Multimodal Detection of Drivers Drowsiness and Distraction

Kapotaksha Das
takposha@umich.edu
Computer and Information Science
University of Michigan-Dearborn
USA

Salem Sharak
sharak@umich.edu
Computer and Information Science
University of Michigan-Dearborn
USA

Kais Riani
kriani@umich.edu
Computer and Information Science
University of Michigan-Dearborn
USA

Mohamed Abouelenien
zmohamed@umich.edu
Computer and Information Science
University of Michigan-Dearborn
USA

Mihai Burzo
mburzo@umich.edu
Mechanical Engineering
University of Michigan-Flint
USA

Michalis Papakostas
mpapakos@umich.edu
Computer Science and Engineering
University of Michigan
USA

ABSTRACT

Considering the ever-growing presence of automobiles around the world, ensuring the safety of those on and near roadways is of great importance. From the causes of accidents, drowsiness and distractedness are among the most consequential. In this paper, we use a multimodal dataset consisting of 11 recorded channels over 45 subjects to model driver's drowsiness and distraction. Our work puts forward the application of this dataset by using segmented windows as features, resulting in four main contributions. We explore the performance of each individual modality and specify which signals and features have a better capability of detecting drowsiness and different kinds of distractions. In addition, we analyze the effects of early fusion on the classification of the driver's state using multiple physiological and thermal channels. Finally, we use cascaded late fusion and test three voting strategies to evaluate the performance of our proposed approach. Our results confirm the effectiveness of utilizing a multimodal approach in detecting both drowsiness and distraction as two separate factors influencing the driver and provide guidelines on which signals are appropriate for detecting different driver's states.

CCS CONCEPTS

• **Information systems** → **Task models**; *Personalization*; • **Applied computing** → Transportation; • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**.

KEYWORDS

multimodal dataset, driver alertness, action units, classification, machine learning , thermal imaging

ACM Reference Format:
Kapotaksha Das, Salem Sharak, Kais Riani, Mohamed Abouelenien, Mihai Burzo, and Michalis Papakostas. 2021. Multimodal Detection of Drivers Drowsiness and Distraction. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479890

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479890

1 INTRODUCTION

Around the world, it is a mutually held fact that driving accidents pose a danger to drivers, occupants, and their surroundings, with the Center for Disease Control and Prevention (CDC) estimating about three million related non-fatal injuries yearly in the United States alone [20]. In 2020, the World Health Organization (WHO) found that global road traffic deaths are estimated at 1.35 million annually [28]. Looking past the significant number of injuries and loss of life, it is also worth considering the great costs caused by motor vehicle accidents. The CDC estimated in 2017 that motor vehicle accidents imparted an estimated collateral cost of $75 billion in the form of medical care costs and productivity losses. At a wider scale, Chen et al. [13] estimated a global macroeconomic burden of $1.8 trillion linked to road-related injuries between 2015 and 2030.

Two of the driver-centric causes most attributed to traffic accidents are distracted driving and driver's drowsiness. According to a recent National Highway Traffic Safety Administration (NHTSA) report, an estimated 400,000 people were injured in motor vehicle crashes, with an additional 2,841 reported deaths involving distracted driving in 2018 [2]. In addition, the NHTSA reported that nearly one-tenth of all teens who died in motor vehicle crashes were killed in crashes that involved distracted driving [4]. However, such figures are likely underreported due to the difficulty involved in detecting such behavior after a crash; the National Safety Council (NSC) reported in 2016 that more than half of US states lacked any fields or codes for police to record texting or hands-free cell phone usage in crash reports [5].

Regarding drowsy driving, an estimated seven percent of traffic accidents are attributed to this behavior in the United States as estimated in 2017, including 16 percent of fatal accidents [39]. Other studies, however, indicate that drowsiness might also be significantly under-reported, with data showing drowsiness as a contributing factor in 20% of all crashes [1]. Drowsy driving is consequential as it results in impaired judgement, decision making, attention, reaction times, and mental processing, as reported by the NHTSA [3]. Moreover, prolonged drowsiness has been shown to

result in driving behavior similar to driving under the influence of alcohol, as demonstrated in earlier work [9, 43].

