Multimodal Classification of Driver Glance

Daniel Baumann
The Computer Laboratory
University of Cambridge, UK
ndebaumann@gmail.com

Marwa Mahmoud
The Computer Laboratory
University of Cambridge, UK
marwa.mahmoud@cl.cam.ac.uk

Peter Robinson
The Computer Laboratory
University of Cambridge, UK
peter.robinson@cl.cam.ac.uk

Eduardo Dias
Jaguar Land Rover
Coventry, UK
edias@jaguarlandrover.com

Lee Skrypchuk
Jaguar Land Rover
Coventry, UK
lskrypch@jaguarlandrover.com

***Abstract*—This paper presents a multimodal approach to in-vehicle classification of driver glances. Driver glance is a strong predictor of cognitive load and is a useful input to many applications in the automotive domain. Six descriptive glance regions are defined and a classifier is trained on video recordings of drivers from a single low-cost camera. Visual features such as head orientation, eye gaze and confidence ratings are extracted, then statistical methods are used to perform failure analysis and calibration on the visual features. Non-visual features such as steering wheel angle and indicator position are extracted from a RaceLogic VBOX system. The approach is evaluated on a dataset containing multiple 60 second samples from 14 participants recorded while driving in a natural environment. We compare our multimodal approach to separate unimodal approaches using both Support Vector Machine (SVM) and Random Forests (RF) classifiers. RF Mean Decrease in Gini Index is used to rank selected features which gives insight into the selected features and improves the classifier performance. We demonstrate that our multimodal approach yields significantly higher results than unimodal approaches. The final model achieves an average F_1 score of 70.5% across the six classes.**

1. Introduction

Driver distraction has emerged as one of the leading causes of vehicle related accidents [10], [13], [15], [16], [19]. The increasing adoption of digital interactions inside the vehicle threatens to aggravate this issue [7]. Driver distraction can be modelled by considering factors such as orientation, duration and history of the driver's gaze. This can then be used for notification scheduling and other safety related applications.

Liang et al. [13] published research on the 100-Car Naturalistic Driving Study comparing 24 driver distraction algorithms. These algorithms relied heavily on three characteristics: glance location, duration and history. Although based on limited manually-labelled data, the research highlights the theoretical importance of differentiating the glance locations as categories rather than raw angles. This is since driving related glances such as side mirrors are less risk inducing than glances at in-vehicle systems [13]. The research demonstrates the importance of considering a short window (less than three seconds) of off-road glances to predict crash risk. This motivates the need for a qualitative classifier versus a quantitative one.

We propose a qualitative 6-class driver glance classification system. The labels Forward, Down, Left, Right, Left Blind Spot, and Right Blind Spot are chosen as useful descriptors of the driver's glance. A baseline glance classifier is first implemented that classifies the glance direction based on the yaw and pitch of the driver's head. This baseline model assumes a perfect head tracker and linearly separable regions. However due to poor calibration, visual noise and limitations of the tracker, more predictors are required to determine the driver's glance accurately. We propose a multimodal approach combining visual cues with non-visual car parameters to classify driver glances. Visual features such as head position, orientation, confidence and eye gaze orientations are extracted. Statistical methods are implemented to filter, calibrate and extrapolate the visual information. Non-visual car parameters such as steering wheel angle and indicator position are extracted from CAN (Controller Area Network) data collected by a RaceLogic VBOX system.

The main contributions of this paper can be summarised as follows:

1) Proposing a multimodal approach for driver glance classification that combines driver's facial features with car parameters.
2) Demonstrating that our multimodal approach performs significantly better than single modalities.
3) Presenting a ranking of predictive importance of the visual and non-visual features considered, which informs future work in this field.

