Distinguishing patterns in drivers’ visual attention allocation using Hidden Markov Models

Mauricio Muñoz a,b,c, Bryan Reimer a,*, Joonbum Lee a, Bruce Mehler a, Lex Fridman a

a Massachusetts Institute of Technology AgeLab & New England University Transportation Center, Cambridge, MA, USA
b TU Munich, LMU Munich, Munich, Germany
c University of Augsburg, Augsburg, Germany

ABSTRACT

Driving is an intricate task where different demands compete for the driver’s attention. Current interface designs present novel multi-modal interactions that extend beyond traditional visual-manual modalities. These new interaction paradigms have given rise to additional subtask elements which call upon varying degrees of cognitive, auditory, vocal, visual, and manual resources. The draw on a larger number of resources has made demand assessment and optimization challenging. How these elements impact the driver’s visual behavior may provide insight into the degree to which a vehicle’s user interface influences attentional focus. This report addresses this question by approaching the problem from a computationally predictive perspective. Data were drawn from two studies that captured visual behaviors of drivers during a series of radio tuning tasks using a traditional manual interface and a multi-modal voice enabled interface during highway driving. Manual annotations of glance times and targets were compiled for each task period and then used to train a predictive model. A statistical machine learning approach (Hidden Markov Model) showed that manual radio tuning, voice-based radio tuning, and “just driving” behaviors result in fundamentally and predictably different strategies of visual attention allocation. We report classification accuracies of over 95% for detecting the correct task modality within a 3 class classification framework, extending prior work to show that time series of glance allocations contain highly descriptive information that generalizes well across drivers of different ages, genders, and driving experience. Results suggest that differences in glance allocation strategies serve as an effective evaluator of the visual demand of a vehicle interface, providing an objective methodology for demonstrating that voice-based technologies allow drivers to maintain a broader distribution of visual attention than the traditional manual interface.

1. Introduction

Driving has long been recognized as a set of dynamic and complex interactions with the environment, composed of information processing, action planning, and low level execution stages (Michon, 1985). New ways of interacting with the vehicle challenge our traditional understanding of interface demand optimization that was based primarily on visual-manual interaction. Results from a naturalistic driving study at the Virginia Tech Transportation Institute (VTTI) conducted for the
National Highway Traffic Safety Administration (NHTSA) suggest that performing visual-manual tasks while driving increases the probability of a safety-critical event by a factor of three (Fitch et al., 2013). The logical observation that drivers cannot react on their own to events in the roadway if they are not looking at the road in the first place has been a major argument for the development of voice-based interface options in the vehicle. The availability of voice interfaces is increasing rapidly - in 2012 37% of vehicles contained an embedded implementation of a voice recognition system, with projections indicating that this figure may rise to 55% by 2019 (IHS Technology, 2013). However, the extent to which this interaction paradigm is inherently more advantageous has been called into question. Work by Lee, Caven, Haake, and Brown (2001) and Strayer et al. (2013) reported increased cognitive load and longer driver reaction times for experimental implementations of in-vehicle voice-based systems. Fitch et al. (2013), Reimer, Mehler, Dobres, and Coughlin (2013), Mehler et al. (2014), Reimer, Mehler, Reagan, Kidd, and Dobres (2016), and Mehler, Kidd et al. (2016), all show that common voice-based interfaces contain visual-manual subtask components. It is thus unclear how the visual-manual subtasks of a voice-based interface relate to traditional visual-manual interfaces that have been linked to increased likelihood of a crash. He, Zong, and Wang (2012) provide evidence to support the notion that there are intrinsic behavioral differences (e.g. driver distraction profiles) between visual-manual and voice-based interactions modalities, even after accounting for parameters such as task duration. Furthermore, ongoing technological trends that seek to supersede traditional visual-manual interactions with interfaces that also draw upon auditory-vocal resources fail to provide an interaction alternative that compares well with naturalistic driving in terms of driver performance (Just, Keller, & Cynkar, 2008; Sawyer, Finomore, Calvo, & Hancock, 2014). In short, these insights seem to suggest that the observable behavioral differences caused by these types of interactions with the vehicle are not clear-cut and as of yet not comprehensively understood or defined.

Traditionally, human-machine interface (HMI) demand has been measured in a limited number of ways. These strategies include measuring the total time visual attention is directed away from the forward roadway (National Highway Traffic Safety Administration, 2013) or to elements specific to the operation of the HMI (Driver Focus-Telematics Working Group, 2006). These distraction indicators (total time off road, total glance time to device) have been used extensively in the literature (Chiang, Brooks, & Weir, 2004; Donmez, Boyle, & Lee, 2009; Tijerina, 1999; Victor, Bärgman, Dozza, & Rootzén, 2013). While these measures are intuitive and have some demonstrated utility, a shortage common to both approaches is that they aggregate behavioral information over a given time span to a single number, disregarding useful information sources such as transitions of glance allocations. They do not characterize the impact a given interaction has on how a driver distributes their attention across a broader driving environment over time, i.e. taking into account glances that may support the driver’s wider situational awareness. Specifically, this includes assimilating information from such locations as the side mirrors, blind spots, rearview mirror, and the instrument cluster. These differences in glance distributions can be qualitatively observed in the glance transition matrices presented in Muñoz, Reimer, and Mehler (2015) in a more information rich form than is readily accessible in classic summary statistics such as total eyes off-road time for the source datasets (Mehler et al., 2014; Reimer et al., 2013; Reimer et al., 2014).

