Detection of Distracted Driver using Convolutional Neural Network

Bhakti Baheti Suhas Gajre Sanjay Talbar
Center of Excellence in Signal and Image Processing,
SGGS Institute of Engineering and Technology, Nanded, Maharashtra, India
{bahetibhakti, ssgajre, sntalbar}@sggs.ac.in

Abstract

Number of road accidents is continuously increasing in last few years worldwide. As per the survey of National Highway Traffic Safety Administrator, nearly one in five motor vehicle crashes are caused by distracted driver. We attempt to develop an accurate and robust system for detecting distracted driver and warn him against it. Motivated by the performance of Convolutional Neural Networks in computer vision, we present a CNN based system that not only detects the distracted driver but also identifies the cause of distraction. VGG-16 architecture is modified for this particular task and various regularization techniques are implied in order to improve the performance. Experimental results show that our system outperforms earlier methods in literature achieving an accuracy of 96.31% and processes 42 images per second on GPU. We also study the effect of dropout, L2 regularization and batch normalisation on the performance of the system. Next, we present a modified version of our architecture that achieves 95.54% classification accuracy with the number of parameters reduced from 140M in original VGG-16 to 15M only.

1. Introduction

According to the World Health Organization (WHO) survey, 1.3 million people worldwide die in traffic accidents each year, making them the eighth leading cause of death and an additional 20-50 millions are injured/ disabled. As per the report of National Crime Research Bureau (NCRB), Govt. of India, Indian roads account for the highest fatalities in the world. There has been a continuous increase in road crash deaths in India since 2006. The report also states that the total number of deaths have risen to 1.46 lakhs in 2015 and driver error is the most common cause behind these traffic accidents.

The number of accidents because of distracted driver has been increasing since few years. National Highway Traffic Safety Administrator of United States (NHTSA) reports deaths of 3477 people and injuries to 391000 people in motor vehicle crashes because of distracted drivers in 2015 [2]. In the United States, everyday approximately 9 people are killed and more than 1,000 are injured in road crashes that are reported to involve a distracted driver [1]. NTHSA describes distracted driving as "any activity that diverts attention of the driver from the task of driving" which can be classified into Manual, Visual or Cognitive distraction [2] [1]. As per the definitions of Center for Disease Control and Prevention (CDC), cognitive distraction is basically "driver's mind is off the driving". In other words, even though the driver is in safe driving posture, he is mentally distracted from the task of driving. He might be lost in thoughts, daydreaming etc. Distraction because of inattention, sleepiness, fatigue or drowsiness falls into visual distraction class where "drivers's eyes are off the road". Manual distractions are concerned with various activities where "driver's hands are off the wheel". Such distractions include talking or texting using mobile phones, eating and drinking, talking to passengers in the vehicle, adjusting the radio, makeup etc.

Nowadays, Advanced Driver Assistance Systems (ADAS) are being developed to prevent accidents by offering technologies that alert the driver to potential problems and to keep the car's driver and occupants safe if an accident does occur. But even today's latest autonomous vehicles require the driver to be attentive and ready to take the control of the wheel back in case of emergency. Tesla autopilot's crash with the white truck-trailor in Williston, Florida in May 2016 was the first fatal crash in testing of autonomous vehicle. Recently in March 2018, Uber's self driving car with an emergency backup driver behind the wheel struck and killed a pedestrian in Arizona. In both of these fatalities, the safety driver could have avoided the crashes but evidences reveal that he was clearly distracted. This makes detection of distracted driver an essential part of the self driving cars as well. We believe that distracted driver detection is utmost important for further preventive measures. If

the vehicle could detect such distractions and then warn the driver against it, number of road crashes can be reduced.

In this paper, we focus on detecting manual distractions where driver is engaged in other activities than safe driving and also identify the cause of distraction. We present a Convolutional Neural Network based approach for this problem. We also attempt to reduce the computational complexity and memory requirement while maintaining good accuracy which is desirable in real time applications.

