Real-Time Driver's Focus of Attention Extraction and Prediction using Deep Learning

Pei-heng Hong[1], Yuehua Wang[2*]
Department of Computer Science
Texas A&M University-Commerce
Commerce, Texas 75428 USA
*Corresponding Author

***Abstract*—Driving is one of the most common activities in our modern lives. Every day, millions drive to and from their schools or workplaces. Even though this activity seems simple and everyone knows how to drive on roads, it actually requires drivers' complete attention to keep their eyes on the road and surrounding cars for safe driving. However, most of the research focused on either keeping improving the configurations of active safety systems with high-cost components like Lidar, night vision cameras, and radar sensor array, or finding the optimal way of fusing and interpreting sensor information without considering the impact of drivers' continuous attention and focus. We notice that effective safety technologies and systems are greatly affected by drivers' attention and focus. In this paper, we design, implement and evaluate DFaep, a deep learning network for automatically examining, estimating, and predicting driver's focus of attention in a real-time manner with dual low-cost dash cameras for driver-centric and car-centric views. Based on the raw stream data captured by the dash cameras during driving, we first detect the driver's face and eye and generate augmented face images to extract facial features and enable real-time head movement tracking. We then parse the driver's attention behaviors and gaze focus together with the road scene data captured by one front-facing dash camera faced towards the roads. Facial features, augmented face images, and gaze focus data are then inputted to our deep learning network for modeling drivers' driving and attention behaviors. Experiments are then conducted on the large dataset, DR(eye)VE, and our own dataset under realistic driving conditions. The findings of this study indicated that the distribution of driver's attention and focus is highly skewed. Results show that DFaep can quickly detect and predict the driver's attention and focus, and the average accuracy of prediction is 99.38%. This will provide a basis and feasible solution with a computational learnt model for capturing and understanding driver's attention and focus to help avoid fatal collisions and eliminate the probability of potential unsafe driving behavior in the future.**

***Keywords*—Driving; attention; interesting zones; deep neural network; deep learning; models**

I. Introduction

While being a basic enabler of our modern society, road vehicle transportation has become a major source of societal concerns. In 2016 alone, over 330, 655 people [1]–[3] were killed in vehicle accidents in the world, which is far greater than the number of US Soldiers killed in action in Vietnam (58,220 fatalities). Comparing to 19 high-income countries [3] including Canada, Germany, France, Spain, United Kingdom, and other 14 countries, the United States has highest vehicle accident death rate, where about 90 people die each day in vehicle accidents. The U.S. National Security Council (USNSC) reports that vehicle accidents cause estimated 40,200 fatalities, a 6% rise from 2015 (i.e., a 14% rise from 2014), which makes 2016 the deadliest year on American roads in nearly a decade. To prevent vehicle accidents, injuries, and deaths [4], we have enforced the use of seat belts, car seats and booster seats for children through at least age 8, restricted alcohol-impaired driving [5] and speeding, and suggested to avoid distracted driving (such as texting, talking on the phone, eating or doing something else that occupies the senses you need to drive). In reality, those are far from enough for safe driving.

A wide variety of approaches [6]–[8], systems [9]–[11], and self-driving cars [12]–[15] have been developed for vehicle surrounding environment perception, including object detection, tracking, localization, navigation. In the market, most of automotive companies like Tesla, Mercede-benz, Audi, BMW, Rolls-Royc, GM, Ford, and Honda have offered diverse advanced driver-assistance systems (ADASs) with various high-cost components including Lidar, high resolution cameras, radar sensor array, sonar, GPS, odometry, and inertial measurement unit. ADASs are electronic systems that assist the human driver while driving or during parking and have significantly improved comfort and safety in nowadays driving.Waymo [13], [14] originated by Google's self-driving car, Navya automated bus driving system [16], Uber driverless car [17], and Apple car [15] are promising autonomous driving vehicle projects to promote fully personal self-driving vehicle and related technology development. Since there are still many technological difficulties and regulatory issues that are need to be addressed, it would be close to a decade before self-driving cars are ready to use safely in a large number [18].

The emergence of artificial intelligence and rapid development of computer vision technology provide more opportunities to improve driving safety without strongly relying on the high-cost components or external information surrounding vehicles. We observe that driver's attention and focus have significant impact on driving safety and can lead to vehicle accidents or even terrible disasters. It, therefore, is extremely important to investigate and understand driver's attention and focus taking into consideration the complexity and dynamics of driver's behaviors under realistic driving conditions. Unfortunately, this topic is under-investigated and the lack of realistic experimental system does not help.

Motivated by the fact that driver's attention and focus are always changing over time in an uncertain, dynamic, and

continuous environment, we propose to capture and examine driver attention and focus using two low-cost dash cameras in real-time. Instead of using expensive, commercial eye tracker/tracking devices, dual dash cameras are placed on the windshield inside the vehicle. One camera is facing front for car-centric view and the other is facing towards the driver for driver-centric view. Based on the raw streaming data from the dash cameras, we provide an approach to extract facial features with face and eye detection, augment face images, track and predicate gaze focus zones during driving. Our first contribution is to build our drive attention database by driving from/to our campus to/from Dallas, Texas. The underlaying ideas is to travel different types of roads including campus road, rural road, suburban road, highways, and metropolitan roadways to collect enough driver-centric and car-centric data under various driving circumstances.

The second contribution is to extract the driver's facial features and generate augmented face images using open source toolkit, Openface [19], [20]. The main visual cues that we are analyzing are head and eye directions that are selected based on the insights that drivers who are paying attention on driving will have a tendency to look forward and keep eyes on roads. To enable to locate the driver's focus, the driver's view is grided into 11 grids within 5 gaze zones. The a feature vector can be formed with the facial features and gaze zone to label the data. The third contribution is the introduction of a data-driven deep learning network called DFaep. This network uses the detected features and augmented face images as the input and output accurate the focus of driver's attention. If the focus is not within the main zones, a warning alert will be issued to bring it to the driver attention. We evaluate DFaep on two datasets, DR(eye)VE [6] and our own dataset. On DR(eye)VE, we substantially improve the predictions of the driver's focus with accurate gaze data on the real road scene. On our dataset, we test the trained model and compare our DFaep with other deep learning networks with the same set of images and extracted features in terms of accuracy, loss, and network complexity to identify the network that is best for real-time systems and applications. We outperform all other networks and show the results that DFaep can achieve highest prediction accuracy, 99.38% with lowest validation loss, 0.018.

The paper is organized as follows: Section II provides a brief literature review about computer vision systems and artificial intelligence relate to driver attention behaviors; Section III details the design and structure of the proposed deep learning network DFaep for modeling the driver attention behaviors and tracking focus; Section IV presents the datasets used in experiments. The numerical performance evaluation results and our finding are then reported; and Section V concludes the paper and describes our future research plan.

