Light-weight Convolutional Neural Network for Distracted Driver Classification

Duy-Linh Nguyen, Muhamad Dwisnanto Putro, Xuan-Thuy Vo, and Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering,
University of Ulsan,
Ulsan, Korea
ndlinh301@mail.ulsan.ac.kr, dwisnantoputro@mail.ulsan.ac.kr, xthuy@islab.ulsan.ac.kr, and acejo@ulsan.ac.kr

IECON 2021 - 47th Annual Conference of the IEEE Industrial Electronics Society | 978-1-6654-3554-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IECON48115.2021.9589212

***Abstract*—Driving is an activity that requires the coordination of many senses with complex manipulations. However, the driver can be affected by a several factors such as using a mobile phone, adjusting audio equipment, smoking, drinking, eating, talking to a passenger or drowsy. Therefore, the development of assistant applications to warn distracted driver is very necessary. Because of the limited space and mobility, the equipment also requires compact, energy-saving and efficient. This paper proposes a light-weight Convolutional Neural Network for a distracted driver warning system. The method is built based on a combination of standard convolution and Depthwise Separable Convolution operation to optimize the network parameters but still ensure the important information and speed. The network was trained and evaluated on two datasets, AUC (the American University in Cairo) and StateFarm dataset from Kaggle's competition. As a result, the evaluation accuracy reached 95.36% and 99.95%, respectively.**

***Index Terms*—Assistant application, Convolutional Neural Network (CNN), Global Average Pooling, Depthwise Separable Convolution, Distracted Driver Classification, Driver warning system.**

I. Introduction

Traffic accidents are becoming common and serious all over the world. Among them, road traffic accidents account for the majority of accidents. According to a report of the World Health Organization (WHO), every year more than 1.35 million people die from road traffic accidents and 90% are caused by drivers [1]. The main cause from the driver is still the distraction created by the influence of many factors related to the driver's vehicle operation. A few seconds of distraction can have dire consequences. There are different definitions of distracted driving. According to [2], [3], distracted driving is any action by a driver that strays away from the driving task. The article in [4] divide distracted driving into three main categories: manual, visual and cognitive distractions. Manual distractions include actions that are not related to driving, such as drinking, eating, smoking, and using mobile phones. Visual distractions appear when the driver does not keep his eyes on the road while driving. Cognitive distractions involve the driver's mind when not fully focused on driving when talking to passenger or drowsing. These are the main possible causes of traffic accidents and collisions. Therefore, for early warning of distraction while driving, it becomes necessary to develop assistant applications for drivers to reduce risks. With the development of GPS (Global Positioning System) technology, sensors and mobile devices, safe driver assistant applications have been integrated on modern vehicles. However, it requires expensive assembly costs and it is difficult to implement in older vehicles. Technology companies have also developed self-driving cars with complex devices, but they are only in the experimental stage and they have not been widely applied in practice. In addition, wearable devices using sensors are also widely developed to monitor the physiological status of the driver. These devices are uncomfortable for the driver and it is subject to some natural phenomena in the human body that can cause signal interference. Since, these applications are still limited. The method used in this paper focuses on driver behavior recognition based on light-weight and efficient Convolutional Neural Network. This network is built on the advantages of standard convolution, Depthwise Separable Convolution operation, Residual connection and the replacement of fully connected layers by the Global Average Pooling layer to optimize network parameters, increasing applicability on low-computing device like CPU device.

The main contributions of this paper are as follows:
1. Proposed a light-weight and efficient Convolutional Neural Network architecture for distracted driver recognition.
2. Developed the distracted driver classifier which can run on low-computing device without ignoring the accuracy.

The rest of the paper is organized as follows: Section II presents the previous technologies relative to driver behavior monitoring systems. Section III explains the detail of proposed methodology. Section IV describes and analyzes experiment results. And, Section V concludes the paper and introduce several future works.

II. Related Work

This section will summary several methodologies implement in driver behavior monitoring systems. These methodologies can be divided into Machine learning and CNN-Based methodologies.