2. Related Work

This section summarises review of some of the relevant and significant work from literature for distracted driver detection. The major cause of manual distractions is usage of cellphones [2]. Motivated by the same, some of the researchers worked on cell phone usage detection while driving. Zhang *et al*. [19] created a database using a camera mounted above the dashboard and used Hidden Conditional Random Fields model to detect cell phone usage. It basically operates on face, mouth, and hand features. In 2015, Nikhil *et al*. [5] created a dataset for hand detection in the automotive environment and achieved average precision of 70.09% using Aggregate Channel Features (ACF) object detector. Seshadri *et al*. [14] also created their own dataset for cell phone usage detection. Authors used Supervised Descent Method, Histogram of Gradients (HoG) and an AdaBoost classifier and achieved 93.9% classification accuracy. The system could operate in a near real-time speed (7.5 frames per second). Le *et al*. achieved higher accuracy than state-of-art methods i.e 94.2% by training Faster-RCNN on the above dataset. Their approach is based on face and hands segmentation to detect cell phone usage. The system could operate on 0.06 FPS and 0.09 FPS for cell phone usage and hands on the wheel detection respectively [7].

UCSDs Laboratory of Intelligent and Safe Automobiles has done significant contribution in this domain but they dealt with only three types of distractions viz. adjusting the radio, adjusting mirrors and operating gear. Martin *et al*. [8] presented an vision-based analysis framework that recognizes in-vehicle activities using two kinect cameras that provided frontal and back views of the driver to provide "hands on the wheel" information. Ohn-bar *et al*. [12] proposed a fusion of classifiers where the image is to be segmented into three regions: wheel, gear and instrument panel to infer actual activity. They also presented a region-based classification approach to detect presence of hands in certain pre-defined regions in an image [11]. A model was learned for each region separately and joined using a second-stage classifier. Authors extended their research to include eye cues to previously existing head and hands cues [13]. However, it still considered only three types of distractions.

Zhao *et al*. [21] designed a more inclusive distracted driving dataset with side view of the driver considering four activities: safe driving, operating shift lever, eating and talking on cellphone. Authors achieved 90.5% accuracy using contourlet transform and random forest. Authors also proposed a system using PHOG and multilayer perceptron that yields accuracy of 94.75% [20]. In 2016, Yan *et al*. [18] presented a Convolutional Neural Network based solution that achieved a 99.78% classification accuracy.

The earlier datasets concentrated on only limited set of distractions and many of them are not publicly available. In April 2016, StateFarm's distracted driver detection competition on Kaggle defined ten postures to be detected (Safe driving + nine distracted behaviours) [3]. This was the first dataset to consider wide variety of distractions and was publicly available. Many approaches proposed by the

(a) C0: Drive safe (b) C1: Text left (c) C2: Talk left (d) C3: Text right (e) C4: Talk right

(f) C5: Adjust radio (g) C6: Drink (h) C7: Hair and makeup (i) C8: Reaching behind (j) C9: Talk to passanger

Figure 1: Ten Classes of Driver Postures from the Dataset

Figure 2: Original VGG-16 architecture that uses 3x3 convolutions throughout and fully connected layers of dimension 4096

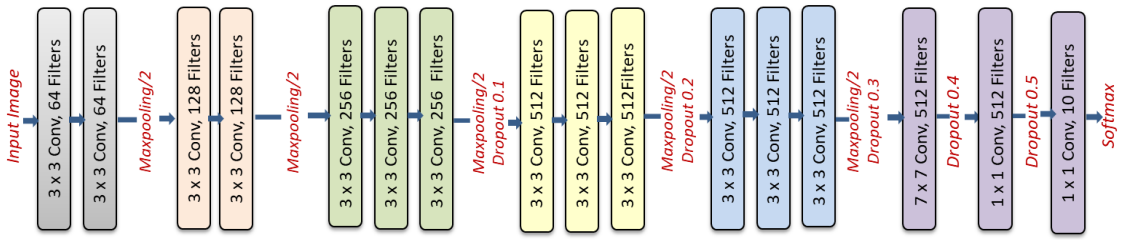


Figure 3: Fully convolutional VGG-16 Architecture where FC layers are replaced by convolutional layers. L2 regularization with λ =0.001 and batch normalisation is applied to all Conv and FC layers. Linearly increasing dropout is applied from 3rd max-pooling layer to FC layers.

researchers were based on traditional hand crafted feature extractors like SIFT, SURF, HoG combined with classical classifiers like SVM, BoW, NN. However CNNs proved to be the most effective techniques achieving high accuracy [9]. But as per the rules and regulation, use of dataset is restricted to competition purpose only.

