actuators

Article

A Multi-Semantic Driver Behavior Recognition Model of Autonomous Vehicles Using Confidence Fusion Mechanism

Hongze Ren, Yage Guo, Zhonghao Bai * and Xiangyu Cheng

The State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, Hunan University,
Changsha 410082, China; hnurhz@hnu.edu.cn (H.R.); guogeya@hnu.edu.cn (Y.G.); hnucxy@hnu.edu.cn (X.C.)
* Correspondence: baizhonghao@163.com

Abstract: With the rise of autonomous vehicles, drivers are gradually being liberated from the traditional roles behind steering wheels. Driver behavior cognition is significant for improving safety, comfort, and human–vehicle interaction. Existing research mostly analyzes driver behaviors relying on the movements of upper-body parts, which may lead to false positives and missed detections due to the subtle changes among similar behaviors. In this paper, an end-to-end model is proposed to tackle the problem of the accurate classification of similar driver actions in real-time, known as MSRNet. The proposed architecture is made up of two major branches: the action detection network and the object detection network, which can extract spatiotemporal and key-object features, respectively. Then, the confidence fusion mechanism is introduced to aggregate the predictions from both branches based on the semantic relationships between actions and key objects. Experiments implemented on the modified version of the public dataset Drive&Act demonstrate that the MSRNet can recognize 11 different behaviors with 64.18% accuracy and a 20 fps inference time on an 8-frame input clip. Compared to the state-of-the-art action recognition model, our approach obtains higher accuracy, especially for behaviors with similar movements.

Keywords: intelligent electric vehicles; driver behavior recognition; multi-semantic description; confidence fusion

Citation: Ren, H.; Guo, Y.; Bai, Z.; Cheng, X. A Multi-Semantic Driver Behavior Recognition Model of Autonomous Vehicles Using Confidence Fusion Mechanism. *Actuators* **2021**, *10*, 218. https://doi.org/10.3390/act10090218

Academic Editors: Peng Hang, Xin Xia and Xinbo Chen

Received: 27 July 2021
Accepted: 29 August 2021
Published: 31 August 2021

1. Introduction

Driver-related factors (e.g., distraction, fatigue, and misoperation) are the leading causes of unsafe driving, and it is estimated that 36% of vehicle accidents can be avoided if no driver engages in distracting activities [1,2]. Secondary activities such as talking with cellphones, consuming food, and interacting with in-vehicle devices lead to the significant degradation of driving skills, and increases in reaction times in emergency events [3]. With the rise of autonomous vehicles, drivers are gradually being liberated from the traditional roles behind steering wheels, thereby more freedom may contribute to complex behaviors [4]. As full automation could be decades away, driver behavior recognition is essential for autonomous vehicles with partial or conditional automation, where drivers have to be ready for requests for intervention [5].

With the growing demand for analyses of driver behaviors, driver behavior recognition has rapidly gained attention. Previous studies mainly adopted machine learning algorithms, such as random forest [6], Adaboost [7], and support vector machine [8], to detect distracted drivers. Deep learning technology hastens the parturition of outstanding driver behavior recognition models due to its powerful studying and generalizing ability. A typical pipeline of driver behavior recognition models based on deep learning is presented in Figure 1. First, driver movements are captured by cameras and fed into the data processing part in sequences of frames. The next step is to extract deep features and assign corresponding labels to these features. During this process, classification accuracy is critical to the model's performance. In [9], the multi-scale Faster-RCNN [10] is employed in driver's cellphone usage detection with the fusion approach based on features and geometric information. Streiffer et al. [11] propose a deep learning solution for distracted driving detection by means of aggregating the classification results of frame-sequence and IMU-sequence. Baheti et al. [12] adapted the VGG-16 [13] with various regularization techniques (e.g., dropout, L2 regularization, and batch normalization) to perform distracted driver detection.

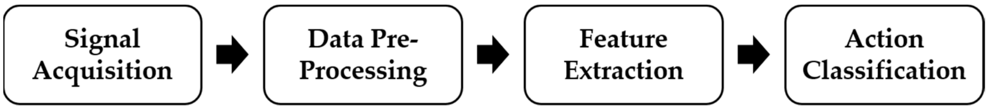


Figure 1. A typical pipeline of driver behavior recognition models based on deep learning.

The 3D-CNN is widely utilized for driver behavior recognition in order to aggregate the deep features from both spatial and temporal dimensions. Martin et al. [14] introduced the large-scale video dataset Drive&Act and provided benchmarks by adopting prominent methods for driver behavior recognition. Reiß et al. [15] adopted the fusion mechanism based on semantic attributes and word vectors to tackle the issue of zero-shot activity recognition. In [16], an interwoven CNN is used to identify driver behaviors by merging the features coming from multi-stream inputs.