A. Machine learning methodologies

The first study of distracted driving was the use of a cellphone while driving. The study in [5] considers the position between cellphone, hands, face, and mouth using Hidden

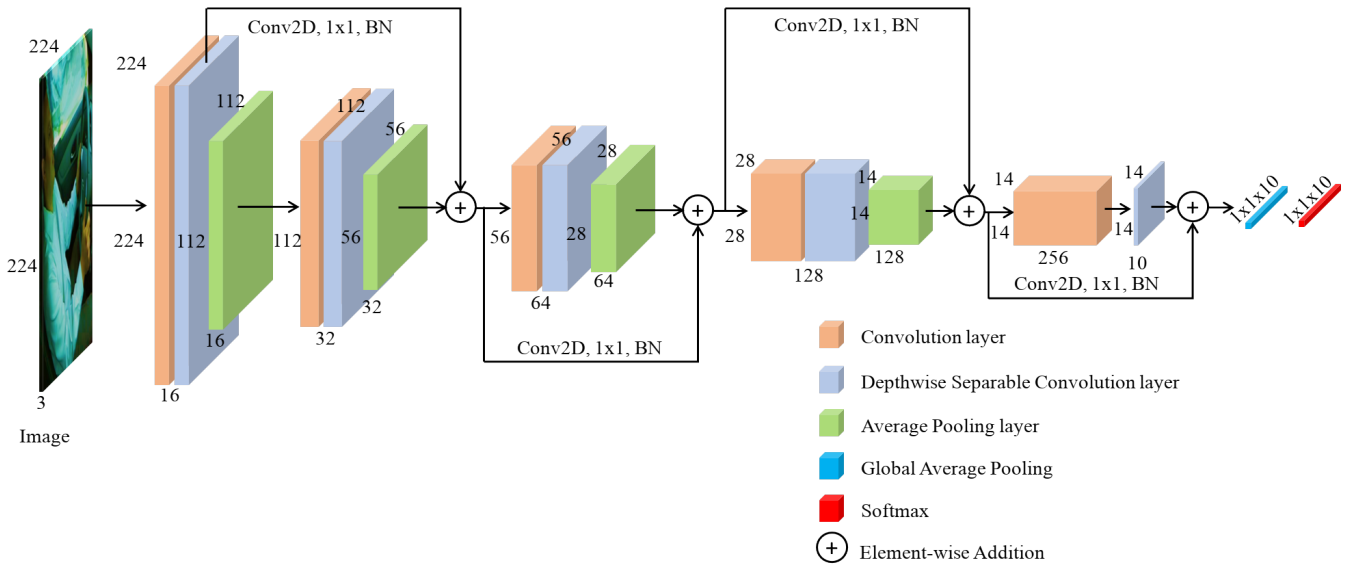

Fig. 1. The proposed distracted driver classification network. It consists of three modules: Stem, Residual connection, and classification module.

Conditional Random Fields (Hidden CRF). The authors in [6] proposes a method based on Histogram of Oriented Gradients (HOG), Supervised Desent Method and Adaboost Classifier to detect the driver's cellphone usage. Another study focused on driver hand detection with Aggregate Channel Features (ACF) method [7]. Besides, many studies also exploit facial, head and eye features to recognize distracted driving by manual feature extraction methods such as Histogram of oriented Gradients (HoG), Haar-like, Local Binary Pattern (LBP) with Support Vector Machines (SVM), Adaboost classifier. In general, these methods handle with several distracted driver with low accuracy.

B. CNN-Based methodologies

Along with the widespread application of Convolutional Neural Network architectures in the field of computer vision, distracted driver recognition and warning has been simplified with enhanced accuracy. In order to detect hand movements on the steering wheel and cellphone use, Le *et al.* [8] used Faster-RCNN network architecture as detectors. The image segmentation method is also applied in [9] to separate the image area into the steering wheel, dashboard, and gear lever, then propose the network to detect the corresponding hand area on the segmented image areas. For more distracted driver behaviors, [10] and [11] proposed a large dataset for ten distracted driver behaviors. Since then, the researchers have focused on developing Convolutional Neural Networks to detect such behaviors with high accuracy, suitable for deployment in all types of cars. Also from [10], the authors used the ensemble training method with five different Convolutional Neural Networks. However, this method produces a heavy weight that makes it difficult to implement in real-time applications. Many other works also develop distracted driver recognizers based on famous CNN architectures such as VGG [12], [13], DenseNet [14], GoogleNet [15]. Nowadays, the design of light-weight network architectures suitable for low-computation and edge devices attracts much attention like SqueezeNet [16], MobileNet [17] and its variants, NASNetMobile [18]. Inspired by the advantages of Depthwise Separable Convolution operation [17] and Residual Network [19], this paper proposes a light-weight Convolutional Neural Network with just under five hundred thousand parameters but with an accuracy comparable to another state-of-the-art network in this field.

