Multimodal Classification of Driver Glance

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 by tracking software, 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. We evaluate our approach on a dataset of 13 drivers, each consisting of 60 seconds at 30 frames per second. 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 68.3% across the six classes.

Keywords
Multimodal Classification; Glance Classification; Driver Distraction; Vehicle CAN data

1. INTRODUCTION

Driver distraction has emerged as one of the leading causes of vehicle related accidents [1]. The increasing adoption of digital interactions inside the vehicle threatens to aggravate this issue. 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. [1] 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 [1]. The research demonstrates the importance of considering a short window (less than 3 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 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.
Figure 1: Overview of the process in generating the driver glance classifiers.

2. RELATED WORK

The problem of modelling driver distraction using multiple modalities such as visual cues, audio, CAN data, biometric sensors and GPS has been investigated widely in recent years [1, 2, 3, 5, 6, 7]. However, attempts to classify driver glance generally depend on specialised equipment. For example, Sodhi et al. [7] used a head-mounted eye-tracking device to evaluate driver distraction based on eye positions and pupil diameters. Kutila et al. [6] 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. [2] 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 new 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 90 minutes. Sixty seconds of naturalistic driver footage from each of 13 drivers was selected at 30 frames per second and processed at a resolution of 640x360 where significant head turning was naturally exhibited. 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. A Rear View Mirror class was considered, but removed from our taxonomy due to difficulty in distinguishing it from the Forward class.

This yielded a total of 23 400 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 Baltrusaitis et al. [4] 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 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 system provides .vbo files for each set of readings which were then converted to .csv files. 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, 20.4% had confidence levels of less than 50% and 6.6% had confidence levels less than 1%. 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) and Random Forests (RF) classifiers are trained. The individual class F1 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 12 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 F1 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. Specifically, the Left and Right Blind Spot labels contain 288 and 662 instances respectively out of the total of 23,400.

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.

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 is tabulated in Table 1.

The RF and SVM classifiers produced comparable results except for the Left Blind Spot label. The following variables were considered and their variable importances were ranked in order of mean decrease in RF Gini Index (most significant to least significant): head yaw, head pitch, horizontal head position, eye orientation X2, head roll, eye orientation X1, head confidence level, vertical head position, eye orientation Y1, eye orientation Y2, change in head yaw, head depth, eye orientation Z1, change in head roll, change in head pitch, eyes detected flag. The classifier performed best when the features after the vertical head position were disregarded. 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.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Modalities</th>
<th>Forward</th>
<th>Down</th>
<th>Left</th>
<th>Right</th>
<th>Left Blind Spot</th>
<th>Right Blind Spot</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Visual</td>
<td>68.4</td>
<td>54.8</td>
<td>3.2</td>
<td>43.5</td>
<td>6.85</td>
<td>33.5</td>
<td>35.0</td>
</tr>
<tr>
<td>SVM</td>
<td>Visual</td>
<td>84.3</td>
<td>59.4</td>
<td>88.0</td>
<td>47.4</td>
<td>60.1</td>
<td>52.3</td>
<td>65.2</td>
</tr>
<tr>
<td>RF</td>
<td>Visual</td>
<td>81.5</td>
<td>57.0</td>
<td>88.3</td>
<td>53.5</td>
<td>19.3</td>
<td>49.4</td>
<td>58.2</td>
</tr>
<tr>
<td>SVM</td>
<td>Non-Visual</td>
<td>66.5</td>
<td>38.6</td>
<td>18.9</td>
<td>17.8</td>
<td>0.0</td>
<td>18.0</td>
<td>26.6</td>
</tr>
<tr>
<td>RF</td>
<td>Non-Visual</td>
<td>66.1</td>
<td>33.6</td>
<td>36.9</td>
<td>20.7</td>
<td>22.2</td>
<td>9.2</td>
<td>31.5</td>
</tr>
<tr>
<td>SVM</td>
<td>Multimodal</td>
<td>82.5</td>
<td>65.5</td>
<td>90.4</td>
<td>45.1</td>
<td>71.5</td>
<td>54.7</td>
<td>68.3</td>
</tr>
<tr>
<td>RF</td>
<td>Multimodal</td>
<td>80.8</td>
<td>64.2</td>
<td>86.9</td>
<td>55.2</td>
<td>49.6</td>
<td>51.3</td>
<td>64.7</td>
</tr>
</tbody>
</table>

Table 1: F₁ scores [%] for each glance class for the baseline, SVM and RF classifiers.

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. The variables considered were calculated to have a variable importance ranking as follows (most significant to least significant): Steering Wheel Angle, Velocity, Brake Pressure, Gear Lever Position, Accelerator Pedal Position, Indicator Position. The RF classifier yielded the best performance which was achieved using only the first three features.

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. In this case the SVM performs better than the RF classifier. The variables considered were calculated to have a variable importance ranking as follows (most significant to least significant): Head yaw, horizontal head position, head pitch, head confidence, head roll, eye component X1, vertical head position, steering wheel angle, gear lever position. The remaining features did not improve the cross validated optimisation function.

By inspecting the confusion matrix for the SVM classifier in Table 2, it is possible to interpret the ambiguities between the chosen classes. As can be seen from the confusion matrix, neighbouring classes are often confused, whereas glances that are distant are less likely to be mistaken.

4.6 Discussion

As can be seen from the Table 1, the SVM multimodal approach surpasses all of the other classifiers considered. The significance is verified with a paired samples t-test. The multimodal SVM results for each driver were significantly more accurate than the unimodal visual SVM results (t(12) = 1.87, p = 0.0428).

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. A second camera at the adjacent front corner of the car could be incorporated to further raise the performance of the Right and Right Blind Spot classes. The size of the training set was another limitation. It is expected that a larger dataset would significantly improve the classifiers.

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 modality alternatives. Visual features such as head and eye gaze orientation are successfully combined with non-visual features such as steering wheel angle and gear lever position. The multimodal SVM classifier returns F₁ scores of 82.5%, 90.4% and 71.5% for the Forward, Left and Left Blind Spot regions respectively. The shortcoming of the other labels are primarily due to the large angle between the driver glance and the camera principle axis. This can be ameliorated by incorporating a similar second camera at the adjacent front corner of the car.

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

6. ACKNOWLEDGMENTS

[Space held for acknowledgements after anonymous review.]
7. REFERENCES