Given a more mature driver behavior model, the quality of a vehicle’s interface might be measured by how well the driver’s behavior pattern remains consistent with a baseline attentive driving style, unencumbered by non-driving related activities. In essence, an optimal interface design may be one that minimizes the divergence of attention from how the driver would be expected to perform if no secondary task was being engaged in. This report extends prior work by providing such a capable driver behavior model and presenting an approach through which the differences between any given set of driver behaviors may be quantified (i.e. calculating the probability that an observed behavior is consistent with baseline driving or any other model of visual demand). More practically, this allows for a quantitative comparison of vehicle interfaces on the degree to which they impact a driver’s overall allocation of attention, e.g. how compatible a particular HMI is with attentive driving from the perspective of visual demand.

2. Background

This report builds on extensive research in the field of driver modeling using Hidden Markov Models, as well as on previous work with glance-based measures of driver attentiveness. The primary contribution presented here consists in applying the former modeling technique to the latter quantification of driver behavior. In order to place this contribution in a historical context, a brief overview is given on Hidden Markov Modeling in the driving context as well as on glance-based behavioral features.

Several studies including Victor, Bärgman, Dozza, and Rootzén (2013), Tijerina (1999), Chiang et al. (2004), Donmez et al. (2009), Liang, Lee, and Yekhshatyan (2012), and Tivesten and Dozza (2014) have focused on glance allocations as behavioral features, aggregating them over time to analyze glance location distributions and durations in the context of risk assessment and situational awareness. Only a small subset (Liang et al., 2012; Victor et al., 2013) have studied them in a time-sensitive context. Concretely, Victor et al. (2013) use glance behavior histories to explain risk from secondary tasks and perform analytic work on the relationship between glance duration/location and vehicle crash risk. Corresponding metrics for risky and less risky glances are introduced, particularly in relation to the timing of lead vehicle closing kinematics. Liang et al. (2012) predict crash/near-crash risk using different combinations of glance duration, history, and location features as inputs to hand-crafted function-like feature detectors which output driver distraction estimates. Linear regression models were used.
to arrive at model estimates, which allowed for an examination of the effects of each input on the final prediction result. The report found that glance history provided little benefit to the overall result (under the linear model).

Hidden Markov Models are statistical graphical models with a long history of time series data analysis. They work by quantizing the complexity within a system into a set of discrete states and a transition relationship between them (Oliver & Pentland, 2000). A thorough overview of Hidden Markov Models is given by Rabiner (1989). In the field of behavior inference alone there have been multiple studies reporting their suitability for the problem (e.g. Ji, Wang, Li, & Wu, 2013; Mendoza & De La Blanca, 2007; Yamato, Ohya, & Ishii, 1992) and notably also within the context of driving by using vehicle performance and position data (He et al., 2012; Krumm, 2008; Kuge, Yamamura, Shimoyama, & Liu, 2000; Kumagai, Sakaguchi, Okuwa, & Akamatsu, 2003; Liu & Pentland, 1997; Mitrovic, 2001; Oliver & Pentland, 2000; Pentland & Liu, 1999). These works fall within the more general area of research that utilizes driver behavior models and machine learning approaches to infer specific driving events and maneuvers such as lane changing, passing, and stopping (Chandrasiri et al., 2012; Chong, Abbas, Flintsch, & Higgs, 2013; Khasionskram et al., 2008; Maye, Triebel, Spinello, & Siegwart, 2011; Qian, Ou, Wu, Meng, & Xu, 2010; Solovey, Zec, Garcia Perez, Reimer, & Mehler, 2014; Tango, Botta, Minin, & Montanari, 2010; Tchankue, Wesson, & Vogts, 2013).

In one of the earlier studies to apply Hidden Markov Models to the driving context, Pentland and Liu (1999) break down a driver’s behavior into different prototypical behaviors, each corresponding to an individual dynamic model. An unknown behavior is then classified by choosing the model yielding the highest posterior probability. In other work, Kuge et al. (2000) build models using raw steering data (steering angle, velocity, and force) fed into a 2 level hierarchical HMM structure in order to infer the driver’s intended maneuver (lane keeping, emergency lane change, and regular lane change). The bottom level of the hierarchical structure is composed of maneuver-specific (in terms of the number of states and transitions) left-to-right HMMs, which were then linked together using a higher-level grammar (in terms of the number of sub-HMMs). A similar 2 level model is found in He et al. (2012), where the lower level encodes short term and immediate driving techniques, which are then organized in the higher level into long term driving strategies. Results were classified both offline and in real-time by feeding steering data into the sub-HMMs according to the grammar, and picking the composite model that maximizes the probability of the input. The high recognition rates of this approach suggest that HMMs can infer complex behaviors using unfiltered raw data directly, though other studies exist (Pentland & Liu, 1999) that show that an abstraction layer over raw data may be necessary for a high recognition rate.

Mitrovic (2001) presents another approach for driving event (turning, stopping, changing speed, etc.) detection. Filtered values from on-road driving periods, including average speed, average lateral and longitudinal acceleration, as well as acceleration slope, were transformed into discrete values using a learned codebook. An HMM was built for each driving event (for a total of seven). High classification accuracy was achieved on telemetry inputs alone by picking the HMM that best explained each test sequence. This lies in contrast to the claims laid out by Kuge et al. (2000) and Oliver and Pentland (2000), which state that the context of the input sequences is vital to the classification efforts.

Finally, Kumagai et al. (2003) suggest that HMMs may be considered as a tool for predicting driving events, much like Oliver and Pentland (2000), but with several important differences. Vehicle speed, acceleration, and brake pedal strokes were measured in an on-road experiment to train an HMM for anomaly detection. The goal is to predict the stop probability at an intersection on the order of 2–4 s instead of <1 s as in Oliver and Pentland (2000), using a hand-crafted probability function above the HMM layer. Though failing to specify traditional performance measures for anomaly detection such as F1 scores or receiver operating characteristic (ROC) curves, the study provides an insightful correspondence between the actual and the prediction stop probability rates, even if only for a single participant.