III. Proposed Methodology

A. Proposed Architecture Network

The network architecture proposed in this paper can be separated into three modules: Stem, Residual connection, and Classification module. The Stem module of the network is used as a multi-level feature extractor. This module is designed by four main convolution blocks and two convolution layers. Each main block consists of a Standard Convolution, a Depthwise Separable Convolution layer with kernel sizes varying from 7×7, 5×5 to 3×3, and a Average Pooling layer. The use of large kernels at the beginning of the processing increases the receptive area and captures the best information from the image and serves as the input of the first Residual connection. Then, the kernel size is gradually reduced to 3×3 in the next two main blocks and in the final Standard Convolution and Depthwise Separable Convolution layer. Large kernels increase network parameters, so Depthwise Separable Convolution interleaved into the extractor is an effective solution. With computational flexibility compared to Standard Convolution, Depthwise Separable Convolution can optimize network parameters and increase application ability in low-computational devices. From the $224 \times 224 \times 3$ input image, a $14 \times 14 \times 10$ feature map is obtained after go through extractor. The number of channels corresponds to the number of classes

in the dataset, here using 10 classes. Inspired by the unique feature of Residual connection which was invented in ResNet architecture, this network uses four Residual connections with different feature map level of $56 \times 56 \times 32$, $28 \times 28 \times 64$, $14 \times 14 \times 128$, and $14 \times 14 \times 10$. When going deep into the network, through different levels, the obtained feature map will reduce a lot of important information. Therefore, combining information from feature maps at previous levels will maintain and enrich the information at the current level. In particular, the high-level feature map gets information from the lower-level by the Element-wise Addition operation. Such sequence with different levels makes the whole network ensure information extraction from beginning to end. In order to create the Residual connection, this work apply Standard Convolution operation with kernel size of 1×1 and number of channel adaptive with each feature map level follow by a Batch Normalization technique. The working principle of Residual connection is shown in Fig. 2.

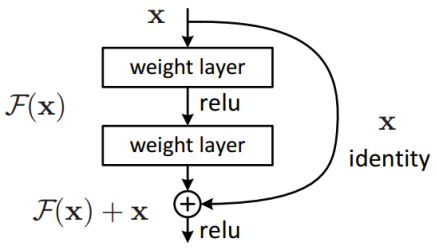

Fig. 2. Single Residual connection [19].

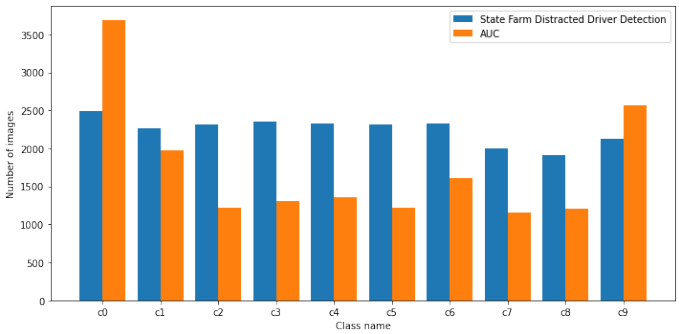

Fig. 3. Dataset distribution.

In the popular Convolutional Neural Network architectures for image classification, the authors will use the Fully connected layer after feature extraction along with the Softmax function to calculate the occurrence probability of the classes. However, this study completely replaces the fully connected layers with only one Global Average Pooling layer which significantly reduces the network parameters. The feature map of $14 \times 14 \times 10$ is reshaped to $1 \times 1 \times 10$ and then the Softmax function is also applied in classification module. Fig. 1 and TABLE I detail the architecture of the proposed network, where Conv2D: Standard convolution, AvgPool: Average Pooling, SepConv2D: Depthwise Separable Convolution, Add: Element-wise Addition, BN: Batch Normalization, ReLU: ReLU activation function.

TABLE I
THE DETAILS THE ARCHITECTURE OF THE PROPOSED NETWORK.

