Received 20 September, 2022; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/OJITS.2022.1234567

Driver Visual Attention Estimation using Head Pose and Eye Appearance Information

SUMIT JHA, STUDENT MEMBER, IEEE, NAOFAL AL-DHAHIR, FELLOW, IEEE, AND CARLOS BUSSO, FELLOW, IEEE

[1]Electrical and Computer Engineering Department, The University of Texas at Dallas, Richardson, TX 75080, USA

CORRESPONDING AUTHOR: Carlos Busso (e-mail: busso@utdallas.edu).

This work was supported by SRC / Texas Analog Center of Excellence (TxACE), under task 2810.014.

ABSTRACT In autonomous, as well as manually operated vehicles, monitoring the driver visual attention provides useful information about the behavior, intent and vigilance level of the driver. The gaze of the driver can be formulated in terms of a probabilistic visual map representing the region around which the driver's attention is focused. The area of the estimated region changes based on the level of confidence of the estimation. This paper proposes a framework based on *convolutional neural networks* (CNNs) that takes the head pose and the eye appearance of the driver as inputs, and creates a fusion model that estimates the driver's gaze on a 2D grid. The model contains upsampling layers to create estimations at multiple resolutions. The model is trained using data collected from 59 subjects with continuous recordings where the subject looks at a moving target in a parked car, and glances at a set of markers inside the car while driving the vehicle and while the car is parked. Our fusion framework provides superior performance than unimodal systems trained exclusively with head pose or eye appearance information. It estimates the gaze region with the target location lying within the 75% confidence region with an accuracy of 92.54%.

INDEX TERMS *convolutional neural networks* (CNN), Driver monitoring systems, Visual Attention, Gaze Estimation

I. Introduction

CURRENT driving systems place the driver at the center of the controls. The driver takes responsibility for ensuring that the entire system runs smoothly. If there is a sudden change in the environment, such as a pedestrian crossing the road, or a vehicle suddenly stopping in front of the ego car, it is up to the driver to take necessary actions to avoid potential accidents. While advances in in-vehicle technologies have helped drivers with safety features, such as collision avoidance, lane assist, and adaptive cruise control, the responsibility is still on the driver. Hence, it is only natural to design systems that can monitor drivers, inferring if they are aware of the driving environment [1]. The transition from manually controlled vehicles to fully autonomous systems is going to be gradual, given the complexity of mixed-autonomy traffic [2]. A semi-autonomous system needs to have synergy between the autonomy of the car and the human driver. When the driving system is unable to make important maneuvering decisions, it must transfer control to the driver. A distracted driver takes longer to resume control of the vehicle [3]. Hence, the system should be able to understand the actions, intents, and behaviors of the driver before transferring control of the car.

Various studies have addressed different aspects of driver attention, such as vehicle state information [4], physiological signals [5]–[7], cognitive distractions [8]–[10] and changes in emotional states [11]. The 100-car naturalistic driving study by the *National Highway Traffic Safety Administration* (NHTSA) concluded that in 80% of crash events and 65% of near-crash events, the driver was looking away from the incoming road just before the event [12]. A driver relies on vision to gather information from the environment, including road signs [13], and pedestrians [14]. Driving tasks such as mirror checking actions [15], [16] and lane change [17], [18] also require the driver's visual attention. Knowledge of the driver's visual attention can be helpful in understanding their intent and their knowledge of the environment [19]. Correlating driver visual attention with visual saliency of the scene can be helpful in gaining insights about what the driver is attending to [20], [21]. This information can be useful in cloning behavioral models onto autonomous systems to replicate safe driving patterns [22]. These examples illustrate the importance of estimating the driver's visual attention, and its potential applications. As the technology evolves towards autonomous vehicles, from level 1 to level 4, the relevance of this task also increases. In non-autonomous vehicles (L1 and L2), the drivers are involved in controlling the operations of the car, so the knowledge of their attention helps in maintaining safe driving conditions. In level 3, it is crucial to have synergy between the driver and the car, as most activities will be shared between them. While the driver might not need to be always attentive, at crucial times during handovers, the system needs to check that the driver is paying attention before transferring control of the car. With full autonomy (L4), the knowledge of a user's visual attention can be helpful for infotainment and navigation systems (e.g., displays information on buildings visually attended by the driver, and resolve ambiguities for road-related driver's commands).

Different studies have tried to use the driver's gaze to estimate distractions. Studies have binarized the problem by considering the duration of gaze-off-the-road events [10], [23]. Other studies have associated head pose and gaze estimation directly with driving activities [4], [17], [24]. An alternative approach is to divide the driver visual attention into gaze zones [25]–[27]. While these methods provide a good coarse estimate of the driver's visual attention, some applications may require a finer estimation of where the driver is directing her/his visual attention. While commercial head-mounted eye trackers such as Tobii and faceLAB can provide an accurate estimate of the driver's gaze, and are useful for research, due to their invasive nature, they provide limited usage in real-world applications. These methods do not provide information about the driver's awareness of particular objects on the road such as other vehicles and pedestrians. Acknowledging the fact that fine details about the driver's gaze may not always be available, we design a model that uses both head pose and eye appearance information to estimate the gaze in real naturalistic driving recordings. We define eye appearance as the image around the eyes of a person.

Visual attention is a broader concept in cognitive science that deals with the study of a person's awareness of the visual world. However, This study restricts our definition of visual attention as the heat map representing the distribution of the driver's gaze (i.e., the area that a driver is looking at while operating the car). We quantify visual attention with a probability map that can be easily projected onto the road, mapping the driving environment with the driver's attention. This formulation also allows us to assign an intensity to the direction of the driver's vision. For example, a driver fixating on a location will have a map with high intensity in a small region. When a driver is exploring the driving environments, the saccade movements will create a map with low intensity in a large area.

In Jha and Busso [28], [29], we proposed that providing probabilistic maps is a practical and effective way to estimate visual attention. Our earlier efforts relied only on the head pose, but driver visual attention is a function of both head pose and eye movement [30]. In this work, we build a fusion model with two branches, each of which takes the head pose and the eye image as inputs and learns a combined representation of the visual attention of the driver. The proposed model contains three parts: a head pose encoder, an eye encoder, and a decoder. The head pose encoder takes the six parameters describing the head pose (position and orientation) as the inputs, which are passed through fully connected networks, followed by reshaping to obtain a low-dimensional map. The eye encoder takes the eye patch image as the input and sends it through a neural network to extract discriminative information. The decoder of our model concatenates the outputs of the encoders, creating a unified feature map that is sent through an upsampling network to obtain the probabilistic map at different resolutions. We classify the driver's visual attention on a 2D grid which is learned by using upsampling with *convolutional neural networks* (CNNs). This representation is purely learned from the data and, hence, is non-parametric. Therefore, it does not require imposing parametric distributions as in our previous model [28], increasing the flexibility of our formulation.

We use the *multimodal driver monitoring* (MDM) dataset [31] to train and evaluate our model. The recordings used for training are a combination of continuous data, where the driver follows a target marker that is moved around in front of a parked car, and discrete data points, where the driver briefly glances at target markers inside the car, while driving the vehicle and while the car is parked. These datasets provide us with a diverse set of data in terms of how the subject approaches a gaze target. We first design simple models that use only one input modality (head pose only or eye appearance only) and compare them with respective baseline models. We observe that our models show superior performance to the baseline models. Our fusion model that takes both head pose and eye appearance as inputs shows better accuracy when compared to a simple model based solely on the head pose or eye appearance. For the eye appearance, we use the face camera and for the head pose we use Fi-Cap labels. We demonstrate potential applications of the model by projecting the probability map onto the road. For example, the visual map that includes 75% of the probability overlaps with the true target point 92.54% of the time.

This paper is organized as follows. Section II discusses related studies in the field of monitoring driver visual attention. Section III describes the portion of the MDM dataset that we use for our proposed model. Section IV describes in detail the proposed model architecture. Section V discusses the experimental settings, including baselines and implementation details. Section VI evaluates our proposed model by comparing its performance in different settings with various baselines. It also discusses projections of the probabilistic gaze map on the road, evaluating the estimation for cases when the subjects were looking at landmarks on the road. Section VII provides the concluding remarks and future research directions.

II. Related Work

Driving is a challenging task requiring a high level of vigilance from the driver. Therefore, many researchers have studied the factors associated with driver behaviors and their effects on vehicle operation [32]. In their review of human behavior in an intelligent vehicle environment, Ohn-Bar and Trivedi [33] noted that multiple studies have analyzed drivers in terms of their intentions, behaviors, and actions for maneuvering control. These studies mostly include the gaze, eye appearance, hands, and head movements of the driver, and their interactions with objects inside the car. This section reviews relevant studies on driver visual attention, emphasizing the relation between head pose and eye appearance for gaze estimation.

A. Driver Visual Attention estimation

Given the importance of visual attention in studying driver vigilance, many studies have proposed methods for using head pose and/or gaze to estimate the driver's visual attention. Dong et al. [34] presented a review of various monitoring systems for driver distraction. They discussed methods based on subjective reports [35]–[37], physiological methods [38]–[41], physical methods [42]–[44], and driving performance measures [45]–[47]. Dong et al. [34] reports many studies that have used the driver's eye and head movements to estimate distraction [42], [48] and fatigue [43], [44]. Tracking eye movement does not require invasive sensors like other physiological signals, such as *electroencephalogram* (EEG) and *electrocardiogram* (ECG). They concluded that using driving performance measures in conjunction with eye and head movements is the most reliable solution for monitoring driver distractions.

1) Relationship between Head Pose and Gaze

The eyes and head move when we glance at a given target location. The gaze-eye relationship is non-trivial, depending on several factors including the cognitive load. Muñoz et al. [49] analyzed the head rotations of a driver during forward glances and glances to the center console of the vehicle. They recognized these two glance patterns using a temporal model based on a *hidden Markov model* (HMM). They observed that head pose is a strong indicator of gaze location. They also observed that there are differences in the patterns observed across individual drivers. Talamonti et al. [50] studied head pose and eye gaze dynamics by asking various subjects to look at different locations in the car in a simulated driving environment. They observed that there is very little head movement when looking at locations such as the odometer and rear-view mirror. However, looking at locations such as the center console and the left mirror require more head movement from the driver. Jha and Busso [30] noted that while there is a strong relationship between head pose and gaze location, the relationship is not one-to-one. They observed that the variability in the gaze changes based on the direction that the driver is directing her/his visual attention to. They also observed a significant difference in gaze patterns when driving a vehicle, where glancing is one of several driving tasks, and when the car is parked, where glancing is the only task and the driver has more time.