Overall, as noted by He et al. (2012), current research in driver behavior inference very much revolves around the problem of detecting driving patterns and higher-order events given observed low-level telemetry data. The present work extends efforts in this general area by considering the allocation of visual attention, in the form of glance allocation sequences, as an input to a predictive framework. An HMM classifier is used to infer the driver’s behavior in terms of their interaction state and modality of task engagement with the vehicle. From the literature that has been explored, this is a novel application of HMMs as predictive models in the driving context. As a case study designed to assess these concepts, this report examines data drawn across three key task types (a period of “just driving”, visual–manual radio tuning, and voice–based radio tuning) to assess differences in predictive performance. The work draws upon series of glance allocations during these periods of interest encoded as transitions of glance allocations and durations of allocations. Predictions based upon Hidden Markov Modeling are provided to assess three key questions: (1) Do key HMI design variables (such as task type) influence glance allocation (transition probabilities/glance location sequencing) and thus the driver’s overall visual behavior pattern? (2) How well can different behavior patterns not only be quantified but consistently predicted given glance allocations as input features (that is, how well can visual glance patterns discriminate between behaviors and generalize across broad demographic samples)? (3) How valuable are glance durations in addition to glance location features?

3. Classifying task modality and visual demand

There are three key elements in any machine learning framework that heavily influence its performance: the quantity of the data (Banko & Brill, 2001), the descriptive potential of the chosen features, and the model chosen to extract this discriminative potential. This report explores how well the combination of glance allocation features and Hidden Markov Models...
works in order to infer driver visual behavior, drawing upon data from two studies with over 156 participants and 21,820 glance transitions harvested from several distinct driving conditions. A three-class classification scheme is employed to distinguish behaviors that are characteristic of “just driving” (baseline), visual-manual interactions with the vehicle’s interface (visual-manual radio tuning tasks), and interactions with the vehicle’s HMI using a voice interface (auditory-vocal radio tuning tasks).

3.1. Methods

Data were drawn from two on-road studies conducted by the MIT AgeLab to study the effects of production voice interfaces in automobiles. The studies were conducted in a 2010 Lincoln MKS with factory installed voice-command systems (Ford SYNC™ for voice control of the phone and media connected by USB and the “next-generation navigation system” with Sirius Travel Link). In each study, participants were asked to complete a series of visual-manual and voice-based tasks while driving on the highway (see Mehler et al., 2014; Reimer et al., 2013; Reimer, Mehler et al., 2014 for full details on experimental procedures and tasks). In brief, the first study examined a sample of younger (20–29) and older (60–69) drivers who were given structured training in how to complete each task in the most efficient manner supported by each interface; the default settings for the voice interface were used and these provided extensive audio prompting and required verbal confirmation of commands. The second study set examined a sample of drivers equally distributed across four age groupings: 18–24, 25–39, 40–54, and 55+, conforming to NHTSA’s recent recommendations for test samples (National Highway Traffic Safety Administration, 2013). This sample was composed of three subgroups: structured task training with default voice system settings, self-guided task learning with default voice system settings, and structured task training with “expert” mode voice settings. In the structured training, participants were taken step-by-step through the most efficient method of completing each task in a practice session in a parking lot prior to the actual driving based assessment. For the self-guided group, during the parking lot training period, participants were provided with examples of the tasks they would be asked to complete during the drive and given the opportunity to explore on their own the respective voice and visual-manual interfaces for completing each task type. The “expert” voice mode removed some of the verbose prompting provided in the default mode and removed verbal confirmation of most commands, shortening the overall duration of interaction. Table 1 summarizes the combinations of tasks and participant types for each study.

Each study included two trials of an “easy” visual-manual preset selection task (pressing a single pre-set station button located on the center stack), two trials of a “hard” visual-manual radio tuning task (multiple button presses and multiple rotations of a tuning knob to select a specified frequency using controls located on the center console), and engagement with the same task end-goals using the voice-based interface (each voice-task required pressing a speech interface activation button on the steering wheel, speaking one or more commands, and, except for the Study 2 expert mode subgroup, providing verbal confirmations). Note, when interacting with the radio through the voice interface, changes would occur on the video display screen in the center console coinciding with selection of the radio and changing stations. Functionally, drivers did not have to look at this screen, although they were free to do so. Thus, as is the case with most, if not all, current production voice-command interfaces, “voice” HMIs are perhaps most appropriately classified as voice-based multimodal interfaces that include some visual-manual components as opposed to more traditional interface designs that are completely visual-manual in nature. The terms “voice” and “manual” are used here for ease of distinction.

Glance data were manually reduced based upon a frame-by-frame review of video from an in-vehicle camera directed at the driver’s face according to the taxonomy and procedures outlined in Reimer et al. (2013) (Appendix F). The coding process involved the use of custom software (Reimer et al., 2013) that was developed (and subsequently open sourced, Reimer, Gruevski, & Coughlin, 2014) to classify glances into different bins. For Study 1 this included ‘left mirror’, ‘instrument cluster’, ‘forward road’, ‘rearview mirror’, ‘center stack’, ‘right mirror’, and ‘other’; apparent blind spot checking and other head rotations in excess of approximately 90° were included in “other”. For Study 2, the bins consisted of ‘left blind spot’, ‘left mirror’, ‘instrument cluster’, ‘forward road’, ‘rearview mirror’, ‘center stack’, ‘passenger seat’, ‘right mirror’, ‘right blind spot’, ‘other’, and ‘unknown’ (for video segments deemed uncodeable due to glare, video quality, etc.). The ‘other’ and ‘unknown’ categories were infrequent and not included in the modeling. For analyses that merged data from both studies (see Section 3.2), the ‘left blind spot’ and ‘right blind spot’ locations coded in Study 2 were not used in the modeling stage.