In summary, it is ambitious to achieve high accuracy while maintaining runtime efficiency for driver behavior recognition. Existing research mostly analyzes driver behaviors by relying on the movements of upper-body parts, which may lead to false positives and missed detections due to the subtle changes among similar behaviors [17,18]. To tackle this problem, an end-to-end model is proposed, inspired by the human visual cognitive system. When humans understand complex and similar behaviors, our eyes capture not only the action cues, but also the key-object cues, in order to obtain more complete descriptions of behaviors. The example in Figure 2 illustrates our inspiration. Therefore, two parallel branches are presented to perform action classification and object classification, respectively. The action detection network, called ActNet, is used to extract spatiotemporal features from an input clip, and the object detection network called ObjectNet is used to extract key-object features from the key frame. Then, the confidence fusion mechanism (CFM) is introduced to aggregate the predictions from both branches based on the semantic relationships between actions and key-objects. Figure 3 illustrates the overall architecture of the proposed model. Our contributions can be summarized as follows:

- An end-to-end multi-semantic model is proposed to tackle the problem of accurate classification of similar driver behaviors in real-time, which can both characterize driver actions and focus on the key-objects linked with corresponding behaviors;
- The category of Drive&Act in the level of fine-grained activity is adapted to establish the clear relationships between behaviors and key-objects based on hierarchical annotations;
- Experiments implemented on the modified version of the public dataset Drive&Act demonstrate that the MSRNet can recognize 11 different behaviors with 64.18% accuracy and a 20 fps inference time on an 8-frame input clip. Compared to the state-of-the-art action recognition model, our approach obtains higher accuracy, especially for behaviors with similar movements.

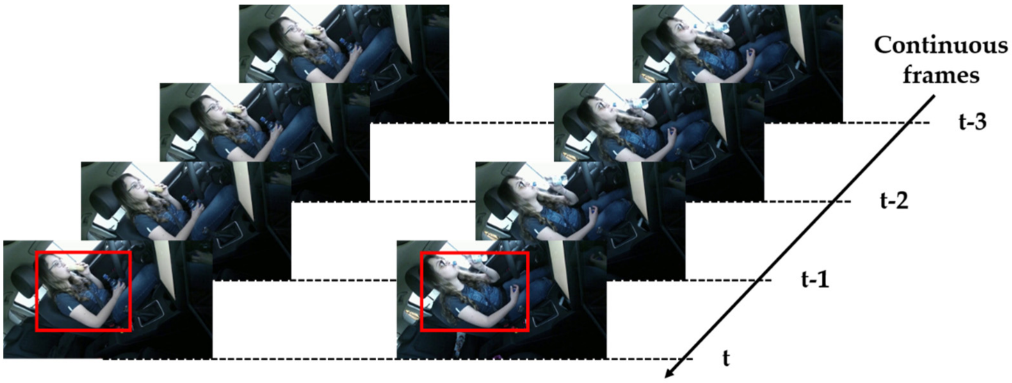


Figure 2. Drinking water or consuming food? Although the region of interest can be effectively obtained, it may not be possible to identify the driver action positively using only action cues. The key-object cues, such as food and bottles, should be integrated to classify which behavior the driver is taking on correctly.

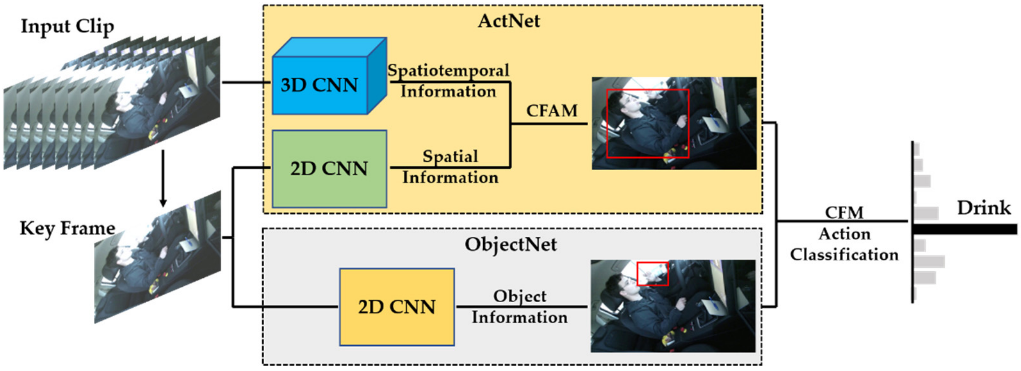


Figure 3. The overall architecture of the proposed model. ActNet is used to extract spatiotemporal features from an input clip and the ObjectNet is used to extract key-object features from the key frame. The predictions from both branches are fed into the CFM to perform confidence fusion and action classification based on the semantic relationships between actions and key objects.