2) Using Head Pose to Estimate Driver Visual Attention

Given the strong relation between head pose and gaze, many studies have used head pose to estimate the driver visual attention. Fridman et al. [25] proposed a method to estimate gaze zones with only head pose. They extracted facial landmark features that provide strong cues for head pose. They used these features as inputs to a *random forest* (RF) classifier to categorize the driver's visual attention into seven gaze zones. They trained and evaluated the model on a dataset collected from 50 subjects with two different vehicles. Yuen et al. [51] used a dataset collected in a naturalistic driving setting to perform face localization, landmark detection, and head pose estimation using *deep neural network* (DNN) architectures. They observed higher performance compared to models trained using public datasets collected in indoor settings. Jha and Busso [28] proposed an alternative probabilistic method to model visual attention of a driver. Instead of dividing the visual area into gaze zones, they use heat maps to represent a probability distribution of the driver's gaze conditioned on the head pose. They use *Gaussian process regression* (GPR), which provides a Gaussian distribution of the gaze as a function of the driver's head pose.

3) Incorporating Eye Appearance Information

Head pose provides important information about the visual attention of the driver. However, the estimates are coarse. To provide finer details, many studies have incorporated information from the eyes. Tawari et al. [52] used both head and eye cues to classify the driver's visual attention into preset gaze zones. They extracted high level geometrical eye features from the eye location and used them as inputs along with the head pose into a RF classifier. They observed that adding eye cues increases the performance of their model, reaching 94.9% accuracy when compared to a head pose only model which had an accuracy of 79.8%. Fridman et al. [53] studied the gain in performance of a model when using head pose alone versus when using information about the head pose and eye appearance. They observed that while there is an improvement in performance, this improvement is user specific. They defined a metric called *owlness*, which quantifies how much the driver depends on head movement alone to complete a glance. The study concluded that if a driver's owlness score is low, there is a larger improvement in the model when adding the eye appearance information. Most of these approaches classify the driver's gaze into discrete zones. Bergasa et al. [43] proposed a model based approach for this problem. They estimate factors related to eye closure, head pose, and gaze to determine the drivers level of vigilance. *Near-infrared* (IR) illuminators were used, which create bright reflections on the eye, making it easier to detect the pupil. They used these factors to estimate the driver's fatigue level and inattentiveness using preset rules in a fuzzy logic design. Vora et al. [26] proposed a generalized framework for estimating driver gaze zones using CNNs. They performed an analysis using different models and images with different parts of the face, concluding that a squeezenet model that takes the top half of the face as input provided the best results with an accuracy of 95.18%. Ewaisha et al. [27] proposed a CNN based multitask learning approach that simultaneously performs gaze estimation and head pose estimation. The model performs regression on both tasks, followed by a clustering step where the data samples are assigned to different gaze zones. Performing head pose estimation as an auxiliary task increased the robustness of the model against head pose variations. They achieved 78.2% accuracy on cross-subject evaluations.

B. Application of Visual Attention

Ahlstrom et al. [54] studied the effect of a visual distraction warning system on the driver's behavior. They used a gaze zone classification system to generate a warning when the driver looked away for more than a certain amount of time. They observed that using this warning system reduced the amount of time the driver spent looking off-the-road, even when involved in a secondary distracting activity. Liang et al. [55] analyzed the crash risk associated with driver gaze variables (i.e., glance history, glance duration, and glance location). They analyzed 24 different algorithms that estimate risk and concluded that the off-the-road glance duration was the most important factor for estimating crash risk. Li and Busso [10] studied various multimodal features from the CAN-Bus data, road scene, and driver's face to estimate the perceived visual and cognitive distraction during driving. The distraction space formed by visual and cognitive distractions was divided into four clusters and a classifier was trained to distinguish between these classes using multimodal features. They noted that the model classified the data as distracted when the drivers were performing secondary tasks characterized by high cognitive and/or visual load.

Sodhi et al. [56] analyzed the eye positions and pupil diameter using an eye tracking device to study the effect of a secondary task on the driver's visual attention. They observed an increase in the amount of off-the-road fixation because of competing visual demands on tasks. They observed that the impact varied based on the task. For example, tuning the radio required more off-the-road fixation compared to checking the odometer. They also noted that cognitive tasks decreased the amount of eye movement, since the standard deviation for the fixation displacement reduced when performing mental calculations. Similarly, Reimer et al. [9] studied the effect of cognitive demand on visual attention. They used gaze concentration as a metric, which is a measure of reduction in gaze variability. As the amount of cognitive demand increased, they found that the amount of gaze concentration also increased, implying that drivers tend to fixate at a point when cognitively distracted. They observed that while the amount of gaze concentration increased when moving from low to moderate difficulty tasks, the difference was less defined than the difference from moderate to high difficulty tasks. Martin and Trivedi [17] used the gaze dynamics of the driver to estimate lane based maneuvers in a real world driving scenario. They designed models that could estimate lane change 600 milliseconds before the event with an accuracy of 85%.

There has also been interest in estimating visual saliency based on the road scene to estimate where the driver is expected to look. Palazzi et al. [20] collected the Dr(eye)ve dataset that uses an eye-tracker to track the driver's fixation on the road view. They used this data to train a temporal model that estimates the visual attention of the driver based on the road scene image. They trained a three branch network that takes the original RGB image, the optical flow image, and a semantic segmentation from the scene view to estimate the focus of attention. Lv et al. [21] improved this model by using reinforced attention. This method uses deep reinforcement learning as a regulatory mechanism that increases the density of the estimation.

C. Relation to Prior Work

Our model is inspired by the model proposed in our previous conference papers [29], [57]. In Jha and Busso [29], we proposed a method that uses CNN with upsampling to create a 2D grid representation of the visual attention learned from the head pose. This model has important limitations, since we exclusively rely on head pose without considering eye information. As a result, the model created large maps of visual attention with low certainty. Additionally, the model was only trained on discrete gaze points which made the model overfit to those specific locations. In Jha and Busso [57], we proposed a method to estimate gaze maps from eye patch images using CNN with maxpooling and upsampling. This model was trained with the MSP-Gaze corpus [58], which was collected in an indoor laboratory setting. Therefore, the model did not address the challenges observed in naturalistic driving conditions [59].

In this study, we propose a model that incorporates both head pose and eye appearance with a novel encoder-decoder

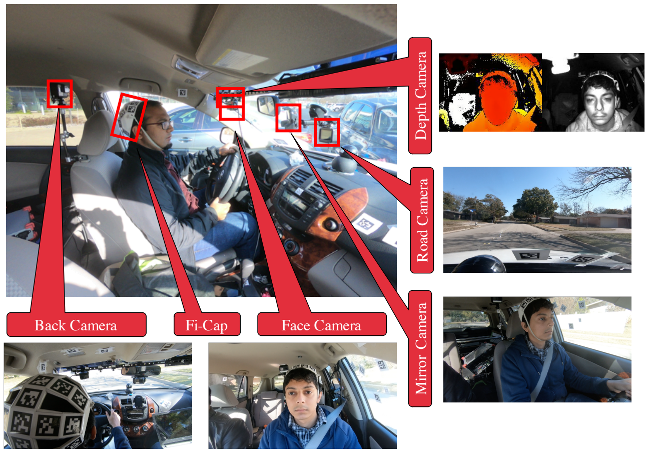


FIGURE 1. **Sensors included in the MDM corpus. This study relies on the face camera. It also uses the back camera to estimate the head pose of the drivers.**

approach that relies on downsampling and upsampling with CNN. We also use data collected with continuous gaze sequences and with discrete gaze locations making the data more representative of gaze in a vehicle. This approach provides better representation, leading to a model that can estimate a small area of gaze region with high accuracy. The contributions of this study are:

- We propose a principled gaze detection approach to fuse head and eye movements. The approach uses two encoders that take inputs from the head pose and eye appearance and fuses the representations with a single decoder that generates the visual map at different resolutions.
- We propose a loss function that uses marginal distribution in the horizontal and vertical directions with Gaussian filtering so that estimations which are further away from the ground truth get higher penalties.
- We create a loss term for each resolution and optimize them in parallel to make sure that the correct representation is learned at each resolution.
- We conduct an exhaustive evaluation using natural driving recordings, demonstrating the performance of the proposed fusion approach, which achieves better performance than alternative baseline methods.
- The approach is trained such that it works even when one modality is missing, making this approach more practical for real world applications.
- We demonstrate the potential of the proposed approach by projecting the estimated probabilistic gaze map onto the road to identify gaze targets outside the vehicle.

III. Database

We use the *multimodal driver monitoring* (MDM) dataset [31] to train and evaluate our model. The MDM dataset is a real world driving dataset collected with 59 subjects, where the drivers are asked to perform various actions in a parked car as well as while driving. The objective of the dataset is to collect naturalistic data with reliable ground truths for the head pose and gaze of the driver. Each subject wore the Fi-Cap helmet [60] during the data collection, which is a cap like structure that contains 23 AprilTags that can be easily tracked in an image (Fig. 1). By tracking these AprilTags, we can establish reliable information about the driver's head pose. Figure 1 shows the sensors used in the data collection. The data is recorded using four GoPro RGB cameras: frontal face, profile face (near rearview mirror), back, and road. The data also includes a PMD picoflexx depth camera, which is not used in this study. This depth camera uses *time-of-flight* (TOF) technology. This dataset provides annotation of the driver's gaze in the 3D space in a variety of situations. Since the data is collected in both driving and parking settings, it provides a diverse set of data to train robust driver-independent models. Another reason for using the MDM corpus for our experiments is the labels available in the corpus for the gaze and head pose.