Layer (type)	Output Shape
Conv2D-1, $7 \times 7 \times 16$, BN, ReLU	$224 \times 224 \times 16$
Conv2D-2, $7 \times 7 \times 16$, BN, ReLU	$224 \times 224 \times 16$
AvgPool-1, 2×2	$112 \times 112 \times 16$
Conv2D-4, $5 \times 5 \times 16$, BN	$112 \times 112 \times 32$
SepConv2D-1, $5 \times 5 \times 16$, BN, ReLU	$112 \times 112 \times 32$
Conv2D-3, $1 \times 1 \times 32$, BN (Residual-1)	$56 \times 56 \times 32$
AvgPool-2, 2×2	$56 \times 56 \times 32$
Add-1	$56 \times 56 \times 32$
Conv2D-6, $3 \times 3 \times 64$, BN	$56 \times 56 \times 64$
SepConv2D-2, $3 \times 3 \times 64$, BN, ReLU	$56 \times 56 \times 64$
Conv2D-5, $1 \times 1\times$, BN (Residual-2)	$28 \times 28 \times 64$
AvgPool-3, 2×2	$28 \times 28 \times 64$
Add-2	$28 \times 28 \times 64$
Conv2D-8, $3 \times 3 \times 128$, BN	$28 \times 28 \times 128$
SepConv2D-3, $3 \times 3 \times 64$, BN, ReLU	$28 \times 28 \times 128$
Conv2D-7, $1 \times 1\times$, BN (Residual-3)	$14 \times 14 \times 128$
AvgPool-4, 2×2	$14 \times 14 \times 128$
Add-3	$14 \times 14 \times 128$
Conv2D-10, $3 \times 3 \times 256$, BN	$14 \times 14 \times 256$
SepConv2D-4, $3 \times 3 \times 64$, BN	$14 \times 14 \times 10$
Conv2D-9, 1×1, BN (Residual-4)	$14 \times 14 \times 10$
Add-4	$14 \times 14 \times 10$
GlobalAveragePooling	$1 \times 1 \times 10$
Softmax	$1 \times 1 \times 10$

B. Loss function

The proposed network uses Categorical cross-entropy loss function to calculate loss during the training phase. This loss function is shown as follows:

$$L_{cls} = -\sum_{c=1}^{10} t_c.log(p_c), \tag{1}$$

where c is number of classes (in this case $c = 10$), t is target indicator ($t = 0$ or $t = 1$), log is natural logarithm function, and p is predicted probability.

IV. EXPERIMENTS

A. Dataset preparation

This classifier was trained and evaluated on two datasets: State Farm Distracted Driver Detection (StateFarm) [10] and American University in Cairo (AUC) [11] dataset. The StateFarm dataset comes from a competition on Kaggle's website and it contains 22,424 color images with a resolution of 640×480 pixel. It is divided into ten classes corresponding to directories marked from c0 to c9. The ten classes in the StateFarm dataset are safe driving (c0), texting - right (c1), talking on the phone - right (c2), texting - left (c3), talking on the phone - left (c4), operating the radio (c5), drinking (c6), reaching behind (c7), hair and makeup (c8), talking to passenger (c9). The American University in Cairo dataset was recorded by the SUS ZenPhone (Model Z00UD) rear camera

Fig. 4. The qualitative result of network on StateFarm and AUC evaluation set. The first two rows for the StateFarm Dataset and the following two rows for the AUC Dataset

with 31 participants from seven countries including 22 males and 9 females. The videos were processed and selected 17,308 single images of 1080×1920 pixel resolution. The number of classes in the AUC dataset is also divided and labeled similarly to the StateFarm dataset. The distribution of images in both datasets is described in Fig. 3. Follow another experiment and properly compare the performance of the proposed classifier, this work divides the datasets into 75% for training and 25% for evaluation phase.

B. Experimental setup

The entire classifier network is trained and evaluated on a GeForce GTX 1080Ti GPU. In the testing phase on real-time video, this experiment also reuses this GPU and another CPU Intel Core I7-4770 CPU @ 3.40 GHz, 8GB of RAM (Personal Computer). During training, the network is set up with several basic parameters such as Adam optimization method, learning rate is 10^{-4}, batch size of 16 and epochs of 200. The input images is resize to 244×224 pixel in training and evaluating process.

C. Experimental results and analysic

This experiment was performed in two phases. Training and evaluation phase on two datasets with configurations mentioned above. Then the obtained results are tested on the video which recorded a driver behaviors. As a result, the network achieved 99.95% of accuracy on the StateFarm and 95.36% on the AUC dataset with a network parameter of only approximately five hundred thousand. This result is comparable to popular classifier networks for low-computational devices and state-of-the-art networks in the same field. TABLE II shows the comparison results of accuracy between the proposed network and the state-of-the-art-network on StateFarm and AUC dataset.