2. Materials and Methods

In this section, the distribution of the fine-grained activity groups in the modified Drive&Act is introduced firstly, in order to facilitate the design, training, and evaluation of the proposed model. Subsequently, an end-to-end model with two parallel branches, called MSRNet, is employed to perform driver behavior recognition. Inspired by the intuition of human vision, the proposed model focuses on both the actions and the objects involved in the actions to derive holistic descriptions of driver behaviors. ActNet is used to extract spatiotemporal features from input clips, which can capture the action cues of driver behaviors. ObjectNet is utilized to extract key-object features from key frames, which mainly concentrates on object cues. The predictions from both branches are merged via the confidence fusion mechanism, based on the semantic relationships between actions and key objects. Overall this ensemble demonstrably improves model accuracy and robustness for driver behavior recognition. Finally, the implementation of MSRNet is described briefly.

2.1. Dataset

In this paper, experiments are conducted on the modified version of the public dataset Drive&Act [14], which collects data on the secondary activities of 15 subjects for 12 h (over 9.6 million frames). Drive&Act provides the hierarchical annotations of 12 classes of coarse tasks, 34 categories of fine-grained activities, and 372 groups of atomic action units. In contrast to the first (coarse task) and the third (atomic action unit) levels, the second level (fine-grained activity) can provide sufficient visual details while maintaining clear semantic descriptions. Therefore, the categories of Drive&Act at the level of fine-grained activity are adapted to establish clear relationships between behaviors and key objects based on hierarchical annotations. First, the classes involved in driving preparation activities (e.g., entering/exiting cars, fastening belts) are excluded due to the fact that the solution only focuses on the secondary activities in the running process of autonomous vehicles. In addition, the integrity of behaviors in the temporal dimension is preserved to simplify the correspondence between actions and key objects. For example, the actions of opening bottles, drinking water, and closing bottles are considered as the different stages of the same action. Finally, the 34 categories of Drive&Act are restructured into 11 classes, including nine semantic relationships between behaviors and key objects. Figure 4 illustrates the distribution of the fine-grained activity groups in the modified dataset.

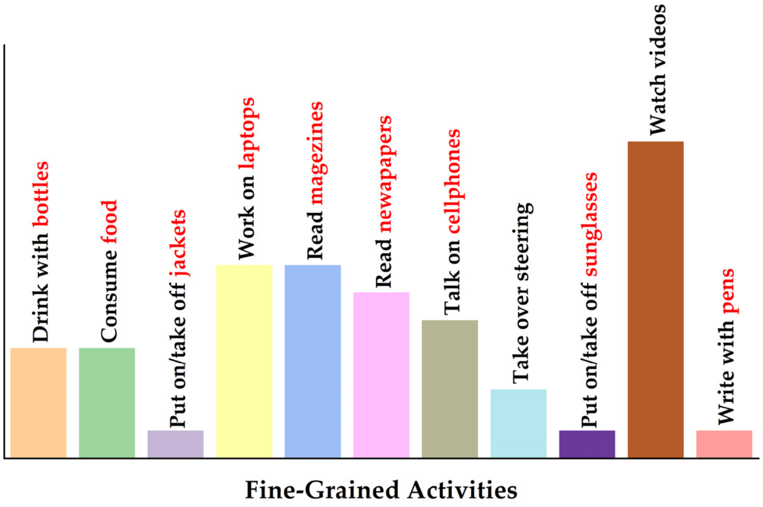


Figure 4. The distribution of the fine-grained activity groups in the modified dataset. The groups are: (1) drink from bottle; (2) consume food; (3) put on or take off jacket; (4) work on laptop; (5) read magazine; (6) read newspaper; (7) talk on cellphone; (8) take over the steering wheel; (9) put on or take off sunglasses; (10) watch videos; (11) write with a pen. The key objects corresponding to actions are colored in red.

2.2. ActNet

Since contextual information is crucial for understanding driver behaviors, the proposed model uses 3D-CNN to extract spatiotemporal features, which is able to capture motion information encoded in multiple consecutive frames. The 3D-CNNs form a cube by stacking multiple consecutive frames, and then apply 3D convolution not only in the space dimension, but also in the time dimension. The feature maps in the convolutional layer are related to the multiple adjacent frames in the upper layer to obtain motion information. YOWO [19] is the state-of-the-art 3D-CNN architecture for real-time spatiotemporal action localization in video streams. In YOWO, a unified network called ActNet is used to obtain the information on driver actions encoded in multiple contiguous frames. ActNet is made up of three major parts. The first part, the 3D branch, extracts spatiotemporal features from an input clip via 3D-CNN. The ResNext-101 is used as the 3D backbone of the 3D branch due to its good performance on kinetics and UCF-101 [20]. The second part, the 2D branch, extracts spatial features from the key frame (i.e., the last frame of an input clip) via 2D-CNN to address the spatial localization issue. Darknet-19 [21] is applied as the 2D backbone of the 2D branch. The concat layer merges the feature maps from the 2D branch and the 3D branch, and feeds them into the third part, the channel fusion and attention mechanism (CFAM), to aggregate the features smoothly from the two branches above.

