Exploring the Situational Awareness of Humans inside Autonomous Vehicles

Akshay Rangesh\(^1\), Nachiket Deo\(^1\), Kevan Yuen\(^1\), Kirill Pirozhenko\(^1\), Mohan M. Trivedi\(^1\), Heishiro Toyoda\(^2\) and Pujitha Gunaratne\(^2\)

Abstract — With increasing automated driving capabilities of commercial vehicles, the study of safe and smooth occupant-vehicle interaction and control transitions is key. In this study, we focus on the development of contextual, semantically meaningful representations of driver and vehicle states, which can then be used to determine the appropriate timing and conditions for transfer of control between driver and vehicle. To this end, we lay out the specifications of the vehicle platform required to conduct such a study, and explore some of the sensors and algorithms that may be needed to produce useful and observable high level cues (features) to make such decisions. These features encode different aspects of the driver state, pertaining to the face, hands, foot and upper body of the driver. Finally, we evaluate these features on their capability of capturing the state of a driver, and demonstrate a strong agreement between these features and a humans’ notion of situational awareness.

I. INTRODUCTION

Motivations for studying driver behavior in the context of highly automated vehicles can be found aplenty in human factors studies. It is widely regarded that as soon as the level of cognitive stimulation falls below a persons’ own comfortable “set point”, the person will seek out alternate/additional sources of information, leading to distraction. This makes the intermediate levels of automation (as per NHTSA or SAE defined degrees of automation) very dangerous, causing problems such as inattention, trust, skill atrophy, complacency etc. [3]. The authors in [3] postulate that rising levels of automation will lead to declining levels of awareness. They also state that most problems are expected to arise in systems that take the driver out of the loop, yet these are the very systems that drivers want, because they free the driver to do something else of interest. Elsewhere, the authors in [16] emphasize the irony of automation, whereby “the more advanced a control system is, the more crucial may be the contribution of the human operator”. They also acknowledge that decades of research has shown that humans are not particularly good at tasks that require vigilance and sustained attention over long periods of time.

1 Laboratory for Intelligent and Safe Automobiles, UC San Diego
2 Toyota Collaborative Safety Research Center (CSRC)
3 Equal contribution from authors

Fig. 1. An example illustration which showcases the importance of understanding and modeling driver behavior in order to ensure safe and smooth transfer of control between automation and humans.

All above points seem to suggest that a drop in attention is inherent in human behavior. Thus, it is not a matter of if, but when the driver will resort to non-ideal behavior. This makes the safe and smooth handling of control transitions extremely important and timely. Consider the scenario illustrated in Fig. 1. This figure indicates that the transition of control from an autonomous agent to the human driver should be a function of the state of the driver. A system that takes the state of the driver into account can decide between handing over control if the driver is ready, versus coming to a safe and smooth halt if not. Driver state can also dictate how and when a takeover alert must be supplied to ensure an uneventful transition of control. In this study, we explore the platform, sensors and algorithms that would be part of such a system and justify our choices with careful reasoning and quantitative evaluations.

II. RELATED RESEARCH

The last few decades have produced numerous studies on the estimation and understanding of driver vigilance, activity, attention, inattention, distraction and awareness. In addition to this, phenomena such as driver fatigue, drowsiness etc. are often attributed as factors that influence driver behavior, and thus fall within the purview of these studies. We consider all topics above to be related to our research, and use the phrase “driver behavior studies” to collectively refer to such work. However, since our major focus is on control transitions between a vehicle and its driver, we ignore studies primarily focused on driver behavior/intent prediction due to the implication that the driver is already under control. For this same reason, we deem methods based on driving performance (methods that monitor how well a driver controls the vehicle) irrelevant to our application and focus entirely on methods that monitor the physiological
state of a driver. Table I lists representative studies that fall under this category.

Based on the related research in the following ways: First, we conduct the study under the assumption of a highly automated testbed, where the human drives intermittently. This promotes secondary activities of humans inside the vehicle to an extent that has not been studied before. We additionally consider an eventuality where any human within a vehicle may assume control of the vehicle, and design our algorithms so that they may be easily extended to cover multiple persons of interest. Next, we note that prior studies mostly focus on a limited set of features to capture the driver state. Although we make an effort to cover most traditional features of interest (e.g. PERCLOS), we additionally explore a new set of features that have not have been produced or analyzed in this context before. We do this by overcoming previous algorithmic challenges, making use of different sensor modalities, and by mounting sensors in previously untested locations within the vehicle cabin. All put together, we propose a pipeline that reliably generates these features of interest, and carry out the analysis of these features on a large dataset comprising of naturalistic driving data in a variety of conditions, captured using a unique testbed.