In 2017, Abouelnaga *et al*. [4] created a new dataset similar to StateFarm's dataset for distracted driver detection. Authors preprocessed the images by applying skin, face and hand segmentation and proposed the solution using weighted ensemble of five different Convolutional Neural Networks. The system achieved good classification accuracy but is computationally too complex to be real time which is utmost important in autonomous driving.

3. Dataset Description

In this paper, we use the dataset created by Abouelnaga *et al*. [4]. The dataset consists of ten classes viz. safe driving, texting on mobile phones using right or left hand, talking on mobile phones using right or left hand, adjusting radio, eating or drinking, hair and makeup, reaching behind and talking to passenger. Sample images of each class from the dataset are shown in fig. 1. The data was collected from thirty one participants from seven different countries using four different cars and incorporated several variations of the drivers and driving conditions. For example, drivers are exposed to different lighting conditions like sunlight and shadows. The dataset consists of 17308 images divided into training set (12977) and test set (4331). We follow the same data distribution as in [4] for true performance comparison.

4. Technical Approach

Deep Convolutional Neural Network is basically a type of Artificial Neural Network (ANN) which is inspired by the animal visual cortex. Since last few years, CNNs have shown impressive progress in various tasks like image classification, object detection, action recognition, natural language processing and many more. The basic building blocks of a CNN based system include Convolutional filters/ layers, Activation functions, Pooling layer and Fully Connected (FC) layer. A CNN is basically formed by stacking these layers one after the other. Since 2012, there has been very rapid progress in CNNs because of availability of large amount of labeled data and the computing power. Various architectures like AlexNet, ZFNet, VGGNet, GoogLeNet, ResNet have established benchmarks in computer vision. In this paper, we explore the VGG-16 architecture proposed by Simonyan and Zisserman [16] and modify it for the task of distracted driver detection.

4.1. Original VGG-16 Architecture

VGG Net is one of the most influential CNN architecture from literature. It reinforced the idea that networks should be deep and simple. The architecture is shown in fig. 2. It worked well on both image classification as well as localization task. VGG uses 3×3 filters in all thirteen convolutional layers, ReLU activation function, 2×2 max pooling with stride 2 and categorical cross-entropy loss func-

tion. We use the pre-trained ImageNet model weights for initialisation and then fine tune all the layers of network with our dataset. As a preprocessing step, all the images are resized to 224×224 and per channel mean of RGB planes is subtracted from each pixel of the image. This has the geometric interpretation of centering the cloud of data around the origin along each dimension. Initial layers of the CNN act as feature extractor and the last layer is softmax classifier which classifies the images into one of the predefined categories. However the original model has 1000 output channels corresponding to 1000 object classes of ImageNet. Hence the last layer is popped and is replaced with softmax layer with 10 classes. Here, the cross entropy loss function is used for performance evaluation.

4.2. VGG-16 with Regularization

From experimentation using original VGG-16 network, it was observed that model is overfitting to the training data. It performs well on the training set achieving almost 100 % accuracy but fails to generalise on the unknown test data. Hence we perform various regularization techniques to reduce the generalization error. Also, LeakyReLU activation function is used instead of ReLU. Following are the main changes from original VGG-16:

- *LeakyReLU Activation Function*
  The Rectified Linear Unit (ReLU) activation function has become very popular in the past couple of years because of efficiency and faster convergence. But as the ReLU function sets output value to zero for all inputs less than zero, weights of some neurons may never get updated and it may result in dead neurons. LeakyReLU overcomes this problem by introducing a small slope in the negative region to keep the updates alive.

- *Dropout*
  Dropout is an efficient way of reducing overfitting by randomly dropping out i.e ignoring some neurons in training phase [17]. It helps to reduce interdependent learning amongst the neurons. We apply linearly increasing dropout in few convolutional as well as fully connected layers.

