# Estimation of Driver's Gaze Region From Head Position and Orientation Using Probabilistic Confidence Regions

Sumit Jha, *Member, IEEE*, and Carlos Busso, *Senior Member, IEEE*

***Abstract*—Visual attention is one of the most important aspects related to driver distraction. Estimating the driver's visual attention can help a vehicle understand the awareness state of the driver, providing important contextual information. While estimating the exact gaze direction is difficult in the car environment, a coarse estimation of the visual attention can be obtained by tracking the head pose. Since the relation between head pose and gaze direction is not one-to-one, this paper proposes a formulation based on probabilistic models to create salient regions describing the driver's visual attention. The area of the estimated region is small when the model has high confidence, which is directly learned from the data. We use *Gaussian process regression* (GPR) to implement the framework, comparing the performance with different regression formulations such as linear regression and neural network based methods. We evaluate these frameworks by studying the tradeoff between spatial resolution and accuracy of the probability map using naturalistic recordings collected with the UTDrive platform. We observe that the GPR method produces the best result creating accurate estimations with localized salient regions. For example, the 95% confidence region is defined by an area covering 3.77% region of a sphere surrounding the driver.**

***Index Terms*—In-vehicle safety, advanced driver assistance system, driver visual attention, gaze detection.**

## I. Introduction

ROAD safety is a major concern in today's world. The main cause of road accidents is the negligence of distracted drivers [1]. Therefore, monitoring the driver's actions can be useful for estimating their behaviors, creating warnings to avoid impending mistakes due to lack of awareness. Smart vehicles today are equipped with multiple sensors, which provide relevant real-time information inside and outside the vehicle. The challenge is incorporating heterogeneous information to provide high-level knowledge to understand the driver, the vehicle, and the road. Monitoring the driver's behaviors can also serve as a tool to design advanced user interfaces for infotainment and navigation systems where the drivers naturally interact with the car, without using manual resources [2] (e.g., interpreting commands such as "what is the address of this building?," while the driver briefly glances towards the target location). With semi-autonomous cars, monitoring the driver behavior can also be helpful in negotiating hand-over control from the vehicle to the driver, or vice-versa.

Visual attention is a major factor when modeling the driver's intentions. The majority of the tasks involved while driving require visual cues. The direction of the driver's gaze strongly depends on the primary driving task and the road condition. Implementing a robust gaze detection system for cars can be helpful in signaling the cognitive state [3], [4], situational awareness [5]–[7], and attention level [8], [9] of the driver. These systems can also be helpful in enhancing in-car dialog systems [2].

In *human-computer interaction* (HCI), the gaze of a subject is estimated by locating the pupil using various appearance based and feature based techniques [10]–[12]. However, these techniques are not practical in a vehicle environment with challenging situations such as varying lighting conditions, high degree of head rotations, and possible occlusions [13]. Moreover, in the detection of the driver's attention is often more important to achieve robustness across conditions rather than high performance under restricted conditions. A coarse estimation of the driver's visual attention is usually enough for many applications. Following this strategy, studies have proposed the use of head pose to infer the driver's gaze [14]–[17]. Head pose has a strong correlation with gaze, but the relationship is not deterministic [18]. Taking the eyes-off-the-road during longer periods significantly increases chances of accidents. Therefore, drivers tend to have short glances, which involve head and eye movements. This relationship changes according to the driver, primary driving task, secondary driving task, and the traffic condition. Therefore, the head orientation cannot uniquely determine the exact gaze direction.

Instead of aiming to detect the precise gaze direction, this paper proposes to estimate a probabilistic visual map describing the region of visual attention where the driver is most likely to direct her/his gaze. Building upon our previous work [19], [20], we propose to create this probabilistic visual map using a two

Manuscript received February 1, 2021; revised August 6, 2021; accepted December 20, 2021. This work was supported by Semiconductor Research Corporation (SRC) Texas Analog Center of Excellence (TxACE) under Grant 2810.014. *(Corresponding author: Carlos Busso.)*

The authors are with the Erik Johnson School of engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: sumit.jha@utdallas.edu; busso@utdallas.edu).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board at the University of Texas at Dallas under Application No. 15-63, and performed in line with Expedited Review under 45 CFR 46.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TIV.2022.3141071.

Digital Object Identifier 10.1109/TIV.2022.3141071

dimensional Gaussian distribution that is directly learned from data. The formulation relies on *Gaussian process regression* (GPR) to estimate the distribution of the gaze given a certain position and orientation of the driver's head. The proposed model provides not only the probabilistic visual map, but also confidence regions, which can be extremely useful for HCI applications for infotainment and navigation systems, and *advanced driver assistance systems* (ADAS). The size of the salient region decreases when the confidence of the model increases, learning all the parameters of the models directly from the data. We train and evaluate the system with recordings from real driving scenarios using affordable equipment that can be easily installed on regular cars.

The experimental evaluation demonstrates the effectiveness of the proposed GPR system, analyzing the tradeoff between accuracy and spatial resolution of the probabilistic visual map. We compare our proposed solution with alternative machine learning methods to estimate the visual maps, including simple regression techniques, deep neural networks and *mixture density networks* (MDNs). The results indicate that our proposed model offers the best accuracy and spatial resolution in estimating the probabilistic region of the driver's gaze. For example, 95% of the target markers lie inside the probabilistic region estimated by the system, where its temporal resolution includes 3.77% of a sphere surrounding the user's range of vision. Finally, we demonstrate the benefit of the proposed probabilistic model by mapping the probabilistic visual map to areas on the road, allowing us to identify coarse regions outside the car that the driver is directing her/his gaze to.

This study is organized as follows. Section II discusses related studies about the importance of visual attention when studying driver's behavior. Section III describes the data collection procedure that we followed to train and evaluate our algorithms. Section IV describes the proposed method to obtain the probabilistic salient visual map to represent visual attention. It also introduces the baseline methods. Section V discusses the results obtained from different models, comparing the tradeoff between spatial resolution and accuracy of the probabilistic salient visual maps. Finally, Section VI concludes the study, suggesting future research directions.

## II. Related Work

### A. Visual Attention of the Driver

Maintaining visual attention while driving a vehicle is important to reduce hazard scenarios. Drivers obtain most information through vision, which is important to maintain road awareness and to complete driving maneuvers [5]. Therefore, several studies have considered the visual patterns of the driver, creating useful automatic tools for intelligent vehicle systems. Liang and Lee [21] conducted experiments by inducing visual distraction, cognitive distraction and a combination of both by asking subjects to perform distracting tasks while operating a driving simulator. They observed that the driving performance was worse when the subjects were performing visually distracting tasks compared to the performance when performing a combination of visual and cognitive distracting tasks.

Robinson *et al.* [22] studied the visual search patterns of a driver by looking at her/his head movements during lane changes and when entering a highway after a stop sign. They observed longer search times at a stop sign, where the drivers had to observe the whole scene before making a decision. In contrast, for lane change actions the search time was shorter since the driver had to make quick decisions. Underwood *et al.* [23] used eye trackers to study the eye movement behavior of experienced and novice drivers in three different types of roads: rural, suburban and divided highways. They analyzed the most common sequences of fixation in various regions of the road to compare the driving behavior. They observed that a novice driver tends to change her/his fixation more often, while an experienced driver tends to use peripheral vision to pick up subtle information such as the demarcation of lanes.

Understanding visual attention can also help us infer information about visual and cognitive distractions. Sodhi *et al.* [24] used a head-mounted, eye-tracker in a vehicle, where they asked multiple subjects to drive a predetermined route while performing tasks that stimulate distractions. They used the eye-tracker to obtain the position and diameter of the pupil. They studied the impact of infotainment systems on driving by stimulating various cognitive and visual distractions. The study observed that the eye movement patterns changed when the driver was distracted by a secondary task. Kutila *et al.* [25] recorded the face of the driver with stereo cameras in a naturalistic driving scenario. They used head and gaze information along with lane position and *controller area network* (CAN)-Bus data to detect visual and cognitive distraction. Gaze data was obtained using a gaze tracker. The driver's visual attention is inferred using eyes-off-the-road duration. The eye movement is fused with cognitive workload inferred from the driving data to obtain cognitive distractions. Liang *et al.* [26] designed a *support vector machine* (SVM) classifier that used measures of driving performance such as steering angle and lane position, and features from eye movement data such as fixations and saccades to detect cognitive distraction. They obtained an accuracy of 96.1% in a simulated environment. Murphy *et al.* [27] implemented a real-time system to track the six degrees of freedom of the head pose of a driver. They designed an appearance based particle filter to design a 3D model of the face in augmented reality. Rezaei and Klette [17] monitored both the driver and the road to find possible hazard situations. They designed an asymmetric *active appearance model* (AAM) to estimate the driver's head pose, which was used in conjunction with features extracted from the vehicles detected on the road to design a fuzzy logic based system to estimate the risk level of the driving situation.

Understanding the driver's behavior is even more relevant with the advances in autonomous cars. Information and datasets derived from drivers in naturalistic conditions, including their visual attention, can be instrumental in the design of autonomous cars [28], following the ideas of behavior cloning [29]. Likewise, cars that are aware of the driver's visual attention can more effectively negotiate hand over situations. Zeeb *et al.* [30] compared the take-over time and quality between a distracted driver and an attentive driver. The drivers were asked to be involved in secondary tasks such as watching videos and writing emails. They observed that while a driver could quickly resume control of the car when prompted, the quality of the take over was worse when the driver was distracted.