The qualitative result of network on StateFarm and AUC evaluation set shows as in Fig. 4. On the other hand, the video testing results with resolution of 640×480 pixel show that it can reach speed up to 26.99 FPS (Frames per second) and 12.50 FPS on GPU and CPU, respectively. All factors of high accuracy, compactness, and high speed prove that this classifier can be deployed on low-computational and edge devices. The confusion matrix in Fig. 5 shows predictive ability over ten classes of the proposed network. In which, the prediction rate of the classes is relatively equal for both datasets but for AUC Dataset the behaviors unrelated to other objects such as ”safe driving”, ”reaching behind”, and ”hair and makeup”, the prediction rate is weaker (93% to 94% of accuracy). For StateFarm dataset, the prediction rate of each class is almost absolute. If observing this result, it can be seen that the proposed method achieves an accuracy on the AUC dataset of 95.36%, it is almost equal to the modified VGG models [12]–[14] but with only nearly five hundred thousand network parameters. In addition, it achieves superior accuracy compared to current mobile architectures such as MobileNet,

TABLE II
THE COMPARISION RESULTS WITH STATE-OF-THE-ART METHODS ON STATEFARM AND AUC DATASET.

Model	Number of parameters	Accuracy (%)	Dataset
MobileNet	4.2	64.2	StateFarm
MobileNet V2	3.5	34.7	StateFarm
SqueezeNet	0.26	11.18	StateFarm
Mobile VGG [13]	2.2	99.75	StateFarm
Proposed network	**0.46**	**99.95**	StateFarm
AlexNet [10]	62	93.65	AUC
Inception V3 [10]	24	95.17	AUC
Majority Voting ensemble [10]	120	95.77	AUC
GA weighted ensemble [10]	120	95.98	AUC
Original VGG [12]	140	94.44	AUC
VGG with Regularization [12]	140	96.31	AUC
Modified VGG [12]	15	95.54	AUC
DenseNet+Latent Pose [12]	8.06	94.20	AUC
MobileNet [13]	4.2	94.67	AUC
MobileNet V2 [13]	3.5	94.74	AUC
NasNet Mobile [13]	5.3	94.69	AUC
SqueezeNet [13]	1.25	93.21	AUC
Mobile VGG [13]	2.2	95.24	AUC
Proposed network	**0.46**	**95.36**	AUC