II. Literature

In this section, we review the existing technologies and studies used to capture and detect driver's attention and focus, which can be classified based on eye/gaze tracking to capture humans attention and detection and prediction methods driven by sensing data.

A. Eye Tracking

Eye tracking data is widely held to be a good window into attention or non-attentive states for learners or people working on tasks [21], [22]. Eye tracking refers to the careful measurement of the movements of the eyes and/or body when the participants are either positively or negatively interacting with the learning environments in a time-varying manner. The measurement devices, hardware platforms, and systems are commonly known as eye trackers. Based on the hardware setup, we classify the existing eye tracking systems into four main types: tower-mounted eye-tracker [23]–[25], screen-based eye tracker [26], [27], head mounted/wearable eye tracker [28]–[30], and mobile eye tracker [31]–[33]. An detailed discussion on those four types of eye tracking devices and platforms can be found in our papers [34]–[36]. Eye trackers can be either well-established commercial devices/systems or low-cost, portable systems designed by educators or researchers. For commercial eye trackers, they are still expensive in the current market, but well developed and maintained by companies like Tobii, SMI, ISCAN, LC Technologies, EyeTech, Oculus, Ergoneers, SmartEye, and Mirametriz [36]. With the requirements of necessary purchasing, only authorized users can use the purchased hardware and software with warranties, documentations, reachable technical support. Comparing to the commercial eye trackers, open source eye tracking devices and systems [29], [30], [37]–[39] have unique abilities to support both and low-cost eye trackers, easily alter experiments to specific scenarios, quickly prototype ideas and enable eye tracking research and applications without major restrictions. Given that, in this study, we propose an active safe driving behavior detection system with two low-cost car dash cameras, ZEdge-Z3 in $100.

B. Attention Detection and Prediction

There are existing studies have been conducted with the eye tracking or camera data to detect and predicate driver's attention. For instance, Tran et al. [40] proposed an assisted-driving testbed with a driving simulator in a laboratory to simulate the driving environment and control the simulator's behavior using a script. Ten distracted driving behaviors like drinking, operation the radio, talking, texting, reaching behind, and making-up were defined. Based on the definition, four deep learning models, CNN, VGG-16, AlexNet, and GoogleNet are trained and tested on the simulator for driving distraction detection by distinguishing the defined distracted behaviors. The results showed that the VGG-16 achieved fastest frequency (14 Hz) while the ResNet model yielded highest accuracy of 92% with highest complexity which needed to have longest time for data processing and model calculation.

In [41], a CNN(Convolution Neural Network) is proposed and trained to mimic the driver based on training data from human driver's driving, it builds a model which takes as an input raw data and map it directly to a decision, using minimum training data and minimum computation, the car learns to drive on roads with or without human-design features. The idea of end to end self-driving car were implemented by NVIDIA [42]. They trained CNN to map raw pixels from one single front camera directly to steering commands. The network architecture is shown in Fig. 2. that consists nine layers which includes five convolutional layers, one normalization

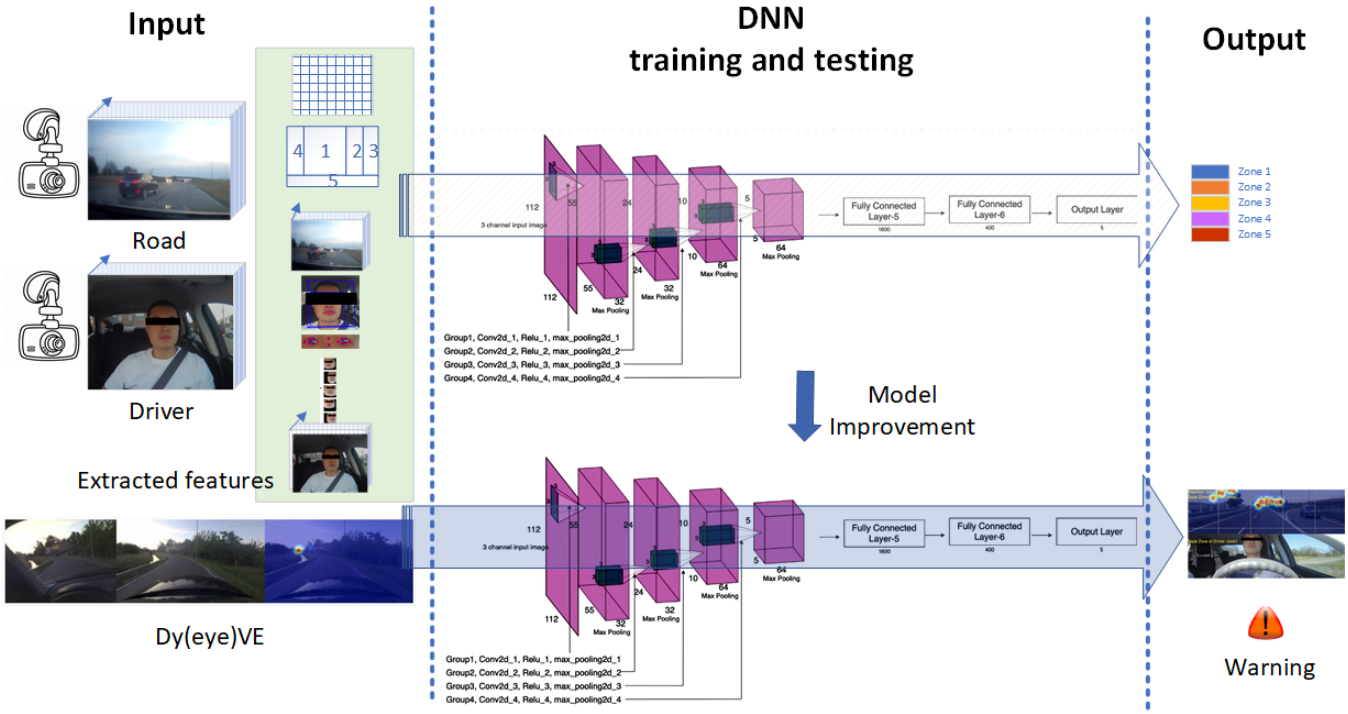
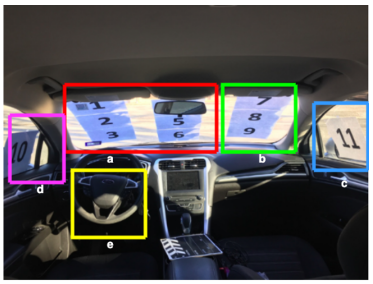


Fig. 1. Flowchart of Driver Attention System.

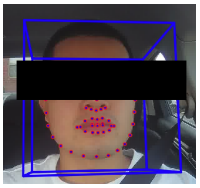
(a) 11 Grids of windshield

(b) Internal dual cameras

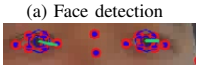


(c) 5 Main gaze zones

Fig. 2. Internal Setting of our Experimental Vehicle.