Each task period of interest was independently coded by two different members of the research staff. Any discrepancies between the two coders—the identification of conflicting glance targets, missed glances, or glance timings that differed by more than 200 ms—were mediated by a third staff member (see Smith, Chang, Glassco, Foley, and Cohen (2005) for

<table>
<thead>
<tr>
<th>Study</th>
<th>Baseline driving (V/M)</th>
<th>“Easy” tasks (V/M)</th>
<th>“Hard” tasks (V/M)</th>
<th>Participants (Total)</th>
<th>Trained/Self-guided participants</th>
<th>Default/“Expert” settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 (1/1)</td>
<td>4 (2/2)</td>
<td>4 (2/2)</td>
<td>60</td>
<td>60/0</td>
<td>60/0</td>
</tr>
<tr>
<td>2</td>
<td>2 (1/1)</td>
<td>4 (2/2)</td>
<td>4 (2/2)</td>
<td>96</td>
<td>64/32</td>
<td>64/32</td>
</tr>
</tbody>
</table>
arguments on the need for two or more independent raters). The duration of each glance was computed as a post-processing step by taking the absolute time difference between its timestamp and that of the next glance.

3.2. Procedures

The final dataset was then constructed by mapping each mediated glance to the index of its bin, resulting in a list of discrete glance allocation sequences for each participant and task. From these, sequences pertaining to all 4 visual-manual radio tuning tasks and all 4 auditory-vocal radio tuning tasks were extracted and aggregated into manual and voice classes accordingly. “Just driving” data for each participant were taken from 2-min baseline reference periods preceding the visual-manual and auditory-vocal radio tasks.

Training and test data were derived from this dataset for the following combinations of the data in Study 1 and Study 2 in order to test the effects of input data size on classification accuracy, as well as to test the model’s ability to generalize across studies:

1. Training and testing data were derived from Study 1.
2. Training and testing data were derived from Study 2.
3. Training data were derived from Study 1 and validated with testing data from Study 2.
4. Training data were derived from Study 2 and validated with testing data from Study 1.
5. Training data were derived from Studies 1 and 2 and validated with testing data from Studies 1 and 2.

Concretely, training data were collected in each respective case by randomly sampling 80% of the participants from the data pool and using all of their corresponding glance sequences. This was carried out in two different ways. In the first instance, glance sequences were created by matching each participant with all of their glances according to type of task (baseline, manual, or voice). For example, in the case of the four trials (2 easy and 2 hard) of visual-manual radio tuning, all glances across these trials were combined into a single glance allocation sequence per participant. In the second approach, each participant was matched with a particular, individual task trial. In this case, four distinct sequences resulted per participant. This results in an evident tradeoff: building longer, fewer sequences results in samples that are intuitively more expressive and thus potentially easier to classify. Building shorter, more numerous sequences results in more realistic samples. Results are presented for both variants, primarily to illustrate the theoretical feasibility of the problem in the former case, as well as the classifier’s performance in a more realistic scenario in the latter case.

In the same fashion, the remaining 20% of participants were used for validation. In order to improve the quality of the training and test data, a threshold of 11 glances was set as the minimum sequence size to exclude sequences with little or no information content. To counteract potential artifacts associated with randomly sampling of the participants used for training from the participant pool, the procedure was repeated according to a Monte Carlo-like sampling method, where each classification procedure was run for 100 iterations, after which the average over all accuracies was computed to produce an estimate of the true classification accuracy. Fig. 1 illustrates this procedure for all cases detailed in Section 3.2.

3.3. Modeling

The classification problem is modeled as a classic instance of HMM inference. This is partly due to the fact that unlike He et al. (2012) and Kuge et al. (2000), there is little that can be assumed about the temporal structure of visual behavior as a whole - the framework presented here does not confine driver attentional strategies to a finite set of possibilities. For

Fig. 1. Monte-Carlo sampling scheme, illustrating all training and test pairs drawn from studies 1 and 2. The numbering corresponds to the test case description in Section 3.2. Training and testing steps are explained in detail in Section 3.3. The 5 individual training and testing data set pairs are listed at the beginning of Section 3.2.
instance, whereas Kuge et al. (2000) model lane changes as a composition of necessarily sequential events (expressed with left-to-right HMMs, so called non-ergodic models in Rabiner, 1989), no such restriction is made here, allowing the model to naturally learn any such pattern during training. Another reason for the simple architectural choice is to provide a baseline model for future studies that might expand in this direction.

In the present study, one HMM was trained per class. Experimenting with different model topologies showed $n = 2$ to be the optimal number of states. One common problem often encountered during HMM training is estimating probabilities for uncommon emissions and transitions that might not be represented in the training set (i.e., just because the input data contains no glance allocations to the ‘right mirror’ does not mean these allocations will certainly never happen). Most implementations of the training algorithm provide a way to counteract this effect by specifying prior pseudo-emissions and pseudo-transition counts, i.e., letting the model know that even extremely unlikely observations may occur in the future. Experiments found performance was most stable by specifying minimal, non-zero counts for emissions that did not appear in the training data. A uniform distribution was chosen for the pseudo-transitions prior. This corresponds to a Laplace smoother, which has been shown to work well in Bayesian inference systems (He & Ding, 2007). This was done for the following reason: each HMM is modeled with a fixed number of states that do not carry an application-specific semantic value, i.e., the number of states is simply another parameter of the model which is tuned with a validation set. This corresponds to the traditional framework for training HMMs for multi-class classification (Kuge et al., 2000; Oliver & Pentland, 2000; Rabiner, 1989). The transition patterns between these artificial states are therefore unknown a priori, hence the additive, unbiased smoothing.