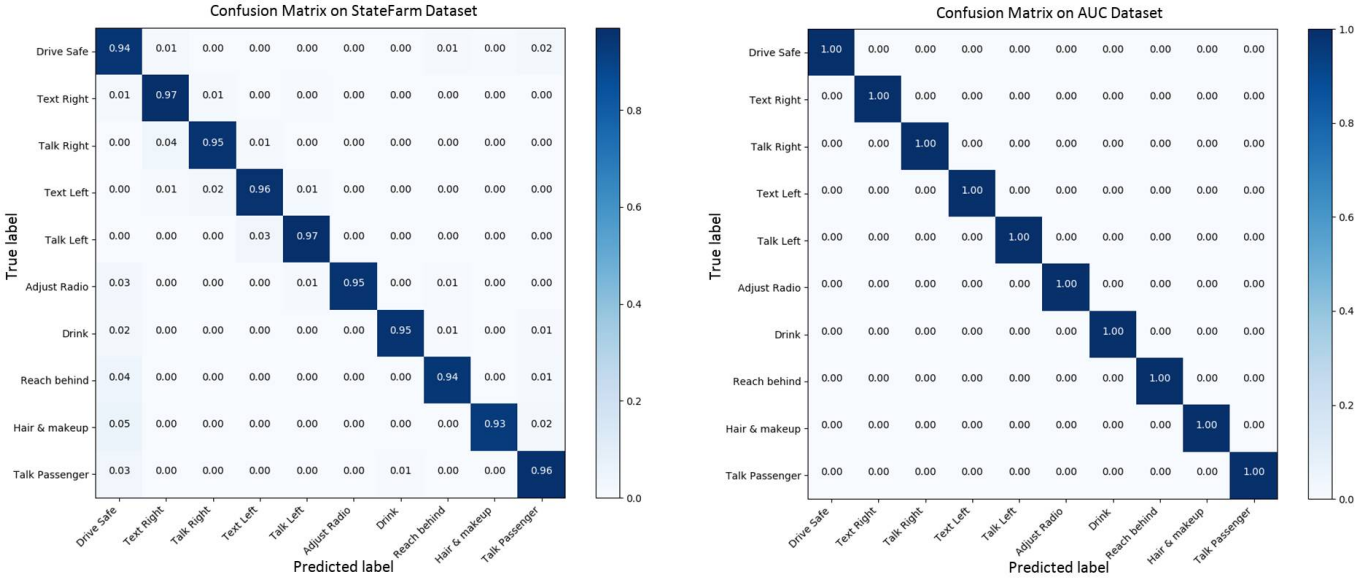

Fig. 5. Confusion Matrix.

NasNet Mobile, SqueezeNet [13]. For the StateFarm dataset, this work also trains the mobile network architectures from scratch for comparison, and as a result, the proposed method has almost perfect accuracy with 99.95% and outperforms all.

V. CONCLUSION AND FUTURE WORK

This paper has proposed the distracted driver classifier with a light-weight Convolutional Neural Network. This network is built on the advantages of Convolution and Depthwise Separable Convolution operation to extract important features. Along with the use of Residual connections to maintain and enrich those features. Besides, replacing Fully Connected layers by Global Average Pooling layer reduces the network parameters significantly. In addition, this work also applies optimization methods in the training process. All of the above factors make up the compactness and efficiency of the proposed network. In the future, this work continues to develop this work with a human body detector to increase the ability to classify distracted driver behaviors when deployed in a real-time system.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the government(MSIT).(No.2020R1A2C200897212)

REFERENCES

[1] "New who report highlights insufficient progress to tackle lack of safety on the world's roads." https://www.who.int/news/item/07-12-2018-new-who-report-highlights-insufficient-progress-to-tackle-lack-of-safety-on-the-world's-roads. Accessed: 2021-08-03.

[2] M. Regan, C. Hallett, and C. Gordon, "Driver distraction and driver inattention: definition, relationship and taxonomy.," *Accident; analysis and prevention*, vol. 43 5, pp. 1771–81, 2011.

[3] S. Kaplan, M. A. Guvensan, A. G. Yavuz, and Y. Karalurt, "Driver behavior analysis for safe driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3017–3032, 2015.

[4] "Distracted driving." https://www.cdc.gov/motorvehiclesafety/distracted-driving/. Accessed: 2021-08-03.

[5] X. Zhang, N. Zheng, F. Wang, and Y. He, "Visual recognition of driver hand-held cell phone use based on hidden crf," in *Proceedings of 2011 IEEE International Conference on Vehicular Electronics and Safety*, pp. 248–251, 2011.

[6] K. Seshadri, F. Juefei-Xu, D. K. Pal, M. Savvides, and C. P. Thor, "Driver cell phone usage detection on strategic highway research program (shrp2) face view videos," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 35–43, 2015.

[7] N. Das, E. Ohn-Bar, and M. M. Trivedi, "On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pp. 2953–2958, 2015.

[8] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides, "Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 46–53, 2016.

[9] E. Ohn-Bar, "Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies," *Journal of Electronic Imaging*, vol. 22, pp. 1119–, 10 2013.

[10] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," *CoRR*, vol. abs/1706.09498, 2017.

[11] "State farm distracted driver detection." https://www.kaggle.com/c/state-farm-distracted-driver-detection/data. Accessed: 2021-07-14.

[12] B. Baheti, S. Gajre, and S. Talbar, "Detection of distracted driver using convolutional neural network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1145–11456, 2018.

[13] B. Baheti, S. Talbar, and S. Gajre, "Towards computationally efficient and realtime distracted driver detection with mobilevgg network," *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 4, pp. 565–574, 2020.

[14] A. Behera and A. H. Keidel, "Latent body-pose guided densenet for recognizing driver's fine-grained secondary activities," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, 2018.

[15] W. Sheng, D. Tran, H. Do, h. Bai, and G. Chowdhary, "Real-time detection of distracted driving based on deep learning," *IET Intelligent Transport Systems*, vol. 12, 07 2018.

[16] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," *CoRR*, vol. abs/1602.07360, 2016.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[18] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," *CoRR*, vol. abs/1707.07012, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.