This paper will begin with a discussion of the related

978-1-5386-0563-9/17/$31.00

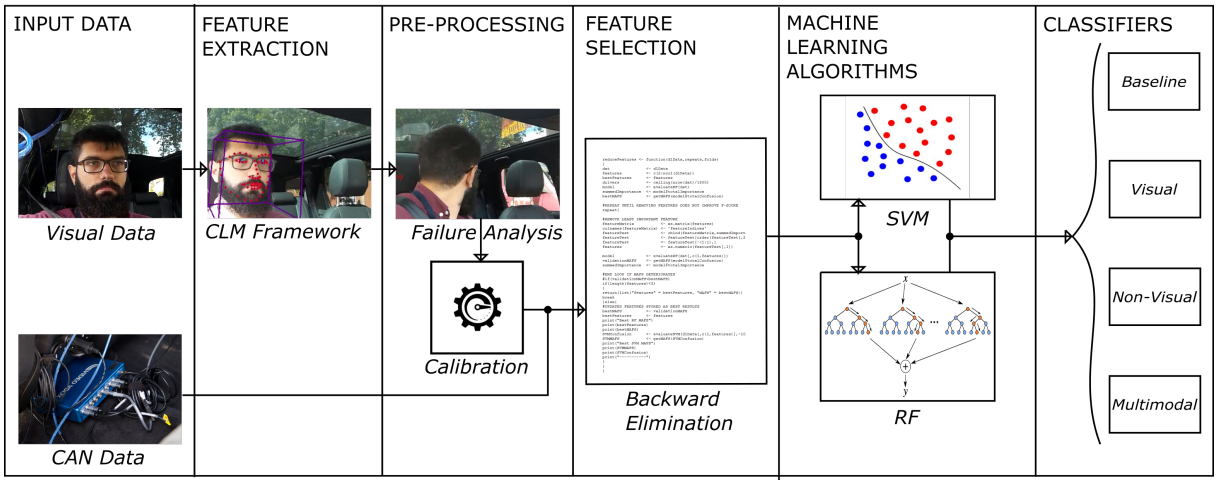

Figure 1. Overview of the process in generating the driver glance classifiers.

work in Section 2. In Section 3, the construction and preprocessing of the dataset is described. In Section 4, a comparison of four designed classifiers will be given followed by a discussion. Lastly in Section 5, conclusions will be drawn from the results and suggestions for future work will be given.

2. Related Work

The problem of modelling driver distraction using multiple modalities such as visual cues [9], [13], [17], audio [1], CAN data [11], biometric sensors [18] and GPS [4] has been investigated widely in recent years. However, attempts to classify driver glance generally depend on specialised equipment. For example, Sodhi et al. [17] used a head-mounted eye-tracking device to evaluate driver distraction based on eye positions and pupil diameters. Kutila et al. [11] utilised a stereo camera set-up combined with lane-tracking data in order to predict cognitive workload. They classified driver glance into four categories: “Road Ahead”, “Left Mirror”, “Right Mirror”, “Windscreen”. However the total cost of this system was reported to be around €35 000. Ji et al. [9] used a configuration of infrared LEDs to illuminate the driver’s pupils for better head and gaze tracking. Their system also consisted of two cameras and extracted additional visual cues such as eyelid movement. The use of infrared LEDs enabled accurate eye gaze tracking, but the range of head movement was still limited. Our proposed approach focuses on using a single, low-cost camera without depth information combined with CAN-Bus data.

3. Methodology

In this section the dataset, feature extraction methods, and evaluation methods are described.

3.1. Dataset

Trials were performed in a Land Rover Discovery fitted with a standard, uncalibrated USB Mobius camera monitoring the driver’s face and a RaceLogic VBOX. A driving route was specified of about 60 minutes. Multiple sixty seconds of naturalistic driver footage from 14 drivers were selected at 30 frames per second and processed at a resolution of 640x360 where significant head turning was naturally exhibited. Segments were selected where head turning was naturally exhibited, as drivers predominantly faced forward which was not useful in training the classifiers. The VBOX system recorded car information at only 10 frames per second. The frames were annotated with CowLog 3.0.2 [8] serving as the ground truth. The class labels Forward, Down, Left, Right, Left Blind Spot, Right Blind Spot relative to the camera’s perspective were chosen as significant regions which can then be further expanded on by other vehicle applications. The front car pillars and base of the windscreen were chosen as the separating boundaries. These regions are visualised in Figure 2. It must be noted that some classes were difficult to disambiguate during fast head turns as well as slight eye changes near the boundaries. As noted by Langton et al. [12], the head orientation is expected to have a large influence on the perception of a person’s glance location.

