Gaze Fixations and Dynamics for Behavior Modeling and Prediction of On-road Driving Maneuvers

Sujitha Martin and Mohan M. Trivedi

Abstract—From driver assistance in manual mode to takeover requests in highly automated mode, knowing the state of driver (e.g. sleeping, distracted, attentive) is critical for safe, comfortable and stress-free driving. Since driving is a visually demanding task, driver’s gaze is especially important in estimating the state of driver; it has the potential to derive what the driver has attended to or is attending to and predict future actions. We developed a machine vision based framework to model driver’s behavior by representing the gaze dynamics over a time period using gaze fixations and transition frequencies. As a use case, we explore the driver’s gaze patterns during maneuvers executed in freeway driving, namely, left lane change maneuver, right lane change maneuver and lane keep. It is shown that mapping gaze dynamics to gaze fixations and transition frequencies leads to recurring patterns based on driver activities. Furthermore, using data from on-road driving, we show that modeling these patterns show predictive powers in on-road driving maneuver detection around a few hundred milliseconds a priori.

I. INTRODUCTION

Intelligent vehicles of the future are that which, having a holistic perception (i.e. inside, outside and of the vehicle) and understanding of the driving environment, make it possible for occupants to go from point A to point B safely, comfortably and in a timely manner [1], [2]. This may happen with the human driver in full control and getting active assistance from the robot, or the robot is in partial or full control and human drivers are passive observers ready to take over as deemed necessary by the machine or humans [3], [4]. Therefore, the future of intelligent vehicles lies in the collaboration of two intelligent systems, one robot and another human.

Driver’s gaze is of particular interest because if and how the driver is monitoring the driving environment is vital for driver assistance in manual model and for take-over requests in highly automated mode for safe, comfortable and stress-free driving [5], [6]. Literary works have addressed the problem of estimating driver’s awareness of the surround in a few different ways. Ahlstrom et al. [7], with an underlying assumption that the driver’s attention is directed to the same object as the gaze, developed a rule based 2-second ‘attention buffer’ which depleted when driver looked away from the field relevant to driving (FRD); and it starts filing up when the gaze direction is redirected toward FRD. One of the reasons for a 2-second buffer is because eyes off the road for more than 2 seconds significantly increases the risk of a collision by at least two times that of normal, baseline driving [8]. Tawari et al. [9], on the other hand, developed a framework for estimating the driver’s focus of attention by simultaneously observing the driver and the driver’s field of view. Specifically, the work proposed to associate coarse eye position with saliency of the scene to understand on what object the driver is focused at any given moment. Li et al. [10], under the assumption that mirror-checking behaviors are strongly correlated with driver’s situation awareness, showed that the frequency and duration of mirror-checking reduced during secondary task performance versus normal, baseline driving. Furthermore, mirror-checking actions were used as features (e.g. binary labels indicating presence of mirror checking, frequency and duration of mirror checking) in driving classification problems (i.e. normal versus tasks/maneuver recognition). However, the classification problem had as it’s input, features from CAN signal and external cameras, whereas the classification and recognition problem addressed in this work is purely based on looking at the driver’s gaze. Work by Birrell and Fowkes [11] is most similar to our work in terms of using glance duration and transition frequencies; however it differs in its definition of representation, in its study on the effects of using in-vehicle smart driving aid and in its lack of predicting modeling.

In this work, we develop a vision based system to model driver behavior and predict maneuvers from gaze dynamics alone. One of the key aspects in modeling driver behavior is in the representation of driver gaze dynamics using gaze fixations and transition frequencies; this can be likened to smoothing out high frequency noise and retaining the fundamental information. For example, when interacting with the center stack (e.g. radio, AC, navigation), one may perform mirror-checking actions were reduced during secondary task performance versus normal, baseline driving. Furthermore, mirror-checking actions were used as features (e.g. binary labels indicating presence of mirror checking, frequency and duration of mirror checking) in driving classification problems (i.e. normal versus tasks/maneuver recognition). However, the classification problem had as it’s input, features from CAN signal and external cameras, whereas the classification and recognition problem addressed in this work is purely based on looking at the driver’s gaze. Work by Birrell and Fowkes [11] is most similar to our work in terms of using glance duration and transition frequencies; however it differs in its definition of representation, in its study on the effects of using in-vehicle smart driving aid and in its lack of predicting modeling.