Given this set of trained models (one each for baseline driving, auditory-vocal radio tuning, and visual-manual radio tuning), a new sequence of glances is assigned to one of these classes by testing each HMM and picking out the model that produces the highest log probability for the input sequence. Intuitively, the model is chosen that best interprets (and is thus most likely to have generated) the new test sequence. Performance metrics are given in accuracy as the percentage of correctly classified sequences.

Classification performance is also reported for glance duration features. In the duration-independent case, only discrete glance locations are given to the classifier with no information regarding how long the respective participant spent in a particular glance. The second case feeds the duration of the glance to the classifier without information regarding the location where the glance was allocated. Whereas the model previously learned a stepwise probability density of each glance per state, in the duration-dependent case it learns a classical continuous density of glance duration per state as a Gaussian distribution as in Rabiner (1989). Determining the probability of an input sequence then involves finding the conjugate prior of the normal distribution of glance duration (this implies computing a normal-inverse Wishart distribution for each state). In theory, however, any distribution model is possible. For the present task, leveraging domain-level knowledge leads to an alternative model of the glance duration variable. Parallel to the Gaussian representation above, duration values were discretized according to a binning scheme. The basic idea consists in building a custom histogram function, thus mapping real duration values to discrete histogram bin indices, similar to the approach in Mitrovic (2001). Fig. 2 illustrates the distribution of glance duration across all task types in Study 1. Clearly, most glances are localized at or below the 4 s mark. Practically, this suggests good performance may be achieved by adding higher resolution in this area. The final implementation defined evenly spaced bins between 0 and 1 s, at 0.1 s intervals, and 1 s onwards at 1 s intervals.

It was expected that using both glance location and duration features jointly would result in an additional performance boost. Glance location and duration (using the custom histogram) features were merged by taking the model that maximized the sum of the probabilities using each feature individually. This corresponds to the simple approach in Alexandre, Campilho, and Kamel (2001) on combining classifier outputs.

![Fig. 2](image-url) Distributions of the glance duration variable for a set of training subjects in Study 1 for the baseline driving task (a), visual-manual tasks (b), and auditory-vocal tasks (c).
4. Results

Table 2 shows classification results (average accuracy) for Study 1, Study 2, and both studies combined for 100 iterations of the Monte Carlo estimation. Glance location, duration, and merged features were used. Two duration models are presented - the standard approach based on a Gaussian distribution of duration, as well as the histogram binning approach detailed above. Sequences were formed by taking glances according to the type of task (baseline, voice, and manual task types). Table 3 presents the corresponding classification results, but using individual task trials to build sequences instead.

As can be seen in the tables, using glance location features results in high accuracy values, especially when building sequences from task types (Table 2). Over 95% accuracy is achieved using data from Study 1. As expected, compromising on sequence length (Table 3) results in a slight drop in performance, implying that the HMM classifier benefits from lengthier, more expressive sequences. High accuracy values on the cross-study data sets are highly suggestive of general behavior patterns, independent of driver demographic or training. Duration features also show strong performance well above that of a random classifier. Leveraging domain-level knowledge using histogram-based duration features improves performance in every case, sometimes quite significantly. This strongly supports the idea that glance duration is also a key signal of the driver’s visual behavior and that task modality may also be characterized by how long the driver decides to allocate attention regardless of where. As expected, merging features leads to an additional performance boost which for the most part outperforms glance location features alone, particularly when aggregating glances by individual tasks (Table 3). This suggests that supplementing these features with each other leads to a stronger predictive signal which the HMM is able to pick up and respond to.

Figs. 3 and 4 visualize the accuracy distribution for each combination of train and test sets (see Fig. 1) for both feature types and each respective aggregation strategy, showing additional statistics such as standard deviation and 95% confidence intervals. High classification accuracies and narrow confidence intervals in Fig. 3 are strongly indicative of the fundamental differences in the visual behaviors of each task modality. Aggregating glances by single task instances (Fig. 4), in comparison to aggregating them by task types (Fig. 3), results in accuracy distributions that are somewhat more condensed, although the general pattern across data sets remains fairly constant. For both glance duration and location features, average accuracy drops by approximately 2–5%, which corresponds to about 3–4 sequences for the single-study data sets, when glances are aggregated from single task instances.

In order to determine the origin of the misclassifications above, confusion matrices were computed on the classifier’s label assignments during testing. Fig. 5, which illustrates these matrices for the multi-study case (with data from Study 1 & Study 2), reveals that baseline driving was, as expected, the easiest of the behaviors to distinguish. Using glance location features and building sequences from task types, baseline sequences were identified perfectly. Manual radio tuning was also clearly distinguishable from baseline driving at every instance. Only on limited occasions was it confused with auditory-vocal radio tuning. Clearly, most of the classification difficulty stems, as expected, from the auditory-vocal radio tuning sequences. This correlates well with early explorations that show, for instance, that the probability of transitioning to the center stack from any other location in the vehicle during auditory-vocal radio tuning is evenly placed between the same probability during baseline driving and visual-manual radio tuning. From all auditory-vocal misclassifications, most sequences were confused with baseline driving when task types were used to build sequences. Fig. 5b shows these patterns also hold for glance duration features, albeit with a slightly higher number of misclassifications.