This yielded a total of 50 400 samples. In order to synchronize car information with the driver camera, the audio of the two devices were matched.

3.2. Feature Extraction

3.2.1. Visual Features. Over the past decade, many head and eye tracking methods have been proposed. In this paper, the Cambridge Face Tracker developed by Baltrušaitis et al. [2] was selected for visual feature extraction. This was chosen due to its real-time operation and focus on performance in the presence of poor lighting conditions, occlusions and extreme poses. The method locates facial landmarks as a function

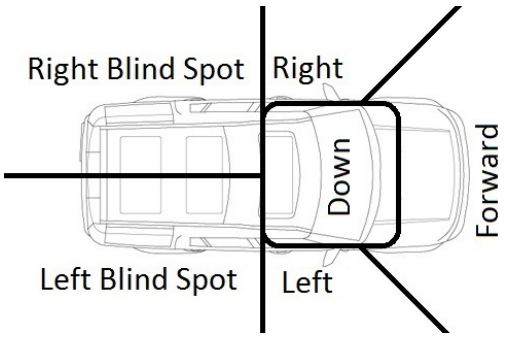

Figure 2. Top view allocation of driver glance regions. The directions are relative to the camera positioned to the right of the right-seated driver.

of spatial relationships and certainties of other detected landmarks. The system works from a single simple camera without a depth sensor, generalises well to different faces and returns a confidence measure which is useful in our subsequent analysis.

The framework extracted the following features from the data: Head Position (Horizontal, Vertical, Depth), Head Orientation (Pitch, Yaw, Roll), Head Orientation Confidence, Eye Orientations (X1,X2,Y1,Y2,Z1,Z2) and an Eyes Detected Flag.

3.2.2. Non-Visual Features. The non-visual data was captured at 10 frames per second with a RaceLogic VBOX. Each VBOX instance was duplicated twice in order to match the visual frame rate. The following features were selected from this system: Velocity, Accelerator Pedal Position, Brake Pressure, Steering Wheel Angle, Indicator Position, Gear Lever Position. These features are hypothesized to be correlated with driver glance behaviour.

3.3. Tracker Failure Analysis

Despite the tracker's robustness, it often lost track of its target. The success rate is further limited by the detectable range of head orientations. The tracker provided a probabilistic confidence measure for each head tracking estimation. Out of all the recorded frames, 16.6% had confidence levels of less than 50%. A low pass filter was implemented to determine the local head rotational velocity in order to extrapolate missing and low confidence (tuned to a threshold of 9%) values. The low pass filter was required to smooth out the noise in the readings that could otherwise forecast spurious values.

3.4. Calibration

Each driver was positioned differently relative to the camera. In order to regularize different driver positions and rotations, the average head position and rotation for each driver is subtracted from the readings. This average is acquired by considering a sliding window of 60 seconds from which to calibrate the data. Furthermore only frames with a high head tracking confidence (tuned to a threshold of 60%) are considered in the calibration calculations to compensate for spurious readings. This ensures that the calibration is user-independent and adjusts with driver position changes.

4. Evaluation

In this section, the implementation details are discussed followed by a comparison of the different constructed non-linear classifiers. Visual and non-visual modalities are first considered separately. The multimodal approach then considers the combination of these features. For each classifier, both Support Vector Machines (SVM) [6], [14] and Random Forests (RF) [3], [5] classifiers are trained. The individual class F_1 scores are reported as well the arithmetic mean.

4.1. Implementation Details

Two methods were chosen to train the classifiers. Random Forests were chosen as they provide variable importance techniques and are simpler to optimise. Support Vector Machines were chosen as they generally perform better on smaller datasets.

In order to design a classifier that generalises to unseen drivers, the system is trained with Leave-One-Out Cross Validation. For each combination, the classifier is trained on 13 drivers and tested on the remaining one.

The number of features was optimised by stepwise backward elimination feature selection guided by Random Forest Mean Decrease in Gini Index. In both classifiers the objective function was chosen to be the cross-validated arithmetic mean of the test class F_1 scores. This function ensures that the size of the classes do not affect their weighting in the optimisation function. Class sub-sampling is performed before training the classifiers due to the largely imbalanced label frequencies.