A. Protocol

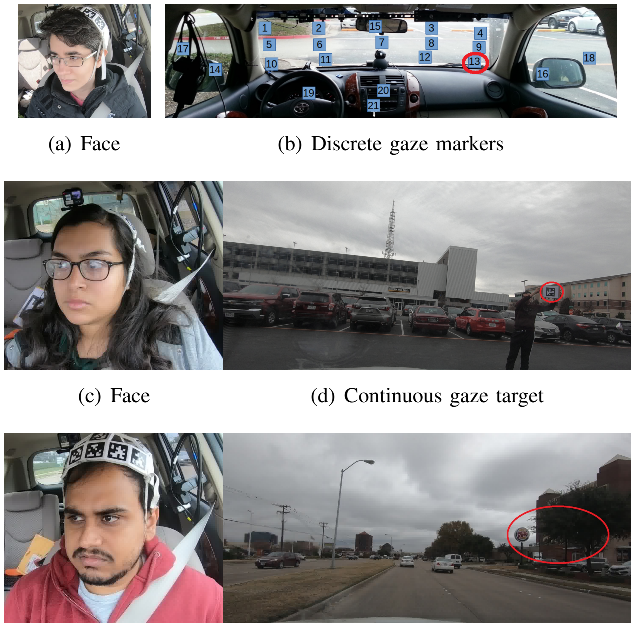
While the data collection protocol involved multiple primary and secondary tasks, we limit our discussion in this paper to the sections that are relevant to the current study. The readers are referred to Jha et al. [31] for more details about the corpus.

1) Discrete Gaze Markers

A goal from this corpus is to have data to train gaze-based models. There are 21 differently numbered markers placed at different locations inside the car (Fig. 2(b)). The numbered markers are placed on the following locations: 1 to 13 on the windshield, 14 to 16 on the mirrors, 17 and 18 on the left and right windows, respectively, 19 on the speedometer panel, 20 on the dashboard, and 21 near the gear of the car. The subjects are asked to look at each of the numbered markers in a random order multiple times. For example, Figure 2(a) shows an example of a subject looking at marker number 13. This step is repeated for two conditions: when the car is parked, and when the participant is driving on a straight road. The location of each of the markers with respect to the back camera is known, which is used as the ground truth gaze location for the frames when the subject is looking at these target markers.

2) Continuous Gaze Target

The corpus also includes continuous gaze data collected while the vehicle was parked. In this step, a researcher conducting the experiment holds a large board with an AprilTag [61] printed on it. The researcher walks in front of the vehicle (Fig. 2(d)). The AprilTag is used so that its 3D location can be tracked from the road camera image. The researcher moves this target and the subject is asked to follow the target with her/his gaze. The data is collected in 3 to 5 sessions of about one minute each. This part of



(a) Face (b) Discrete gaze markers

(c) Face (d) Continuous gaze target

(e) Face (f) Gaze target landmark

FIGURE 2. Protocol to collect ground truth for gaze data. (a) Discrete gaze markers, where drivers are asked to look at markers inside the car, (b) continuous gaze target, where drivers are asked to follow a board held by a researcher when the car is parked, and (c) gaze target landmark, where the driver is asked to look at landmarks outside the car.

the recordings provides us with continuous data with ground truth gaze annotated for each subject. This protocol cannot be implemented when the driver is operating the vehicle, resulting in limited diversity in terms of appearance and changes in illumination. However, it provides very rich data to be augmented with the marker data.

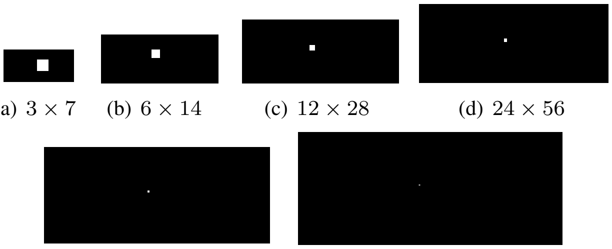
3) Gaze Target Landmark

In this part of the protocol, the subjects are asked to look at various landmarks on the road and answer questions about them (Fig. 2(f)). For example, we ask them to identify stores on the side of the road. This data is collected when the subject is driving. The landmark location is captured on the road camera. Since the target information is limited to the 2D projections on the road, we use this data to validate our model in real world scenarios by projecting our model estimation onto the road (Sec. F).

B. Data Preparation

The proposed model takes as input the head pose of the driver and an image of her/his eyes to generate a 2D grid that represents the horizontal and vertical gaze directions.

The head pose is obtained using the Fi-Cap reading. While the model can work with head pose obtained from automatic computer vision algorithms, we use the Fi-Cap for this purpose because of its reliability in providing head pose in all six degrees of freedom for almost all the frames (position and orientation). The Fi-Cap is in the back of the head so it does not occlude important facial features. Hence, an accurate head pose estimation algorithm that relies on facial images will also work on the MDM dataset. Using the Fi-Cap, the head orientation angles are obtained using the multiple local reference frame calibration framework explained in Hu et al. [62]. These angles are with respect to the depth camera that is placed alongside the face RGB camera. Local reference frames are picked from each of the videos with near-frontal face orientation. The relative rotations of these frames from a global reference frame are calculated using the *iterative closest point* (ICP) algorithm. The angles for each frame are calculated based on the rotation of the Fi-Cap with respect to these local reference frames. The position of the head is approximated with the position of the center of the Fi-Cap. The back camera (Fig. 1) is used as the reference location to ensure that the gaze targets and the face lie on the same side of the coordinate system.



(a) 3×7 (b) 6×14 (c) 12×28 (d) 24×56

(e) 48×112 (f) 96×224

FIGURE 3. Example of the process to convert the ground truth gaze angles into grid. The figure shows the maps as we increase the resolution from 3×7 to 96×224. The true angles in this example are $\theta = -0.15$ and $\phi = 0.21$.

We obtain the eye patch from the *face alignment network* (FAN) algorithm [63] on images captured by the face camera. The landmarks surrounding the eye region are used to create a bounding box for the eyes. We add an extra margin by including 10 to 40 pixels around the eye, picking the actual number at random. The image is resized to 112×224 pixels while maintaining the aspect ratio by adding blank spaces in the boundaries of the larger dimension. We add random augmentations such as scaling, illumination variation, noise, and compression to increase the variability in the images.

The ground truth gaze location during discrete gaze is obtained using the absolute location of the markers, which is estimated before the recordings. The ground truth location during continuous gaze is obtained using the location of the AprilTag tracked with the road camera. The locations are all transformed into the back camera coordinate system. The gaze vector ($\mathbf{g}$) is obtained by connecting the head location ($\mathbf{h_{pos}}$) and the target gaze location ($\mathbf{t_{gaze}}$). The horizontal and vertical gaze angles are obtained from the vector $\mathbf{g} = [g_x, g_y, g_z]$ using the following equations:

$$\mathbf{g} = \mathbf{t_{gaze}} - \mathbf{h_{pos}} \tag{1}$$

$$\hat{\mathbf{g}} = \frac{\mathbf{g}}{\|\mathbf{g}\|} \tag{2}$$

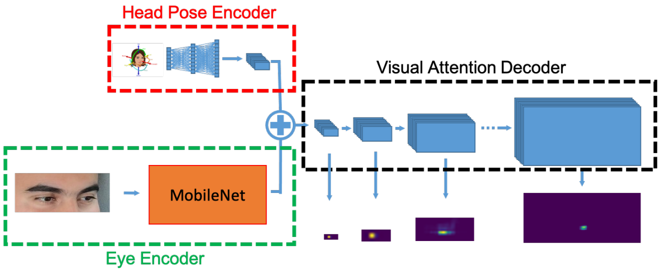


FIGURE 4. **Our proposed architecture to fuse head pose and eye information to estimate visual attention. The model contains three parts: head pose encoder that takes 6D head pose information, eye encoder that takes an eye patch image from the face camera**

, and visual attention decoder that generates visual attention maps at multiple resolutions.

$$\theta_{gaze} = \arctan\left(\frac{\hat{g}_x}{-\hat{g}_z}\right) \tag{3}$$

$$\phi_{gaze} = \arctan\left(\frac{\hat{g}_y}{\sqrt{\hat{g}_x^2 + \hat{g}_z^2}}\right) \tag{4}$$

The gaze angle values are truncated within the range of $-1.75 \leq \theta_{gaze} \leq 1.75$ and $-0.5 \leq \phi_{gaze} \leq 1.5$. The resulting gaze angles are binned into 2D grids to create maps of different resolutions (3×7, 6×14, $\ldots$, 96×224), which are used to train our model. Figure 3 shows an example for the angle $\theta = -0.15$ and $\phi = 0.21$.

IV. Proposed Model

Our proposed approach uses the driver head pose and eye information to obtain a map representing the visual attention of the driver. Figure 4 shows the entire model, which consists of three blocks: the head pose encoder, the eye encoder and the visual attention decoder. The motivation behind this design is to obtain a common representation of the visual attention from both head pose and eye information at low resolution and then gradually upsample to refine the information. The head pose encoder and the eye encoder take the head pose and eye patch information as inputs, respectively, and generate multiple 2D maps that are concatenated. The visual attention decoder uses upsampling to create high resolution 2D maps that represent the direction of the driver visual attention. This section presents these blocks in details.

A. Head Pose Encoder

The goal of the head encoder is to transform the head pose information into a tensor, which is used to upsample the visual attention representation. We have observed good performance by using CNNs to upsample the representation obtained from the head pose [29]. For this study, the CNN-based upsampling is conducted by the visual attention decoder (Sec. C). The head pose information, represented by a six dimensional vector, is passed through two *fully connected* (FC) layers. The first FC layer has 512 nodes and the second FC layer has 672 nodes. The output of the FC layers is a 672D feature representation that is reshaped to transform this representation into a $3 \times 7 \times 32$ tensor. This tensor is then used to obtain 2D maps to represent the visual attention of the driver.

While predicting a high-resolution 2D map from just six numbers may appear as an under-defined problem, we can effectively represent this gaze distribution by training the method per stage, presenting the target ground truth gaze area at different resolutions, as demonstrated by the experiments in Section A.

B. Eye Encoder

The appearance of the eyes, including the relative position of the iris, gives valuable information about the target gaze direction [57], [58], [64], [65]. Our goal is to incorporate this information into the model. The eye encoder takes the eye image and generates a tensor which serves as a representation to estimate the gaze area. We can use several networks to obtain a discriminative feature representation from the eye patch provided by the FAN algorithm. This study relies on the MobileNet network [66], without the FC layers. This network relies on depth-wise separable convolutions, which provide a competitive performance with a reasonable size. The architecture has been found useful in various computer vision tasks such as image classification, detection, and segmentation. The output of the eye encoder is a representation with dimension $3 \times 7 \times 1,024$ (e.g., 1,024 channels with a 3×7 resolution).