- *L2 Weight regularization*
  Weight regularization also called weight decay strongly relies on the implicit assumption that a model with smaller weights is somehow simpler than a network with large weights [10]. It is implemented by penalizing the squared magnitude of all the parameters directly in the cost function. We add the term $\frac{1}{2}\lambda w^2$ to the cost function considering every weight w in the network, where λ is the regularization strength. The choice of λ is a hyperparameter and is set to 0.001.

- *Batch Normalisation*
  Batch normalisation helps to improve the performance and stability of neural networks by explicitly forcing the activations through a layer of network to follow a unit Gaussian distribution [6]. It reduces strong dependence on weight initialisation, improves gradient flow through the network as well as allows higher learning rates. In our work, activations of all convolutional and fully connected layers are normalised.

4.3. Modified VGG-16

The major drawback of VGG-16 is total number of parameters which counts to be nearly 140M. Fully connected layers are computationally too expensive and also consume most of these parameters. Also, the network with fully connected layer can be applied to input of fixed size only. Replacing fully connected layer with convolution layer saves the number parameters and it can be applied to varying input size [15]. So, we build a fully convolutional neural network by replacing dense layers with 1×1 convolutions.

Table 1: Confusion Matrix using Original VGG-16 Architecture with Regularizers

Actual Label \ Predicted Label	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
C0	882	2	2	2	0	5	4	7	3	15
C1	0	316	5	0	3	0	2	0	0	0
C2	0	14	327	0	0	0	0	0	0	0
C3	9	2	0	473	5	2	1	0	1	1
C4	0	0	0	10	295	0	1	0	0	0
C5	6	0	0	0	0	298	1	0	0	0
C6	3	0	0	1	0	2	394	0	2	1
C7	10	0	0	0	0	0	0	288	1	2
C8	9	0	0	0	0	0	4	2	273	2
C9	10	0	0	0	0	0	7	0	1	625

Table 2: Class-wise Accuracy using Original VGG-16 Architecture with Regularizers

Class	Total Samples	Correct Predictions	Incorrect Predictions	Accuracy (%)
Safe Driving	922	882	40	95.66
Texting Using Left Hand	326	316	10	96.93
Talking on Phone Using Left Hand	341	327	14	95.89
Texting Using Right Hand	494	473	21	95.75
Talking on Phone Using Right Hand	306	295	11	96.40
Adjusting Radio	305	298	7	97.70
Drinking	403	394	9	97.77
Hair and Makeup	301	288	13	95.68
Reaching Behind	290	273	17	94.14
Talking to Passanger	643	625	18	97.20

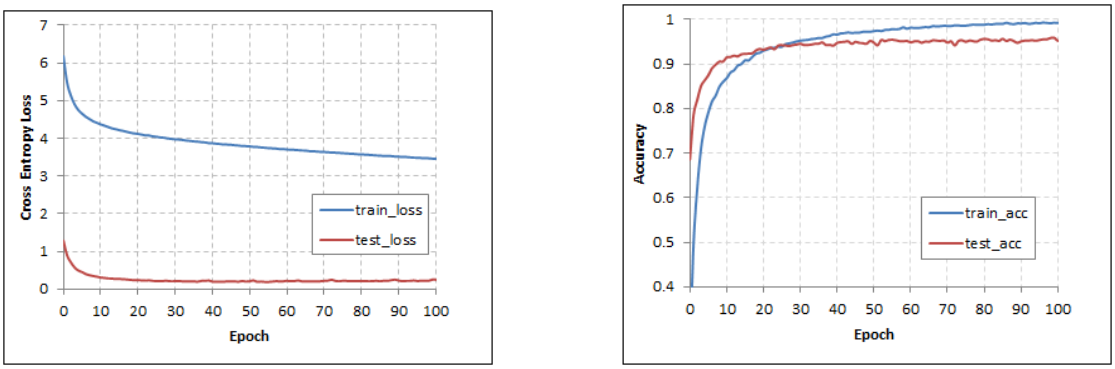
The modified network architecture is shown in fig. 3. Number of parameters are reduced to 15M that is only 11% of the original VGG-16 parameters. All the regularization parameters remain unchanged as in previous section.