Fig. 6, which presents analogous results for sequences generated from individual task instances, visualizes another interesting property: when building glances per task, auditory-vocal tuning sequences are much more frequently misclassified as visual-manual tuning, yet the reverse does not hold. This resonates well with the finding that auditory-vocal interactions often build on subtasks with strong visual-manual components (Fitch et al. 2013; Mehler et al., 2014; Mehler, Kidd et al., 2016; Reimer et al., 2013; Reimer et al., 2016). Considering that Fig. 6 presents results based on shorter sequences (in comparison to Fig. 5), it is likely that these visual-manual subcomponents correspond to very localized glance transition patterns that become smoothed out in longer sequences over time, thus decreasing their likelihood of detection. In the vast majority of cases, however, auditory-vocal interactions are identified as their own type of visual behavior with their own unique visual demand profile. Finally, Fig. 7 gives the confusion matrices for classification on sequences built from task types and individual tasks using merged location and duration features.

<table>
<thead>
<tr>
<th>Train set</th>
<th>Test set</th>
<th>Glance location (%)</th>
<th>Duration (Gaussian) (%)</th>
<th>Duration (Histogram) (%)</th>
<th>Merged (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1 + 2</td>
<td>Study 1 + 2</td>
<td>90.71</td>
<td>77.81</td>
<td>79.71</td>
<td>90.65</td>
</tr>
<tr>
<td>Study 1</td>
<td>Study 2</td>
<td>87.38</td>
<td>76.14</td>
<td>76.96</td>
<td>89.00</td>
</tr>
<tr>
<td>Study 2</td>
<td>Study 1</td>
<td>91.53</td>
<td>67.94</td>
<td>81.99</td>
<td>93.53</td>
</tr>
<tr>
<td>Study 1</td>
<td>Study 1</td>
<td>95.70</td>
<td>76.93</td>
<td>79.85</td>
<td>91.72</td>
</tr>
<tr>
<td>Study 2</td>
<td>Study 2</td>
<td>86.75</td>
<td>73.57</td>
<td>77.72</td>
<td>86.96</td>
</tr>
</tbody>
</table>
5. Discussion

This report presents a methodology for assessing the quality of an HMI in terms of the visual demands it imposes on driver behavior. This assessment builds on the logical assumption that the driver has the potential to be most aware of his or her environment when not engaged in other tasks.

### Table 3

Classification accuracies combined for 100 iterations of the Monte Carlo estimation when building sequences from single task instances and using with the following feature combinations: glance location, Gaussian-distributed glance duration, histogram binned glance duration, and merged glance location & histogram-binned glance duration.

<table>
<thead>
<tr>
<th>Train set</th>
<th>Test set</th>
<th>Glance location (%)</th>
<th>Duration (Gaussian) (%)</th>
<th>Duration (Histogram) (%)</th>
<th>Merged (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1 + 2</td>
<td>Study 1 + 2</td>
<td>83.39</td>
<td>73.51</td>
<td>76.22</td>
<td>88.65</td>
</tr>
<tr>
<td>Study 1</td>
<td>Study 2</td>
<td>83.67</td>
<td>66.37</td>
<td>73.61</td>
<td>87.64</td>
</tr>
<tr>
<td>Study 2</td>
<td>Study 1</td>
<td>84.49</td>
<td>58.11</td>
<td>78.37</td>
<td>91.11</td>
</tr>
<tr>
<td>Study 1</td>
<td>Study 1</td>
<td>86.89</td>
<td>73.85</td>
<td>78.11</td>
<td>91.11</td>
</tr>
<tr>
<td>Study 2</td>
<td>Study 2</td>
<td>81.79</td>
<td>70.15</td>
<td>71.99</td>
<td>85.27</td>
</tr>
</tbody>
</table>

Fig. 3. Accuracy distributions for every train and test set combination using (a) glance location features and (b) glance duration features (histogram) when using task types to build sequences. Average number of trials (sequences) is given for each train and test set pair. 1 standard deviation is represented by the larger rectangular box, the 95% confidence interval by the lighter internal box, and the mean by the horizontal line.

Fig. 4. Accuracy distributions for every train and test set combination using (a) glance location features and (b) glance duration features (histogram) when using individual tasks to build sequences. Average number of trials (sequences) is given for each train and test set pair. 1 standard deviation is represented by the larger rectangular box, the 95% confidence interval by the lighter internal box, and the mean by the horizontal line.

5. Discussion

This report presents a methodology for assessing the quality of an HMI in terms of the visual demands it imposes on driver behavior. This assessment builds on the logical assumption that the driver has the potential to be most aware of his or her environment when not engaged in other tasks.
her environment in the baseline (“just driving”) scenario as opposed to periods of increased sensory, manual, or cognitive workload. In this sense, one might hypothesize that a good HMI would support visual behaviors that are indistinguishable from, or quite similar to, the “just driving” case, regardless of the subtask being completed by the driver.