The Radial Basis Function was chosen as the SVM kernel. The model parameters were chosen by an iterative logarithmic grid search to optimize the cross-validated objective function. The RF model was tuned by first maximising the cross-validated objective function by adjusting the allocated number of features per candidate split, then adjusting the number of decision trees in the forest via a linear search.

4.2. Baseline algorithm

As a baseline we chose a Rule-based classifier using the visual yaw and pitch of the drivers head, since head pose is a good proxy for gaze direction. Without knowledge of the shortcoming of available trackers, one might expect this classifier to be sufficient. For example, if the head pitch exceeded a threshold, the frame would be labelled as Down. The process was then continued for the remaining classes using the head yaw. Threshold values were chosen by inspection from annotated training footage and class distributions. Calibration is omitted, but failure analysis is performed so that blind spots can be extrapolated. The performance of the classifier is tabulated in Table 1.

TABLE 1. F_1 SCORES [%] FOR EACH GLANCE CLASS FOR THE BASELINE, SVM AND RF CLASSIFIERS.

Algorithm	Modalities	Forward	Down	Left	Right	Left Blind Spot	Right Blind Spot	Mean
Baseline	Visual	71.1	44.1	72.8	19.2	14.3	22.7	40.7
SVM	Visual	72.4	54.9	86.0	54.7	65.8	48.0	63.9
RF	Visual	76.0	63.7	87.7	**61.3**	69.4	48.0	67.7
SVM	Non-Visual	57.5	27.4	35.6	11.9	20.4	33.6	31.1
RF	Non-Visual	57.2	31.8	34.8	14.0	14.1	37.9	31.6
SVM	Multimodal	74.8	59.3	83.2	56.8	63.7	47.9	64.3
RF	Multimodal	**76.9**	**65.0**	**88.9**	60.7	**75.0**	**56.7**	**70.5**

The baseline approach struggles to differentiate between the Left and Left Blind Spot given the head yaw and pitch. It assumes a linear separable boundary and does not account for the noise present in the features. The other approaches use machine learning methods that build more sophisticated, non-linear boundaries between the classes that account for the presence of noise and the additional available features.

4.3. Visual Classifier

For this classifier additional visual information is incorporated from the tracker such as the remaining head rotation value, eye information, confidence values and head angular velocities. Failure analysis is included to extrapolate missing values and calibration is applied to account for driver posture changes. Feature selection is performed to remove insignificant features. The resulting performance of the classifiers are tabulated in Table 1.

As expected from the baseline approach assumptions, the head yaw and pitch are the most important variables. The eye orientation X1 component also scores highly as expected since the glance classification can shift solely on the change in eye positions. The horizontal head position scores well on the list due to a tendency of drivers to translate horizontally when turning their head. The confidence level aids the classification since it is likely to be inversely correlated to the angle between the drivers glance and the camera's principle axis.

4.4. Non-Visual Classifier

This classifier is trained purely on non-visual VBOX data. The performance of the classifiers are tabulated in Table 1. The non-visual classifier performs surprisingly well given the features and limited training data. This indicates the importance of non-visual CAN data in glance estimation. Furthermore, these features do not require any failure analysis or calibration and contain negligible noise.

4.5. Multimodal Classifier

This classifier merges the non-visual information from the CAN data with the visual information. The performance of the classifiers are tabulated in Table 1.

The variables considered and their variable importances were ranked in order of RF mean decrease in Gini index as shown in Figure 3. The larger the mean decrease, the more valuable it is in constructing accurate decision trees. The features following the Gear Lever Position did not improve the cross-validated optimisation function.

TABLE 2. CONFUSION MATRIX FOR THE MULTIMODAL CLASSIFIER. ROWS REPRESENT THE PREDICTED VALUES. COLUMNS REPRESENT THE GROUND TRUTH VALUES.

P \ GT	F	D	L	R	LB	RB	Precision
F	**3329**	175	227	416	9	44	79.3
D	635	**1841**	100	427	11	155	58.1
L	183	64	**3523**	72	109	30	88.5
R	300	383	33	**2070**	12	473	63.3
LB	0	0	110	0	**413**	0	79
RB	0	15	0	443	25	**661**	57.8
Recall	74.9	74.3	88.2	60.4	71.3	48.5	

4.6. Discussion

As can be seen from the Table 1, the RF multimodal approach surpasses all of the other classifiers considered. The significance is verified with a paired samples t-test. The multimodal RF results for each driver were significantly more accurate than the unimodal visual RF results ($t(13) = 1.95, p < 0.05$) with a Cohen's D of 0.52.