Considering the gravity of the problem, there is significant motivation among the research community to address the issue, especially when linked with motor vehicle crashes. This research includes earlier work by Brookhuis and Waard [12] as well as work by Reimer and Mehler [33]; these works aimed to model the driver's state to include an indistinct combination of driver drowsiness and distractedness. More recent work further focused on modeling driver's alertness built on sensor data, such as electroencephalogram (EEG), respiration rate and skin temperature sensors, among others [11, 29, 30]. Finally, research studies including those by Naqvi et al. [27], Raorane et al. [32], Lopez et al. [25], and Kiashari et al. [23], among others, modeled driver's drowsiness using visual, near infra-red (NIR), and thermal imaging.

In this paper, we present four main contributions. Firstly, we collected a novel multimodal dataset of 45 subjects, across seven recorded channels: one visual RGB, one NIR, one thermal channel, and four physiological signals, including heart rate, skin conductance, respiration rate, and skin temperature. Secondly, we use these channels to model alertness, drowsiness, and distraction as three separate driver's states, looking into the modalities and features that describe each state the best. Thirdly, we highlight the usage of segmented windows to split the dataset into smaller units of time that can be used for drowsy and distraction detection without needing the entire recording. Fourthly, we explore cascaded late fusion as a way to leverage this multimodal dataset in order to generate predictions that are more accurate than seen in unimodal segmented classification while testing various strategies to optimize a voting system for detection.

2 RELATED WORK

With an increased variety of sensors being integrated into modern cars, researchers have been investigating the use of physiological sensors to model driver's behavior [37]. The review performed by Chowdhury et al. [15] discussed existing proposals which aimed at using physiological signals to detect drowsiness. As the community was primarily using signals, such as heart rate, respiration rate, skin conductance, ECG, EEG, and skin temperature, the modalities in use were generally similar across the papers in review.

In addition to drowsiness detection, physiological data was evaluated to classify distractors. McDonald et al. [26] addressed the benefits of ensemble learners in modeling driver's activity and classifying distractors based on physiological indicators. The study published by Papakostas et al. [29] utilized a deep learning method with four physiological signals to detect driver's alertness.

In spite of the good performance shown by such physiological features, the need for sensor intrusiveness remained a significant complication when considering the viability of these approaches. With that in mind, researchers used the visual modality to tackle this problem. D'Orazio et al. [18] made an early attempt to implement a neural classifier to identify driver's sleepiness by analyzing recorded image sequences and identifying when subjects' eyes were closed. Several works have since been published, using novel and sophisticated computational methods to address the problem.

Deep learning methods were also used to detect distractions. The dataset proposed by Abouelnaga et al. [8] was used in the works presented by Kose et al. [24] and Rao et al. [31] to classify video segments into 10 target classes by utilizing convolutional neural networks. In their work, they went beyond distraction detection to distraction recognition, however, their methods were limited to other types of passive distractors which relate to frustration, anxiety or even verbal interaction.

In addition to these visual methods, a Near Infrared (NIR) camera system directed at the vehicle's front windshield was used by Artan et al. [10] to detect the driver's cell phone usage. They used SVMs on a dataset of 1500 images collected on a public roadway. Naqvi et al. [27] proposed a deep learning-based gaze detection system for automobile drivers, which utilizes an NIR camera sensor to capture the driver's head and eye expression. More recently, Janveja et al. [21] presented two approaches for driver monitoring using a smartphone RGB camera along with an NIR LED. Their setup showed great potential for detecting driver's alertness in low-light with an accuracy and F1-score of 85% and 93.8%, respectively.

Researchers also studied thermal imaging as a possible non-invasive modality. Lopez et al. [25] developed a Therm-App mobile thermal camera which was used to identify fatigued individuals. Their procedure consisted of three primary steps. The first included the following substeps: detection, segmentation, and alignment of thermal facial images. The second step utilized two distinct convolutional neural networks to produce fixed-length deep feature vectors extracted from the facial images and regions. The third step then used these features with a Support Vector Machine (SVM) to classify a subject's state as fatigued or resting.