(a) Face detection



(b) Eye detection

Fig. 3. Driver Facial Detection.

layer and three fully connected layers, they used NVIDIA DevBox, Torch 7 and NVIDIA DRIVETM PX self-driving car computer for training and determining the steering angle. However, in their work, they did not consider about the throttle factor into training, they only focus on steering angle control, Therefore, based on their work, adding speed throttle factor and implementing it with RC car, can pose new challenges.

In this study, we compare the internal and external environment of the vehicle and propose a method that leverages deep learning network to predict the driver's attention in the environment inside and outside the vehicle. For the driver warning part, we have added scenes outside the vehicle as warning auxiliary factors to determine whether the driver is distracted by the driving of the vehicle.

III. Proposed System and Methodology

A. Overview of Our Driver Attention Database

The overview of this research is shown in Fig. 1. We use a camera to capture the raw images of the the driver's frontal view based on the corresponding windshield inside the vehicle, each zone for 1 minute. Then, the driver's facial features in raw images are extracted through the OpenFace [19], [20]. After that, each image of driver facial feature is cropped to 112×112 pixels for input DNN for training. We divide the data into training set, validation set and testing set. Here, our driver attention system will classify the driver's attention zone based on probability score of DNN model for each zone on the windshield. In DR(eye)VE database [6], we leverage saliency maps in the DR(eye)VE database to simulate our driving scene and label each frame of images, then we perform a synchronous comparison between the estimation zone of our driver attention detection classification system and attention zone of the DR(eye)VE database for evaluating the method

that we proposed. Finally, the synchronous comparison result is obtained, the system will give a warning to driver if the final label comparison result is not matched.

B. Detecting Facial Landmarks

The experimental environment with 11 grids within five gaze zones on the windshield of internal vehicle in this research is shown in Fig. 2. The reason we divided it into five gaze zones is to obtain more realistic driving scenarios. In this research we leverage a standard web camera to accomplish this task. A camera sensor is mounted in front of the driver. Since the size of the web camera is moderate (8 cm × 5 cm × 1.5 cm), it could be mounted in the vicinity of the windshield, and can continuously capture the driver's facial. Power to the web camera was supplied by the car using a universal serial bus (USB) line, connecting to the camera. The captured images by web camera is successively transmitted to the computer with graphics processing unit (GPU) for training via memory card of the web camera. The configuration of the web camera and the gaze zone on the windshield are shown in Fig. 2. We simulate driver's driving attention to different zone of windshield and their driving habits as if driving in real world, and save the videos recorded by the web camera, for each videos save the correspondent zone area performed by the driver. The dataset consists of 9,000 frames divided in five gaze zones, each of which is 1 minutes long, as shown in Fig. 4. The frames resolution size is 1920 × 1080 pixels.

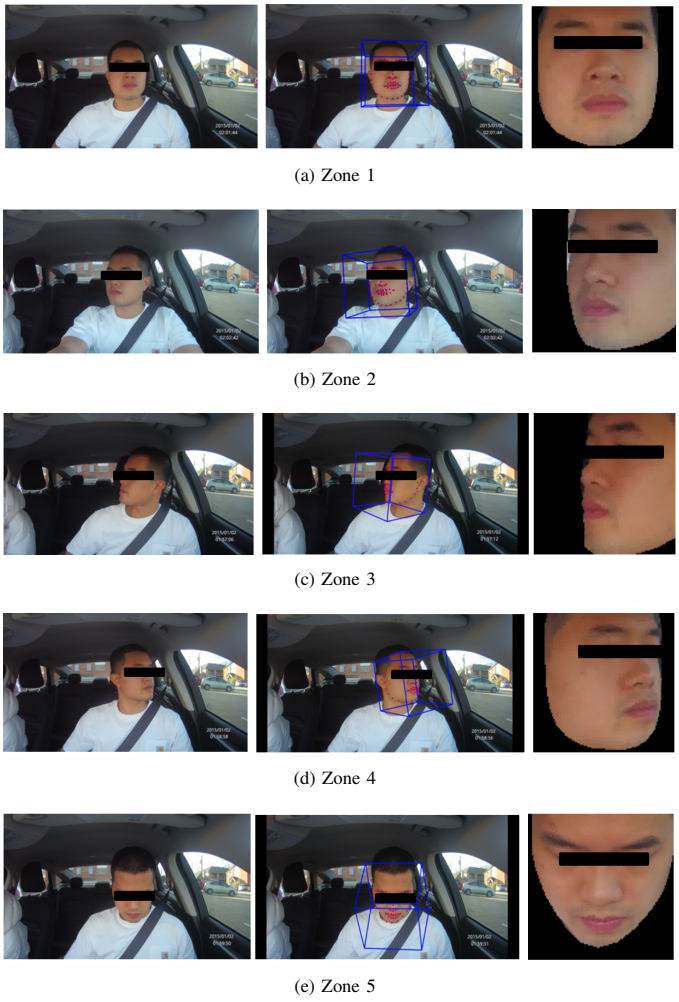
C. Driver's Data Pre-processing

In our research, in order to improve the training efficiency and reduce unnecessary interference factors, our measures are to capture the driver's facial information and reduce the size of the original input image. In here, we leverage Openface to implement this goal. Openface was proposed by Tadas Baltrušaitis et al. [19], [20]. There are a lot of tools that can implement the face feature detection in images or videos. However, most of them did not provide the source code which makes it very difficult to reproduce experiments on different datasets. In Openface, it includes facial landmarks, head pose estimation, eye gaze estimation and the most importantly it is opens source, free and a tool that provides source code.

In our experiment, the facial feature of driver was captured based on Openface [19], [20]. Openface is a tool designed for computer vision and machine learning researchers, it leveraged Multi-task Convolutional Neural Network (MTCNN) facial feature tracker [43] to capture 68 face landmarks, eyelids, the iris, eyelids and the pupil are detected by a Constrained Local Neural Field (CLNF) landmark detector [44], and the head pose is extracted through 3D representation of facial landmarks of Convolutional Experts Constrained Local Model (CE-CLM) [45]. In the end, we obtained the appearance extraction face alignment images and cropped to 112×112 pixels for input to DFaep as shown in Fig. 5.