C. Visual Attention Decoder

The 3×7 maps obtained from the head pose and the eye encoders are concatenated, forming a $3 \times 7 \times 1056$ tensor, which is used as input to the decoder network. The decoder network consists of convolution layers along with upsampling operations to increase the resolution of the probabilistic visual map. Figure 5 shows the structure of the upsampling block, which consists of two convolutional layers with a 3×3 kernel. The number of filters, n, varies according to the layer. The convolution function is followed by batch normalization and dropout. We also add a shortcut connection that bypasses the convolution function to enable residual learning. The output before the upsampling operation is passed through a single 3×3 convolution kernel followed by a softmax operation to generate the output. This approach has also several advantages including the flexibility to define the resolution of the mask by adding or subtracting convolutional layers. For our experiments, we run six stages of upsampling as we observe that the performance saturates at this point.

D. Loss

Figure 6 illustrates the approach that we use to estimate the loss. The loss is separately calculated in the horizontal and vertical directions. This approach helps in reducing the number of operations by allowing us to obtain losses in

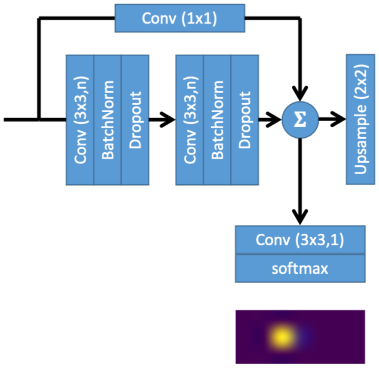


FIGURE 5. Detailed description of one upsampling block in the visual attention decoder.

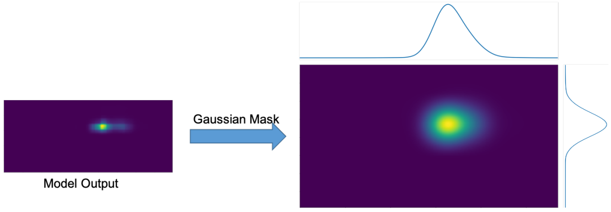


FIGURE 6. Post processing of the model output and ground truth to estimate loss function. The model separately estimates the losses for the horizontal and vertical axes.

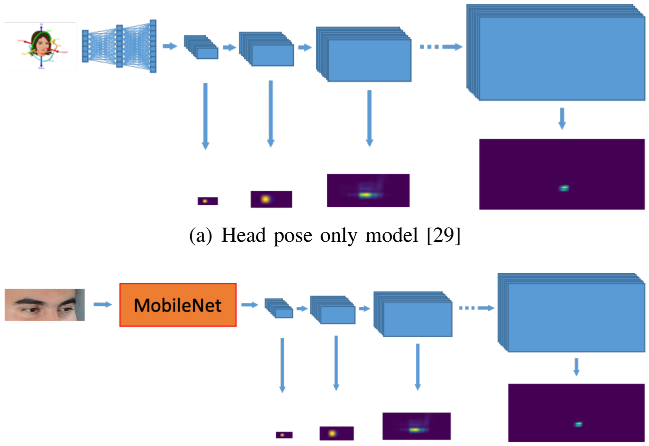
two separate 1D spaces (horizontal and vertical directions) instead of looking at each point in a common 2D space. The visual map output and the ground truth are first filtered with a Gaussian mask to create a neighborhood of influence around each point. This approach starts rewarding the model when the estimations are getting spatially closer to the ground truth value, facilitating the learning process (i.e., overlap between the estimated and Gaussian masks). Then, the marginal distributions in both the horizontal and vertical directions are calculated by adding each column and row, respectively. The final loss is the sum of the cross entropy and the mean absolute error loss between the estimated and the ground truth vectors in each direction,

$$L_h = \sum_{l=0}^{5} cc(p_{true(hr)}, p_{pred(hr)}) + mae(p_{true(hr)}, p_{pred(hr)}) \tag{5}$$

$$L_v = \sum_{l=0}^{5} cc(p_{true(vr)}, p_{pred(vr)}) + mae(p_{true(vr)}, p_{pred(vr)}) \tag{6}$$

$$L = L_h + L_v \tag{7}$$

where, $p_{true(hr)}$ and $p_{pred(hr)}$ are the marginal distributions of the ground truth and the estimated vectors in the horizontal direction, and $p_{true(vr)}$ and $p_{pred(vr)}$ are the marginal distributions of the ground truth and the estimated vectors in the vertical direction. The term $cc(true, pred)$ is the cross entropy function between the ground truth and the estimated vector and the term $mae(true, pred)$ is the mean absolute difference between the ground truth and the estimation vector.



(a) Head pose only model [29]

(b) Eye only model [57]

FIGURE 7. Network architecture for two baselines used to evaluate our proposed approach.

V. Experimental Settings

A. Baseline Models

We compare our approach with implementations of our approach relying on either head pose or eye appearance. We also compare with competitive alternative approaches for systems implemented with either head pose or eye appearance information.

1) Upsampled Neural Networks Using Head Pose

This model follows the same architecture as our main model without the eye encoder. The model estimats the visual attention of the driver with only the head pose. Figure 7(a) shows the architecture, which follows a similar structure as the one presented in our preliminary work [29]. We refer to this method as *HP upsample NN*.

2) GPR Model with Head Pose Input

To compare our models with regression-based methods, we train a *Gaussian process regression* (GPR) model to study the performance of the HP upsample NN. The model takes the head pose as the input, and provides Gaussian distributions of the horizontal and vertical gazes as the output [28]. The model uses a linear basis function with a squared exponential kernel function with automatic relevance determination. Equation 8 shows the expression for the kernel function, where σ_f represents the amplitude defining the autocovariance of the data point and l_i represents the length scale parameter which determines the covariance in each dimension with respect to the distance between the points. Jha and Busso [28] provides the details of this implementation.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{i=0}^{d}\frac{\|x_i - x_i'\|^2}{l_i^2}\right) \tag{8}$$

3) Upsampled Neural Networks Using Eye Appearance

We train an eye only model without the head pose encoder. Figure 7(b) shows the architecture of this baseline model, which is similar to the one presented in our preliminary work [57]. We refer to this method as *eye upsample NN*.

4) Regression Model With Eye Image Input

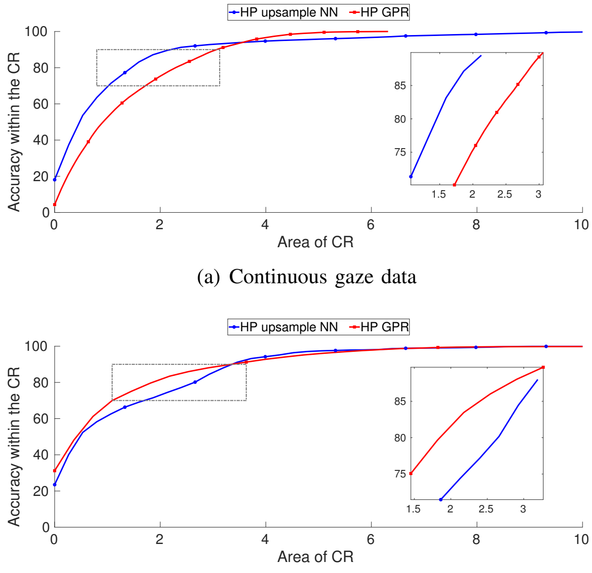
We train a regression model to compare the performance of the eye upsample NN. For this purpose, we design a model that uses the same architecture as the eye decoder. The output is then connected to a *global average pooling* (GAP) layer followed by a fully connected layer to give the horizontal and vertical gazes as the outputs. The mean squared error on the development set is taken as the variance of the model to construct a Gaussian distribution of the gaze around the mean value estimated by the model.

B. Implementation Details

All the CNN based models are trained using Tensorflow. We use a subject independent partition for training, development and testing. Out of the 59 subjects, data from 39 drivers are used for the train set, data from 10 drivers for the development set, and data from 10 drivers for the test set. Each partition is balanced in terms of gender and whether or not the subjects are wearing glasses. The training data consists of both the continuous gaze set (Sec. 2) as well as the discrete gaze set (Sec. 1). Since there are more samples in the continuous part, we oversample the discrete gaze set by five. We obtain the final training set by combining both sets. We use the ADAM optimizer with a learning rate set to 10^{-3}. The models were trained on an NVIDIA RTX 2080 Ti.

VI. Experimental Results

This section presents the extensive evaluations conducted to study the proposed model, which we refer to as *fusion upsample NN*. Since our model estimates a *confidence region* (CR) of visual attention as opposed to a single gaze location, we study the accuracy of our model as a function of the area within which the accuracy is obtained. A model can have good accuracy, but poor spatial resolution (e.g., a confidence region that includes the entire windshield). It can also have a good spatial resolution, but poor accuracy (e.g., a small confidence region that is incorrectly located). The ideal case is when a model achieves high accuracy within a small area. We present the performance with curves describing the tradeoff between accuracy and spatial resolution. The accuracy is measured by estimating the proportion of the gaze targets in the test set that are inside of the estimated confidence region. Since the estimations are in the horizontal and vertical angles, the spatial resolution is represented as a fraction of a sphere $(360°(\theta) \times 180°(\phi))$. As we increase the spatial resolution, we increase the accuracy as more target gazes are included. In the curves, the most relevant parts are small areas with high accuracy. We provide a zoomed snippet of these regions in Figures 8 to 14 to compare the results in more detail.