## III. Testbed

### A. Platform

Since our emphasis is on the collection and analysis of driver behavior data in a naturalistic driving scenario, our testbed must be based on a real production vehicle available for purchase by the consumer. Also, as the project focuses on the behavior and interactions of a driver with a highly automated vehicle, a vehicle with state-of-the-art automotive functionality is desired. Our testbed **LISA-T**, built on top of the Tesla Model S, best satisfies the above requirements.

Based on available specifications and claims made by Tesla, we are confident that this testbed would continue to represent the state-of-the-art in commercially available autonomous/highly automated vehicles for the next few years, with constant software updates to add or improve existing autonomous functionality.

### B. Sensors

Our choice of sensors was directed by two main factors. First, the sensors need to be non-intrusive and should not inhibit the normal behavior of the driver in any manner. This precludes the use of EEG, PPG and other sensors attached directly to the driver. Second, the sensors must...
be relatively cheap and easily available. With these constraints in mind, we adopt a sensor suite made up of RGB cameras, a depth camera, and small form factor infrared sensors. The placement of each sensor inside the vehicle cabin can be found in Fig. 2. We describe each sensor modality and its purpose below.

1) **RGB Cameras:** Since most vision algorithms are dependent on the quality of images captured, we chose the GoPro Hero 4 Black as our vision sensor. This provides us many benefits such as high resolution real time image capture, large fields of view (FoVs), large depth of field, small form factor, all without the need for additional compute resources.

All GoPros are specified to record with $1920 \times 1440$ resolution at 30fps, with a wide FoV. Although we capture high resolution images, we downsample all images by at least a factor of 4 before processing them. Having a higher resolution offers us the flexibility in case we require larger images in the future.

The RGB cameras are positioned to monitor specific aspects of driver behavior. We categorize them as follows:

- **Dashboard camera:** captures the information and instructions relayed by the testbed to the driver at any given instant. This includes the vehicle speed, take-over requests, cruise control and auto-steering status indicators, estimated lane markers, detected vehicles, and auto-lane changing status.
- **Cabin camera** (Fig. 3): used to carry out facial analysis of the driver and passenger.
- **Hand camera:** used for analyzing the hands of the driver and passenger.
- **Pose camera** (Fig. 6): used to calculate the upper body pose of the driver.
- **Foot camera** (Fig. 5): used for coarsely localizing the feet of the driver.

2) **Depth Camera:** In addition to the RGB cameras, we use a single depth camera for localizing and analyzing the hands in 3D (see Fig. 4). For this study, we choose a Kinect v2 sensor for its ease of use, excellent community support and high-resolution depth images.

3) **Infrared Sensors:** To measure the distance of each foot to the gas and break pedals, we make use of a time-of-flight range finder for the VL6180 infrared (IR) emitter and receiver (manufactured by SparkFun). These sensors are mounted on the backside of each pedal cover, with a small hole drilled in the front to let IR light pass through. They provide accurate range measurements up to 25cm, and are used in conjunction with the foot camera for accurate foot localization.

IV. ALGORITHMS FOR DRIVER BEHAVIOR ANALYSIS

A. Face and Gaze Analysis

The algorithm workflow for facial analysis is depicted in Fig. 3. The input to this block is an image from the cabin camera, and the outputs (features) include the gaze zone, blink frequency, PERCLOS, eye closure duration, mouth opening and head pose. The facial landmarks are obtained using OpenPose [26], and the gaze zone is obtained using the network proposed in [31].

B. Hand Analysis

For analyzing the hands of the driver, we use the recently proposed HandyNet [23]. This network takes in an image from the depth sensor and outputs the segmentation mask of each hand instance, and the object held in each hand. The segmentation output can then be used to obtain accurate 3D distances of each hand from control elements like the steering wheel. Alternatively, one could use RGB images from the hand camera and adopt an approach similar to the one proposed in [35]. The features generated by this block include, the 3D distance of each hand to the wheel, and the object held in each hand. This process is illustrated in Fig. 4.

C. Foot Analysis

The algorithm workflow for foot analysis consists of two independent processes (seen in Fig. 5). First, we carry out a coarse localization of the nearest foot of the driver using images from the foot camera. We achieve this by training a convolutional neural network (CNN) designed for classifying a foot camera image directly into an activity class (listed in Fig. 5). We use the “lightweight” SqueezeNet [11] architecture, and finetune it for the proposed task. In addition to the classification output above, we also use the distance of the nearest foot
We intend to estimate the situational awareness of the occupants of the cabin based on the analysis of video feed obtained from cameras viewing the driver and co-passengers. However, to train supervised learning algorithms to estimate situational awareness, we require ground-truth that is faithful to the task. We propose to use mean ratings provided by human raters, observing only the video feed from the cameras, as the ground-truth. We term this the observable situational awareness (OSA) index. We evaluate the consistency of human ratings and the correlation of OSA with various outputs of the driver behavior analysis algorithms described.