Although a large number of studies used a single modality to detect driver's alertness [41, 42], few studies exploited the potential of multimodal data for detecting drowsiness or distractions [17, 34, 35]. Moreover, few studies have looked specifically at how different common driving distractions affect specific physiological and behavioral responses [45]. Table 1 highlights key papers that approached the problems of driver distraction and drowsiness along with their performance metrics.

Table 1: Performance Metrics of past research in this field

Authors (Year)	Modality(s) Used	Detection Task	Performance Metrics
M.F.Valstar et al. (2015) [40]	Visual	Drowsiness	73%
M.B.Lopez et al. (2017) [25]	Thermal	Drowsiness	81.5%
M.Papakostas et al. (2020) [29]	Physiological	Distraction, Drowsiness	77% (AUC)
M.Abouelenein et al. (2015) [6]	Physiological, Vehicular Data	Drowsiness	86%
K.Riani et al. (2020) [34]	Thermal, Visual & Physiological	Distraction, Drowsiness	82%
Y.Du et al. (2018) [16]	Visual, Audio & Vehicular Data	Distraction	56% (F1-Score)

3 DATASET

For our experiments, we gathered a novel multimodal dataset that consists of thermal, NIR, visual, and physiological data from 45

subjects. The dataset was obtained using a driver's simulator from 30 male and 15 female participants, all between the ages of 20 and 33 years. The recordings took place during two different sessions. The earlier recording was typically taken before 11 a.m., while the other recording was usually taken between 4 p.m. and 8 p.m. We requested all participants to arrange the earlier recording as the first task in their day in order to simulate driver's alertness. For the other recording, the participants were scheduled for the session usually before going home, and were asked not to nap the whole day before the recording time. The majority of the participants were graduate and undergraduate students who took part in the late recording after spending long hours at school.

Every recording lasted an average of 45 minutes and consisted of three separate sub-recordings: 'baseline,' 'freedriving', and 'distraction.' This was repeated for both recording sessions, with a rearranged order of the distractors. During each session, and for all distractions and freedriving segments, the subjects began driving on a low-traffic highway and were then free to continue on the highway or divert to city-like driveways. No pedestrians were included and clear weather conditions under daytime were selected for our simulations. During the baseline recording, the participants were asked to sit still and breathe naturally during the first half of the baseline. In the second half, they were asked to follow with their gaze a target displayed on the screens. Each half lasted approximately two minutes and thirty seconds. Before starting the freedriving recording, we explained the virtual environment and the basic operating controls to the participants. We then let them drive around the environment for a few minutes to familiarize themselves with the simulator. During the recording, the subjects were instructed to drive for approximately 15 minutes without any interruption. The final part was the recording of 'distractions'. This recording consisted of four different subsections simulating various types of common distractors present when driving.

Below are the four different distractors to which participants have been exposed:

Physical: We asked the participants to type a short text message into their personal mobile devices. The text was an 8-word message which the experiment supervisor dictated on the fly to the participant. However, text incorporates all three distraction classes identified by NHTSA and the CDC, which are the manual, visual and cognitive. Through using predefined messages, we tried to mitigate the negative cognitive effect of forming texts and focused on the physical disengagement. The mobile device was placed on a flexible device next to the driver.

Cognitive: The N-Back test was the second distractor. This distraction was meant exclusively to challenge the short term memory skills of the driving subjects. N-Back is a cognitive task commonly used in the field of psychology and cognitive neuroscience for the measurement of working memory [22]. The subjects were asked to determine whether a letter matched another letter from n steps in a prerecorded 50 letter sequence. In our experiment, we set n=1.

Emotional: In this distractor, participants were asked to listen to an audio clip from the news and then share their opinions about an emotionally provocative case. In particular, we had two recordings related to a potential active shooter event and a fatal road accident due to the use of cell phone while driving.