HMMs have been used extensively in the past for analysis of time series data. They are especially well suited for the problem of behavior modeling over time, since they intrinsically model the semantic difference between observable data (here: the sequence of glance allocations) and unobservable data (here: the modality of the task as a representation of the driver’s visual demand), inferring the latter from the former. The HMM’s classification error rate is proposed as a measure for how similar these behaviors truly are; a higher error rate corresponds to more confusion between these classes, suggesting higher similarity between them and vice versa.
This quantification is relevant for the development of future attentional demand and safety policies. This approach suggests different metrics and design evaluation methods that may be useful in evaluating indicators of distraction in modern mixed mode HMIs, as they represent a higher order measure for summarizing an HMI’s impact on visual attention than the standard total time eyes off road or total glance time to device metrics (e.g., Chiang et al., 2004; Donmez et al., 2009; Driver Focus-Telematics Working Group, 2006; Mehler et al., 2014; National Highway Traffic Safety Administration, 2013; Reimer et al., 2013; Tijerina, 1999; Victor et al., 2013). As an example, an interface designer or other interested party could record a set of user glance sequences for a given established HMI or target glance allocation sequence. The visual allocation characteristics of a new HMI can then be compared to the target through the HMM’s predictive capability. Acceptable thresholds of comparison (i.e., how similar or dissimilar) might well be defined by a design philosophy or other reference point. HMMs essentially quantize a given behavior pattern into a set of states and transitions, making them largely independent of factors such as task time, given a reasonable amount of training data. This provides a comfortable platform for comparing the behavior patterns behind low-level measures such as total glance time instead of comparing these measures directly. This is particularly an issue as some HMI interactions such as address entry into a navigation system or point-of-interest selection vary significantly in temporal characteristics from tasks such as basic radio tuning (Mehler, Reimer, Dobres, Foley, & Ebe, 2016; Reimer et al., 2013) that significantly influenced the development of earlier design guidelines (Driver Focus-Telematics Working Group, 2006; NHTSA, 2013). In a real-time scenario, driving assistance systems may greatly benefit from these efforts by harvesting information from and dynamically reacting to inferred driver behavioral data. For instance, given a reliable eye tracker or other method of inferring glance locations in a production vehicle (Fridman, Langhans, Lee, & Reimer, 2016; Fridman, Lee, Reimer, & Victor, 2016), an HMM-based driver assistance system could continuously process new glance allocations and provide suitable warnings if it predicts the driver to be engaged in a pattern of visual allocation that is known to be associated with delayed response times or is inconsistent with design specifications. The latter may be potentially important in the maintenance of situational awareness in automated vehicles that are designed to require some driver monitoring of vehicle/environmental status information as part of a supervisory role.

This report extends on previous work in the field of driver modeling based on time series analysis methods by bringing glance features together with an established machine learning approach to analyze their predictive signal as part of a traditional classification task. This effort is notoriously missing in the available literature. Liang et al. (2012) have a similar objective, albeit with two fundamental differences. This work approaches the problem from a classification perspective, providing traditional machine learning performance measures instead of strictly low-level prediction measures such as $R^2$ and maximal Odds Ratio. As a result, glance history features are not limited to a linear fit, but rather enjoy the flexibility of a more appropriate statistical model (Hidden Markov Model).

Previous efforts in HMM modeling have focused on tailoring models to the classification task at hand. Kuge et al. (2000), He et al. (2012), and Pentland and Liu (1999) for instance provide hierarchical, staged models for driver behavior. The research presented here, however, operates under the assumption that visual behavior may vary considerably across participants and vehicle interfaces, and thus must not necessarily conform to any sort of staged development model, as is the case with driving maneuvers (Oliver & Pentland, 2000). As such, this effort restricts itself to a flat, standardized HMM structure for classification on the raw, discrete glance allocation data.
One common thread with previous works (He et al., 2012; Mitrovic, 2001; Oliver & Pentland, 2000; Rabiner, 1989) may be found in the classification framework itself. In the present study, one HMM was built for each class, where a class corresponds to an interaction modality, i.e. “just driving”, manual interaction (visual-manual radio tuning), and voice interaction (auditory-vocal radio tuning). A new test sequence was assigned to a class by picking out the HMM that was most likely to produce it. However, in addition to providing the probability of an unseen observation sequence, HMMs also provide posterior probability estimates for the probability of being in a particular state at a given time \( t \). A classification framework based on this probability as a predictive measure would work with a single HMM, where the states (in contrast to the standard approach above) take application-specific semantic values that correspond to the classes in the classification scheme. Within the driving context, this approach models a slightly different problem, namely how a driver weaves in and out of different behaviors over time. This is interesting for future work, as more data are collected for subtask periods within more complex (and presumably heterogeneous) tasks.

Data were collapsed in two variants, according to the type of task, and in parallel according to individual task trials. These experiments showed that HMMs are particularly sensitive to the amount of training and testing data provided to them. In contrast to other modeling approaches that work on single observation samples, HMMs classify entire observation sequences, thus increasing the amount of necessary data needed for a fair accuracy estimate, and are thus a relatively expensive modeling technique. Large amounts of data (i.e. collapsing according to task type) are necessary to build realistic feature distributions at each state, minimizing the problem of zero probability estimates and thus avoiding having to manually specify pseudo transition or emission counts as detailed in Section 3.3. In terms of testing data, a higher number of sequences (i.e. collapsing according to individual task trials) yields a more realistic understanding of the model’s performance. The implication is that a tradeoff must be made between the number of sequences and their individual expressivity when applying HMMs to any data set of fixed size. In this particular study, both ends of the spectrum resulted in superior classification performance significantly beyond that of random guessing. One key finding from this report is that glance allocation features generalize well across driver demographics, as a total of 21,280 transitions were drawn from heterogeneous pool of 156 participants. In contrast, some previous works (Kumagai et al., 2003) gathered data from a single participant.

Given that a near-optimal classification accuracy was observed with only \( n = 2 \) states (thus resulting in relatively simple models) the question must be asked whether the Hidden Markov Modeling technique used is in fact the most appropriate approach to model driver visual behavior. Future work should look to compare this modeling technique with other time series approaches.