### A. Data

We consider eight 1-minute video clips of naturalistic drive data consisting of 6 different drivers. The clips were captured with the vehicle running on auto-pilot. Different driver behaviors that lead to distracted driving, such as, texting, operating the instrumentation panel, talking to a co-passenger, and eating or drinking were
included in the data. Each video clip was rated by 5 different raters. Each clip was divided into snippets of 2 seconds, each of which was assigned a rating from 1 to 5 by the raters, where 1 corresponds to the lowest level of situational awareness and 5 corresponds to the highest. We used feed from a single view in order to not overwhelm the raters. We limited the experiment to the pose camera since it captures the most information, including facial, hand, and pose cues.

**B. Normalizing for rater bias**

One source of variance in the human ratings is rater bias and spread. Raters can be strict or lax, and use a varying range of values. We posit that the trend in the rater values is the more invariant quantity, and a better representation of the ground truth. This requires normalization of the ratings. We normalize for rater bias and spread by using percentiles. First, we calculate the percentile value of each rating within the same raters ratings. This rating is then reassigned to the global rating, pooled across all raters, corresponding to the same percentile value.

**C. Inter-rater agreement**

We use intraclass correlation co-efficients (ICCs) as formulated by McGraw et. al. [21], to evaluate inter-rater agreement. We model the human ratings as a two-way mixed-effect model without interaction, assuming n observations and k raters. Under this model, the rating \( x_{ij} \) assigned by rater \( j \) to clip \( i \) can be expanded as,

\[
x_{ij} = \mu + r_i + c_j + e_{ij},
\]

(1)

where, \( \mu \) is the global average rating, \( r_i \)’s are the deviations due to the different clips which are independent and are normally distributed with mean 0 and variance \( \sigma^2_r \), \( c_j \)’s are fixed deviations corresponding to the rater bias, with

\[
\sum_j c_j = 0
\]

(2)

and

\[
\sum_j c^2_j/(k-1) = \theta^2_c.
\]

(3)

And finally, \( e_{ij} \) is the normally distributed measurement error with zero mean and variance \( \sigma^2_e \).

We report the following ICC values for the normalized and unnormalized ratings, each of which is defined in [21].

- **ICC(C,1):** This is given by the expression

\[
ICC(C,1) = \frac{\sigma^2_r}{\sigma^2_r + \sigma^2_e}.
\]

(4)

We report the following ICC values for the normalized and unnormalized ratings, each of which is defined in [21].

<table>
<thead>
<tr>
<th>Rating type</th>
<th>ICC(C,1)</th>
<th>ICC(A,1)</th>
<th>ICC(A,k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnormalized</td>
<td>0.556</td>
<td>0.428</td>
<td>0.789</td>
</tr>
<tr>
<td>Normalized</td>
<td>0.565</td>
<td>0.560</td>
<td>0.864</td>
</tr>
</tbody>
</table>

and can be interpreted as the degree of consistency of the rating values. This is independent of the rater bias, and has a high value if the trend of ratings across raters is consistent.

- **ICC(A,1):** This is given by the expression

\[
ICC(A,1) = \frac{\sigma^2_r}{\sigma^2_r + \theta^2_c + \sigma^2_e}.
\]

(5)

This is the degree of absolute agreement of rater values. This has a high value only if the raters are in agreement in terms of the actual value of the ratings.

- **ICC(A,k):** This is given by the expression

\[
ICC(A,k) = \frac{\sigma^2_r}{\sigma^2_r + \theta^2_c + \sigma^2_e/k}.
\]

(6)

This is the degree of absolute agreement of the average rating.

All ICC values are bounded between 0 and 1. The \( \sigma \) values are estimated using two-way analysis of vari-ances (ANOVA). Table II shows the ICC values for the normalized and unnormalized ratings. We note from the ICC(C,1) and ICC(A,1) values for the unnormalized ratings that there is considerable consistency among the raters, although the absolute values themselves may not agree. This is fixed by using the percentile based normalization, which considerably improves the ICC(A,1) metric. Finally, we note that there is a high degree of absolute agreement for the average value of the annotator ratings, especially true for the normalized ratings. For subsequent experiments, we define the OSA index for each second snippet of video clips to be the mean value of the normalized ratings of all the human raters for that snippet.