(a) Continuous gaze data

(b) Discrete gaze markers

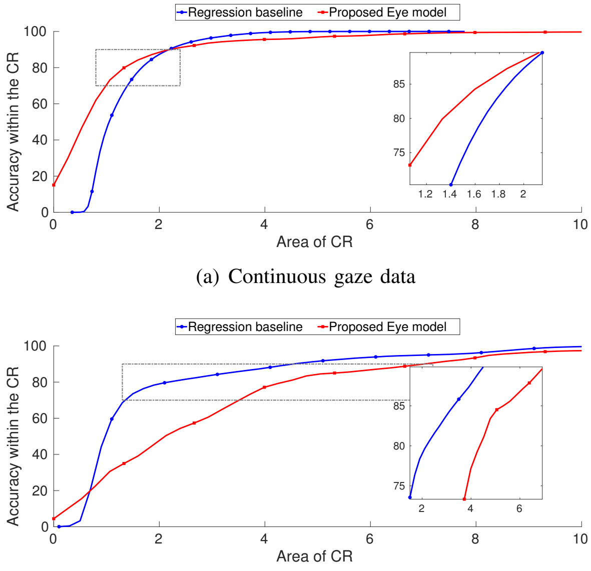
FIGURE 8. Comparison of the head pose only model using the upsampled neural network (Fig. 7(a)) and the GPR model. While the models are trained with discrete and continuous gaze data, we separately report the performance of the models in the test set for both sets.

The continuous data and the discrete marker data pose different sets of challenges in the estimation. The continuous gaze target cover less range in terms of angles, because the recordings are limited to the front of the vehicle where the road camera can see the marker. In the vertical direction, the range is further limited by the extent to which the researcher can move the marker. The discrete markers have higher range, as we have placed the target markers on the side windows, mirrors and the gear of the car. However, the points are grouped only around the markers, providing a more sparse coverage of possible gaze directions. For this purpose, we separately report our test results for the continuous and the discrete gaze markers. The training, however, is done on a combination of both datasets, with the exception of the results in Section E.

A. Performance of the Proposed Approach

Before we present the results of our proposed gaze model, we present the models implemented with only head pose or eye appearance data. We compare these models with alternative baselines.

We start our analysis with the performance of the head pose only model. For this purpose, we compare the HP upsample NN with the GPR model. Figures 8(a) and 8(b) show the performance for the continuous gaze targets and the discrete gaze markers. In the continuous gaze targets, we observe that the HP upsample NN shows an improvement in performance compared to the GPR model. Looking at



(a) Continuous gaze data

(b) Discrete gaze data

FIGURE 9. Comparison of the eye only model using the upsampled neural network (Fig. 7(b)) and the regression models using eye appearance. While the models are trained with discrete and continuous gaze data, we separately report the performance of the models in the test set for both sets.

the discrete gaze markers, while the performance of the two models are very close, the GPR model has a slightly better performance. The GPR model is a regression function that learns a representation of the gaze distribution, while the HP upsample NN is a classification function that learns the target distribution purely based on data. Therefore, the model learns a rich representation in the continuous space where the gaze distribution is denser but with a limited range. The HP upsample NN achieves a non-parametric map that does not make assumptions about the distribution of gaze. This architecture also helps us in obtaining a representation that can be fused with the eye patch information to obtain a single model. In contrast, the GPR model learns better when the data is limited and the learning depends on the extrapolation of the available data. The parametric nature of the GPR model is suitable for this case.

We also analyze the model using only eye appearance information. We compare the eye upsample NN with the regression model with eye image input. Both models have an identical encoder architecture. Figure 9 shows the result of the evaluation. We observe a clear performance gain in continuous gaze targets (Fig. 9(a)) when the area of CR is under 2%. The regression model performs better for the test data associated with the discrete gaze markers (Fig. 9(b)). The eye upsample NN model shows poor performance for discrete gaze data because of the spatial sparsity of the data points, where the data is centered around few markers.

After evaluating the models trained with either head pose or eye appearance, we consider our proposed approach that combines both types of information (fusion upsample NN).

TABLE 1. Accuracy for all the models for a given confidence region (1%, 2% and 4% of the sphere surrounding the driver's head).

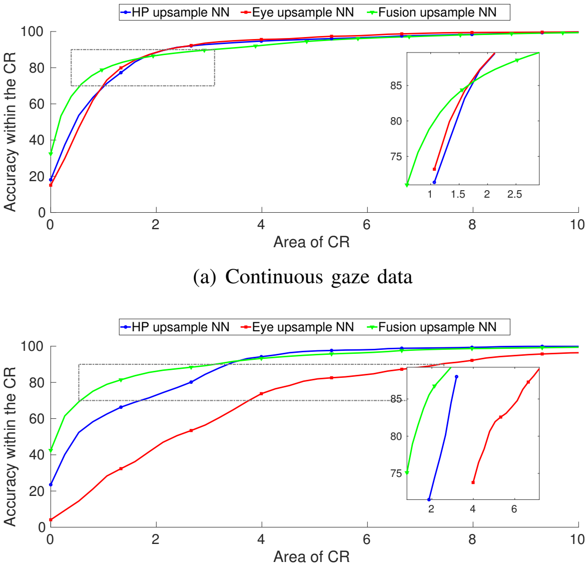
Method	Continuous gaze data			Discrete gaze data		
	1% [%]	2% [%]	4% [%]	1% [%]	2% [%]	4% [%]
HP GPR	53.22	74.95	96.73	69.99	83.46	92.78
HP upsample NN	71.33	89.53	94.70	62.55	74.40	**94.20**
Eye regression	48.27	87.35	**99.42**	45.77	78.72	87.21
Eye upsample NN	73.18	**89.61**	95.60	28.30	46.95	73.78
Fusion upsample NN	**78.79**	86.49	92.53	**78.96**	**86.74**	93.28

Figure 10 shows the performance, which demonstrates the clear advantages of combining head pose and eye appearance information. We observe clear improvements for continuous gaze data and discrete gaze data. The head pose can only provide limited information about the gaze. The information provided by the appearance of the eye is also needed, which is clearly observed in the figure. Likewise, head position also provides complementary information about the gaze. The models can better interpret gaze information inferred from the eye appearance by having the location and orientation of the head. Hence, combining both sets of information helps our proposed model provide a better estimation of the driver visual attention.

We compare all five models in Table 1. We obtain the accuracy of each model by fixing the size of the confidence region. We use three sizes corresponding to 1%, 2% and 4% of the sphere around the driver's head. The table separately reports the results for the continuous and discrete data. For the discrete gaze data, the fusion model shows the best performance for all the confidence regions, with the exception of the 4% CR condition, where the HP upsample NN performs marginally better. For the continuous gaze data, our proposed approach shows high accuracy for small CRs (1%). At higher CRs, some of the baseline models show superior performance, particularly the regression model based on eye patch. Overall, Table 1 shows that our proposed approach offers the best tradeoff between accuracy and spatial resolution for small CR, which is appealing for practical applications.

B. Fusion Model Performance at Different Resolutions

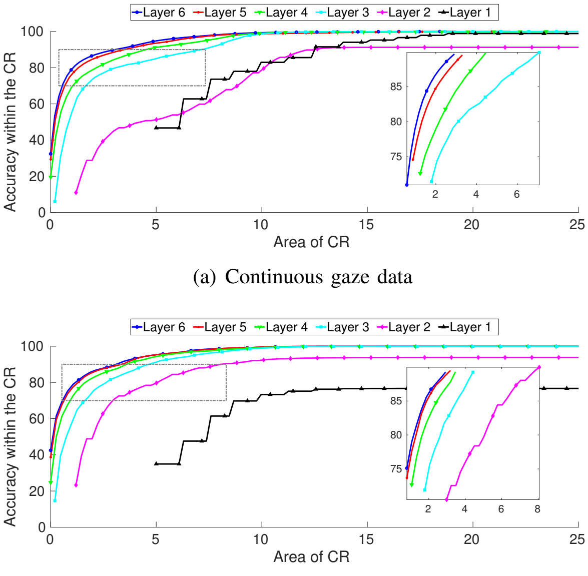
This section discusses the performance of our model at different resolutions. The output can be obtained at any desired resolution, since an output layer is connected after every upsampling stage. The resolution of the output at layer 1 is 3×7. The resolution is doubled after each upsampling step. The resolution at the final layer is 96×224. Figure 11 shows the performance achieved at different layers. The results in the continuous gaze data show that the performance gets consistently better as the resolution increases. The performance saturates at higher resolutions. We observe a reduced improvement when we compare the performances at layer 5 and layer 6. We observe a similar pattern for the



(a) Continuous gaze data

(b) Discrete gaze data

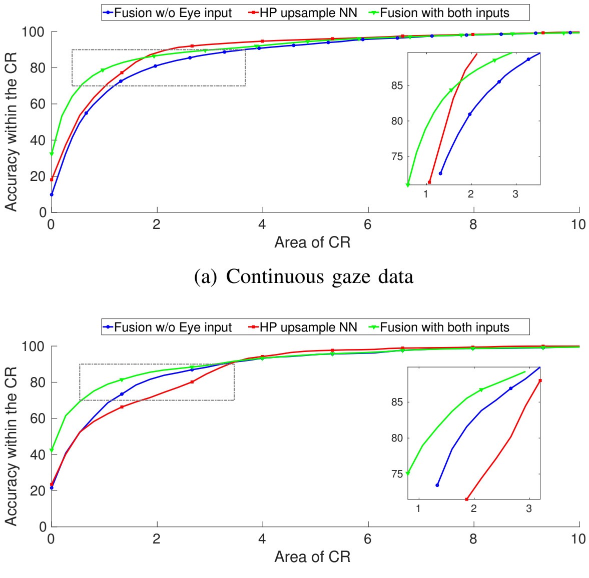
FIGURE 10. Comparison of the proposed gaze model (Fig. 4) that uses head pose and eye appearance information and the upsampled neural networks using either head pose (Fig. 7(a)) or eye appearance (Fig. 7(b)). While the models are trained with discrete and continuous gaze data, we separately report the performance of the models in the test set for both sets.



(a) Continuous gaze data

(b) Discrete gaze data

FIGURE 11. Comparison of the performance of the proposed gaze model (Fig. 4) at different resolutions. While the models are trained with discrete and continuous gaze data, we separately report the performance of the models in the test set for both sets.



(a) Continuous gaze data

(b) Discrete gaze data

FIGURE 12. Comparison of the proposed gaze model (Fig. 4) when the eye information is missing (i.e., missing a modality). While the models are trained with discrete and continuous gaze data, we separately report the performance of the models in the test set for both sets.