The prior mechanism proposed in [21] is utilized to bound box regression localization. The final outputs are resized to $[5 \times (11 + 4 + 1) \times H \times W]$, indicating five prior anchors, 11 categories of activities, four coordinates, a confidence score, and the height and width of the images in the grid, respectively. The smooth $L1$ loss [22],

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise,} \end{cases} \tag{1}$$

is adopted to calculate the loss of bounding box regression, where x is the difference in the elements between the bounding box and the groundtruth. The focal loss [23],

$$\text{FL}(p_\text{t}) = -(1 - p_\text{t})^\gamma \log(p_\text{t}), \tag{2}$$

is applied to determine classification loss, where p_t

$$p_\text{t} = \begin{cases} p, & \text{if } y = 1 \\ 1 - p, & \text{otherwise,} \end{cases} \tag{3}$$

is the variation in cross-entropy loss, and $(1 - p_\text{t})^\gamma$ is a modulating factor in cross-entropy loss, with a tunable focusing parameter $\gamma \geq 0$.

2.3. ObjectNet

ActNet is able to capture the action cues of driver behaviors from input clips directly, and provide accurate predictions in most situations. However, driver behaviors may be so subtle or similar that they lead to false positives and missed detections. Therefore, ObjectNet is proposed to capture the key-object cues involved in driver actions, such as bottles for drinking, food for eating, and laptops for working. ObjectNet is expected to further filter the predictions of ActNet in order to classify subtle or similar actions. YOLO-v3 [24] is one of the more popular algorithms used for generic object detection, and is successfully adapted to many recognition problems. YOLO-v3 is employed as the basic framework of ObjectNet due to its excellent trade-off between accuracy and efficiency. In order to enhance the performance to detect small objects, ObjectNet extracts features from multiple scales of the key frame, following the same guideline as the feature pyramid network [25]. In detail, the multi-scale outputs of different detection layers are merged to derive the final predictions using non-maximum suppression.

2.4. Confidence Fusion Mechanism

The outputs of ActNet and ObjectNet are reshaped to the same dimension (i.e., class index, four coordinates, and confidence score). For a specific class, the confidence score for each box is defined as

$$\Pr(\text{Class}_i\,|\text{Object})*\Pr(\text{Object}) * \text{IOU}^{\text{truth}}_{\text{pred}} = \Pr(\text{Class}_i) * \text{IOU}^{\text{truth}}_{\text{pred}}, \tag{4}$$

which reflects both the probability of the class appearing in the box and how well the predicted box fits the object [26]. To utilize the complementary effects of different items of semantic information, the Confidence Fusion Mechanism (CFM) is introduced to aggregate predictions from both ActNet and ObjectNet based on the semantic relationships between actions and key-objects. The CFM is a decision fusion approach that combines the decisions of multiple classifiers into a common decision about driver behavior. This grounds independence from the type of data source, making it possible to aggregate the information derived from different semantic aspects.

In order to illustrate the implications of the CFM, we consider a simple scenario: there are two binary classifiers (S1 and S2) used to detect whether drivers are drinking water or not. It performs one detection using S1 and S2, and there will be four possible situations, as shown in Table 1. If the results of S1 and S2 are in agreement, it is reasonable to conclude on whether drivers are drinking water or not. Otherwise, the results of the classifier with greater confidence will be preferably accepted.

Table 1. The possible situations of driver drinking detection by two binary classifiers.

Possible Situations	S1	S2
(0, 0)	not_drink (0)	not_drink (0)
(0, 1)	not_drink (0)	drink (1)
(1, 0)	drink (1)	not_drink (0)
(1, 1)	drink (1)	drink (1)

Expanding the simple scenario to our task, ActNet performs driver behavior recognition on a given clip, and outputs N predictions. In general, we can conclude which actions drivers engage in by reference to the maximum confidence score. Figure 5 illustrates the algorithm flowchart of the CFM. First, the N predictions are sorted in order of confidence scores from largest to smallest. Afterwards, the top m predictions are fed into the decision in turn to examine whether they match with the correspondences between actions and key-objects. In this paper, we set m as 3, because the confidence scores of these predictions are generally lower than the threshold when m is beyond 3. If the prediction (i) is compatible with the key-object detected by ObjectNet, it is assumed that the prediction (i) is accurate, and the circulation is ended. Otherwise, this process will continue until all the top m predictions have been examined. In addition, there is a possible situation wherein none of the top m predictions match with the key-object. In this case, the original results of ActNet will be adopted.