In this work, we develop a vision based system to model driver behavior and predict maneuvers from gaze dynamics alone. One of the key aspects in modeling driver behavior is in the representation of driver gaze dynamics using gaze fixations and transition frequencies; this can be likened to smoothing out high frequency noise and retaining the fundamental information. For example, when interacting with the center stack (e.g. radio, AC, navigation), one may perform mirror-checking actions were reduced during secondary task performance versus normal, baseline driving. Furthermore, mirror-checking actions were used as features (e.g. binary labels indicating presence of mirror checking, frequency and duration of mirror checking) in driving classification problems (i.e. normal versus tasks/maneuver recognition). However, the classification problem had as it’s input, features from CAN signal and external cameras, whereas the classification and recognition problem addressed in this work is purely based on looking at the driver’s gaze. Work by Birrell and Fowkes [11] is most similar to our work in terms of using glance duration and transition frequencies; however it differs in its definition of representation, in its study on the effects of using in-vehicle smart driving aid and in its lack of predicting modeling.

In this work, we develop a vision based system to model driver behavior and predict maneuvers from gaze dynamics alone. One of the key aspects in modeling driver behavior is in the representation of driver gaze dynamics using gaze fixations and transition frequencies; this can be likened to smoothing out high frequency noise and retaining the fundamental information. For example, when interacting with the center stack (e.g. radio, AC, navigation), one may perform mirror-checking actions were reduced during secondary task performance versus normal, baseline driving. Furthermore, mirror-checking actions were used as features (e.g. binary labels indicating presence of mirror checking, frequency and duration of mirror checking) in driving classification problems (i.e. normal versus tasks/maneuver recognition). However, the classification problem had as it’s input, features from CAN signal and external cameras, whereas the classification and recognition problem addressed in this work is purely based on looking at the driver’s gaze. Work by Birrell and Fowkes [11] is most similar to our work in terms of using glance duration and transition frequencies; however it differs in its definition of representation, in its study on the effects of using in-vehicle smart driving aid and in its lack of predicting modeling.
TABLE I
DESCRIPTION OF ANALYZED ON-ROAD DRIVING DATA.

<table>
<thead>
<tr>
<th>Trip No.</th>
<th>Duration Full drive [min]</th>
<th>No. of Events Left Lane Change</th>
<th>Right Lane Change</th>
<th>Lane Keeping</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81.2</td>
<td>11</td>
<td>16</td>
<td>104</td>
</tr>
<tr>
<td>2</td>
<td>88.5</td>
<td>14</td>
<td>15</td>
<td>92</td>
</tr>
<tr>
<td>3</td>
<td>89.6</td>
<td>16</td>
<td>17</td>
<td>112</td>
</tr>
<tr>
<td>4</td>
<td>93.9</td>
<td>14</td>
<td>15</td>
<td>179</td>
</tr>
<tr>
<td>All</td>
<td>353.3</td>
<td>55</td>
<td>49</td>
<td>487</td>
</tr>
</tbody>
</table>

The contributions of this work are two fold. First is the automatic gaze dynamics analyzer which takes a video sequence as input and outputs context based gaze fixations and transition frequencies (as illustrated in Fig. 1). Second is in the modeling of gaze dynamics for activity prediction. Lastly, using data from on-road driving, we show quantitatively the ability to predict maneuvers reliably around a few hundred milliseconds in advance.

II. NATURALISTIC DRIVING DATASET

A large corpus of naturalistic driving dataset is collected using subjects’ personal vehicles over the span of six months. Individual vehicles are instrumented with four Hero4 GoPro cameras: two for looking at the driver’s face, one for looking at the driver’s hands and one for looking at the forward driving direction. In this study, the focus is in analyzing the driver’s face, while the forward view provides context for data mining; the hand looking camera is instrumented for future studies of holistic modeling of driver behavior.

All cameras are configured to capture data with 1080p resolution at 30 frames per second (fps). With exact camera configuration and similar camera placements, the subjects captured data using their instrumented personal vehicles during their long commutes. The drives consisted of some driving in urban settings, but mostly in freeway settings with lanes ranging from a minimum of two up to six lanes.

The data is especially collected in their personal vehicles in driving settings, but mostly in freeway settings with lanes ranging from a minimum of two up to six lanes.

This study analyzes data from a subset of the large corpus, to be exact four drives, whose details are as given in Table I. From the collected data, with a special focus on freeway settings, events were selected when the driver executes a lane change maneuver, by either changing left or right, and when the driver keeps the lane. Table I shows the events considered and their respective counts.

