovska, B.; Atanasova-Pacemska, T.; Corizzo, R.; Mignone, P.; Lameski, P.; Zdravevski, E. Aerial scene classification through fine-tuning with adaptive learning rates and label smoothing. *Appl. Sci.* **2020**, *10*, 5792. [CrossRef]
31. Zhao, Z.; Luo, Z.; Li, J.; Chen, C.; Piao, Y. When Self-Supervised Learning Meets Scene Classification: Remote Sensing Scene Classification Based on A Multitask Learning Framework. *Remote Sens.* **2020**, *12*, 3276. [CrossRef]
32. Izadpanahkakhk, M.; Razavi, S.M.; Taghipour-Gorjikolaie, M.; Zahiri, S.H.; Uncini, A. Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning. *Appl. Sci.* **2018**, *8*, 1210. [CrossRef]
33. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
34. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
35. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First Aaai Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
36. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
37. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
38. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
39. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
40. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 21–37.