### B. Estimation of Visual Attention

Several studies have worked on estimating the visual attention of the driver, realizing the importance of visual attention in monitoring the behaviors of the driver. The most common approach is to partition the gaze region of the driver into different gaze zones. Then, the problem is formulated as a classification problem to identify the area that the driver is directing her/his attention. Tawari and Trivedi [14] video recorded the face of the driver from two different angles in the car to capture the head pose. The two cameras increased the angular range of the head pose estimation. The task was to classify the driver's gaze into eight different gaze zones. They used annotations obtained from human experts as the ground truth for the target gaze zone, training a random forest classifier with the head pose as features. The zone estimations had high confusion between adjacent zones such as looking forward and looking at the speedometer. Lee *et al.* [15] suggested a robust method to estimate the yaw and pitch of the head. The method relied on simple edge features from the face, making the approach robust to rotation and illumination, and fast enough to be run in real time. They used the estimated yaw and pitch angles to identify one of the 18 predefined gaze zones using an SVM classifier. Chuang *et al.* [16] designed a gaze estimation system using a smartphone camera. They placed the smartphone on the dashboard to record the driver's face. They used the location of the eyes, nose and mouth regions as features to classify the gaze among eight different zones. Vora *et al.* [31] tried a generalized approach to classify gaze zones which is subject invariant. They used *convolutional neural networks* (CNNs) to obtain the gaze zone from the driver's facial image. The best network achieved a 93.36% accuracy while performing a seven class classification task (six gaze zones plus a class for eye closure)

While the gaze zone provides useful information about the visual attention of the driver, this information is too coarse for several applications. However, it is challenging to design gaze estimation methods with high precision that work well inside a vehicle. Although there is a strong relationship between head movement and gaze direction, the relation is not one-to-one [18]. In naturalistic driving scenarios, the driver relies not only on head movements to direct her/his gaze toward a target location, but also on eye movement. The interplay between head and eye movements depends on the cognitive load of the driver and the underlying driving task. A feasible alternative to gaze zone estimation or unreliable gaze algorithms that do not work in a vehicle is the definition of a probabilistic salient visual map describing the visual attention of the driver. This probabilistic salient visual map can be used to define spatial confidence regions describing the direction of the driver's gaze. This study pursues this novel formulation, creating models that capture the relationship between head pose and gaze, creating a probability distribution of the gaze given the orientation and position of the driver's head.

### C. Relation to Prior Work

The formulation of creating a probabilistic salient visual map to model driver attention is novel. To the best of our knowledge, the only relevant study is our preliminary work [19], [20], which provided initial evidences of the benefits of using this promising formulation. Jha and Busso [19] used the GPR framework to estimate the gaze distribution conditioned on the head pose. Jha and Busso [20] explored a nonparametric approach to create this probabilistic salient visual map. This paper builds upon these preliminary studies providing better modeling capabilities, which are evaluated with exhaustive experiments. The contributions of our paper with respect to prior work are:

- We improve the modeling capability of the GPR framework by exploring multiple configuration including implementing the basis function with a neural network, and using *automatic relevance determination* (ARD) for the kernel function. These approaches increase the capacity and flexibility of the models, leading to better performance.
- We explore multiple regression-based frameworks and compare them to our method to establish the superiority of the proposed approach in comparison to other methods.
- We demonstrate the application of using confidence maps with our method to project the angular distribution onto the road scene, moving us closer to deploy our solution in practical applications.

Instead of relying on methods that assume a one-to-one relationship between head pose and gaze, which has been the predominant approach in previous studies, our method creates a probability distribution that takes in consideration the uncertainty in the predictions. This approach represents a novel formulation from a theoretical perspective. The implementation of the approach in a real system is feasible, but not the primary goal of this paper.

## III. Data Collection

This study uses recordings from real driving scenarios collected with the UTDrive platform [32], [33], which is a vehicle equipped with multiple sensors (Fig. 1(a)). The UTDrive has been successfully used to study driver behaviors [34]–[36]. Instead of using the specific sensors from this car, we decided to only use the commercially available dash camera Blackvue DR-650GW-2ch (Fig. 1(b)), which can be easily installed in any regular car. The device features two cameras along with a *global positioning system* (GPS) and accelerometer sensors. The front camera was used to record the road view, while the rear camera was used to record the face of the driver. The system is currently implemented offline.

### A. Data Collection Protocol

For the analysis, we require data where we know the ground truth information about the direction of the gaze. We achieve this goal by asking the driver to look at predefined markers. We place 21 numbered markers on the windshield (#1-#13), mirrors (#14-#16), side windows (#17-#18), speedometer panel (#19), radio (#20), and gear (#21) (Fig. 1(c)). Then, we ask

(a) UTDrive platform (b) Dash camera (c) Layout of Markers

Fig. 1. (a) Vehicle used for data collection. (b) Dash camera (Blackvue DR-650GW-2ch) used to record the face of the driver (primary camera) and the road (secondary camera). (c) Markers placed on the windshield (1-13), mirrors (14-16), side windows (17-18), speedometer panel (19), radio (20), and gear (21). The subjects were asked to look at these markers.

the subjects to look at these markers multiple times, where we carefully annotated the corresponding timing information. We recruited 16 students (10 males, 6 females) with valid US driver licenses from the University of Texas at Dallas. We designed a three-phase protocol:

*Phase 1:* The first phase is recorded when the vehicle is parked. The subject is asked to sit in the driver seat, looking at different markers. The numbers are called out in random order and the driver is asked to look at the corresponding points. Each number was repeated five times in random order. We did not provide any further instruction. The goal of this phase is to estimate and model the gaze-head relationship when our subjects are not driving. They have plenty of time to complete this task without worrying about visual, manual and cognitive demands associated with the driving tasks. The drivers can also get familiar with the task in a safe environment.

*Phase 2:* The second phase consists of the same task while the subject is driving the vehicle. The subject is asked to drive on a straight road with low traffic. Following the protocol approved by the *institutional review board* (IRB) at UT Dallas, we carried out the data collection during the day, avoiding peak hours to reduce the cognitive load of the driver in traffic conditions. A passenger reads the numbers, pointing to the target location reducing the cognitive demand of the task. The numbers are requested only when the driver does not have to perform any maneuver. The safety of the subject is our first priority. We do not provide any additional instruction on how to look at the markers. We use this phase to estimate and model the gaze-head relationship while the subject is driving.

*Phase 3:* During the third phase, we ask the driver to park the car and perform the same task again. This time, the driver is asked to look at each marker directing her/his head toward the point. In this controlled condition, the gaze of the driver is the same as the head pose, without the bias added by the movement of the eye. To enforce this requirement, we request the driver to wear a glass frame with a low power laser mounted at the center (Fig. 2). The driver is asked to point the laser towards the marker. The windows of the car are covered during this phase to reduce the lighting inside the car to make the laser more visible for the subject to complete the task. This approach also prevents the laser beam to project outside the car. Each number was repeated three times at random for each marker. This phase provides valuable data, where the gaze is exactly aligned with the head orientation. While this phase is not used in the experiments discussed in this study, it provides valuable calibration information for other studies [18].

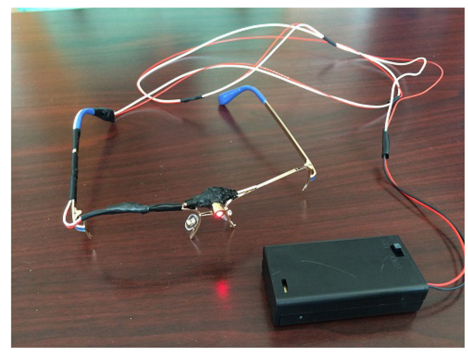


Fig. 2. Laser pointer mounted on a glass frame for the controlled head pose condition during phase 3 of the data collection.

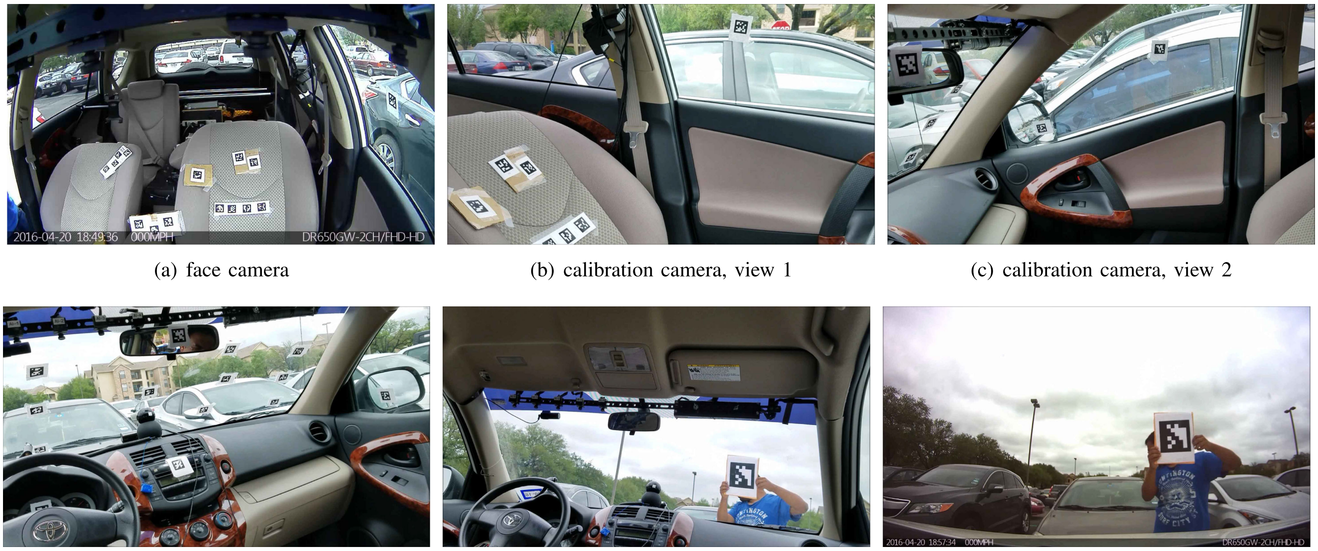
(a) AprilTag (b) Headband with AprilTags

Fig. 3. (a) Example of a AprilTag, (b) Headband with AprilTags for robust head pose estimation.