III. METHODOLOGY

This section describes a vision based system to model driver behavior from gaze dynamics and predict maneuvers. There are three main components: gaze zone estimation, gaze dynamics representation and gaze-based behavior modeling.

A. Gaze-zone Estimation

Two popular approaches to gaze zone estimation are an end-to-end CNN [12] and building blocks leading up to higher semantic information [13], [14]. The latter approach is employed here because when designing each of the semantic modules leading up to the gaze zone estimator, these intermediate representations of gaze can be used for other studies, such as tactical maneuver prediction [15] and pedestrian intent detection [16]. Key modules in our gaze zone estimation system include, face detection using deep convolutional neural networks [17], landmark estimation from cascaded regression models [18], [19], head pose from relative configuration of 2-D points in the image plane to 3-D points in the head model [20], horizontal gaze surrogate based on geometrical formulation of the eye ball and iris position [13], vertical gaze surrogate based on openness of the upper eye lids [21] and appearance descriptor, and finally, a 9-class gaze zone estimation from naturalistic driving data driven random forest algorithm. While the focus of this work is in

Let the vector $G = [g_1, g_2, \ldots, g_N]$ represent the estimated gaze for an arbitrary time period of $T$, where $N = fps(\text{frames per second}) \times T$, $g_n \in Z$, $n \in \{1, 2, \ldots, N\}$ and $Z$ represent the set of all gaze zones of interest as $Z = \{\text{LeftShoulder, Left, Front, Speedometer, Rearview, Front Right, Center Stack, Right, EyesClosed}\}$. Figure 2 illustrates sample output time segments of the gaze zone estimator. It illustrates multiple 10-second time segments prior to the start of lane change, two from left lane change and two from right lane change. In the figure, the $x$-axis represents time and color displayed at a given time $t$ represents the estimated gaze zone; let $\text{SyncF}$ denote the time when the tire touches the lane marking before crossing into the next lane, which is the ”0-seconds” displayed in the figure. Note how, prior to the lane change, there is some consistency observed across the different time segments within a given event (e.g. left lane change); consistencies such as the total gaze fixations and gaze transitions between gaze zones. In the next section, we define a temporal sequence of gaze dynamics using gaze fixations and transition frequencies, which remove some temporal dependencies but still capture sufficient spatio-temporal information to distinguish between different gaze behaviors.

B. Gaze Dynamics Representation

Gaze fixation is a function of the gaze zone. Given a gaze zone, gaze fixation for that gaze zone is the amount of time driver spends looking at the gaze zone within a time period; which is then normalized by the time window for relative duration calculation. Glance duration then is calculated for each of the gaze zones, $z_j$, where $z_j \in Z$ and $j \in \{1, 2, ..., M\}$, as follows:

$$\text{Glance Fixation}(z_j) = \frac{1}{N} \times \sum_{n=1}^{N} 1(g_i == z_j)$$

where $1(\cdot)$ is an indicator function.
Lane Change of interest: (MVN). A unnormalized MVN is trained for each behaviors tasks or maneuvers, using a multivariate normal distribution. Then, we model the gaze behaviors of respective events, compute the mean of the feature vectors within each class. Left Lane Change For instance, the class labels can be: Y and their corresponding class labels k represent the d f column. In order to remove the order of transition, FGT is first decomposed into upper and lower triangular matrices. The lower triangular matrix is transposed and summed together with the upper triangular matrix, and then normalized to produce the new glance transition matrix:

\[ F'_{GT}(d, k) = \frac{1}{N \times \text{fps}} \times \begin{cases} 0 & d \leq k \\ f_{dk} + f_{kd} & d > k \end{cases} \]

where \( f_{dk} \) is the number of transitions from the gaze zone representing the \( d^{th} \) column to the gaze zone representing the \( k^{th} \) column.

The final feature vector, \( \tilde{h} \), is composed of gaze fixations computed for every gaze zone of interest and upper triangular matrix of the new glance transition frequency matrix, \( F'_{GT} \) in vectorized the form over a time window of gaze dynamics.