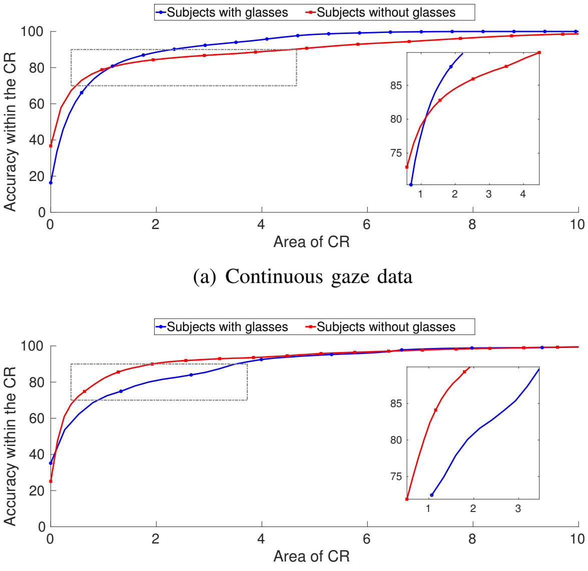
results on the discrete gaze data. The performance saturates, showing similar performances at layers 5 and 6. Since we are looking at discrete points, the output at layer 5 is discriminative enough to provide enough information for gaze estimation. Increasing the resolution even more does not give additional benefits.

C. Performance of Proposed Model Without Eye Information

The fusion model takes both head pose information and an image of the eyes. The eye patch detection might not work in cases when the illumination is not ideal or the head pose is extreme, which are common problems in naturalistic driving conditions [59]. In these cases, our model takes a blank image as input and only uses the head pose to provide information about the visual attention. In this situation, we expect the model to perform with a similar accuracy as a model trained with only head pose information. Figure 12 compares the performance of the proposed model when only head pose is available. For comparison, we include the performance of the model trained with only head pose (Fig. 7(a)), and the proposed approach evaluated with both inputs (head pose and eye appearance information). We observe that the model shows similar performance when we black out the eye input as the model trained with only head pose. For the discrete gaze data, Figure 12(b) shows that the proposed approach with missing eye information can achieve even better performance than the model trained exclusively with head pose information. Figure 12(a) shows that there is small difference between both conditions for the continuous gaze data.

D. Effect of Wearing Glasses on the Performance

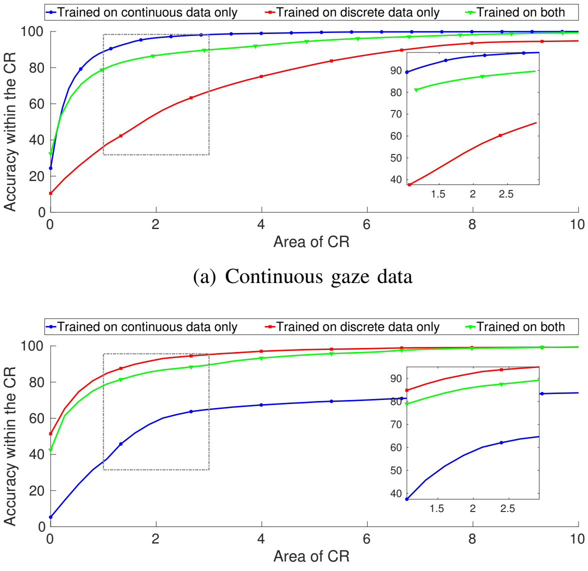
Jha and Busso [59] discussed that the use of glasses is another important challenge in naturalistic conditions. The presence of glasses can affect the model, as our model



(a) Continuous gaze data

(b) Discrete gaze data

FIGURE 13. Analysis of the performance of the proposed approach for data collected from subjects wearing glasses and subjects without wearing glasses. While the models are trained with discrete and continuous gaze data, we separately report the performance of the models in the test set for both sets.



(a) Continuous gaze data

(b) Discrete gaze data

FIGURE 14. Analysis of the performance of the proposed approach trained with either the discrete or continuous gaze data. The performance drops for mismatched train and test conditions, indicating the need for using discrete and continuous gaze data to train the model.

depends on the appearance of the eyes. Five out the ten subjects included in the test set wore eye glasses. This section analyzes the differences in performance observed when the subjects wore or did not wear glasses. Figure 13 shows that the model works slightly better with subjects who did not wear glasses. This difference is minimum because our training data also contain a mix of subjects with and without glasses. Therefore, our models learn the representations well. Methods that compensate for challenges caused by glasses [67] can potentially improve the performance of our model.

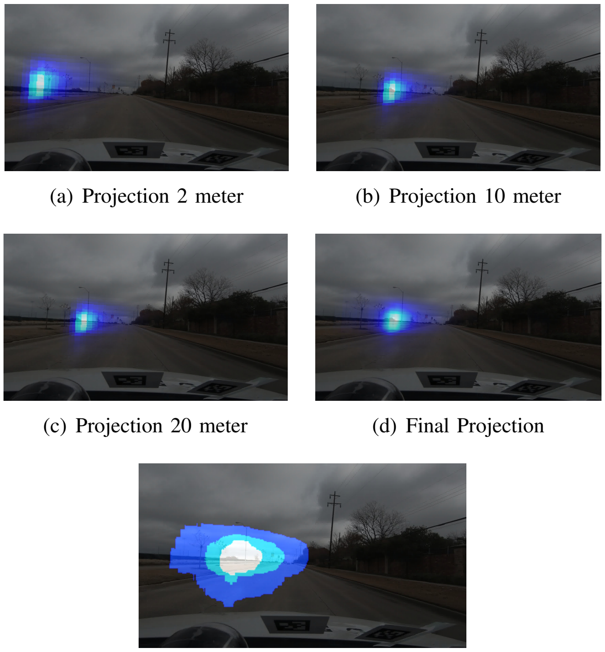
E. Training with Matched and Mismatched Datasets

Our model is trained with data collected using two different data sets: the continuous gaze data and the discrete gaze data. For all the previous results presented in this paper, the training set includes data collected from both sets. This section evaluates the performance of the models when training only with the continuous or discrete gaze data. Figure 14 shows the performance of these different models. We observe that the model trained using the discrete gaze data shows low performance on the continuous gaze data on the test set. Similarly, the model trained using the continuous gaze data shows low performance on the discrete gaze data on the test set. These results show that a model trained on a single type of data tends to overfit on similar data. For example, the model trained only on the discrete data tends to estimate only gaze targets around the numbered markers. Similarly, since the continuous data is limited to the frontal region of the windshield, the model fails when the test samples contain data outside the range of its distribution (e.g., looking at the side windows). The performance of the model trained with both datasets comes very close to the model trained and tested with matched datasets. This evaluation demonstrates the need to train our model with both continuous and discrete gaze data. This is not the case in many related studies, which use broad gaze zones or limited target markers [26], [52].

F. Projecting Probabilistic Maps onto the Road

As an example, we illustrate the benefit of our proposed model by projecting the estimated distribution by our model onto the road to correlate the visual attention with targets on the scene. The model estimates the visual attention in terms of angles with respect to the reference of the driver's head frame. Since we aim to project this region onto the camera view, we make some approximations. First we realize multiple planes at different distances from the camera ranging from 2 meters to 20 meters in one meter intervals. Figures 15(a), 15(b) and 15(c) show examples for projections at 2, 10, and 20 meters. For each of these planes, we calculate the angle subtended at the head for each pixel, which helps us in calculating the gaze given by the model. The distributions obtained for each map at different distances are added and normalized to obtain the final map. Figure 15(d) shows the final probability distribution map after combining the results at different planes.

We obtain continuous estimations for data when the subject looks at landmarks on the road, as described in Section 3, for all the test subjects. The output of the model is a 2D probability distribution. Starting from the mean value, we can define a confidence region by increasing the area until reaching a target probability. For example, we can



(a) Projection 2 meter

(b) Projection 10 meter

(c) Projection 20 meter

(d) Final Projection

(e) 50%,75% and 95% probability regions

FIGURE 15. Illustration of the estimated confidence regions projected on the road. The figure shows the projections at 2, 10 and 20 meters, which are combined to create the final projection. The figure also shows example of 50%, 75% and 95% probability regions estimated from the estimated map. The figure is best viewed in color.

TABLE 2. Accuracy of the model estimation when using 50%, 75% and 95% confidence regions.

	50% [%]	75% [%]	95% [%]
Total Accuracy	80.51	92.54	98.16
Horizontal Accuracy	94.49	98.16	100
Vertical Accuracy	85.29	93.38	98.16

define a 50% probability map, where it is equally likely that the gaze is inside or outside this region. Using this approach, we define 50%, 75% and 95% gaze regions for this analysis. Figure 15(e) shows an example of these regions. The 95% gaze region is naturally larger than the 50% gaze region. The estimation region is evaluated against the ground truth landmark that the subject was asked to look at during the recordings. We consider success if there is an overlap between the ground truth target and the visual attention map. We evaluate a total of 272 examples from the 10 subjects in the test set.

Table 2 shows the accuracy at different regions. We observe that the model could correctly estimate the gaze within the 50% region for most examples (80%). Increasing the area to 95% region gave us an accuracy of 98.16%, Only a few outliers were missed in this region. We also explore if the error is caused by the horizontal or vertical gaze estimations. We separately evaluate the performance for horizontal and vertical directions. We observe that the model failed more often to estimate the vertical gaze direction. Our model correctly estimates the horizontal visual attention with an accuracy of 94.49%, using the 50% gaze region. The accuracy increases to 100% when we use the 95% region for the horizontal gaze direction. Figure 16 shows some examples of correct estimation maps projected onto the road. We highlight with a red circle the target gaze location. Figure 17 shows examples where our algorithm fails. Some of the reasons for failure are highly reflective glasses, anomalous gaze patterns, and some distortion caused by our projection method.