Extending on previous work, the predictive signal in glance duration time series was also analyzed. As before, all glance sequences were left at their original length. A minimum threshold of 11 glances was set during validation in order to filter out samples with low information content. Note, some glances (such as glancing to the forward roadway) have inherently longer durations than others (e.g. glancing to the blind spots). Handling these features requires care. Experiments showed that several input training sequences were unable to adequately fit a Gaussian distribution, resulting in non-positive definite matrices during the likelihood estimation procedure. This was likely due to glance transition sequences of limited length and variance present in the data. These instances were skipped during training. A similar issue was encountered when using both glance location and duration features as part of the same feature vector, resulting in badly scaled, near singular matrices in the normal-inverse Wishart fit. As a remedy, a similar approach to that used in Yamato et al. (1992) was tested, in which each feature vector (composed of a discrete location variable and a real duration value), was mapped to a single discrete value by using k-means clustering to cluster all feature vectors and taking the ID of the corresponding cluster as the input observation to the HMM. Exploring this solution for \(< 6 \) clusters did not, however, lead to performance levels that were significantly different from random guessing. Section 3.3 describes the histogram-based binning approach that allowed proper merging of glance location and duration data.

There is a drop in classification accuracy for Study 2 as compared to the joint data across both studies as well as the data from Study 1 in particular. This suggests that Study 2 contains a subset of especially difficult sequences. This can be traced back to the fact that unlike Study 1, Study 2 was structured along additional criteria, i.e. with subgroups. In addition to participants that had been previously trained in the structure of the task and those that had not, Study 2 also contained drivers that used the vehicle’s “expert” voice mode instead of the default or standard voice mode of the interface. In order to examine the impact these variables have on the classification accuracy, the data from Study 2 were reduced to multiple combinations of participant subgroups, which were then used to independently train and validate the classifier. The subgroups that were examined included all default-mode participants, all default-mode trained participants, and all participants that were not self-trained. From these, the latter group resulted in a misclassification rate of auditory-vocal sequences (35%) that best approximates the corresponding misclassification rate in Study 1 (28%). This correlates well with the fact that all participants in Study 1 were trained to engage in each task in the same manner. This finding suggests that presenting participants with this structured training resulted in a more fine-tuned and systematic visual behavior, and that the modeling technique explored here can pick-up on this difference.

6. Future work

Experiments showed the difficulty of using a single HMM to model discrete and real variables simultaneously. However, the relatively strong predictive signal provided by glance duration features in isolation shows the potential of supplementing
location information with temporal data. A potentially promising area of future work lies in exploring other possibilities to jointly accommodate both discrete and real input variables, for instance by defining new, hand-crafted B parameters (Rabiner, 1989) of the base HMM model. An alternative would look beyond standard generative HMMs to discriminative variants, such as Maximum Entropy Markov Models (McCallum, Freitag, & Pereira, 2000), Factorial Hidden Markov Models (Ghahramani & Jordan, 1997), and Conditional Random Fields (Metzler & Croft, 2005). Yet another alternative could focus on working with features that inherently encode both location and duration information. As noted by Liang et al. (2012), the visual buffer concept illustrated in Kircher and Ahlström (2009) integrates all three characteristics of glance patterns, namely location, duration, and history. This is a potentially good match for the HMM classifier.

Efforts should be made to benchmark these results with that of alternative modeling approaches, including traditional machine learning algorithms such as decision trees, neural nets, dynamic naïve Bayes classifiers (Avilés-Arriaga, Sucar-Succar, Mendoza-Durán, & Pineda-Cortés, 2011) in conjunction with other time series analysis and reduction tools such as dynamic time warping (Müller, 2007).

Expanding beyond the thematic scope of this report, future work should consider alternative class structures. Promising work includes inferring user demographics, identity, emotional state, etc. based on glance behavior and/or driver performance measures. This challenge is accompanied by another familiar issue, namely the amount of data required. Given the rather limited number of participants/sequences overall, any subdivision of the present subject pool is likely to be too small for any meaningful testing. Future work should likewise examine to what extent the presented results hold for larger data collections.

Additionally, several studies have focused on HMMs as tools for real-time classification (Kumagai et al., 2003; Liu & Pentland, 1997; Pentland & Liu, 1999; Tchankue et al., 2013) by predicting near-future events. An additional potential area of future work could look at HMM modeling to predict a driver’s future glance allocations given an online analysis of his visual behavior.

7. Conclusions

This report has examined glance location and duration features as inputs to a Hidden Markov Model-based framework for classification of the visual demand in different task modalities (baseline driving, visual-manual radio tuning, auditory-vocal radio tuning). In contrast to previous approaches, glance features were used directly in a statistical classification framework to show that these modalities result in fundamentally and predictably different driver visual behaviors. These visual behaviors may be traced back to key HMI design variables such as task type. This methodology could likewise be used to analyze task structure (i.e. how the task is delivered). Classification error rate is suggested as a possible metric for how compatible an HMI is with an idealized state (e.g. with attentive single task baseline driving). Data were taken from two on-road driving studies for a total of 156 participants spanning across gender, age, and driving experience. Glance location sequences are shown to robustly generalize across these factors, consistently predicting different visual behavior patterns. Duration features are shown to complement location features to significantly boost classification performance. This report may be considered as a foundation for more intricate modeling approaches, as well as for real-time implementations that extend beyond the qualitative analysis of vehicle HMI design.

Acknowledgments

Support for the this work was provided by the Advanced Human Factors Evaluator for Automotive Demand (AHEAD) Consortium, the US DOT’s Region I New England University Transportation Center at MIT, and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs’ class counsel. Acknowledgement is also extended to The Santos Family Foundation and Toyota’s Collaborative Safety Research Center (CSRC) for providing funding for the studies (Mehler et al., 2014; Reimer et al., 2013) from which the data were drawn.

References