Additionally, we asked the last three of our subjects to look at specific locations on the road including billboards, street signals, and buildings to validate our systems in real-world applications. This data is used to assess the mapping between gaze detection and objects on the roads.

### B. Head Pose Estimation Using AprilTags

It is challenging to use computer vision algorithms in a car environment. In our previous work [13], we demonstrated that the robustness of a state-of-the-art head pose estimation algorithm was low for non-frontal faces rotated more than 45°. For this analysis, we aim to have more robust estimations of head poses regardless of the head orientation. We achieve this goal by using a headband with AprilTags (Fig. 3(a)). For future




(a) face camera (b) calibration camera, view 1 (c) calibration camera, view 2

(d) calibration camera, view 3 (e) calibration camera, view 4 (f) Road camera

Fig. 4. Description of the process to calibration the cameras and define a common reference system. Figures (a)-(d) show the calibration for the markers' locations and the face camera. Figures (e)-(f) show the calibration for the road camera.

work, we can rely on depth cameras to obtain robust head pose estimation [37], [38].

AprilTags [39] are 2D barcodes primarily used for augmented reality applications, robotics and camera calibration. Fig. 3(a) shows an example of an AprilTag. The unique black and white patterns of each AprilTag are easy to automatically detect with computer vision algorithms. From the pattern, it is possible to accurately estimate the position and orientation of the tags. Instead of using a single AprilTag, we rely on many tags to robustly estimate the position and orientation of the head. For this purpose, we designed a headband with 17 square faces ($2 \times 2$ cm each), separated by an angle of $12^{\circ}$. Each of the square faces contains a 1.6 cm $\times$ 1.6 cm unique tag. Fig. 3(b) shows the headband worn by the participants. During the data collection, the subject is asked to wear the band for the entire recording. The selected design allows us to observe multiple tags for each video frame, regardless of the orientation of the driver's head. Therefore, we can robustly infer the position and orientation of the headband.

The AprilTags from the headband are used to obtain the position and orientation of the driver's head. The AprilTag toolkit provides an estimate of the position and orientation of each tag present in an image. The structure of the headband and orientation of the visible bands help us estimate the pose of the headband.

The use of the headband facilitates the analysis of head pose regardless of the orientation of the head or the environmental condition in the vehicle. For real-world applications, the head orientation will be estimated using automatic algorithms using either RGB cameras [40] or depth cameras [37], [38].

### C. *Calibration of Camera and Markers*

A key challenge is to define a common coordinate system. We need the location and orientation of the driver's head along with the location of each enumerated marker in a 3D space with respect to a single coordinate system. Fig. 4(a) shows the view from the rear camera facing the driver, and Fig. 4(f) shows the view from the road camera. It is clear from these two figures that most of the markers are not included in the view of either of the camera. The problem is even more challenging as we aim to map the gaze direction to areas on the road camera. We need a calibration process to find the exact target marker location in the 3D space and the transformation between the cameras to represent all the coordinates in a single reference system. The calibration process relies on AprilTags to find the location of the markers in the 3D space and to find the relative homogeneous transformation between each camera. The proposed solution consists of placing AprilTags in the vehicle. The AprilTags are used to establish a connection between the road and face cameras, which do not have any overlap in their field of view. The calibration process has two steps: create a common reference coordinate system, and create a mapping between objects outside the vehicle.

The first step in the calibration is to establish a reference coordinate system. AprilTags are placed on each of the markers (Fig. 4(d)), and some reference locations in the field of view of the face camera (Figs. 4(a)-4(d)). These tags are only used to calibrate the system, and are removed during the data collection. Then, we use a third camera to take multiple pictures containing subsets of these AprilTags (Fig. 4). This camera captures locations that are not in the field of view of either of the dash cameras. The relationship between frames containing multiple common tags is calculated. The face camera captures a subset of these additional tags. Using the location of these tags, we create homogeneous transformations to obtain the location of all the tags, including the 21 markers, with respect to the coordinate system of the face camera.

The second step in the calibration consists of estimating a mapping between the reference coordinate system and objects outside the vehicle. For this step, the third camera is fixed inside the vehicle such that it records the windshield and the road view (Figs. 4(e) and 4(f) – these two images were simultaneously taken). As a result, we have two cameras facing the road: the road camera used in the data collection and the third camera used for calibration. We printed an AprilTag sign, which is placed in front of the vehicle such that it is in the field of view of both cameras. This process helps us to establish the relationship between the cameras for different points. Finally, the relation between the face camera and the third camera is established using the location of the markers. With this process, we derive all the transformations needed to map everything into the coordinate system of the face camera. The transformation matrix to relate each camera is calculated using the Kabsch algorithm [41].

Since the placement of the headband slightly varies across subjects, we need to obtain a standard reference per subject to estimate the head pose from the AprilTags. For this purpose, we assume that the long-term average of the driver's head pose is consistent across subjects. We calculate the average orientation of the headband in the quaternion space using *spherical linear interpolation* (Slerp). We set the origin of the coordinate system as the average pose of the driver's head by subtracting the average head position per participant and multiplying by the inverse of the average rotation matrix to normalize the orientation. We also subtract the average head position value from each of the target marker's location to apply a similar transformation to the target gaze. Finally, the ground truth gaze vector at a given instant is obtained by subtracting the target gaze location from the head position at the given instant, calculating the horizontal and vertical gaze angles from this vector.

## IV. Methodology

This paper aims to create a probabilistic salient visual map describing the visual attention of the driver. We project this map onto the windshield creating spatial distributions for the gaze direction. Then, we map this visual map on the road camera, defining areas on the road where we estimate the driver is directing her/his gaze. We can estimate confidence regions for the driver's visual attention by creating a probabilistic map, which is an appealing method with more practical applications than methods estimating a single point for the gaze direction. This section proposes our main method as well as three alternative baselines to obtain a probabilistic distribution of the gaze angle from the position and orientation of the driver's head. These methods estimate the probabilistic salient visual map for the horizontal and vertical angles by modeling the mean and variance of the gaze angles as a function of the position and orientation of the head.

### A. Gaussian Process Regression

Our proposed model is based on the original design implemented in our preliminary work [19]. It relies on *Gaussian process regression* (GPR) [42], which models the outputs as a Gaussian process with the co-variance defined by a kernel function. The model assumes that any subset of the output is a joint Gaussian distribution. Using the ground truth of the training data in the vicinity of each point, the model learns the uncertainty in the estimation of any test data. This method provides a promising and effective approach to learn the many-to-many relationship between the head pose and the gaze. It learns a gaze distribution as a function of different head poses presented in the training data that are in the neighborhood of the target head pose.

Let $\mathbf{x} \in \mathbb{R}^d$ be the input of the system, where $d$ is its dimension (in our case $d = 6$). Let $Y$ be a Gaussian random process representing the output. If $\mathbf{y}$ is a vector representing $n$ realizations, as a Gaussian random process, $\mathbf{y}$ follows a joint Gaussian distribution with prior distribution $f_{\mathbf{y}}$,

$$f_{\mathbf{y}} = \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{1}$$

where, $\mathbf{y} = \{y_1, y_2, y_3, \ldots, y_n\} \subset Y$ ($y_i$ is a realization of $Y$). The parameters $\boldsymbol{\mu}$ and $\Sigma$ are functions of $\mathbf{x}$. As shown in (2), the mean provides the deterministic component of the model, where $\boldsymbol{\omega} \in \mathbb{R}^d$ and $\omega_0 \in \mathbb{R}$ are learned while training the models.

$$\boldsymbol{\mu} = \mathbf{x}^T \boldsymbol{\omega} + \omega_0 \tag{2}$$

The probabilistic component is given by $\Sigma$, which is the covariance matrix. The covariance of any point $\mathbf{x}$ calculated jointly with the input points in the training set $\mathbf{x}'$ is given by the kernel $k(\mathbf{x}, \mathbf{x}')$. The covariance is modeled using a squared exponential kernel ( (3)). The correlation is learned with respect to the input data from the training set in the neighborhood of the data of interest. This kernel imposes that the outputs will be more correlated to the points in the training data that are closer to the test input data as the covariance matrix will have higher values for points that are closer.

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(\frac{-\|\mathbf{x} - \mathbf{x}'\|^2}{2\, l^2}\right) \tag{3}$$

In (3), the parameter $\sigma_f$ represents the amplitude of the covariance. This parameter defines the autocovariance of the data points (i.e., $k(\mathbf{x}, \mathbf{x}) = \sigma_f^2$). The parameter $l$ represents the length scale value, which defines how much the distance between the training and estimated data affects the cross-covariance between two data points. If $l$ is high, $k(\mathbf{x}, \mathbf{x}')$ slowly reduces, as the distance between the points increases ($\|\mathbf{x} - \mathbf{x}'\|^2$). These parameters decide the size of the confidence interval of our estimation as the covariance matrix is a function of these parameters. We also explore the use of *automatic relevance determination* (ARD). Using ARD, the kernel learns different length scale parameters for each input variable. The kernel function with ARD is given in (4).

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{i=0}^{d} \frac{\|x_i - x_i'\|^2}{l_i^2}\right) \tag{4}$$

Using different values for $l$ may be useful, since the input to our models include position and orientation of the head, which may have different scales. We learn the values for $\sigma_f$ and $l$ (or $l_i$) while training the models by maximizing the log-likelihood of the ground truth data in the train set.