C. Gaze-based Behavior Modeling

Consider a set of feature vectors \( H = \{\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_N\} \), and their corresponding class labels \( Y = \{y_1, y_2, \ldots, y_N\} \). For instance, the class labels can be: Left Lane Change, Right Lane Change, Merge, Secondary Task. Given \( H \) and \( Y \), we compute the mean of the feature vectors within each class. Then, we model the gaze behaviors of respective events, tasks or maneuvers, using a multivariate normal distribution (MVN). A unnormalized MVN is trained for each behaviors of interest:

\[ M_b(\tilde{h}) = \exp \left( -\frac{1}{2} (\tilde{h} - \tilde{\mu}_b)^T \Sigma_b^{-1} (\tilde{h} - \tilde{\mu}_b) \right) \]

where \( b \in B = \{\text{Left Lane Change, Right Lane Change, LaneKeep}\} \), and \( \tilde{\mu}_b \) and \( \Sigma_b \) represent the mean and covariance computed from the training features vectors for the gaze behavior represented by \( b \). One of the reasons for modeling gaze behavior is, given a new test scanpath \( \hat{h}_{test} \), we want to know how does it compare to the average scanpath computed for each gaze behavior in the training corpus. One possibility is by taking the euclidean distance between the average scanpath, \( \mu_b \), and the test scanpath, \( \hat{h}_{test} \), for all \( b \in B \) and assign the label with the shortest distance. However, this assigns equal weight or penalty to each component of \( \hat{h} \). We want to the weights to be a function of component as well the behavior under consideration. Therefore, we use the Mahalanobis distance, which is the component in the exponent of the unnormalized MVN. By exponentiating the Mahalanobis distance, the range is mapped between 0 and 1. To a degree this can be used to assess the probability or confidence that a certain test scanpath, \( \hat{h}_{test} \) belongs to a particular gaze behavior model.

IV. On-road Performance Evaluation

Every instance of driving on a freeway can be broken or categorized exclusively into one of these three categories, left lane change, right lane change and lane keep. As a point of synchronization, for lane change events, when the vehicle is half in the source and half in the destination it is marked at annotation. A time window of t-seconds before this instance defines the left and right lane change events, where \( t \) is varied from 10 to 0. For the lane keeping events, a lengthy stretch of lane keeping is broken into non-overlapping respective t-sec time windows to create lane keeping events. Table I contains the number of such events annotated and considered for the following analysis.

All evaluations conducted in this study is done with a four-fold cross validation; four because there are four different drives as outlined in Table I. Using the popular leave one out cross-validation, training is done with events from leaving out one drive and tested with events from the remaining drive. With this setup of separating the training and testing samples, we explore the precision and recall of the gaze behavior model in predicting lane changes as a function of time.
Training occurs on the 5-second time window before SyncF as defined in Section III-A. While testing, however, we want to test how early the gaze behavior models are able to predict lane change. Therefore, starting from 5-seconds before SyncF, sequential samples with \( \frac{1}{50} \) of a second overlap are extracted up to 5-seconds after SyncF; note that the time window at 5-seconds before the SyncF encompasses data from 10 seconds before the SyncF up to 5-seconds before the SyncF. Each of the samples are tested for fitness across the three gaze behavior models, namely models for left lane change, right lane change and lane keep. The sample is assigned the label based on the model which procures the highest fitness score and if the label matches the true label the sample is considered a true positive. Note that each test sample is associated with a time index of where it is sampled from drives not included in the training set, precision and recall values are calculated as a function of the time index.

When calculating precision and recall values, true labels of samples were remapped from three classes to two classes; for instance, when computing precision and recall values for left lane change prediction, all right lane change events and lane keep events were considered negatives samples and only the left lane change events are considered positive samples. Similar procedure is observed for computing precision and recall values for right lane change prediction. Table II shows the development of the precision-recall values for both left and right lane change prediction in an interval of 200 milliseconds starting from 2000 milliseconds prior to SyncF up to 0 milliseconds prior to SyncF. Interestingly, there is a plateauing effect in recall for both left and right lane change prediction around a 600 milliseconds before the event. One possibility is that there is a strong indication at that time index of intended lane change using gaze analysis only. Future works in experimenting with the time window for modeling and testing will reveal more information.

With respect to the precision rate, the values are not expected to be as high because even during lane keeping, drivers may exhibit lane change like behavior without lane changing. This is especially observed with the precision rate for left lane change prediction, where checking left rear view mirror is a strong indicator of lane change but also part of driver’s normal mirror checking behavior during lane keep. One of the main cause is the gaze model for lane keep, which encompasses a broad spectrum of driving behavior and future work will consider finer and more separated labeling of classes.