VII. Conclusions

This paper proposed a novel architecture to estimate the driver visual attention using probabilistic maps from head pose and eye appearance information. The approach relies on the use of upsampling based blocks implemented with CNNs. We obtained a driver independent representation of the gaze that can be directly obtained from the eye appearance and the head pose of the driver without any calibration. The maps are non-parametric and, hence, purely learned from the data. The approach provides an efficient way to estimate visual attention, and helps in designing an efficient fusion mechanism. The variance of the maps is non-parametric and depends on the gaze direction that is learned directly from the data. Therefore, the size of the confidence regions will vary according to the underlying uncertainty. Our evaluation results showed that models implemented with either head pose or eye appearance information outperformed their respective regression baselines in estimating continuous gaze targets. The fusion model was found better than the models based on single modalities in both continuous and discrete gaze data, showing the complementary information provided by eye images and head pose values. One crucial advantage of the proposed model is that it can provide a coarse estimation of the visual attention when only the head pose is available, adding finer details when the eye appearance information is added. This feature is important when accurate eye patches cannot be reliably obtained due to challenging naturalistic driving conditions. Therefore, this architecture can have many potential applications in active safety systems.

Every vision-based solution for in-vehicle applications needs to be robust against car movements. Vibrations can lead to imaging problems that may interfere with the instant appearance of the face for a given frame (e.g., blurred images, and head movements that are not associated with visual attention). While our approach is sensitive to these issues, we expect that the results reported in this study are representative given that the MDM corpus was collected in a real car capturing common driving conditions. We predict the gaze direction relative to the driver's head location, instead of predicting the absolute gaze location. Therefore, our model is reasonably robust to upper body movement. The robustness

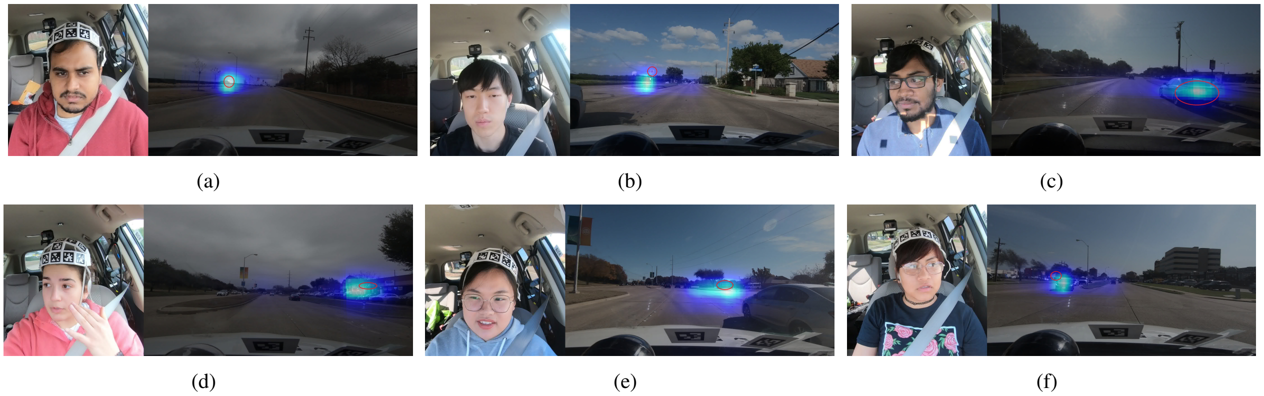


FIGURE 16. **Examples of road projection of the visual attention estimation where the ground truth overlaps with the model estimation. The red circles indicate the target gaze direction. The figure is best viewed in color.**

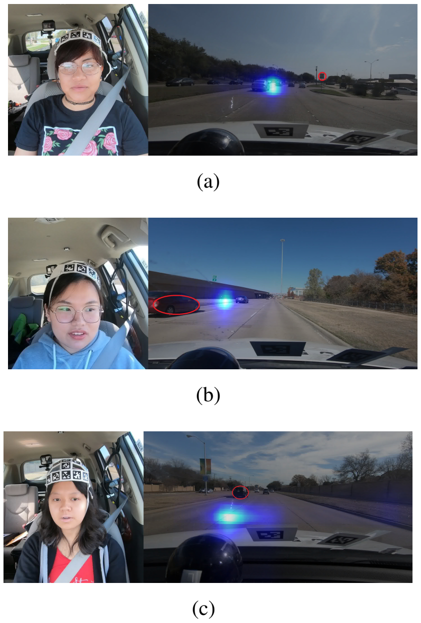


FIGURE 17. **Examples of road projection of the visual attention estimation where the ground truth does not overlap with the model estimation. The red circles indicate the target gaze direction. The figure is best viewed in color.**

shown against glasses provides confidence that our proposed model can deal with some of these challenges.

In this paper, we defined visual attention as a heat map representing the distribution of the driver's gaze. This definition does not capture the broad concept of visual attention (e.g., driving context, saccade movements, time spent looking at a place or object). First, the approach focuses mostly on eye fixations. Saccadic motions are more challenging to track because of the inherent randomness associated with them. Additionally, since the attention is not focused on a specific place, it is challenging to describe the ground truth gaze. Second, the definition does not consider cases of inattention due to cognitive distractions (e.g., seeing but not noticing the driving environment). This contextual information can be added to the attention map provided by the proposed model to obtain a more comprehensive safety system.

There are areas of potential improvement in the algorithm. We observed sub-optimal results for a few subjects. Factors that affect the performance include error in calibration between cameras, appearance variations with subjects wearing highly reflective glasses, and difference in gaze behaviors across subjects. Various methods can be used to improve our algorithm. For example, we can implement adaptation methods to personalize the system, addressing the variability problem across subjects. Some semi-supervised methods can also be used that leverage natural data where the ground truth gaze is not available. Another limitation of our work is that it does not consider temporal information. The discrete gaze marker section of the MDM corpus only considers gaze information when the drivers were looking at the target markers and objects. Since the data only has gaze information for those moments, we do not have continuous gaze information describing the trajectories that led one driver to focus on a particular position. Currently, this option can only be done with the continuous gaze data since the MDM corpus provides gaze labels for only some frames in the discrete gaze data. Lastly, we rely on the Fi-Cap data to obtain the head pose for our method. This limitation can be easily addressed by using automatic head pose estimation algorithms. While current RGB based algorithms do not reliably provide the head pose in all 6 degrees of freedom, the MDM database also has point cloud images collected with a *time of flight* (TOF) camera. This modality was used by Hu et al. [62] to estimate the orientation of the driver's head. This method can be easily enhanced to also estimate the head position of the driver, providing the necessary information our model requires.

REFERENCES

[1] K. Vellenga, H. J. Steinhauer, A. Karlsson, G. Falkman, A. Rhodin, and A. C. Koppisetty, "Driver intention recognition: State-of-the-art review," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 602–616, August 2022.

[2] R. Valiente, B. Toghi, R. Pedarsani, and Y. P. Fallah, "Robustness and adaptability of reinforcement learning-based cooperative autonomous

driving in mixed-autonomy traffic," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 397–410, May 2022.

[3] A. Eriksson and N. Stanton, "Takeover time in highly automated vehicles: noncritical transitions to and from manual control," *Human factors*, vol. 59, no. 4, pp. 689–705, June 2017.

[4] N. Li, J. Jain, and C. Busso, "Modeling of driver behavior in real world scenarios using multiple noninvasive sensors," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1213–1225, August 2013.

[5] Y. Qiu, T. Misu, and C. Busso, "Analysis of the relationship between physiological signals and vehicle maneuvers during a naturalistic driving study," in *Intelligent Transportation Systems Conference (ITSC 2019)*, Auckland, New Zealand, October 2019, pp. 3230–3235.

[6] ——, "Driving anomaly detection with conditional generative adversarial network using physiological and can-bus data," in *ACM International Conference on Multimodal Interaction (ICMI 2019)*, Suzhou, Jiangsu, China, October 2019, pp. 164–173.

[7] ——, "Use of triplet loss function to improve driving anomaly detection using conditional generative adversarial network," in *Intelligent Transportation Systems Conference (ITSC 2020)*, Rhodes, Greece, September 2020, pp. 1–7.

[8] N. Li and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, July 2013, pp. 1–6.

[9] B. Reimer, B. Mehler, B. Mehler, Y. Wang, and J. Coughlin, "A field study on the impact of variations in short-term memory demands on drivers' visual attention and driving performance across three age groups," *Human Factors*, vol. 54, no. 3, pp. 454–468, June 2012.

[10] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 51–65, February 2015.

[11] I. Abdić, L. Fridman, D. McDuff, E. Marchi, B. Reimer, and B. Schuller, "Driver frustration detection from audio and video in the wild," in *International Joint Conference on Artificial Intelligence (IJCAI 2016)*, New York, NY, USA, July 2016, pp. 1354–1360.

[12] T. A. Dingus, S. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. Knipling, "The 100-car naturalistic driving study, phase II - results of the 100-car field experiment," U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC, USA, Technical Report DOT HS 810 593, April 2006. [Online]. Available: https://rosap.ntl.bts.gov/view/dot/37370

[13] L. Simon, J. Tarel, and R. Bremond, "Alerting the drivers about road signs with poor visual saliency," in *IEEE Intelligent Vehicles Symposium (IV 2009)*, Xi'an, China, June 2009, pp. 48–53.

[14] K.-H. Lee and J.-N. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1429–1438, September 2015.

[15] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 980–992, April 2016.

[16] ——, "Driver mirror-checking action detection using multi-modal signals," in *The 6th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems*, Seoul, Korea, September-October 2013, pp. 101–108.

[17] S. Martin and M. M. Trivedi, "Gaze fixations and dynamics for behavior modeling and prediction of on-road driving maneuvers," in *IEEE Intelligent Vehicles Symposium (IV 2017)*, Los Angeles, CA, USA, June 2017, pp. 1541–1545.

[18] G. H. Robinson, D. J. Erickson, G. L. Thurston, and R. L. Clark, "Visual search by automobile drivers," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 14, no. 4, pp. 315–323, August 1972.

[19] A. Doshi and M. M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," in *IEEE Conference on Intelligent Transportation Systems (ITSC 2011)*, Washington, DC, USA, October 2011, pp. 1892–1897.

[20] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "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, July 2019.

[21] K. Lv, H. Sheng, Z. Xiong, W. Li, and L. Zheng, "Improving driver gaze prediction with reinforced attention," *IEEE Transactions on Multimedia*, vol. 23, pp. 4198–4207, 2021.

[22] W. Farag, "Cloning safe driving behavior for self-driving cars using convolutional neural networks," *Recent Patents on Computer Science*, vol. 12, no. 2, pp. 120–127, 2019.