D. Network Structure

In this research, we propose a deep learning network for driver's focus of attention extraction and prediction (DFaep)(including six layers) as shown in Table I. The model used in this research is a simple stack of of four convolutional layers (with ReLU activation [46]), four max pooling layers (with ReLU activation) and two fully connected layers (with ReLU and Softmax [47] activation). This model is modified from the LeNet [48]. We try to develop a DNN model with high accuracy and low parameter computation. Today we have far better models than this model, but this is the lightest one in terms of computation, and it is also the most suitable one for



(a) Zone 1

(b) Zone 2

(c) Zone 3

(d) Zone 4

(e) Zone 5

Fig. 4. Example Images of Driver's Face Landmark (left), after OpenFace Capture Driver's Attention (middle), and Driver's Facial Feature Image (right) while Looking at Distinct Zones of Fig. 2. Cases of Looking at (a) Zone 1; (b) Zone 2; (c) Zone 3; (d) Zone 4; (e) Zone 5.

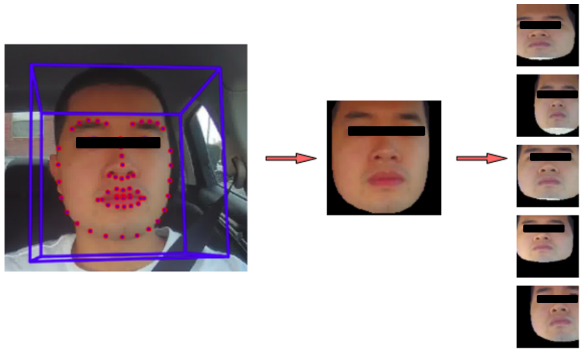


Fig. 5. The Example of Augmented Images.

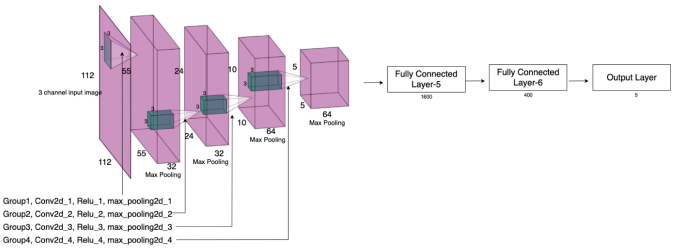


Fig. 6. Network Architecture used for Extracting Driver Gaze Features and Predicating Attention Focus.

TABLE I. STRUCTURE OF DFAEP MODEL

	Layers	# Filters	Size	Kernel	Stride	Padding
	Input layer	-	112 × 112 × 3	-	-	-
Group 1	1st Conv	32	110 × 110 × 32	3 × 3	1 × 1	0 x 0
	ReLU-1	-	110 × 110 × 32	-	-	-
	Pool-1	1	55 × 55 × 32	2 × 2	2 × 2	0 × 0
Group 2	2nd Conv	32	53 × 53 × 32	3 × 3	1 × 1	0 × 0
	ReLU-2	-	53 × 53 × 32	-	-	-
	Pool-2	1	26 × 26 × 32	2 × 2	2 × 2	0 × 0
Group 3	3rd Conv	64	24 × 24 × 64	3 × 3	1 × 1	0 × 0
	ReLU-3	-	24 × 24 × 64	-	-	-
	Pool-3	1	12 × 12 × 64	2 × 2	2 × 2	0 × 0
Group 4	4th Conv	64	10 × 10 × 64	3 × 3	1 × 1	0 × 0
	ReLU-4	-	10 × 10 × 64	-	-	-
	Pool-4	1	5 × 5 × 64	2 × 2	2 × 2	0 × 0
	1st FC	-	400 × 1	-	-	-
	ReLU-5	-	400 × 1	-	-	-
	Output layer	-	400 × 1	-	-	-
	2nd FC	-	400 × 1	-	-	-
	Softmax	-	400 × 1	-	-	-
	Output layer	-	400 × 1	-	-	-

our research. The performance comparison of different DNN models will be discussed in Section 4. In this model of DFaep, we executed the data augmentation in our training dataset. After extracting five features of last fully connected layer, the final attention zone of driver was estimated based on softmax function [47] of our DNN model.

The structure of our DFaep that is presented in Fig. 6 and will be illustrated in Table I. Our DFaep model is constructed of four convolutional layers (with ReLU activation [46]), 4 max pooling layers (with ReLU activation) and two fully connected layers. In the first convolutional layer, there are 32 filters of size 3 x 3 are applied to the input of 112 x 112 x 3. Here, 112 x 112 x 3 stands for width, height, and channel number, separately.Thus, a feature map of 112 x 112 x 32 is acquired. It can be computed according to the following standard:(output height = (input height - filter height + 2 x number of padding)/number of stride + 1) [49]. As shown in Table I, the height of input, height of filter, number of padding, and number of stride are 112, 3, 0, and 1 separately. From that the height of output of the feature map was obtained by the convolution is computed as 110 = ((112 - 3 + 2 x 0)/1 + 1). If the value is positive, it can be valid output used in the next layer. If the value is negative, the output is 0.

E. Activation

The Rectified linear unit (ReLU) activation function [46].

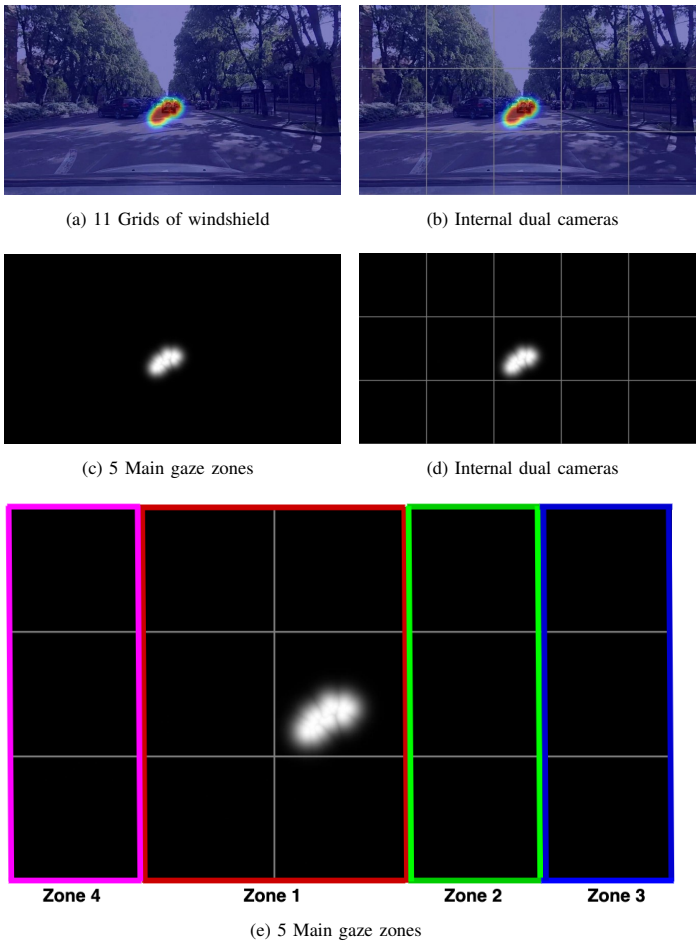
$$ReLU = max(0, x) \quad (1)$$

The ReLU function is not differentiable across the entire interval, but the non-differentiable part can be performed using Sub-gradient. Instead, ReLU is the most frequently used excitation function in recent years. Because of its following characteristics, including: solving the problem of gradient explosions, calculating quite quickly, and converging quickly, it will be analyzed in detail below.