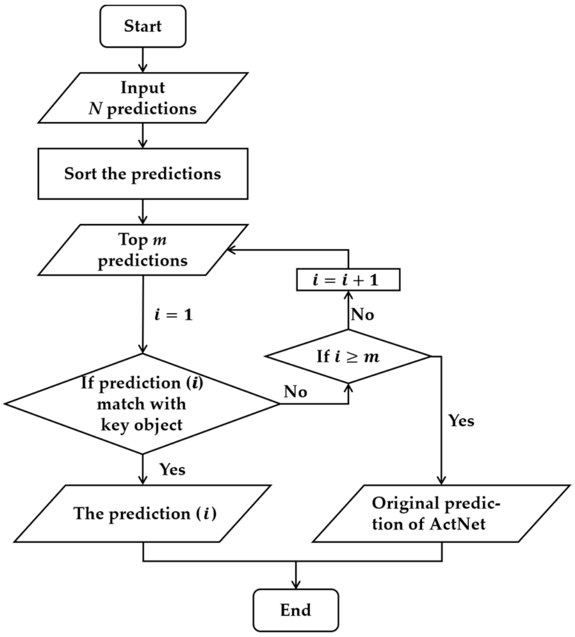


Figure 5. The algorithm flowchart of the confidence fusion mechanism.

2.5. Implementation Details

The publicly released YOLO-v3 [24] model is used for ObjectNet and is fine-tuned on the modified Drive&Act [14] following default configuration. For ActNet, the parameters of the 3D backbone and the 2D backbone are initialized on kinetics [27] and COCO [28], respectively. The training is implemented using stochastic gradient descent with an initial learning rate of 0.0001, which is degraded with a modulating factor of 0.5 after the 30 k, 40 k, 50 k, and 60 k iterations. The weight decay rate is set to 0.0005, and the momentum value is set to 0.9. For the dataset Drive&Act, the training process is converged after five epochs. Both ActNet and ObjectNet are trained and tested using a Tesla V100 GPU with 16 GB RAM. The proposed model is carried out end-to-end in PyTorch.

3. Results and Discussion

In this section, the accuracies of the MSRNet and YOWO are compared to illustrate the improvement in driver behavior recognition by aggregating multi-semantic information. Afterwards, the visualization of the output from different branches is used to determine what is learned by the MSRNet. Finally, some limitations that affect the MSRNet's performance are discussed.

Experiments are implemented on the modified public dataset Drive&Act. As in [14], the datasets for training, validation, and testing are randomly divided based on the identity of subjects; using videos, we assign the data of 10 persons for training, 2 persons for validation, and 3 persons for testing. Each action segment is spilt into 3-s chunks for balancing the various durations of driver behaviors. The standard evaluation metric of accuracy is adopted to measure the performance of the proposed dataset. Table 2 reports the results derived from comparing the accuracy between MSRNet and the state-of-the-art action recognition model YOWO [19]. It is observed that MSRNet performs better in both validation and testing, with significant 4.65% (Val) and 3.16% (Test) improvements in accuracy when recognizing 11 different behaviors on an 8-frame input clip.

Table 2. The results of comparing the accuracy between MSRNet and YOWO.

Model	Val (%)	Test (%)
YOWO	67.71	61.02
Our Model	72.36	64.18

Figure 6 illustrates the activation maps giving a visual explanation of the classification decision made by ActNet and ObjectNet [29]. It can be observed that ActNet mainly focuses on the areas where movements are happening, whereas ObjectNet mainly focuses on the key-objects. Figure 7 gives a precise description of 11 fine-grained activities carried out on the modified Drive&Act by the confusion matrixes. Each row of the confusion matrix represents the instances in an actual label, while each column represents the instances in a predicted label. As can be seen from the confusion matrixes, the proposed model accurately recognizes the majority of classes, with 99% accurate identification of drinking with bottles, 95% accurate identification of working on laptops, and 94% accurate identification of reading magazines. In addition, a significant improvement is made in recognizing similar actions. For example, 16% (drinking with bottles vs. consuming food) and 14% (reading magazines vs. reading newspaper) of the misrecognitions are avoided when using the MSRNet. Our experiments demonstrate the effectiveness of utilizing multi-semantic classification for driver recognition with the confidence fusion mechanism. Although the proposed model shows superiority in solving the problem of interclass similarity, it also suffers from some limitations that degrade its performance. Figure 8 illustrates examples of images for which the MSRNet fails in driver behavior recognition. It is observed that the misrecognition of the proposed model is mainly caused by some challenging situations in Drive&Act, such as occlusion and multi-class visibility.