Frustration: During this step, we asked participants to locate a particular destination on a GPS by interacting verbally. This distractor was intended to induce confusions and frustration to the participants. In this case the 'GPS' was simulated by a researcher in the background via pre-recorded audio, providing the subject with false answers.

The content used for the distractors was different for the two recordings sessions, including using different N-back tests and a different emotional event.

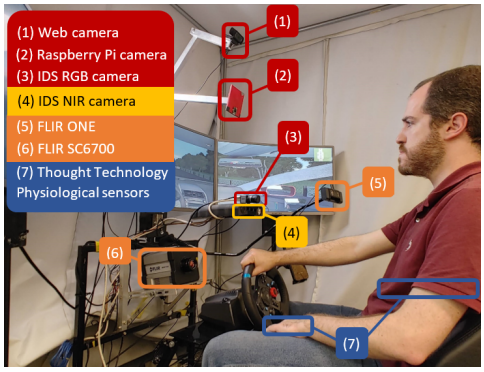

Figure 1: Experimental setup

In addition to the different scenarios, we had a system of cameras and sensors prepared and attached to the driver during each recording as shown in Fig. 1. A set of four physiological sensors were used to record the physiological data at rate of 2048 Hz: a) *Blood Volume Pulse* (BVP), b) *Skin Temperature*, c) *Skin Conductance* and d) *Respiration*. The respiration rate sensor was attached to the torso and the three other sensors were attached to the subject's non-dominant hand in order to reduce the noise in the data. We collected the visual data using three different cameras. A Raspberry-Pi camera placed approximately 38 inches away from the subject, angled downwards at an angle of approximately 18°, recording at 25 FPS, and an RGB camera from IDS placed approximately 28 inches away from the subject, angled upwards at an angle of 12°, recording at 20 FPS were used, each providing a face closeup view. In addition, we used a Logitech HD web camera placed 42 inches away from the subject's face angled downwards at an angle of approximately 28°, recording at 30 FPS, providing a top-down oblique view of the subject. We also used an NIR camera from IDS providing a close-up view, placed 28 inches away from the subject, angled upwards at an angle of approximately 12°, recording at 20 FPS. For our thermal modality, we used a low resolution thermal FLIR ONE camera, recording the subject's face at a slight angle at seven FPS as well as a high resolution FLIR SC6700 thermal camera, placed approximately 42 inches away from the subject's face angled upwards at an angle of approximately 18° capturing the subject's face at 100 FPS with a resolution of 640x512 pixels and 7.2M electrons.

4 METHODOLOGY

4.1 Thermal Modality

For the thermal modality, three major processing steps were conducted to generate the thermal map. The first step involved segmenting the frame into five regions: the whole face, the forehead, the eyes, the cheeks, and the nose. Following that, the thermal-based tracking algorithm proposed in [7] was used to map those regions in the recordings. Ultimately, we created the final thermal feature vectors by generating thermal maps for all Regions Of Interest (ROIs). Looking at these steps in further detail, we started by manually locating the ROI by determining their bounding boxes in the first frame, as automatic facial detection methods did not provide acceptable performance on thermal images. Thereafter, we used a variation of the Shi-Tomasi corner detection algorithm [36] in order to detect points of interest in each ROI by computing the weighted square difference between two consecutive frames. These interesting points represented areas of the skin with sharper change in temperatures.

We tracked the ROI bounding box for the duration of the videos by applying a fast version of the KLT tracking method [38], which provides accurate results when the objects maintain their shape over time. The algorithm predicted the relocation of points of interest from one frame to the next with a slight displacement between pixels, which was ideal for our tracking requirements. Following the tracking and displacement estimation processes, a geometric transformation was used to map the points of interest between frames on the basis of similarity.

In order to take account of probable occlusion including incomplete presence of the subject's face in the frame, we set a threshold of 95% of correctly mapped points between two succeeding frames. In cases of occlusion, the frame was removed and the tracking process continued using the following frame. Finally, in order to derive potentially indicative thermal features of drowsy and distracted behaviors in our five areas of concern, we created a thermal map that illustrated the thermal features in the ROIs. To that end, the following steps had to be performed in this order for each ROI: a) ROI segmentation, b) segment binarization, c) image masking, and d) cropping of each ROI. Figure 2 illustrates this procedure.