For neural networks such as error back transfer operation (BN), gradient calculation considerations are most important when updating weights. Sigmoid function [50] is prone to vanishing gradient problems. When the input value approaches the saturation region (sigmoid function). When the excitation is performed outside [-4, +4], the first-order differential value approaches 0, and the problem of gradient disappearance occurs, which makes the backward transfer of the error calculation impossible, and the weight update cannot be performed effectively, and the neural network layer is deepened. The time is more obvious. Therefore, it is a difficulty encountered in deep neural network training, and the piecewise linear nature of ReLU can effectively overcome the problem of gradient disappearance. In our research, each convolutional layer is followed by one ReLU activation and one max pooling layer in Table I. The size of filter and stride within each max pooling layer is 2 × 2 and 1 × 1, respectively. As shown in Fig. 6, the size of each feature map is reduced by a pooling layer, ReLU-1 (110 × 110 × 32) is decreased to Pool-1 (55 × 55 × 32), ReLU-2 (55 × 55 × 32) to Pool-2 (24 × 24 × 32), ReLU-3 (24 × 24 × 32) to Pool-3 (10 × 10 × 64) and ReLU-4 (10 × 10 × 64) to Pool-4 (5 × 5 × 64).

After the four convolutional layers (with ReLU activation [46]), and 4 max pooling layers (with ReLU activation), we obtained the final feature map size of 5 × 5 × 64 pixels. And final feature map would be processed by the additional two fully connected layers. For each fully connected layers, the feature maps are separately of 400 × 1 and 5 ×1 as shown in Table I. In our driver attention classification system, the driver's attention zone is classified by . And the gaze zones of the vehicle are five as shown in Fig.6, hence after the softmax function' [47] in the last fully connected layer, we will determine which classification among these five is our final result according to the maximum probability calculated by the softmax function. In the last fully connected layer, the softmax activation is adopted as presented in Equation (1). Here, Vi is the output of the classifier's previous output unit; i is the category index; The total number of categories is C. Si stands for the ratio of the index of the current element to the sum of the indices of all elements. (1) From the observation during the deep learning process that the accuracy of training set is very high, but the prediction accuracy is extremely low of validation set. This is due to the over-fitting issue. The over-fitting means that the learning is performed too thoroughly, and all the features of the training set have been learned, so the machine has learned too many local features, and too many fake features due to noise, which caused the model to fail.

In this research, we leverage data augmentation and dropout methods [51] for preventing the over-fitting problem. For data augmentation methods, the details of it will be discussed. In the dropout methods, we decide to dropout probability of 50 per cent after the the 1st fully connected layer. After the four convolutional layers (with ReLU activation),



(a) 11 Grids of windshield (b) Internal dual cameras

(c) 5 Main gaze zones (d) Internal dual cameras

(e) 5 Main gaze zones

Fig. 7. Example Images of Driver's Face Landmark (left), after OpenFace Capture Driver's Attention (middle), and Driver's Facial Feature Image (right) while Looking at Distinct Zones of Fig. 2. Cases of Looking at (a) Zone 1; (b) Zone 2; (c) Zone 3; (d) Zone 4; (e) Zone 5.

and four max pooling layers (with ReLU activation), and 2nd fully connected layers (with ReLU and Softmax activation), the estimation zone of driver attention is obtained.

F. Model Improvement

We have used the DNN model to estimate the driver's gaze zone on the windshield of the vehicle. When the driver does not pay attention on the area where they should be focus, we should give a warning. We use the DR(eye)VE dataset to improve our model based on the attention zones of a real road driving scene. The DR(eye)VE dataset [6], it contains six hours of driving data, including 555,000 images, 74 video sequences of five minutes each. Videos were recorded in various environments, such as highway, downtown and countryside, sunny, rainy, cloudy, daytime and night. Because the DR(eye)VE project is currently the largest dataset of driving scenes with driving attention zones are available, we leverage its saliency maps dataset to compare to the estimation zone of driver to ensure whether or not to make the alarm during the driving.

Here we are using sequence videos in the DR (eye) VE dataset to simulate real road driving scene. In this work, we divide at a rate of 5 frames per second to the videos into 1800 frames. In our research, the saliency maps are divided into the 15 grids. In our case, the driver 's face landmark on the windshield is defined as four zones. As shown in the Fig. 7, the red block is defined as Zone 1, the reason for the larger proportion of this block is because the driver's concentration will usually focus on the front part, so our proportion in this area will be greater than other blocks. The green part It is Zone 2, the blue part is Zone 3, and the purple part is Zone 4.

For performing a matching verification action with our driver's face concentration, we need to label each saliency map. For obtaining the label of our saliency maps, the filter of the size of 1×1(stride number is 1) is used for the input of 1920×1080 pixels. We divide the 1920 by 5, then obtain a number 384 represents for each zone and then scan each pixel in each zone. The pixel of width from 1 to 384 is defined as Zone 4 (purple part of Fig. 7 (c)), 385 to 1152 is defined as Zone 1 (red part of Fig. 7 (c)), 1153 to 1536 is defined as Zone 2 (green part of Fig.7 (c)), and 1537 to 1920 is defined as Zone 3 (blue part of Fig. 7 (c)). Since there are only pure black (0) and white (255) in the saliency map, thus the value of each pixel will be much unsophisticated. Finally, the attention zone whose final sum value is largest within 5 zones of Fig. 7 (c) is defined as the label of this saliency map.

G. Validation of Systems and Algorithms

With the road scene, we have labeled the attention area for each frame by calculating each pixel of frame. Then, we can match the prediction gaze zone of our current driver's facial image of Fig. 3 and the attention zone label of road scenarios of Fig. 7(c). After that, the driver prediction zone and road attention zone were determined whether or not identical based on label comparison. Finally, the label comparison result is obtained, the system will give a warning to driver if the final label comparison result is not matched.

IV. Experiments and Results

A. Experiment Data

In our experiment, our dataset has been collected for the driver's attention classification system through a web camera as presented in Fig. 2. Our database was collected in an actual car instead of in a lab. There are many databases on driver concentration in previous research, such as Robust simultaneous modeling and tracking monitoring video (RS-DMV) dataset [52], The Chinese Face Database (CAS-PEAL) database [53] and Berkeley Deep Drive (BDD) [54], etc. However, these databases may not suitable for the research method we proposed this time due to the label of ground-truth attention zone is not defined. Without this key element, we cannot verify the method that purposed in this research. Therefore, we recorded our own database. In Fig. 4, 5 zones were split on the wind shield and designated to gaze at for this experiment. The size of raw image was $1920 \times 1080 \times 3$. When the experiment was starting, the driver has to perform normally, like they were actually driving in the ordinary day. After that, we cropped the size of images to 112×112 pixels of driver facial through OpenFace for DFaep training and validation as presented in Fig. 3.