One of the motivations behind this work is to estimate the driver’s readiness to take over in highly automated vehicles. Situations where system cannot handle thus requiring takeover include system boundaries due to sensor limitation and ambiguous environment. In such situations, looking inside at the state of the driver and how much time is required to reach readiness to take-over is important. Therefore, in this study we developed a framework to estimate his or her readiness to handle the situation by modeling gaze behavior from non-automated naturalistic driving data. In particular, gaze behavior during lane changes and lane keep is modeled. Figure 3 illustrates the fitness or confidence of the model around left and right lane change maneuvers. The figure shows mean (solid line) and standard deviation (semitransparent shades) of three models (i.e. left lane change, right lane change, lane keep) for two different maneuvers (i.e. left lane change and right lane change), using the events from naturalistic driving dataset described in Table I. The model confidence statistics are plotted 5 seconds before and after the lane change maneuvers, where time of 0 seconds represents when the vehicle is half way between the lanes. Interestingly, during left (right) lane change maneuver fitness of the right (left) lane change gaze model is very low within the 10 second bracket and left (right) lane change model peaks in fitness in a tighter time window around the maneuver. Furthermore, for both maneuvers, the lane keep model as desired is high in the periphery of the lane change maneuvers hinting at how this model performs during actual lane keep situations.

### V. Concluding Remarks

In this study, we explored modeling driver’s gaze behavior in order to predict maneuvers performed by drivers, namely left lane change, right lane change and lane keep. The particular model developed in this study features three major aspects: one is the spatio-temporal features to represent the gaze dynamics, second is in defining the model as the average of the observed instances and interpreting why such a model fits the data of interest, third is in the design of the metric

<table>
<thead>
<tr>
<th>Time before maneuver</th>
<th>Left Lane Change Prediction</th>
<th>Right Lane Change Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frame Milliseconds</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>60</td>
<td>0.4130</td>
<td>0.4318</td>
</tr>
<tr>
<td>54</td>
<td>0.4490</td>
<td>0.5000</td>
</tr>
<tr>
<td>48</td>
<td>0.4912</td>
<td>0.6364</td>
</tr>
<tr>
<td>42</td>
<td>0.5484</td>
<td>0.7727</td>
</tr>
<tr>
<td>36</td>
<td>0.5397</td>
<td>0.7727</td>
</tr>
<tr>
<td>30</td>
<td>0.5692</td>
<td>0.8409</td>
</tr>
<tr>
<td>24</td>
<td>0.5758</td>
<td>0.8636</td>
</tr>
<tr>
<td>18</td>
<td>0.5672</td>
<td>0.8636</td>
</tr>
<tr>
<td>12</td>
<td>0.5672</td>
<td>0.8636</td>
</tr>
<tr>
<td>6</td>
<td>0.5735</td>
<td>0.8864</td>
</tr>
<tr>
<td>0</td>
<td>0.5857</td>
<td>0.9318</td>
</tr>
</tbody>
</table>
for estimating fitness of model. Applying this framework in a sequential series of time windows around lane change maneuvers, the gaze models are able to predict left and right lane change maneuver with an accuracy above 85% around 600 milliseconds before the maneuver.

The overall framework, however, is designed to model driver’s gaze behavior for any tasks or maneuvers performed by driver. To this end, there are multiple future directions in site. One is to quantitatively define the relationship between the time window from which to extract those meaningful spatio-temporal features and the task or maneuvers performed by driver. Another is in exploring and comparing different modeling approaches, including HMM, LDCRF and multiple kernel learning. Other future directions include exploring unsupervised clustering of gaze behaviors, especially during lane keep, and exploring the effects of quantity and quality of events (e.g. same vs. different drives, different drives from same or different time of day) on gaze behavior modeling.

ACKNOWLEDGMENT

The authors would like to thank their colleagues at the Laboratory of Intelligent and Safe Automobiles (LISA) for their valuable comments, especially Sourabh Vora for his assistance with data analysis. The authors gratefully acknowledge the support of industry partners, especially Fujitsu Ten and Fujitsu Laboratories of America.

REFERENCES


Fig. 3. Illustrates the fitness of the three models (i.e. Left lane change, Right lane) during left and right lane change maneuvers. Mean (solid line) and standard deviation (semi-transparent shades) of the three models as applied for to the lane change events described in Table I.