To obtain the posterior distribution from the prior model, the model is conditioned on the given training data. Let, $X_{tr} \in$

$\mathbb{R}^{L\times d}$ be the training dataset and $y_{tr} \in \mathbb{R}^{L}$ be the output Gaussian random variables, where $L$ is the number of frames in the train set. Let $y_*$ be the random variable we are trying to estimate for the input vector $\mathbf{x}_*$. From (1), the joint distribution is given by,

$$\begin{bmatrix} f_{y_{tr}} \\ f_{y_*} \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} X_{tr}^T\boldsymbol{\omega} + \omega_0 \\ \mathbf{x}_*^T\boldsymbol{\omega} + \omega_0 \end{bmatrix}, \begin{bmatrix} \Sigma_{(X_{tr},X_{tr})} & \Sigma_{(X_{tr},\mathbf{x}_*)} \\ \Sigma_{(\mathbf{x}_*,X_{tr})} & \Sigma_{(\mathbf{x}_*,\mathbf{x}_*)} \end{bmatrix}\right) \tag{5}$$

Using (5), the posterior distribution can be calculated with the conditional probability when $y_{tr} = y_{obs}$:

$$f_{y_*|y_{tr}=y_{obs}} = \mathcal{N}(\hat{\mu}_*, \hat{\Sigma}_*) \tag{6}$$

$$\hat{\mu}_* = \mathbf{x}_*^T\boldsymbol{\omega} + \omega_0 + \Sigma_{(\mathbf{x}_*,X_{tr})}[\Sigma_{(X_{tr},X_{tr})}]^{-1} \times (y_{obs} - X_{tr}^T\boldsymbol{\omega} - \omega_0) \tag{7}$$

$$\hat{\Sigma}_* = \Sigma_{(\mathbf{x}_*,\mathbf{x}_*)} - \Sigma_{(\mathbf{x}_*,X_{tr})}[\Sigma_{(X_{tr},X_{tr})}]^{-1}\Sigma_{(X_{tr},\mathbf{x}_*)} \tag{8}$$

where $y_{obs}$ is the ground truth value (i.e., observed $y$). We use four different settings for the deterministic function. The first setting is a GPR model without the deterministic component (i.e. $\boldsymbol{\omega} = 0, \omega_0 = 0$). The mean of the posterior distribution is purely estimated from the kernel function ( (9)).

$$\hat{\mu}_* = \Sigma_{(\mathbf{x}_*,X_{tr})}[\Sigma_{(X_{tr},X_{tr})}]^{-1}y_{obs} \tag{9}$$

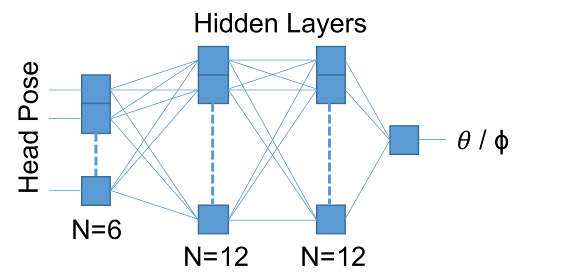
The second setting is with a constant deterministic component. The model learns $\omega_0$ as a single constant mean for the distribution (i.e., $\boldsymbol{\omega} = 0$). The third setting estimates both $\boldsymbol{\omega}$ and $\omega_0$ during training. We refer to this setting as linear model. The fourth setting estimates the deterministic component of the model with a *neural network* (NN). We implement this approach by training $NN(\mathbf{x})$, using back propagation. The network is implemented with two hidden layers, following the architecture used for our second baseline (Fig. 5(a)). Then, we estimate the residual error, $r(\mathbf{x}) = y_{obs} - NN(\mathbf{x})$, which is modeled with the GPR formulation, without the deterministic component ($\boldsymbol{\omega} = 0, \omega_0 = 0$). The conditional mean for this implementation is given by (10).

$$\hat{\mu}_* = NN(\mathbf{x}_*) + \Sigma_{(\mathbf{x}_*,X_{tr})}[\Sigma_{(X_{tr},X_{tr})}]^{-1}(y_{obs} - NN(X_{tr})) \tag{10}$$

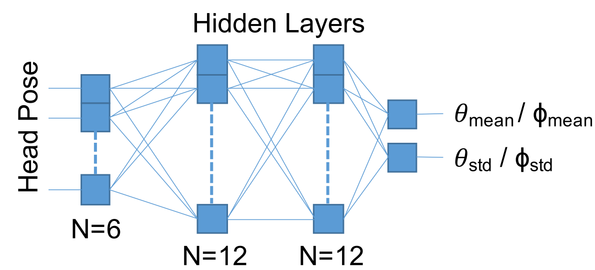
Using this framework, we learn two separate models for the horizontal angle ($\theta$) and the vertical angle ($\phi$). An important feature of our formulation is modeling the output as a heteroscedastic process, where the variance of the output salient map varies depending on the input variables. Therefore, the size of the probabilistic salient visual map increases for regions with higher uncertainty, and decreases when the model is confident in its estimation.

## B. Baseline Methods

We compare the model with three methods. Two of these baselines are based on normal regression functions designed with the mean square error loss. We adapted these regression models to create a probability map as the output by assuming a Gaussian distribution. For the third baseline, we explore a variation of *mixture density network* (MDN) that uses the log-likelihood as the loss function to model the conditional probability density of the gaze given the input head pose. This section provides the details of these baseline models.




(a) Neural Network




(b) Neural Network for Density Estimation

Fig. 5. Architecture of baseline methods to estimate the probabilistic salient visual map describing visual attention. The same architecture is used for both horizontal and vertical angles.

*1) Linear Regression:* The first baseline is the most basic regression model. The gaze is obtained as a linear function of the head pose parameters (orientation and position). The dependent variables are the six degrees of freedom of the head corresponding to its position (x,y,z) and orientation angles ($\alpha, \beta, \gamma$). Two separate models are created for the gaze angle in the horizontal and vertical directions. Equations 11 and 12 show the models, where, $\theta_{gaze}$ and $\phi_{gaze}$ are the horizontal and vertical gaze angles, respectively.

$$\theta_{gaze} = a_0 + a_1x + a_2y + a_3z + a_4\alpha + a_5\beta + a_6\gamma \tag{11}$$

$$\phi_{gaze} = b_0 + b_1x + b_2y + b_3z + b_4\alpha + b_5\beta + b_6\gamma \tag{12}$$

This model is similar to the one trained in Jha and Busso [18], but instead of obtaining the gaze location, we obtain the angles representing the gaze vectors. To create a probability distribution as our estimation, we consider a Gaussian distribution with the mean value provided by the regression models. The variance is obtained from the mean square error estimated on the train data. Notice that this model is homoscedastic, where the variance is constant across the data.

*2) Regression With Neural Network:* For the second baseline, we design a neural network to perform the regression task. Fig. 5(a) shows the model. The neural network contains two fully connected layers, each of them implemented with twelve nodes. The activation used for the hidden layers is the *rectified* *linear unit* (ReLU) activation using a linear function at the output layer. The neural network is optimized to minimize the mean square error between the true gaze angle and the estimated gaze angle from the model. Similar to our previous baseline model, the probabilistic distribution is obtained by assuming a Gaussian distribution for the output, where the mean is the estimated gaze, and the variance is estimated with the mean square error in the train data. This approach is also homoscedastic.

*3) Neural Network for Density Estimation:* The third baseline is inspired by the MDN proposed by Bishop [43]. MDN can directly learn the standard deviation of the output as a non-linear function of the input data. MDNs are used to model the output as a *Gaussian mixture model* (GMM) by optimizing the log-likelihood function in (14).

$$p(y) = \sum_{k=1}^{M} \pi_k \mathcal{N}(y|\mu_k, \sigma_k) \tag{13}$$

$$L_{llk}(y, \pi_k, \mu_k, \sigma_k) = -log(p(y)) \tag{14}$$

The network has $3 \times M$ nodes in the output layer, where $M$ is the number of components. The output represents the component weights $\pi_k$, mean $\mu_k$ and standard deviation $\sigma_k$ with $k \in 1, \ldots, M$. Since we assume that our output is a single Gaussian distribution, we design a model with one component, reducing the number of parameters to two. Therefore, the output layer has two nodes that provide the mean $\mu$ and the standard deviation $\sigma$. Our objective is to estimate a Gaussian distribution that maximizes the probability of the ground truth data. To achieve this, we use as our loss function the negative log-likelihood of the ground truth gaze with respect to the estimated mean and the standard deviation.

$$L_{llk}(y, \mu, \sigma) = -\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{(y-\mu)^2}{2\sigma^2}\right)\right) \tag{15}$$

The variable $\sigma$ is obtained as the exponential of the corresponding output node to avoid the standard deviation from being negative. With this formulation, we estimate not only the mean, but also the variance in each estimation, providing an appropriate scaling to the uncertainty of each output estimation. This baseline is a heteroscedastic method, where the variance changes according to the input data.

Fig. 5(b) shows the network architecture, which has two hidden layers implemented with 12 nodes. The network uses the Adam optimizer [44] with a learning rate of $r = 0.001$, using mini batches of size 32. The neural network is implemented in Keras [45] with Tensorflow [46] as backend. The networks is trained for 1,000 epochs, and the model with minimum loss in the development set is chosen as the final model to be evaluated in the test set.