TABLE II. DESCRIPTION OF TRAINING, VALIDATION AND TESTING DATA IN OUR DATASET

Dataset	Amount
Training	7,000 (1,400 × 5) images for each gaze zone on the wind shield
Validation	2,010 (402 × 5) images for each gaze zone on the wind shield
Testing	1,802 images from the real driving scenarios

In our experiment, we performed three different DNN structures for training and validation. In here, we adopt AlexNet [55], VGG16 [56] and DFaep model. The model structure is shown in Fig. 6. For preventing the over-fitting problem that we adopted some measures for data augmentation which are as follows. Shifting 2 pixels in vertically and horizontally and shear angle is 20 degrees in counter-clockwise direction as presented in Fig. 5. The raw images were used in validation process whereas the augmented images were used for training. In here, we used 7,000 augmented images in our dataset as training set, 2,010 images as validation set, and used completely independent 1800 images as testing set as shown in Table II. For the convolution neural network training, validation and testing, the desktop computer that we used is Intel i7-8700 CPU at 3.20 GHz and 16 GB memory. And the graphics card we used NVIDIA GeForce GTX 1070 (CUDA 10.0 and 8 GB memory). The backend was achieved by Keras (version 2.3.1) [57] with TensorFlow (version 2.1.0) [58]. And the algorithm was achieved by Visual Studio 2013, Dlib (version 19.19) and OpenCV (version 4.2.0) library.

B. Training of DFaep Model

In this experiment, Adaptive Moment Estimation(Adam) [59] optimizer method was used in our DFaep training. Adam is distinct to classical optimizer such as Stochastic gradient descent SGD [60], RMSprop [61] and Momentum [61], etc. Adam can be considered as a combination of RMSprop and Momentum, which uses the first and second moment estimates of the gradient to dynamically adjust the learning rate of each parameter. The main advantage of Adam is that after offset correction, the learning rate of each iteration has a certain range, making the parameters relatively stable as shown in Equations (2), (3) and (4).

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \end{aligned} \tag{2}$$

$$\begin{aligned} \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \end{aligned} \tag{3}$$

$$W \leftarrow W - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{4}$$

Where mt and vt are the weight to estimate of the first moment (the mean) and the second moment of the gradients respectively, g is the weight to gradient on current mini-batch, β_1 is used for decaying the running average of the gradient, β_2 is used for decaying the running average of the square of gradient, ε is the weight to prevent division from zero error. In our DFaep training, the experiment was adopted for the predefined setting. The β_1, β_2, and ε of Equations 2, 3 and 4 were set as 0.9, 0.999, and 10^{-8}, respectively.

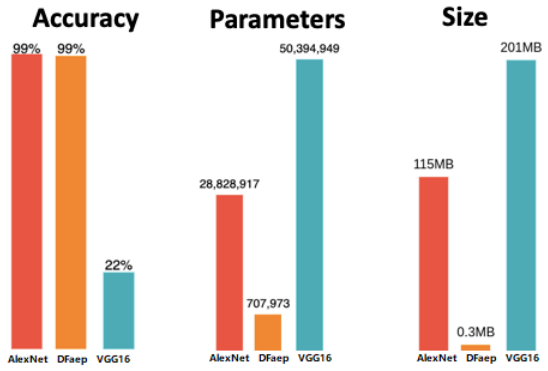


Fig. 8. The Example of Augmented Images.

TABLE III. COMPARISON OF PERFORMANCE IN THREE DIFFERENT CNNS

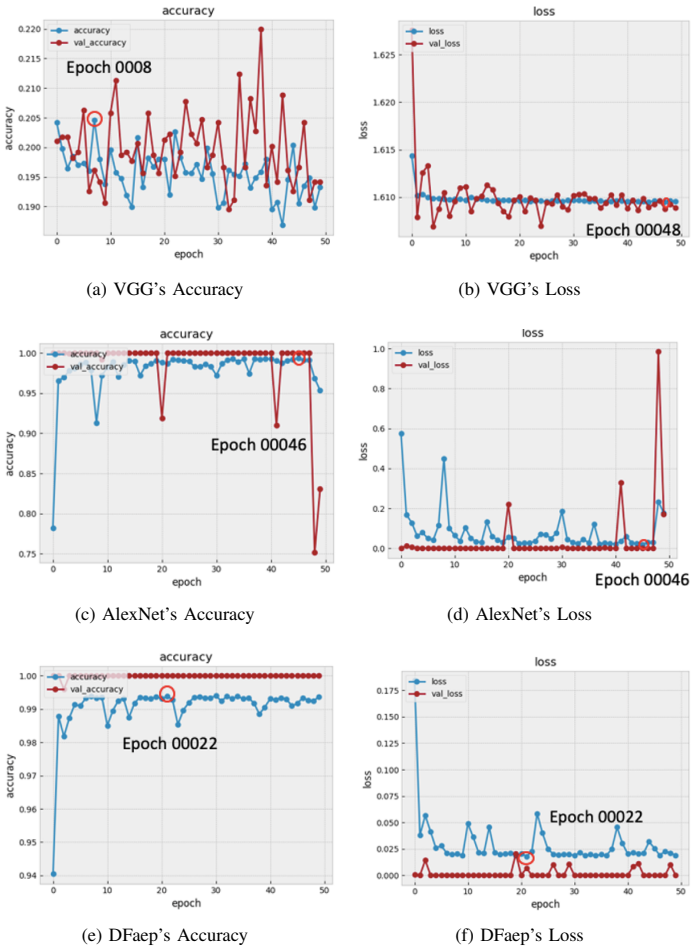
Network	#Parameters	#Layers	Model description	Accuracy
VGG16	50 M	16	13 conv + 3 fc layers	22 %
AlexNet	24 M	8	5 conv + 3 fc layers	99 %
DFaep	0.7 M	6	4 conv + 2 fc layers	99 %

In this research, we attempted AlexNet, VGG16 and DFaep model for training and validation, respectively. As shown in Fig. 8 and Fig. 9, the visualization of the accuracy and loss during training of VGG16, AlexNet and DFaep, respectively. The x-axis stands for epoch of training. The y-axis represents the loss and accuracy of training and validation, respectively. It can be observed from the results of training, the VGG16 model with loss 1.60952 and training accuracies 0.2046, the AlexNet model with loss 0.01979 and training accuracies 0.9934 and our DFaep model with loss 0.01803 and training accuracies 0.9938.