5. Results and Discussion

We design a Convolutional Neural Network based system for distracted driver detection. The pre-trained ImageNet model is used for weight initialisation and concept of transfer learning is applied. Weights of all the layers of network are updated wrt the dataset. After rigorous experimentation, all the hyperparameters are finetuned. The training is carried out using Stochastic Gradient Descent with learning rate of 0.0001, decay rate of 10^{-6} and momentum value 0.9. The batch size and number of epochs are set to 64 and 100 respectively. Training and testing is carried out using NVIDIA P5000 GPU having 2560 CUDA cores with 16 GB RAM. The framework is developed using Keras and Theano.

When original VGG-16 is used as it is for the task of distracted driver detection, it produces 100% accuracy on the training set and 94.44% accuracy on the test set. Performance of the system is significantly improved with the addition of dropout, L2 weight regularization and batch normalisation which results in 96.31% accuracy on the test set. The system processes 42 images per second on an average. Table 1 provides precise and complete metric for analysis of results of the system in the form of confusion matrix. Table 2 depicts class-wise accuracies for each of the ten classes from dataset. Fig. 4 shows the training and testing accuracy and loss curves with varying epochs.

As number of parameters and hence the memory requirement of VGG-16 is high, we present an modified architecture with almost 90% reduction in parameters without much affecting the accuracy. We achieve accuracy of 95.54% on the test set. Table 3 and Table 4 show the confusion matrices and class-wise accuracies with the modified VGG-16 architecture.



(a) Train and Test Loss Plots

(b) Train and Test Accuracy Plots

Figure 4: Summary of Distracted Driver Classification Results using VGG-16 Architecture with Regularization

Table 3: Confusion Matrix using Modified VGG-16 Architecture

Actual Label \ Predicted Label	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9
C0	863	1	3	4	1	11	6	11	2	20
C1	0	315	5	0	4	0	2	0	0	0
C2	1	11	329	0	0	0	0	0	0	0
C3	9	5	0	471	5	2	1	0	0	1
C4	1	0	0	14	291	0	0	0	0	0
C5	7	0	0	0	0	297	1	0	0	0
C6	2	0	0	1	2	5	392	0	1	0
C7	13	0	0	0	0	0	0	284	1	3
C8	11	0	0	0	0	0	4	1	273	1
C9	11	0	0	1	0	0	5	1	2	623

Table 4: Class-wise Accuracy using Modified VGG-16 Architecture

Class	Total Samples	Correct Predictions	Incorrect Predictions	Accuracy (%)
Safe Driving	922	863	59	93.60
Texting Using Left Hand	326	315	11	96.63
Talking on Phone Using Left Hand	341	329	12	96.48
Texting Using Right Hand	494	471	23	95.34
Talking on Phone Using Right Hand	306	291	15	95.09
Adjusting Radio	305	297	8	97.38
Drinking	403	392	11	97.27
Hair and Makeup	301	284	17	94.35
Reaching Behind	290	273	17	94.14
Talking to Passanger	643	623	20	96.89

It is observed from the above confusion matrices that mainly 'safe driving' and 'talking to passenger' postures are confused with each other. It may be because of "hands on the wheel" position in both classes. Also, talking on cellphone is confused with texting on cellphone. Such misclassification can be because of lack of temporal information in the analysis.

The system for distracted driver detection and posture classification proposed by Abouelnaga *et al*. [4] consists of genetically weighted ensemble of five convnets. These five convolutional neural networks are trained on raw images, skin-segmented images, hand images, face images and 'hands + face' images. Authors trained the system with AlexNet and InceptionV3 on above mentioned five image sources. This approach makes system too heavy for real time application which is very much essential in self-driving cars. On the contrary we use a single ConvNet which is less complex and still achieves better accuracy than earlier methods as shown in Table 5.