## V. Experimental Evaluation

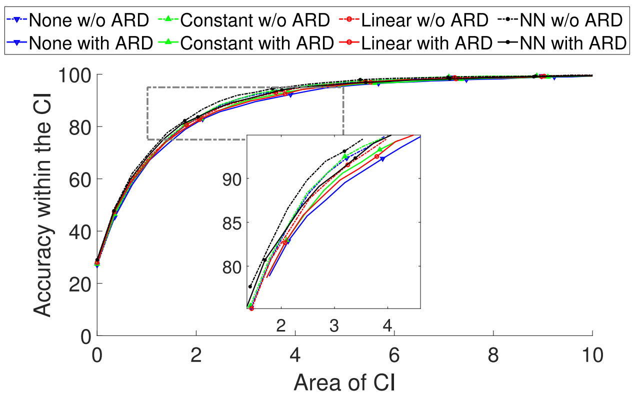
This section evaluates the proposed solution and baselines to estimate the probabilistic salient visual map. The models are separately trained and evaluated for data collected in phase 1 (parked vehicle) and phase 2 (driving condition). The database is partitioned into train, test and development sets using a leave-one-driver-out cross-validation approach. Data from one subject are used for the development set, data from one subject are used for the test set, and data from the remaining fourteen subjects are used for the train set. The development set is used to optimize the hyperparameters and decide on the best model. The best model is evaluated on the test set. This approach is repeated sixteen times, where we report the results across the 16 folds. Note that all the data are, at some point, part of the test set.

We need to analyze the estimated probabilistic salient visual map in terms of accuracy and spatial resolution to evaluate and compare the effectiveness of the baseline and proposed models. Accuracy is measured as the percentage of the target gaze directions included in a given confidence interval. If the majority of the data do not lie within the confidence interval, the model is not accurate. The spatial resolution determines how large the confidence interval is. Likewise, if the spatial resolution is too high, the estimation is not very useful even if most data lies within the interval. To evaluate the spatial resolution of the system, we evaluate the size of the confidence interval created by each model. The outputs of the model are horizontal ($\theta$) and vertical ($\phi$) angles. Therefore, we express the area of the confidence region in terms of the fraction of a sphere surrounding the driver's head. An ideal approach will create a confidence interval that is both accurate and with reduced spatial resolution. To analyze the tradeoff between accuracy and spatial resolution, we present plots with the accuracy of our model at different spatial resolution (Figs. 6, 7, 11).

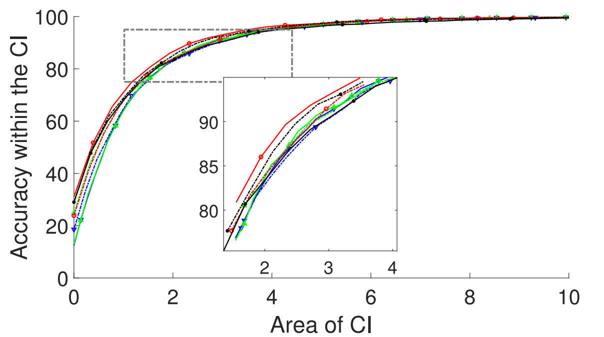
The first evaluation considers different implementations of the GPR model with different parameters to establish the best method for our purpose (Section V-A). Then, we compare the best performing GPR models with the three alternative baselines (Section V-B). Then, we demonstrate the features of the model by projecting the confidence regions onto the windshield (Section V-C), and road camera (Section V-D). Then, we study the performance when we have the orientation of the driver's head, but limited information about the head's position, which is a possible scenarios if regular cameras are used to estimate the head information (Section V-E). We also evaluate the time required for training and inference as a function of the train set size (Section V-F), and the performance as a function of the train set size (Section V-G).

### A. GPR Model Selection

Fig. 6 shows the accuracy of our model within different confidence interval for the different GPR models. Fig. 6(a) reports the results for phase 1 (parked condition) and Fig. 6(b) reports the results for phase 2 (driving condition). We zoom these figures between the 75% and 95% confidence intervals for better visualization. We observe that different models work better for parked and driving conditions. We have shown that the relationship between head movements and gaze changes when a person is driving [18]. There is more uncertainty in the relationship between head pose and gaze, where drivers tend to use more eye movements to glance at a target object. The increase in uncertainty explains the differences in patterns

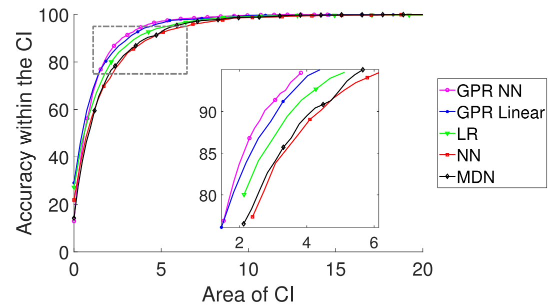



(a) Parked conditions

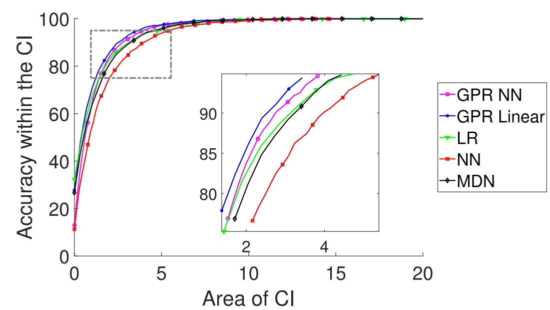



(b) Driving Conditions

Fig. 6. Comparison of the accuracy versus temporal resolution of different implementations of the GPR model. We zoom the plots for better visualization. The results are separately reported for parked and driving conditions. The figure is better viewed in colors. Accuracy is calculated within a *confidence interval* (CI) given by the area.




(a) Parked conditions




(b) Driving Conditions

Fig. 7. Comparison of the accuracy versus temporal resolution of GPR and baseline models. We zoom the plots for better visualization. The results are separately reported for parked and driving conditions. The figure is better viewed in colors. Accuracy is calculated within a *confidence interval* (CI) given by the area.

across phases 1 and 2. During the driving condition (phase 2), implementing the deterministic component with a linear model leads to the best performance. For this case, the use of ARD in the kernel function leads to improvements for all four models. Since the relationship between head pose and gaze is more ambiguous when driving, as noted in Jha and Busso [18], a stronger probabilistic component helps to better describe the relationship ( (3)). Since ARD encodes a separate length scale parameter for each variable (variable $l$ in (3)), this model provides a more sophisticated description of the variance of the gaze random variable. Therefore, it is expected that the best model for the driving condition uses the ARD framework. The deterministic part of the model is dictated by the mean (variable $\mu$ in (2)). The results show that adding a more sophisticated mean model does not provide a gain in performance while increasing the complexity in learning. Therefore, a linear mean function with ARD function provides the best performance in the driving condition. During the parked condition (phase 1), in contrast, the best GPR model is when the deterministic part is implemented with a neural network, and the kernel is implemented without ARD. This result shows that adding a more powerful deterministic function is enough to achieve good performance. We consistently observe lower performance when using ARD, regardless of the implementation of the deterministic function. The gaze has a strong predictability since the variance is reduced compared to the driving condition case. Hence, the deterministic part of the model is more important. A neural network provides a better estimate of the mean. Given the more straightforward relation between head pose and gaze, the kernel function without ARD provides a better estimate of the variance.

For the rest of the evaluation, we will consider the two GPR models that led to the best performance for phase 1 (GPR with neural network model without ARD) and phase 2 (GPR with linear model with ARD). The case when the ego vehicle is static can be used for modeling gaze when the car is stopped at intersections, or traffic signals. In these cases, the driver will take more time to asses the environment compared to cases when driving the car.

### B. Comparison With Baselines

This section compares our proposed models with the three baselines described in Section IV-B: *linear regression* (LR), *neural network regression* (NN) and *mixture density network* (MDN).

*1) Accuracy Versus Spatial Resolution:* Fig. 7 shows the accuracy of our models at different spatial resolutions, comparing with results with the curves of different baseline models. We observe that both GPR models perform better than all the baseline models. They are consistently above other curves showing not only higher accuracies, but also smaller regions. The linear regression baseline is the model with higher performance from the baselines. The values are constantly below our two implementations of the GPR models.

TABLE I
AVERAGE AREA OF THE CONFIDENCE INTERVALS FOR 50%, 75% AND 95% ACCURACY. THIS AREA IS MEASURED AS THE FRACTION OF A SPHERE SURROUNDING THE DRIVER'S HEAD

| Method | Parked – Phase 1 | | | Driving – Phase 2 | | |
|---|---|---|---|---|---|---|
| | 50% [%] | 75% [%] | 95% [%] | 50% [%] | 75% [%] | 95% [%] |
| LR | 0.43 | 1.71 | 5.12 | 0.47 | 1.42 | 4.74 |
| NN | 0.68 | 2.05 | 6.48 | 0.92 | 2.00 | 5.39 |
| MDN | 0.94 | 2.12 | 5.66 | 0.68 | 1.71 | 4.43 |
| GPR NN | 0.38 | 1.13 | 3.76 | 0.61 | 1.38 | 3.98 |
| GPR Linear | 0.37 | 1.46 | 4.39 | 0.34 | 1.37 | 3.77 |

TABLE II
AVERAGE ACCURACY OF THE CONFIDENCE INTERVALS FOR PROBABILISTIC SALIENT VISUAL MAPS OF DIFFERENT SIZES (1%, 2% AND 4% OF THE SPHERE SURROUNDING THE DRIVER'S HEAD)