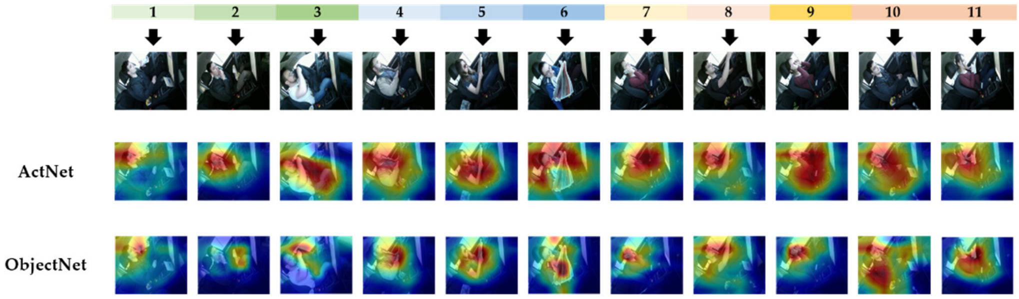
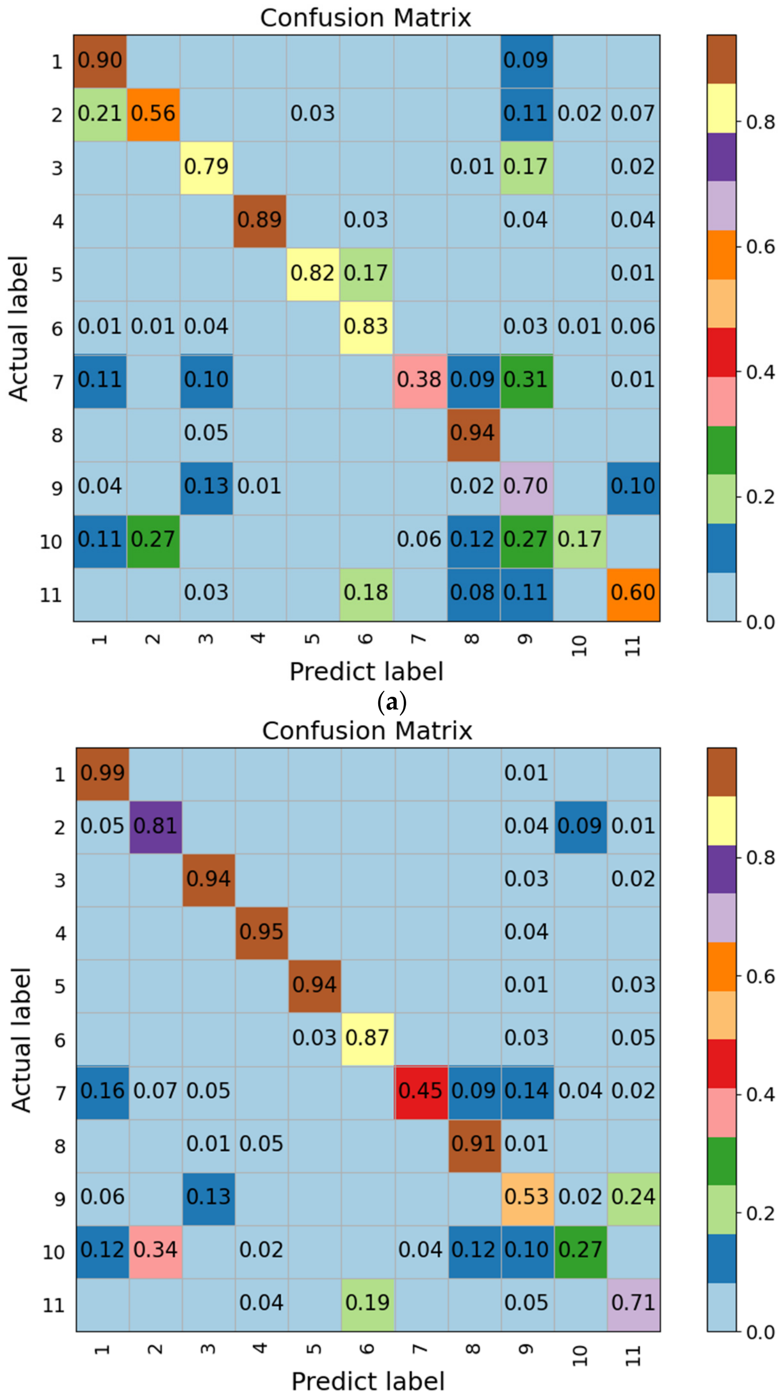


Figure 6. The activation maps giving a visual explanation of the classification decision made by ActNet and ObjectNet.