Table 5: Summary of Distracted Driver Detection Results and Comparision with Earlier Approaches from Literature

Model	Source	Accuracy(%)
AlexNet [4]	Original	93.65
	Skin Segmented	93.60
	Face	84.28
	Hands	89.52
	Face + Hands	86.68
Inception V3 [4]	Original	95.17
	Skin Segmented	94.57
	Face	88.82
	Hands	91.62
	Face + Hands	90.88
Majority Voting ensemble of all 5 [4]		95.77
GA weighred ensemble of all 5 [4]		95.98
Original VGG (140M parameters)	Original	94.44
VGG with Regularization (140M parameters)	Original	96.31
Modified VGG (15M parameters)	Original	95.54

6. Conclusion and Future Work

Driver distraction is a serious problem leading to large number of road crashes worldwide. Hence detection of

distracted driver becomes an essential system component in self driving cars. Here, we present a robust Convolutional Neural Network based system to detect distracted driver and also identify the cause of distraction. We modify the VGG-16 architecture for this particular task and apply several regularization techniques to prevent overfitting to the training data. With the accuracy of 96.31% the proposed system outperforms earlier approaches of distracted driver detection from literature on this dataset as shown in Table 5. The system processes 42 images per second on NVIDIA P5000 GPU with 16GB RAM. We also propose a thinned version of VGG-16 with just 15M parameters and still achieving satisfactory classification accuracy.

As an extension of this work, we are working towards lowering the number of parameters and computation time. Incorporating temporal context may help in reducing misclassification errors and thereby increasing the accuracy. Also, in future, we wish to develop a system that will detect visual and cognitive distractions as well along with manual distractions.

References

[1] Center for disease control and prevention. `https://www.cdc.gov/motorvehiclesafety/distracted_driving/`.

[2] National highway traffic safety administration traffic safety facts. `https://www.nhtsa.gov/risky-driving/distracted-driving/`.

[3] State farm distracted driver detection. `https://www.kaggle.com/c/state-farm-distracted-driver-detection`.

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

[5] 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*, pages 2953–2958, Sept 2015.

[6] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[7] T. H. N. Le, Y. Zheng, C. Zhu, K. Luu, and M. Savvides. Multiple scale faster-rcnn approach to driver cell-phone usage and hands on steering wheel detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 46–53, June 2016.

[8] S. Martin, E. Ohn-Bar, A. Tawari, and M. M. Trivedi. Understanding head and hand activities and coordination in naturalistic driving videos. In *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pages 884–889, June 2014.

[9] R. P. A. S. Murtadha D Hssayeni, Sagar Saxena. Distracted driver detection: Deep learning vs handcrafted features. *IS&T International Symposium on Electronic Imaging*, pages 20–26, 2017.

[10] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 78–, New York, NY, USA, 2004. ACM.

[11] E. Ohn-Bar, S. Martin, A. Tawari, and M. M. Trivedi. Head, eye, and hand patterns for driver activity recognition. In *2014 22nd International Conference on Pattern Recognition*, pages 660–665, Aug 2014.

[12] E. Ohn-Bar, S. Martin, and M. M. Trivedi. Driver hand activity analysis in naturalistic driving studies: challenges, algorithms, and experimental studies. *J. Electronic Imaging*, 22(4):041119, 2013.

[13] E. Ohn-Bar and M. Trivedi. In-vehicle hand activity recognition using integration of regions. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 1034–1039, June 2013.

[14] 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)*, pages 35–43, June 2015.

[15] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, April 2017.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

[18] C. Yan, F. Coenen, and B. Zhang. Driving posture recognition by convolutional neural networks. *IET Computer Vision*, 10(2):103–114, 2016.

[19] 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*, pages 248–251, July 2011.

[20] C. Zhao, B. Zhang, X. Zhang, S. Zhao, and H. Li. Recognition of driving postures by combined features and random subspace ensemble of multilayer perceptron classifiers. *Neural Computing and Applications*, 22(Supplement-1):175–184, 2013.

[21] C. H. Zhao, B. L. Zhang, J. He, and J. Lian. Recognition of driving postures by contourlet transform and random forests. *IET Intelligent Transport Systems*, 6(2):161–168, June 2012.