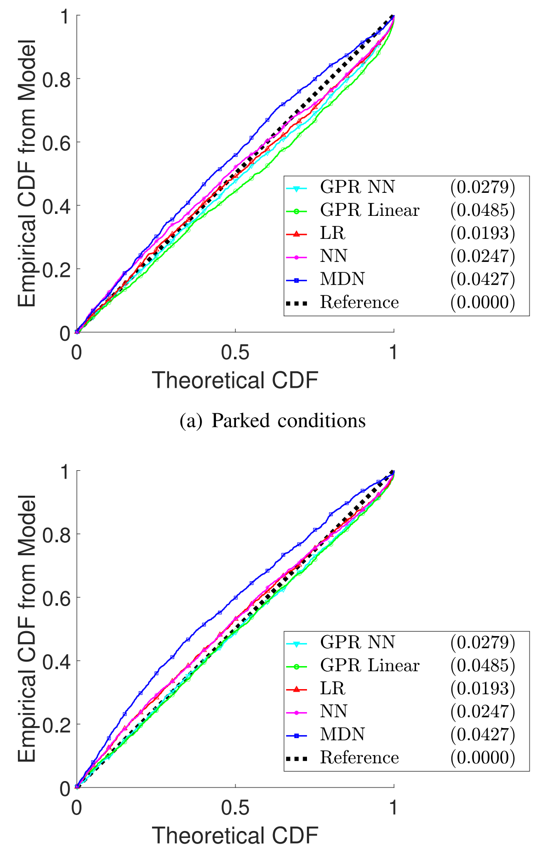
| Method | Parked – Phase 1 | | | Driving – Phase 2 | | |
|---|---|---|---|---|---|---|
| | 1% [%] | 2% [%] | 4% [%] | 1% [%] | 2% [%] | 4% [%] |
| LR | 57.5 | 80.1 | 91.4 | 66.1 | 81.6 | 92.9 |
| NN | 56.8 | 73.0 | 89.0 | 52.4 | 74.8 | 90.2 |
| MDN | 52.4 | 73.2 | 89.5 | 63.9 | 81.2 | 94.0 |
| GPR NN | 71.2 | 83.7 | 95.7 | 66.7 | 83.7 | 95.1 |
| GPR Linear | 68.8 | 80.4 | 94.3 | 71.0 | 86.2 | 96.5 |

To quantify the spatial resolution of the models, Table I lists the area of the confidence interval at 50%, 75% and 95% accuracies for the baseline and GPR models. We observe that the areas of the confidence interval for the GPR models are smaller than the areas for the baseline models. In phase 1, GPR NN has the smallest area for the 95% confidence interval (3.76%). In phase 2, GPR Linear has the smallest area for the 95% confidence interval (3.77%). The ability to provide high accuracy within a small region makes the GPR models more efficient.

To quantify the accuracy of the models, Table II lists the accuracy observed when the fractions of a sphere surrounding the driver's head is 1%, 2% and 4%. This analysis quantifies the performance of the proposed and baseline models when their confidence intervals have consistent area. We observe that we can get 86.2% accuracy in phase 2 within an area of 2% with the GPR Linear model. Similarly on phase 1, we can obtain an accuracy of 83.7% with the GPR NN model.

*2) Theoretical Versus Empirical Cumulative Density Function:* Since our proposed and baseline models assume that the estimated gaze follow a Gaussian distribution, it is important to analyze how well this Gaussian assumption holds with respect to the empirical distribution of the ground truth data around the estimations. For this analysis, we plot the fraction of the data observed within the confidence region (y-axis) as a function of the theoretical *cumulative density function* (CDF) of the region (x-axis). Fig. 8 shows the results for the parked and driving conditions. Ideally, we should observe the curves as close as possible as the reference diagonal curve (black curve). We measure the absolute area between each curve and the reference diagonal curve using (16). The legend in Fig. 8 reports the results.

$$area = \frac{1}{N} \sum |cdf_{theoretical} - cdf_{empirical}| \tag{16}$$



(a) Parked conditions

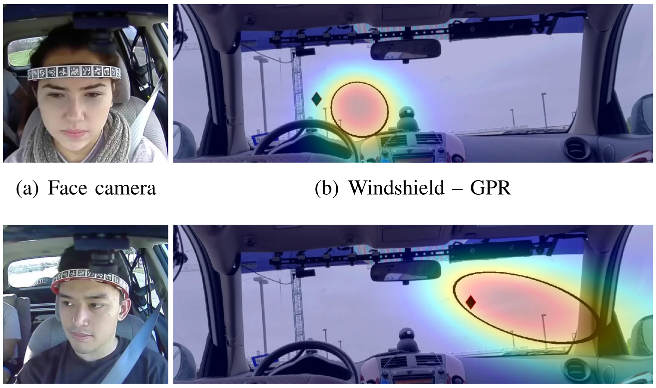
(b) Driving Conditions

Fig. 8. Theoretical versus empirical cumulative distribution function for the GPR and baseline models. This figure evaluates whether the resulting probabilistic salient visual maps cover the target gaze direction as estimated by the Gaussian assumption in the models. The numbers in the legend quantify the fit using (16).

In the parked condition, the LR model is the closest to the reference diagonal curve. Since the distribution of the data is structured, a simple linear regression model with constant mean square error is enough to properly match the theoretical distribution for the confident intervals, although with lower accuracies and spatial resolutions than our proposed models (Tables I and II). In the driving condition, the two GPR models are very close to the theoretical curve. The absolute areas from the reference diagonal curve are smaller than the corresponding absolute area for the baseline models. Therefore, the GPR models not only provide better tradeoff for accuracy and spatial resolution, but also offer confidence intervals that are closer to the theoretical confidence intervals for the most important condition (phase 2).

## *C. Mapping the Confidence Regions Onto the Windshield*

This section projects the estimated confidence regions onto the windshield. We have the marker position, which is used as the ground truth for the gaze. We only use the targets markers from #1 to #13 for this purpose (Fig. 1(c)). We model the windshield as a plane by fitting the best plane containing these thirteen points. Small errors are introduced because the windshield is slightly curved so the points do not exactly lie on a plane. Therefore,



(a) Face camera (b) Windshield – GPR

(c) Face camera (d) Windshield – GPR

Fig. 9. Two examples of projections of the probabilistic salient visual maps into the windshield. The target marker is highlighted with a black diamond. The darkened curves represent 50% confidence intervals.

when we project the original points back to the camera, they do not exactly match the target marker location (Fig. 9). From the gaze angles ($\alpha$ and $\beta$), the gaze direction is obtained by estimating the line from the position of the head ($[x_{hp}, y_{hp}, z_{hp}]$) towards the direction provided by the gaze vector. Equation 17 provides the projection used in the study. We estimate the region where a line meets the windshield plane. The probability density function at each point is calculated based on the probabilistic salient visual map created by the models.

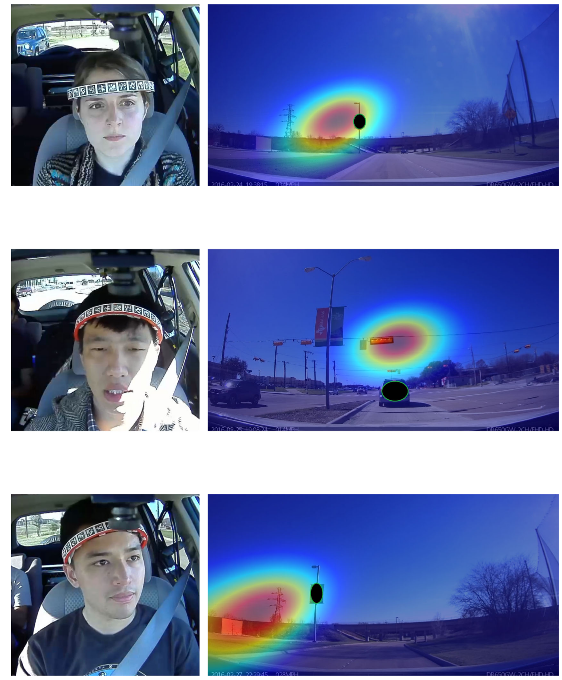
$$[x, y, z] = [x_{hp}, y_{hp}, z_{hp}] + [\sin(\alpha), \cos(\alpha)\sin(\beta), \cos(\alpha)\cos(\beta)] \quad (17)$$

Fig. 9 shows two examples for the confidence regions created with the GPR model. While these are just two examples, they are representative of the probabilistic salient visual map created by the models. These figures also demonstrate how the estimated angles can be mapped onto the real world coordinates. The figure shows that the confidence regions in front of the drivers are smaller than the confidence regions on the side of the windshield, signaling more uncertainty. The size of the regions is learned from the data.

### D. Mapping Confidence Regions Onto the Road

We also projected the estimated confidence regions onto the road view. As explained in Section III-A, we asked three of the subjects to look at multiple targets on the road. For these cases, we approximate the gaze distribution in the road by projecting the confident regions at different distances from the car, ranging from 10 to 200 meters in increments of 10 meters (i.e., 20 different projections). Then, we calculated the unweighted average of the probabilities for each pixel creating a 2D visual map projected on the road camera.

Fig. 10 gives three examples, showing the driver's face, the road view, and the estimated salient visual map created with the GPR models. The target object is highlighted with a black ellipse. We observe that the GPR models perform reasonably well providing an estimation around the target regions attracting the attention of the driver. Notice that in this study we only consider the position and orientation of the head.



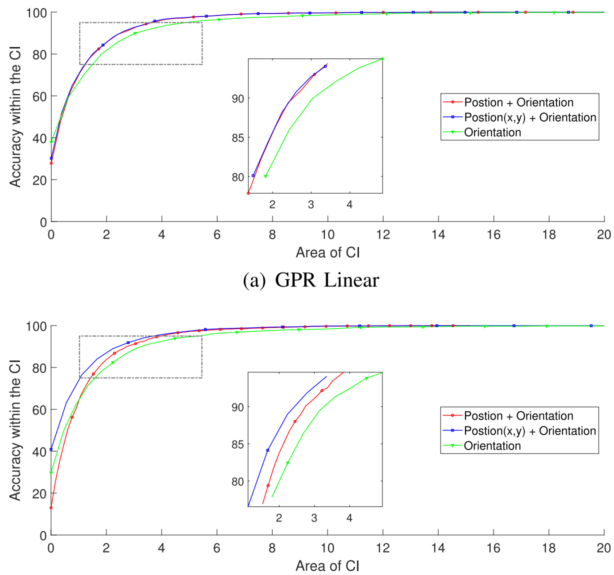
(e) Face camera (f) Mapping on road, GPR

Fig. 10. Three examples of projections of the probabilistic salient visual maps onto the road. These regions are estimated at different distance, combining the results into a single probabilistic map. The target marker is highlighted with a black ellipse.