**D. Correlation of features with OSA**

We report the correlation of some of the features extracted with the algorithms described in section IV with the OSA values.

1) **Gaze zones:** We use the frame-wise probabilities assigned by the gaze-zone classifier (described in IV-A) to obtain a 7 dimensional feature vector for each frame.

2) **Distance to wheel:** We calculate the correlation of the OSA values with the 3D distance of the closest hand to the steering wheel for a given frame. This distance is estimated using the procedure detailed in IV-B.
3) **Object held by the driver**: Similar to the gaze zones, we also consider the frame-wise probabilities assigned by HandyNet (see IV-B) to each object class.

4) **Hand activity classes**: We use frame-wise probabilities for hand activity classes given by the model described in [35].

5) **Joint locations from pose estimation**: We consider the $x$ and $y$ locations of 14 upper body joints given by the pose analysis block outlined in IV-D.

Fig. 7 shows the correlation of each of the features with the assigned OSA. We note that features related to hand activity have a higher correlation with OSA. In particular, the distance of the driver’s hands to the steering wheel has a strong negative correlation with OSA. In terms of hand activity, OSA has significant negative correlation with the driver’s hands being in the air and the driver operating the radio, while it has a positive correlation with the driver’s hands being on the wheel or hovering the wheel. The driver having a cell phone in their hands has a strong negative correlation with the OSA values, while the driver not having any object in their hands is positively correlated with OSA.

**E. t-SNE plots of features**

t-Distributed Stochastic Neighbor Embedding (t-SNE) [20] is a non-linear dimensionality reduction algorithm. It maps high-dimensional data to a low dimensional (typically 2-D) space, while preserving the local similarity structure of points from the high dimensional space. The following sections show 2-D plots of the t-SNE map points corresponding to the described features. Each cluster in the 2-D space corresponds to points that are similar in the feature space. We color the points according to the OSA rating assigned by the raters.

Homogeneously colored clusters imply that that OSA ratings are well separated out in that feature space.

1) **Gaze features**: From Fig. 8, we note that the gaze features do not lead to a good separation of the OSA values. However, we do get some homogeneous clusters. For example, clusters A and B correspond to the driver looking at the speedometer and radio respectively. Both of these are assigned low OSA values by the annotators. We note that a large cluster (cluster C) corresponding to the driver looking forward at the road have mixed OSA values. This suggests that gaze alone cannot be used to determine the driver’s situational awareness.

2) **Hand features**: Fig. 9 shows the t-SNE plots for the hand features. We note that these features lead to a slightly better separation of the OSA values. It also shows plots color coded by the distance of the driver’s hands to the wheel and the most probable hand activity and object classes. Cluster A corresponds to the drivers hands being in the air, while holding a cell phone, with the hands far from the steering wheel. These points are assigned the lowest OSA values by the annotators. Cluster E corresponds to the driver operating the radio/infotainment unit, while not holding any object. This cluster also has low OSA ratings. Conversely, clusters C and D correspond to the driver’s hands hovering or operating the steering wheel. These clusters are assigned a high OSA value by the annotators. Finally, cluster B corresponds to the the driver’s hands being on their lap. This cluster has both low as well as high OSA values. However, the ratings can be explained using the distance to the wheel, with a low distance to the wheel corresponding to a high OSA rating.

3) **All Features**: Finally, Fig. 10 shows the t-SNE plot for the gaze, eye and pose features together. We obtain a greater number of clusters due to the high dimensionality of the pose features. However, we see greater separation in the OSA values than that from any single feature stream.

**VI. CONCLUSIONS AND FUTURE WORK**

In this paper, we have presented a framework for estimating the situational awareness of occupants of an autonomous vehicle using naturalistic drive data. A novel testbed geared towards occupant behavior analysis in autonomous vehicles, with multiple sensor modalities has been described. Algorithms for holistic analysis of occupant behavior and state, including facial, hand, foot and pose based cues have been described. A measure of situational awareness of occupants based on human ratings, termed observable situational awareness (OSA) has been proposed and evaluated in-terms of inter-rater agreement and correlation with features extracted using
the described algorithms. Intra-class correlation coefficients of the human ratings show consistency across human raters. Correlation values and t-SNE plots suggest that the features extracted contain useful cues for estimating the OSA. We are currently involved in expanding the dataset rated by the human raters, and improving the size and diversity of the rater pool. Subsequently, we hope to train supervised learning algorithms to estimate the OSA value.

REFERENCES