[23] Y. Liang and J. Lee, "Combining cognitive and visual distraction: Less than the sum of its parts," *Accident Analysis & Prevention*, vol. 42, no. 3, pp. 881–890, May 2010.

[24] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! no accident!" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, Columbus, OH, June 2014, pp. 129–136.

[25] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 49–56, May-June 2016.

[26] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 254–265, September 2018.

[27] M. Ewaisha, M. Shawarby, H. Abbas, and I. Sobh, "End-to-end multitask learning for driver gaze and head pose estimation," in *International Symposium on Electronic Imaging (IS&T 2020*, Burlingame, California, January 2020, pp. 110:1–5.

[28] S. Jha and C. Busso, "Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 59–72, January 2023.

[29] ——, "Probabilistic estimation of the gaze region of the driver using dense classification," in *IEEE International Conference on Intelligent Transportation (ITSC 2018)*, Maui, HI, USA, November 2018, pp. 697–702.

[30] ——, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 2016, pp. 2157–2162.

[31] S. Jha, M. Marzban, T. Hu, M. Mahmoud, N. Al-Dhahir, and C. Busso, "The multimodal driver monitoring database: A naturalistic corpus to study driver attention," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 736–10 752, August 2022.

[32] Y. Liu and J. H. L. Hansen, "A review of UTDrive studies: Learning driver behavior from naturalistic driving data," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 2, pp. 338–346, August 2021.

[33] E. Ohn-Bar and M. M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *EEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 90–104, March 2016.

[34] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 596–614, June 2011.

[35] A. Craig, Y. Tran, N. Wijesuriya, and P. Boord, "A controlled investigation into the psychological determinants of fatigue," *Biological Psychology*, vol. 72, no. 1, pp. 78–87, April 2006.

[36] M. Ingre, T. Åkerstedt, B. Peters, A. Anund, and G. Kecklund, "Subjective sleepiness, simulated driving performance and blink duration: examining individual differences," *Journal of sleep research*, vol. 15, no. 1, pp. 47–53, March 2006.

[37] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa, "Validation of the Karolinska sleepiness scale against performance and EEG variables," *Clinical Neurophysiology*, vol. 117, no. 7, pp. 1574–1581, July 2006.

[38] I. Chouvarda, C. Papadelis, C. Kourtidou-Papadeli, P. D. Bamidis, D. K. E. Bekiaris, and N. Maglaveras, "Non-linear analysis for the sleepy drivers problem," in *World Congress on Health (Medical) Informatics; Building Sustainable Health Systems (Medinfo 2007)*, ser. Studies in Health Technology and Informatics, K. A. Kuhn, J. Warren, and T.-Y. Leong, Eds. Amsterdam, Netherlands: IOS Press, 2007, vol. 129, pp. 1294–1298.

[39] C.-T. Lin, Y.-C. Chen, T.-Y. Huang, T.-T. Chiu, L.-W. Ko, S.-F. Liang, H.-Y. Hsieh, S.-H. Hsu, and J.-R. Duann, "Development of wireless brain computer interface with embedded multitask scheduling and its application on real-time driver's drowsiness detection and warning," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 5, pp. 1582–1591, May 2008.

[40] M. Yeo, X. Li, K. Shen, and E. Wilder-Smith, "Can SVM be used for automatic EEG detection of drowsiness during car driving?" *Safety Science*, vol. 47, no. 1, pp. 115–224, January 2009.

[41] B. T. Jap, S. Lal, P. Fischer, and E. Bekiaris, "Using EEG spectral components to assess algorithms for detecting fatigue," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2352–2359, March 2009.

[42] L. M. Bergasa, J. M. Buenaposada, J. Nuevo, P. Jimenez, and L. Baumela, "Analysing driver's attention level using computer vision," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2008*, Beijing, China, October 2008, pp. 1149–1154.

[43] L. Bergasa, J. Nuevo, M. Sotelo, R. Barea, and M. Lopez, "Real-time system for monitoring driver vigilance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 1, pp. 63–77, March 2006.

[44] F. Friedrichs and B. Yang, "Camera-based drowsiness reference for driver state classification under real driving conditions," in *IEEE Intelligent Vehicles Symposium (IV 2010)*, La Jolla, CA, USA, June 2010, pp. 101–106.

[45] T. Ersal, H. Fuller, O. Tsimhoni, J. Stein, and H. Fathy, "Model-based analysis and classification of driver distraction under secondary tasks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 692–701, September 2010.

[46] K. Torkkola, N. Massey, and C. Wood, "Driver inattention detection through intelligent analysis of readily available sensors," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2004)*, Washington, WA, USA, October 2004, pp. 326–331.

[47] A. Sathyanarayana, S. Nageswaren, H. Ghasemzadeh, R. Jafari, and J. Hansen, "Body sensor networks for driver distraction identification," in *IEEE International Conference on Vehicular Electronics and Safety (ICVES 2008)*, Columbus, OH, USA, September 2008.

[48] J. Pohl, W. Birk, and L. Westervall, "A driver-distraction-based lane-keeping assistance system," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 221, no. 4, pp. 541–552, June 2007.

[49] M. Muñoz, J. Lee, B. Reimer, B. Mehler, and T. Victor, "Analysis of drivers' head and eye movement correspondence: Predicting drivers' glance location using head rotation data," in *International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, Salt Lake City, UT, USA, June 2015, pp. 203–209.

[50] W. Talamonti Jr, W. Huang, L. Tijerina, and D. Kochhar, "Eye glance and head turn correspondence during secondary task performance in simulator driving," in *human factors and ergonomics society annual meeting (HFES 2013)*, San Diego, CA, USA, September-October 2013, pp. 1968–1972.

[51] K. Yuen, S. Martin, and M. M. Trivedi, "Looking at faces in a vehicle: A deep CNN based approach and evaluation," in *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 2016, pp. 649–654.

[52] A. Tawari, K. H. Chen, and M. M. Trivedi, "Where is the driver looking: Analysis of head, eye and iris for robust gaze zone estimation," in *International IEEE Conference on Intelligent Transportation Systems (ITSC 2014)*, Qingdao, China, October 2014, pp. 988–994.

[53] L. Fridman, J. Lee, B. Reimer, and T. Victor, "'Owl' and 'lizard': patterns of head pose and eye pose in driver gaze classification," *IET Computer Vision*, vol. 10, no. 4, pp. 308–314, June 2016.

[54] C. Ahlstrom, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 965–973, June 2013.

[55] Y. Liang, J. Lee, and L. Yekhshatyan, "How dangerous is looking away from the road? algorithms predict crash risk from glance patterns in naturalistic driving," *Human Factors*, vol. 54, no. 6, pp. 1104–1116, December 2012.

[56] M. Sodhi, B. Reimer, and I. Llamazares, "Glance analysis of driver eye movements to evaluate distraction," *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 4, pp. 529–538, November 2002.

[57] S. Jha and C. Busso, "Estimation of gaze region using two dimensional probabilistic maps constructed using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 3792–3796.

[58] N. Li and C. Busso, "Calibration free, user independent gaze estimation with tensor analysis," *Image and Vision Computing*, vol. 74, pp. 10–20, June 2018.

[59] S. Jha and C. Busso, "Challenges in head pose estimation of drivers in naturalistic recordings using existing tools," in *IEEE International Conference on Intelligent Transportation (ITSC 2017)*, Yokohama, Japan, October 2017, pp. 1624–1629.

[60] ——, "Fi-Cap: Robust framework to benchmark head pose estimation in challenging environments," in *IEEE International Conference on Multimedia and Expo (ICME 2018)*, San Diego, CA, USA, July 2018, pp. 1–6.

[61] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *IEEE International Conference on Robotics and Automation (ICRA 2011)*, Shanghai, China, May 2011, pp. 3400–3407.

[62] T. Hu, S. Jha, and C. Busso, "Temporal head pose estimation from point cloud in naturalistic driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8063–8076, July 2022.

[63] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *IEEE International Conference on Computer Vision (ICCV 2017)*, Venice, Italy, October 2017, pp. 1021–1030.

[64] N. Li and C. Busso, "Evaluating the robustness of an appearance-based gaze estimation method for multimodal interfaces," in *International conference on multimodal interaction (ICMI 2013)*, Sydney, Australia, December 2013, pp. 91–98.

[65] ——, "User-independent gaze estimation by exploiting similarity measures in the eye pair appearance eigenspace," in *International conference on multimodal interaction (ICMI 2014)*, Istanbul, Turkey, November 2014, pp. 335–338.

[66] A. 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," *ArXiv e-prints (arXiv:1704.04861)*, pp. 1–9, April 2017.

[67] A. Rangesh, B. Zhang, and M. M. Trivedi, "Driver gaze estimation in the real world: Overcoming the eyeglass challenge," in *IEEE Intelligent Vehicles Symposium (IV 2020)*, October-November, October-November 2020, pp. 1054–1059.

Sumit Jha(S'16) received the B.Tech degree in Electronics and Communication Engineering form the National Institute of Technology(NIT), Trichy, India, in 2012 and the MS degree in Electrical Engineering from the University of Texas at Dallas (UTD), Texas in 2016. He is currently a PhD candidate at the University of Texas at Dallas. At UTD, he has been a part of the Multimodal Signal Processing (MSP) laboratory since 2015. His research interest lies in machine learning computer vision solutions for driver monitoring and in-vehicle safety systems. He has been a student member of IEEE since 2016.

Naofal Al-Dhahir is Erik Jonsson Distinguished Professor & ECE Associate Head at UT-Dallas. He earned his PhD degree from Stanford University and was a principal member of technical staff at GE Research and AT&T Shannon Laboratory. He is co-inventor of 43 patents, co-author of 490 papers and co-recipient of 5 IEEE best paper awards. He is an IEEE Fellow, Fellow of the National Academy of Inventors, received 2019 IEEE SPCE technical recognition award, and served as Editor-in-Chief of IEEE Transactions on Communications.

Carlos Busso(S’02-M’09-SM’13-F’23) is a professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He received the Best of IEEE Transactions on Affective Computing Paper Collection in 2021 (with R. Lotfian) and the Best Paper Award from IEEE Transactions on Affective Computing in 2022 (with Yannakakis and Cowie). His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, nonverbal behaviors for conversational agents, and machine learning methods for multimodal processing. He is an IEEE Fellow.