We observe that the estimated probabilistic salient visual maps do not always include the true gaze target. These cases are useful to identify some limitations of our model to project the region on the road. First, we add some distortion during the projections, as discussed before. Second, some subjects may depend on subtle eye movements that our models do not capture. Notice that the eye information is not used by our models, which is used in most of the gaze detection system designed for HCI in controlled environment. Third, the inter-driver variability can impact the results, as differences in height and driving behaviors can affect the relationship between head movements and gaze. In spite of these limitations, the results in this paper demonstrate that our models effectively capture the visual attention of the drivers by just modeling their head pose. We include a video with the results as a supplemental document.

### E. Gaze Angle Estimation With Limited Head Pose Information

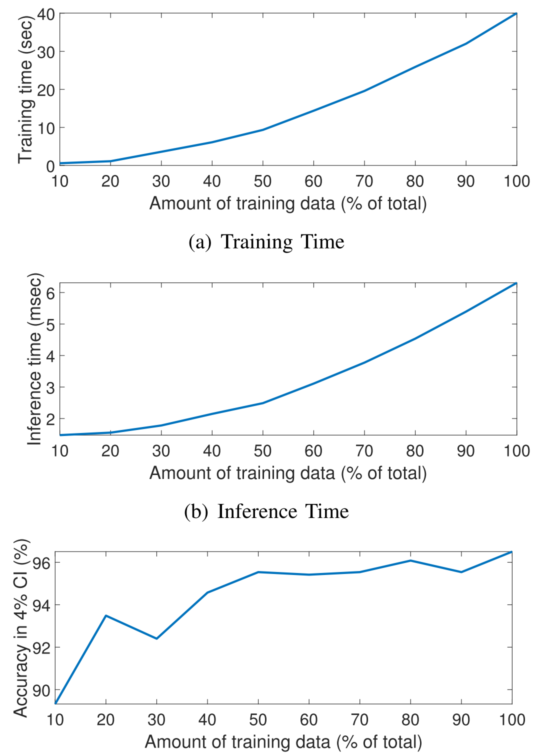
RGB cameras are the most common sensors that are used to capture the driver data in the car. Since regular cameras lack depth information, it is not possible for algorithms to reliably estimate the head position in all three degrees of freedom. Our GPR models require this information to estimate the probabilistic salient visual maps. Therefore, we retrain our GPR models by using only head orientation, or by augmenting head orientation with partial head position. We consider two conditions. The




(a) GPR Linear

(b) GPR NN

Fig. 11. Comparison of the accuracy versus temporal resolution of GPR models implemented with limited head pose information. We zoom the plots for better visualization. The results are separately reported for GPR linear and GPR NN. The figure is better viewed in colors. Accuracy is calculated within a *confidence interval* (CI) given by the area.




(a) Training Time

(b) Inference Time

(c) Accuracy at 4% confidence interval (CI)

Fig. 12. Analysis of the training time, inference time and performance of the model as a function of the training set size. The analysis is implemented with the GPR linear model with ARD kernel using phase 2 of the corpus.

first condition only considers the head orientation (i.e., 3D vector). The second conditions is head orientation plus the $x$ and $y$ position of the head, estimated with the AprilTag-based headband. These models are compared with the GPR models trained with the 6D vector, including full orientation and position of the head.

Fig. 11 presents the results for the GPR models on the test set in phase 2 (driving condition). The GPR linear model with orientation and partial position information achieves results that are very close to the results achieved by the full model (Fig. 11(a)). The GPR NN model implemented with partial head position even outperforms the results of the model with full information (Fig. 11(b)). We conclude that the distance between the driver and the camera is not critical to build an effective model. We hypothesize that this results is due to the reduced head movements along the z-direction observed while a driver is operating a vehicle. The performances of both models drop when they are exclusively trained with head orientation, indicating that some information about the head position is needed.

### F. Training and Inference Time Versus Train Set Size

This section discusses the complexity of the algorithm and its dependency on the train set size. This analysis uses the GPR linear model with ARD kernel using phase 2 of the corpus. We use a single computer with a 64-bit Intel Xeon CPU and 32 GB RAM. We study the training and inference time of our approach when only a portion of the training data is used.

Fig. 12(a) shows the training time when we gradually increase the training data from 10% to 100% of the training data. We randomly select the data that we add to the training set. As expected, there is a consistent increase in the training time when adding more training data. However, Fig. 12(a) shows that our model can be trained in less than one minute, as opposed to deep-learning models that are computationally intensive. Similarly, Fig. 12(b) shows the inference time to evaluate one sample as we increase the size of the training set. We observe a similar trend seen in Fig. 12(a), where the inference time increases as we add more data for training. Since we model a joint Gaussian process with the training data ( (5)), the complexity during inference increases with the training data size. However, a closed form solution is available, so the inference model is practical. We observe that even when using all the training data, the inference of a sample takes about 6 milliseconds. Our approach is 5.56 faster than real-time for a video stream at 30 fps.

### G. Performance Versus Training Set Size

Finally, we study how the performance of the model changes when using part of the training data. For this purpose, we calculate the accuracy within 4% of the sphere around the driver's head (Fig. 12(c)) We observe that, as we increase the training data, the performance of the model gets better, as expected. We observe dramatic changes in the performance gain while adding data to a small fraction of the training set. However, the performance gain is minimum with additional data, when the data is sufficiently large. This result leads us to conclude that while adding a large amount of data with added diversity may increase the performance, the amount of data that we have used in the current experiment is reasonable to demonstrate the benefits of the proposed framework.

## VI. Conclusion

This paper proposed a novel probabilistic model based on GPR to define a salient visual map to estimate the driver's gaze location. The proposed method estimates confidence regions containing the gaze direction of the driver using only the position and orientation of her/his head. The size of the confidence region is determined by the uncertainty of the model in the estimated region (heteroscedastic model). To demonstrate the potential use of the proposed method, we projected this salient visual map onto the windshield and the road images. The results demonstrated reasonable performance, achieving accuracies higher that the baseline models. An appealing feature of having a distribution describing the driver's visual attention is the opportunity to operate with different tradeoffs between accuracy and spatial resolution. For example, the GPR Linear model implemented with ARD can reach a 86.2% accuracy in phase 2 (e.g., driving condition) within an area of 2% of the sphere around the driver's head.

There are open challenges to accurately estimate the six degrees of freedom for the driver's head in real driving conditions. We use AprilTags for this purpose in our analysis. Using a single RGB camera, it can be difficult to track the head pose in a vehicle when the rotation is higher than a given threshold (e.g., when the face is not completely visible [13]). We also need to estimate the distance of the driver's head from the camera, which will affect the gaze angle. To address this challenge, we are working on using depth sensors to reliably estimate the orientation and position of the head [37], [38]. One of the limitations of the head band used in this study is that it covers parts of the face. We are working on an alternative design that addresses this limitation [47]. This type of data collection protocol serve as a valuable resource to train and evaluate head pose algorithms in real driving scenarios, advancing algorithm development in this area.

This study opens various potential areas for research where the predicted driver visual attention can be used as a starting point to improve the safety on the road. The proposed technology can play an important role in vehicle applications for security, infotainment, and navigation. The study relies on a commercial dash camera that can be easily installed on regular vehicles. This setting is ideal for in-vehicle solutions in all cars, regardless of their proprietary built-in sensors. Once the salient visual map is created and projected onto the road scene, we can leverage computer vision algorithms to detect target objects within the highlighted area. For example, patterns of changes in visual attention can be correlated to the external environment and/or driving anomalies detected on the car [48], [49]. We can look more closely at the region and estimate what possible objects the driver is directing her/his gaze (other vehicles, pedestrians, or billboards). We can also determine important objects that a driver fails to attend, creating an appropriate warning. Furthermore, unnecessary warnings to drivers can be avoided if we infer that the driver is aware of specific objects/people on the road (e.g., a pedestrian crossing the street). The probabilistic saliency map can be used by an ADAS to identify cases where the driver is performing a maneuver without paying attention to other vehicles. This approach can also be used for multimodal navigation systems [2], where the predicted probabilistic saliency map is used to understand navigation commands (e.g., queries such as "what is that store?" while looking at a given building). The predicted probabilistic maps can be ideal for these scenarios. Likewise, we expect that a model adapted to a given driver can lead to better performance, as the variance in the relation between head pose and gaze will be reduced. As a future work, we will explore adaptation schemes using unsupervised methods that take unlabeled data from the target driver to adapt the model, leading to better personalized systems. Another possible improvement is to rely on temporal modeling by constraining the network on previous frames. We have seen improvements in head pose estimation by adding temporal modeling [38], so we expect similar improvements for gaze prediction. To achieve this goal, we need continuous gaze movements with ground truth information, which was not collected in our data. We are collecting new recordings with an improved protocol that includes continuous gaze [50].

## References


[1] S. Klauer, T. Dingus, V. Neale, J. Sudweeks, and D. Ramsey, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," Nat. Highway Traffic Saf. Admin., Blacksburg, VA, USA, Tech. Rep. DOT HS 810 594, Apr. 2006.