The main bottleneck for the performance of the classifiers is the success rate of the tracker. Tracking failure increased with the angle of the driver glance to the camera principle axis. Augmenting the visual features with the CAN data helped to achieve better performance in these cases. The size of the dataset was limited given the time consuming task of data collection and labelling. It is expected that a larger dataset could further improve the classifiers.

By inspecting the confusion matrix for our RF multimodal classifier in Table 2, we can see that neighbouring classes are often confused, whereas glances that are distant are less likely to be mistaken.

5. Conclusions and Future Work

We have presented a novel multimodal approach to classify driver glance behaviour into six classes. We demonstrate that augmenting the visual features with vehicle CAN data outperformed the separate single modality alternatives. Visual features such as head and eye gaze orientation are successfully combined with non-visual features such as

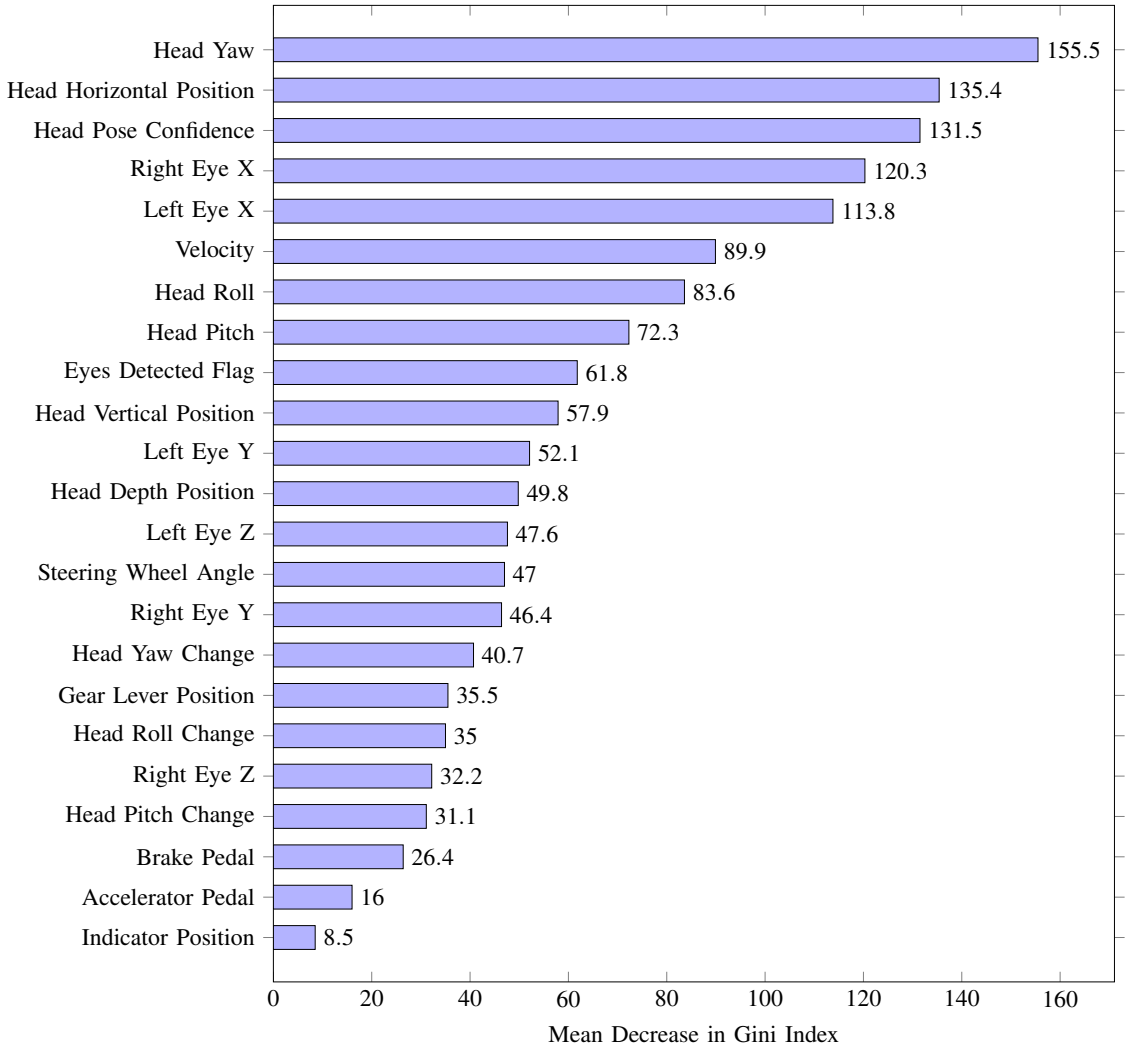

Figure 3. Variable importance ranking determined by the RF mean decrease in Gini index for the multimodal classifier.

steering wheel angle and gear lever position. The multimodal RF classifier returns F_1 scores of 76.9%, 88.9% and 75.0% for the Forward, Left and Left Blind Spot regions respectively. This demonstrates the potential of classifying driver glance direction using a single camera enhanced with the use of auxiliary vehicle CAN data.

For future work, we would like to investigate the addition of a Rear View Mirror class given more data. As for non-visual modalities, we would like to incorporate GPS data in order to consider the expected behaviour given the driver location.

Acknowledgments

The work presented in this paper was funded and supported by Jaguar Land Rover, Coventry, UK.

References

[1] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. H. Hansen, "Utdrive: driver behavior and speech interactive systems for in-vehicle environments," in *Intelligent Vehicles Symposium, 2007 IEEE*, pp. 566–569, IEEE, 2007.

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 354–361, 2013.

[3] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] C. Busso and J. Jain, "Advances in multimodal tracking of driver distraction," in *Digital Signal Processing for In-Vehicle Systems and Safety*, pp. 253–270, Springer, 2012.