It can be concluded from the observation that the accuracy of the VGG16 model is the lowest 0.2046 (20%), and the accuracy of AlexNet and our DFaep model is very similar, their accuracy curves close to 1 (100%). However, compared to DFaep and AlexNet, the depth of the DFaep model structure is only 6 layers which is shallower than the 8 layers used by AlexNet. Moreover, the parameter amount of DFaep is less than one-third of AlexNet and the storage size of DFaep is only 0.3MB, this makes DFaep's training efficiency and time consumption greatly lead AlexNet, and has achieved close to 100% accuracy as shown in Table III. From observation, because the rate of accuracy and amount of parameters can be changed according to the different construction of the DFaep such as depth, dense and kernel size. Therefore, even the input data are the identical, the accuracy will not be proportional to the depth of the model.

C. Validation of Proposed Method

In the next experiment, we have made the comparison of the results of internal and external vehicle driving data as shown in Table III. We have matched the prediction results of driver's gaze zone and the salient point of DR(eye)VE



(a) VGG's Accuracy (b) VGG's Loss

(c) AlexNet's Accuracy (d) AlexNet's Loss

(e) DFaep's Accuracy (f) DFaep's Loss

Fig. 9. The Diagram of Loss and Accuracy of Training and Validation According on the Number of Epoch with Three Models of (a-b) VGG; (b-c) AlexNet; and (e-f) DFaep.

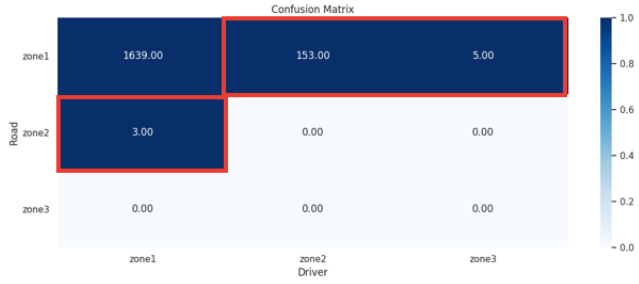


Fig. 10. The Confusion Matrix with the DR(eye)VE Dataset.

database through the confusion matrix (as shown in Fig. 10) for evaluating whether or not the results are consistently as shown in Table III. We compared the prediction results of driver's facial feature inside the vehicle with using external vehicle driving data. Table III presents the comparison results from the data internal and external the vehicle.

"Road" and "Driver" stand for the zone which driver should be attended to and estimation zone of driver, separately. By observing the results of the confusion matrix in Fig. 10, almost

TABLE IV. ESTIMATION ZONES AND ATTENTION AREAS OF DRIVER AND ROAD SCENE

Data	Zone 1	Zone 2	Zone 3	Total
Driver	1787	3	0	1800
Road Scene	1642	153	5	1800

Fig. 11. The Synchronous Comparison of Estimation Zones and Attention Areas.

all the attention zones have concentrated on zone 1. Although the direction of gaze regions is sparse, the method proposed in this research has demonstrated is doable. Later, we obtained the different matched results by confusion matrix for giving appropriate alert to driver as shown in red block of Table IV. These matched gaze zone are mostly concentrate in zone 1. Therefore, we found that it is comply to human driving behaviors in the normal daily driving. In Fig. 11 demonstrates the synchronous comparison of estimation zones and attention areas.

V. CONCLUSION

In this paper, we have proposed a deep learning network to map the driver's attention zone interior of the vehicle environment. For driver gaze zone mapping, driver's face landmark, head position, eye gaze images are captured by the dual low-cost dash cameras. We have performed facial feature tracker with Openface for extracting the images of driver's facial landmark. We have extracted the feature of input image and located the final gaze zone from our network based on the the final score of fully connected layer. The final score of fully connected layer is obtained based on all the extracted features calculated by softmax function to derive the final result. We have compared the losses and accuracies of our idea with three different DNN architecture. Based on the performance of these three models, our DFaep network not only has shown a high scale of accuracy but also performed low level of parameters and storage size. We have demonstrated that such a focus model can detect a large proportion of a driver's focus and is even fast with high acceptable accuracy for detecting distraction and sending a warning or an alert to the driver whenever it is needed. Additionally, we show evidence that such a deep learning network and its trained model provide a feasible way for understanding the driver's attention and focus and making it possible to tame the uncertainty and dynamics of driver's attention and corresponding behaviors.

Empowered by the learnt model trained by the DFaep network, we have conducted multiple test driving. By comparing the estimation zone of driver and the normal attention area of the road scene, warning is issued when the driver has any abnormal driving behavior or distraction during driving (see Fig. 11). From the results, we can see that the method proposed in this study is feasible and effectiveness. The propose network and trained model can be used to offer potential reduction of driving distraction and help drivers be more focus on roads during driving in such a way that road and driving safety can be significantly improved. In the future, we need to further split the zones and improve the accuracy. In our study, we notice that the distribution of driver's attention and focus are highly skewed. To estimate driver attention and focus, the zones defined in this work are needed to be further splitted or refined dynamically by road scene. Moreover, it is noteworthy to introduce road sense semantic segmentation and object detection into our network to estimate driver's interests for better prediction performance.

REFERENCES

[1] Forturn, "2016 was the deadliest year on american roads in nearly a decade?." Available at:http://fortune.com/2017/02/15/traffic-deadliest-year/, 2017.

[2] Icebike, "Real time traffic accident statistics." Available at:http://www.icebike.org/real-time-traffic-accident-statistics/, 2017.

[3] C. for Disease Control and prevention, "Vital signs: Motor vehicle injury prevention — united states and 19 comparison countries." Available at:https://www.textrequest.com/blog/how-much-time-people-spend-mobile-phones-2017/, 2016.

[4] J. Briggs, "How in-dash night-vision systems work." Available at:http://electronics.howstuffworks.com/gadgets/automotive/in-dash-night-vision-system.htm, 2016.

[5] D. A. Owens and M. Sivak, "Differentiation of visibility and alcohol as contributors to twilight road fatalities," *Human Factors*, vol. 38, no. 4, pp. 680–689, 1996.

[6] A. Palazzi, D. Abati, F. Solera, R. Cucchiara, *et al.*, "Predicting the driver's focus of attention: the dr (eye) ve project," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1720–1733, 2018.

[7] J. Xu, J. Min, and J. Hu, "Real-time eye tracking for the assessment of driver fatigue," *Healthcare technology letters*, vol. 5, no. 2, pp. 54–58, 2018.

[8] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th international conference on software engineering*, pp. 303–314, 2018.

[9] K. A. Brookhuis, D. De Waard, and W. H. Janssen, "Behavioural impacts of advanced driver assistance systems–an overview," *European Journal of Transport and Infrastructure Research*, vol. 1, no. 3, 2001.