[2] T. Misu, "Visual saliency and crowdsourcing-based priors for an in-car situated dialog system," in *Proc. Int. Conf. Multimodal Interact.*, Seattle, WA, USA, 2015, pp. 75–82.

[3] N. Li and C. Busso, "Predicting perceived visual and cognitive distractions of drivers with multimodal features," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 51–65, Feb. 2015.

[4] H. Koma, T. Harada, A. Yoshizawa, and H. Iwasaki, "Detecting cognitive distraction using random forest by considering eye movement type," *Int. J. Cogn. Informat. Natural Intell.*, vol. 11, no. 1, pp. 16–28, Jan.–Mar. 2017.

[5] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 980–992, Apr. 2016.

[6] A. Doshi and M. Trivedi, "Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions," in *Proc. IEEE Intell. Veh. Symp.*, Xi'an, China, 2009, pp. 887–892.

[7] Z. Wang, R. Zheng, T. Kaizuka, and K. Nakano, "Relationship between gaze behavior and steering performance for driver-automation shared control: A driving simulator study," *IEEE Trans. Intell. Veh.*, vol. 4, no. 1, pp. 154–166, Mar. 2019.

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

[9] C. Ahlstrom, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 965–973, Jun. 2013.

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

[11] S. Baluja and D. Pomerleau, "Non-intrusive gaze tracking using artificial neural networks," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-94-102, Jan. 1994.

[12] S. Jha and C. Busso, "Estimation of gaze region using two dimensional probabilistic maps constructed using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, U.K., 2019, pp. 3792–3796.

[13] S. Jha and C. Busso, "Challenges in head pose estimation of drivers in naturalistic recordings using existing tools," in *Proc. IEEE Int. Conf. Intell. Transp.*, Yokohama, Japan, 2017, pp. 1–6.

[14] A. Tawari and M. Trivedi, "Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos," in *Proc. IEEE Intell. Veh. Symp.*, Dearborn, MI, USA, 2014, pp. 344–349.

[15] S. J. Lee, J. Jo, H. G. Jung, K. R. Park, and J. Kim, "Real-time gaze estimator based on driver's head orientation for forward collision warning system," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 254–267, Mar. 2011.

[16] M. C. Chuang, R. Bala, E. A. Bernal, P. Paul, and A. Burry, "Estimating gaze direction of vehicle drivers using a smartphone camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Columbus, OH, USA, Jun. 2014, pp. 165–170.

[17] M. Rezaei and R. Klette, "Look at the driver, look at the road: No distraction! no accident!," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 129–136.

[18] S. Jha and C. Busso, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, Rio de Janeiro, Brazil, 2016, pp. 2157–2162.

[19] S. Jha and C. Busso, "Probabilistic estimation of the driver's gaze from head orientation and position," in *Proc. IEEE Int. Conf. Intell. Transp.*, Yokohama, Japan, 2017, pp. 1–6.

[20] S. Jha and C. Busso, "Probabilistic estimation of the gaze region of the driver using dense classification," in *Proc. IEEE Int. Conf. Intell. Transp.*, Maui, HI, USA, 2018, pp. 697–702.

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

[22] G. H. Robinson, D. J. Erickson, G. L. Thurston, and R. L. Clark, "Visual search by automobile drivers," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 14, no. 4, pp. 315–323, Aug. 1972.

[23] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, and D. Crundall, "Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers," *Ergonomics*, vol. 46, no. 6, pp. 629–646, May 2003.

[24] M. Sodhi, B. Reimer, and I. Llamazares, "Glance analysis of driver eye movements to evaluate distraction," *Behav. Res. Methods, Instrum., Comput.*, vol. 34, no. 4, pp. 529–538, Nov. 2002.

[25] M. Kutila, M. Jokela, G. Markkula, and M. Rue, "Driver distraction detection with a camera vision system," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, Texas, USA, vol. 6, pp. VI-201–VI-204, 2007.

[26] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 340–350, Jun. 2007.

[27] E. Murphy-Chutorian and M. Trivedi, "HyHOPE: Hybrid head orientation and position estimation for vision-based driver head tracking," in *Proc. IEEE Intell. Veh. Symp.*, Eindhoven, The Netherlands, 2008, pp. 512–517.

[28] S. Alletto, A. Palazzi, F. Solera, S. Calderara, and R. Cucchiara, "DR(eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Las Vegas, NV, USA, Jun./Jul. 2016, pp. 54–60.

[29] M. Bojarski *et al.*, "End to end learning for self-driving cars," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS 2016) Deep Learn. Symp.*, Dec. 2016, pp. 1–9.

[30] K. Zeeb, A. Buchner, and M. Schrauf, "Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving," *Accident Anal. Prevention*, vol. 92, pp. 230–239, Jul. 2016.

[31] S. Vora, A. Rangesh, and M. M. Trivedi, "On generalizing driver gaze zone estimation using convolutional neural networks," in *Proc. IEEE Intell. Veh. Symp.*, Los Angeles, CA, USA, 2017, pp. 849–854.

[32] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. Hansen, "UT-Drive: Driver behavior and speech interactive systems for in-vehicle environments," in *Proc. IEEE Intell. Veh. Symp.*, Istanbul, Turkey, 2007, pp. 566–569.

[33] P. Angkititrakul *et al.*, "Getting start with UTDrive: Driver-behavior modeling and assessment of distraction for in-vehicle speech systems," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, Antwerp, Belgium, Aug. 2007, pp. 1334–1337.

[34] J. Hansen, C. Busso, Y. Zheng, and A. Sathyanarayana, "Driver modeling for detection and assessment of driver distraction: Examples from the UTDrive test bed," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 130–142, Jul. 2017.

[35] J. Jain and C. Busso, "Analysis of driver behaviors during common tasks using frontal video camera and CAN-Bus information," in *Proc. IEEE Int. Conf. Multimedia Expo*, Barcelona, Spain, 2011, pp. 1–6.

[36] N. Li and C. Busso, "Analysis of facial features of drivers under cognitive and visual distractions," in *Proc. IEEE Int. Conf. Multimedia Expo.*, San Jose, CA, USA, 2013, pp. 1–6.

[37] T. Hu, S. Jha, and C. Busso, "Robust driver head pose estimation in naturalistic conditions from point-cloud data," in *Proc. IEEE Intell. Veh. Symp.*, Las Vegas, NV, USA, 2020, pp. 1176–1182.

[38] T. Hu, S. Jha, and C. Busso, "Temporal head pose estimation from point cloud in naturalistic driving conditions," *IEEE Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2021.3075350.

[39] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proc. IEEE Int. Conf. Robot. Automat.*, Shanghai, China, 2011, pp. 3400–3407.

[40] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. IEEE Conf. Autom. Face Gesture Recognit.*, Xi'an, China, 2018, pp. 59–66.

[41] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Sect. A.*, vol. 34, no. 5, pp. 827–828, Sep. 1978.

[42] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds., Berlin, Germany: Springer, Oct. 2004, pp. 63–71.

[43] C. Bishop, "Mixture density networks," Aston Univ., Birmingham, U.K., Tech. Rep. NCRG/94/004, Feb. 1994. [Online]. Available: http://www.ncrg.aston.ac.uk/

[44] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 1–13.

[45] F. Chollet, "Keras: Deep learning library for Theano and TensorFlow," Apr. 2017. [Online]. Available: https://github.com/fchollet/keras

[46] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. Symp. Operating Syst. Des. Implementation*, Savannah, GA, USA, 2016, pp. 265–283.

[47] S. Jha and C. Busso, "Fi-CAP: Robust framework to benchmark head pose estimation in challenging environments," in *Proc. IEEE Int. Conf. Multimedia Expo*, San Diego, CA, USA, 2018, pp. 1–6.

[48] Y. Qiu, T. Misu, and C. Busso, "Driving anomaly detection with conditional generative adversarial network using physiological and can-bus data," in *Proc. ACM Int. Conf. Multimodal Interaction*, Suzhou, Jiangsu, China, 2019, pp. 164–173.

[49] Y. Qiu, T. Misu, and C. Busso, "Use of triplet loss function to improve driving anomaly detection using conditional generative adversarial network," in *Proc. Intell. Transp. Syst. Conf.*, Rhodes, Greece, 2020, pp. 1–7.

[50] 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 Trans. Intell. Transp. Syst.*, to be published, doi: 10.1109/TITS.2021.3095462.


**Sumit Jha** (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology, Trichy, India, in 2012, and the M.S. degree in electrical engineering in 2016 from the University of Texas at Dallas (UTD), Richardson, TX, USA, where he is currently working toward the Ph.D. degree. At UTD, he has been a part of the Multimodal Signal Processing (MSP) Laboratory since 2015. His research interests include machine learning computer vision solutions for driver monitoring and in-vehicle safety systems.

**Carlos Busso** (Senior Member, IEEE) received the B.S. and M.S. degrees (high Hons.) in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2008. He is an Associate Professor with the Electrical Engineering Department, University of Texas at Dallas (UTD), Richardson, TX, USA. He was selected by the School of Engineering of Chile as the best Electrical Engineer graduated in 2003, across Chilean universities. From 2003 to 2005, he received a Provost Doctoral Fellowship from USC, and from 2007 to 2008, a Fellowship in Digital Scholarship. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory. He was the recipient of the NSF CAREER Award in 2014, the ICMI Ten-Year Technical Impact Award in 2015, his student received the third Prize IEEE ITSS Best Dissertation Award (N.Li), 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). He is the Co-Author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research focuses on human-centered multimodal machine intelligence and applications. His current research interests include affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the General Chair of ACII 2017 and ICMI 2021. He is a Member of ISCA, AAAC, and a Senior Member of ACM and IEEE.