	1	2	3	4	5	6	7	8	9	10	11
1	0.90								0.09		
2	0.21	0.56			0.03				0.11	0.02	0.07
3			0.79					0.01	0.17		0.02
4				0.89		0.03			0.04		0.04
5					0.82	0.17					0.01
6	0.01	0.01	0.04			0.83			0.03	0.01	0.06
7	0.11		0.10				0.38	0.09	0.31		0.01
8			0.05					0.94			
9	0.04		0.13	0.01				0.02	0.70		0.10
10	0.11	0.27					0.06	0.12	0.27	0.17	
11			0.03			0.18		0.08	0.11		0.60

(**a**)

	1	2	3	4	5	6	7	8	9	10	11
1	0.99								0.01		
2	0.05	0.81							0.04	0.09	0.01
3			0.94						0.03		0.02
4				0.95					0.04		
5					0.94				0.01		0.03
6					0.03	0.87			0.03		0.05
7	0.16	0.07	0.05				0.45	0.09	0.14	0.04	0.02
8			0.01	0.05				0.91	0.01		
9	0.06		0.13						0.53	0.02	0.24
10	0.12	0.34		0.02			0.04	0.12	0.10	0.27	
11				0.04		0.19			0.05		0.71

(**b**)

Figure 7. The confusion matrixes for YOWO (**a**) and MSRNet (**b**). The indexes of rows and columns from 1 to 11 represent: (1) drink from a bottle; (2) consuming food; (3) putting on or taking off a jacket; (4) working on a laptop; (5) reading a magazine; (6) reading a newspaper; (7) talking on a cellphone; (8) taking over the steering wheel; (9) putting on or taking off sunglasses; (10) watching videos; (11) writing with a pen.

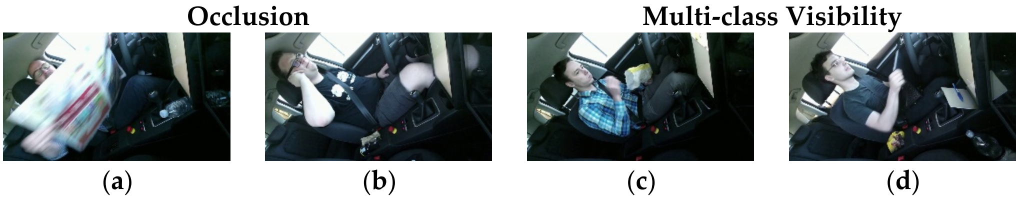


Figure 8. The examples of driver images for which MSRNet fails driver behavior recognition. The challenging situations are: (**a**) the newspaper covers the driver's upper body; (**b**) the cellphone is completely covered by the driver's hand; (**c**) the driver is consuming food while watching a video; (**d**) a bottle, a pen and food are all visible.

4. Conclusions

In this paper, an end-to-end multi-semantic model is proposed for driver behavior recognition, employing a confidence fusion mechanism known as MSRNet. First, the category of Drive&Act at the level of fine-grained activity is adapted to establish the clear relationships between behaviors and key-objects based on hierarchical annotations. This modification facilitates the design, training, and evaluation of the proposed model. Subsequently, MSRNet uses two parallel branches to perform action classification and object classification, respectively. ActNet mainly focuses on areas wherein movements are happening, whereas ObjectNet mainly focuses on key objects. The proposed confidence fusion mechanism aggregates the predictions from both branches based on the semantic relationships between actions and key-objects. The proposed approach can both characterize driver actions and focus on the key-objects linked with behaviors to obtain more complete descriptions of behaviors. Overall, this approach demonstrably improves the model's accuracy and robustness for driver behavior recognition. The experiments have demonstrated that the MSRNet performs better in terms of both validation and testing, with significant 4.65% (Val) and 3.16% (Test) improvements in accuracy when recognizing 11 different behaviors in an 8-frame input clip. The proposed model can perform accurate recognition for the majority of classes, such as 99% accurate identification of drinking from a bottle, 95% accurate identification of working on a laptop, and 94% accurate identification of reading a magazine.

Although the MSRNet shows superiority in solving the problem of interclass similarity, it also suffers from some limitations (e.g., occlusion and multi-class visibility) that degrade its performance. In future work, we would like to try other possible approaches to solving these limitations. As feature extraction from occluded human body parts is rarely possible, it is important to find robust classifiers that can handle the occlusion problem, such as probabilistic approaches. In addition, collecting additional sensor data (e.g., body pose, depth, and infrared) from other sensors mounted on real cars is a potential mitigation strategy. It is considered that this could help in deriving more complete descriptions of driver behavior.