[10] A. Shaout, D. Colella, and S. Awad, "Advanced driver assistance systems-past, present and future," in *2011 Seventh International Computer Engineering Conference (ICENCO'2011)*, pp. 72–82, IEEE, 2011.

[11] J. Piao and M. McDonald, "Advanced driver assistance systems from autonomous to cooperative approach," *Transport reviews*, vol. 28, no. 5, pp. 659–684, 2008.

[12] A. Taeihagh and H. S. M. Lim, "Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks," *Transport reviews*, vol. 39, no. 1, pp. 103–128, 2019.

[13] Waymo, "We are building the world's most experienced drivertm.." Available at: https://waymo.com/, 2021.

[14] P. LeBeau, "Waymo starts commercial ride-share service," *URL: https://www. cnbc. com/2018/12/05/waymo-starts-commercial-ride-share-service. html*, 2018.

[15] AppleInsider, "Apple car." Available at: https://appleinsider.com/inside/apple-car, 2021.

[16] N. Tech, "Navya shuttle brochure." Available at: https://navya.tech/wp-content/uploads/documents/Brochure_Shuttle_EN.pdf, 2021.

[17] S. Shetty, "Uber's self-driving cars are a key to its path to profitability," 2020.

[18] P. LeBeau, "Relax, experts say it's at least a decade before you can buy a self-driving vehicle," 2019.

[19] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, IEEE, 2018.

[20] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, IEEE, 2016.

[21] M. Carrasco, "Visual attention: The past 25 years," *Vision research*, vol. 51, no. 13, pp. 1484–1525, 2011.

[22] J. L. Orquin and S. M. Loose, "Attention and choice: A review on eye movements in decision making," *Acta psychologica*, vol. 144, no. 1, pp. 190–206, 2013.

[23] T. E. Halverson, *An" active vision" computational model of visual search for human-computer interaction*. PhD thesis, University of Oregon, 2008.

[24] S. E. Gaither, K. Pauker, and S. P. Johnson, "Biracial and monoracial infant own-race face perception: An eye tracking study," *Developmental science*, vol. 15, no. 6, pp. 775–782, 2012.

[25] S. Hutt, C. Mills, S. White, P. J. Donnelly, and S. K. D'Mello, "The eyes have it: Gaze-based detection of mind wandering during learning with an intelligent tutoring system.," *International Educational Data Mining Society*, 2016.

[26] A. J. Hornof and A. Cavender, "Eyedraw: enabling children with severe motor impairments to draw with their eyes," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 161–170, 2005.

[27] M. Chau and M. Betke, "Real time eye tracking and blink detection with usb cameras," tech. rep., Boston University Computer Science Department, 2005.

[28] T. Toyama, T. Kieninger, F. Shafait, and A. Dengel, "Gaze guided object recognition using a head-mounted eye tracker," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 91–98, 2012.

[29] M. Kassner, W. Patera, and A. Bulling, "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pp. 1151–1160, 2014.

[30] D. Li, J. Babcock, and D. J. Parkhurst, "openeyes: a low-cost head-mounted eye-tracking solution," in *Proceedings of the 2006 symposium on Eye tracking research & applications*, pp. 95–100, 2006.

[31] E. Miluzzo, T. Wang, and A. T. Campbell, "Eyephone: activating mobile phones with your eyes," in *Proceedings of the second ACM SIGCOMM workshop on Networking, systems, and applications on mobile handhelds*, pp. 15–20, 2010.

[32] X. Zhang, H. Kulkarni, and M. R. Morris, "Smartphone-based gaze gesture communication for people with motor disabilities," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2878–2889, 2017.

[33] E. Wood and A. Bulling, "Eyetab: Model-based gaze estimation on unmodified tablet computers," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 207–210, 2014.

[34] Y. Wang, S. Lu, and D. Harter, "Eye tracking and learning analytics for promoting proactive teaching and learning in classroom a survey," in *2020 The 4th International Conference on Education and E-Learning*, pp. 156–160, 2020.

[35] Y. Wang, S. Lu, and D. Harter, "Towards a collaborative and intelligent framework for pervasive and proactive learning," *Submit to International Journal of Engineering Education (IJEE)*, 2021.

[36] Y. Wang, S. Lu, and D. Harter, "Multi-sensor systems and infrastructure for capturing student attention and understanding engagement in learning: A review," *Submit to IEEE Sensors Journal*, 2021.

[37] T. Santini, W. Fuhl, D. Geisler, and E. Kasneci, "Eyerectoo: Open-source software for real-time pervasive head-mounted eye tracking.," in *VISIGRAPP (6: VISAPP)*, pp. 96–101, 2017.

[38] E. S. Kim, A. Naples, G. V. Gearty, Q. Wang, S. Wallace, C. Wall, J. Kowitt, L. Friedlaender, B. Reichow, F. Volkmar, *et al.*, "Development of an untethered, mobile, low-cost head-mounted eye tracker," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 247–250, 2014.

[39] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A dual-cameras-based driver gaze mapping system with an application on non-driving activities monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4318–4327, 2019.

[40] D. Tran, H. M. Do, W. Sheng, H. Bai, and G. Chowdhary, "Real-time detection of distracted driving based on deep learning," *IET Intelligent Transport Systems*, vol. 12, no. 10, pp. 1210–1219, 2018.

[41] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[42] NVIDIA, "Designed from the ground up for the largest hpc and ai workloads." Available at:https://www.nvidia.com/en-us/, 2017.

[43] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[44] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 354–361, 2013.

[45] A. Zadeh, Y. Chong Lim, T. Baltrusaitis, and L.-P. Morency, "Convolutional experts constrained local model for 3d facial landmark detection," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2519–2528, 2017.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[47] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5528–5531, IEEE, 2011.

[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[49] A. Karpathy, "Stanford university cs231n: Convolutional neural networks for visual recognition," *url: http://cs231n. stanford. edu/syllabus. html*, 2018.

[50] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *International Workshop on Artificial Neural Networks*, pp. 195–201, Springer, 1995.

[51] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8609–8613, IEEE, 2013.

[52] J. Nuevo, L. M. Bergasa, and P. Jiménez, "Rsmat: Robust simultaneous modeling and tracking," *Pattern Recognition Letters*, vol. 31, no. 16, pp. 2455–2463, 2010.

[53] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The cas-peal large-scale chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 1, pp. 149–161, 2007.

[54] "Berkeley DeepDrive." Available at: https://bdd-data.berkeley.edu/.

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[57] "Keras api reference-utilities-backend utilities." Available at: https://keras.io/api/utils/backend_utils/.

[58] "Tensorflow, "tensorflow/tensorflow," github." Available at: https://github.com/tensorflow/tensorflow.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[60] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186, Springer.

[61] "Keras optimizers.." Available at: https://keras.io/optimizers/.