[5] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 617–624, IEEE, 2011.

[6] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2348–2355, 2004.

[7] P. Green, "Crashes induced by driver information systems and what can be done to reduce them," in *Sae Conference Proceedings p*, pp. 27–36, SAE; 1999, 2000.

[8] L. Hänninen and M. Pastell, "Cowlog: Open-source software for coding behaviors from digital video," *Behavior Research Methods*, vol. 41, no. 2, pp. 472–476, 2009.

[9] Q. Ji and X. Yang, "Real-time eye, gaze, and face pose tracking for monitoring driver vigilance," *Real-Time Imaging*, vol. 8, no. 5, pp. 357–377, 2002.

[10] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, D. J. Ramsey, *et al.*, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," 2006.

[11] M. Kutila, M. Jokela, G. Markkula, and M. R. Rué, "Driver distraction detection with a camera vision system," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 6, pp. VI–201, IEEE, 2007.

[12] S. R. Langton, H. Honeyman, and E. Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction," *Perception & psychophysics*, vol. 66, no. 5, pp. 752–771, 2004.

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

[14] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection," in *Computer vision and pattern recognition, 1997. Proceedings., 1997 IEEE computer society conference on*, pp. 130–136, IEEE, 1997.

[15] C. J. Patten, A. Kircher, J. Östlund, and L. Nilsson, "Using mobile telephones: cognitive workload and attention resource allocation," *Accident analysis & prevention*, vol. 36, no. 3, pp. 341–350, 2004.

[16] D. A. Redelmeier and R. J. Tibshirani, "Association between cellular-telephone calls and motor vehicle collisions," *New England Journal of Medicine*, vol. 336, no. 7, pp. 453–458, 1997.

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

[18] E. T. Solovey, M. Zec, E. A. Garcia Perez, B. Reimer, and B. Mehler, "Classifying driver workload using physiological and driving performance data: two field studies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 4057–4066, ACM, 2014.

[19] D. L. Strayer, J. M. Cooper, and F. A. Drews, "What do drivers fail to see when conversing on a cell phone?," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 48, pp. 2213–2217, SAGE Publications, 2004.