Author Contributions: Conceptualization, H.R. and Z.B.; methodology, H.R. and Y.G.; software, H.R. and Y.G.; validation, H.R. and X.C.; formal analysis, H.R.; resources, H.R. and Z.B.; data curation, H.R.; writing—original draft preparation, H.R.; writing—review and editing, Y.G.; visualization, H.R. and X.C.; supervision, Y.G. and Z.B.; project administration, Z.B.; funding acquisition, Z.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Foundation of Hunan Province, grant number 2020JJ4184.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Detailed data are contained within the article. More data that support the findings of this study are available from the author R.H. upon reasonable request.

Acknowledgments: The authors appreciate the reviewers and editors for their helpful comments and suggestions in this study.

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

References

1. Singh, S. *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*; National Highway Traffic Safety Administration: Washington, DC, USA, 2015.
2. Dingus, T.A.; Guo, F.; Lee, S.; Antin, J.F.; Perez, M.; Buchanan-King, M.; Hankey, J. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 2636–2641. [CrossRef] [PubMed]
3. Dindorf, R.; Wos, P. Analysis of the Possibilities of Using a Driver's Brain Activity to Pneumatically Actuate a Secondary Foot Brake Pedal. *Actuators* **2020**, *9*, 49. [CrossRef]
4. Naujoks, F.; Purucker, C.; Neukum, A. Secondary task engagement and vehicle automation–Comparing the effects of different automation levels in an on-road experiment. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *38*, 67–82. [CrossRef]
5. International, S. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*; SAE; SAE International: Washington, DC, USA, 2018.
6. Ragab, A.; Craye, C.; Kamel, M.S.; Karray, F. A Visual-Based Driver Distraction Recognition and Detection Using Random Forest. In *Proceedings of the Transactions on Petri Nets and Other Models of Concurrency XV*; Springer Science and Business Media LLC: Berlin/Heidelberg, Germany, 2014; pp. 256–265.
7. Seshadri, K.; Juefei-Xu, F.; Pal, D.K.; Savvides, M.; Thor, C.P. Driver cell phone usage detection on Strategic Highway Research Program (SHRP2) face view videos. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Columbus, OH, USA, 23–28 June 2014; IEEE: New York, NY, USA, 2015; pp. 35–43.
8. Liu, T.; Yang, Y.; Huang, G.-B.; Yeo, Y.K.; Lin, Z. Driver distraction detection using semi-supervised machine learning. *Ieee Trans. Intell. Transp. Syst.* **2015**, *17*, 1108–1120. [CrossRef]
9. Le, T.H.N.; Zheng, Y.; Zhu, C.; Luu, K.; Savvides, M. Multiple Scale Faster-RCNN Approach to Driver's Cell-Phone Usage and Hands on Steering Wheel Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2016; pp. 46–53.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
11. Streiffer, C.; Raghavendra, R.; Benson, T.; Srivatsa, M. Darnet: A deep learning solution for distracted driving detection. In Proceedings of the 18th Acm/Ifip/Usenix Middleware Conference: Industrial Track, Las Vegas, NV, USA, 11–15 December 2017; pp. 22–28.
12. Baheti, B.; Gajre, S.; Talbar, S. 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, UT, USA, 18–22 June 2018; pp. 1145–1151. [CrossRef]
13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
14. Martin, M.; Roitberg, A.; Haurilet, M.; Horne, M.; ReiB, S.; Voit, M.; Stiefelhagen, R. Drive&Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2019; pp. 2801–2810.
15. Reis, S.; Roitberg, A.; Haurilet, M.; Stiefelhagen, R. Activity-aware Attributes for Zero-Shot Driver Behavior Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June2020; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2020; pp. 3950–3955.
16. Zhang, C.; Li, R.; Kim, W.; Yoon, D.; Patras, P. Driver behavior recognition via interwoven deep convolutional neural nets with multi-stream inputs. *Ieee Access* **2020**, *8*, 191138–191151. [CrossRef]
17. Wang, H.; Zhao, M.; Beurier, G.; Wang, X. Automobile driver posture monitoring systems: A review. *China J. Highw. Transp* **2019**, *2*, 1–18.
18. Wharton, Z.; Behera, A.; Liu, Y.; Bessis, N. Coarse Temporal Attention Network (CTA-Net) for Driver's Activity Recognition. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), New York, NY, USA, 5–9 January2021; IEEE: New York, NY, USA, 2021; pp. 1278–1288.
19. Köpüklü, O.; Wei, X.; Rigoll, G.J.A.P.A. You Only Watch Once: A Unified Cnn Architecture for Real-Time Spatiotemporal Action Localization. *ArXiv* **2019**, arXiv:1911.06644.
20. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8–23 June 2018; IEEE: New York, NY, USA, 2018; pp. 6546–6555.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: New, York, NY, USA, 2017; pp. 6517–6525.
22. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]

23. Lin, T.-Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
24. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
26. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 7–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 779–788.
27. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
28. Tsung-Yi, L.; Maire, M.; Belonge, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 740–755.
29. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.