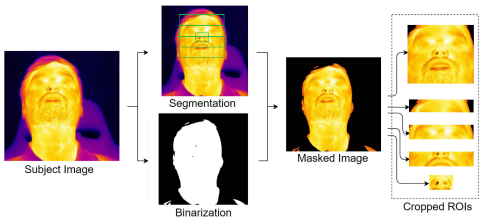

Figure 2: The process of segmenting, binarizing, masking and cropping the thermal faces

For frames captured with the thermal camera, the pixel values represented temperatures in Fahrenheit. The map consisted of extracting the mean pixel values within the ROI, the minimum pixel value, the maximum pixel value representing the highest temperature, the difference between the maximum and the minimum, and a 20-bin histogram that represents the temperature distribution in that region, forming a total of 24 features per feature vector. Certain recordings were eliminated for the thermal modality due to errors in segmenting these videos, resulting in a total of 421 recordings; 210 alert recordings and 211 drowsy recordings, as well as 85 undistracted and 336 distracted recordings, respectively.

4.2 RGB and NIR Modalities

To process the data generated by the visual cameras, we employed the OpenFace library to extract features from the RGB and NIR cameras. Such features include facial landmarks, eye gaze, head pose, and a series of detected Action Units (AUs) as defined by the Facial Action Coding System (FACS) [19]. In extracting these features, we deployed a Convolutional Experts - Constrained Local Model (CE-CLM) within OpenFace, as demonstrated by Zadeh et al. [44].

For this paper, the system extracted the facial landmarks of the face at every frame of the video, then approximated the 3D positioning of the head, along with a derived feature vector defining the head's pose. The disparate facial features were evaluated as the difference between the current facial state and a neutral expression, then abridged into AUs. These AUs, which are specified along with the detected intensity, abstract facial features into related groups, summarizing the feature vector coming from individual facial deformations while maintaining critical information. Examples of such AUs are shown in Fig. 3. Finally, the system produced an approximated gaze direction using facial landmarks in the region associated with the eyes by determining the intersection of the pupil's center and the eyeball sphere.

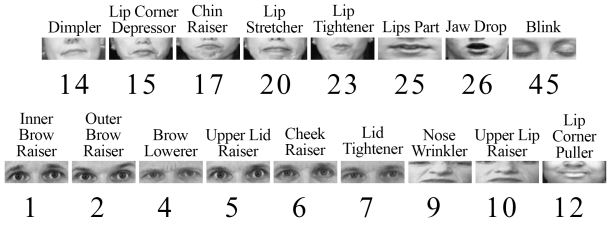

Figure 3: AU definitions and their numbering as defined by FACS.

After the system extracted the 709 features across each frame, we employed the segmentation process described in Section 4.4. Finally, the results were normalized by dividing against each subject's respective baseline recording. In total, the 45 subjects' recordings resulted as follows: For the NIR modality, there were a total of 427 recordings, with 209 alert recordings and 218 drowsy recordings, as well as 87 undistracted and 340 distracted recordings, respectively. As for the visual modality, there were a total of 420 recordings, with 205 alert recordings and 215 drowsy recordings, as well as 86 undistracted and 334 distracted recordings, respectively.

4.3 Physiological Modality

Using the data collected by the aforementioned physiological sensors, statistical features were extracted over four signal channels as

follows: Blood Volume Pulse (BVP), Skin Conductance, Skin Temperature, and Respiration Rate. The BVP had the maximum sample rate at 2048 Hz. The other three modalities were sampled at 256 Hz and then upscaled up to 2048 Hz to allow for consistency.

From these four signals 73 features were derived, out of which 49 belonged to the BVP, consisting of time domain statistical features, such as the mean, minimum, maximum and standard deviation which described the overall behavior of the signal and also the relation between consecutive inter-beat interval (IBI), Normal to Normal heartbeat interval (NN) and features associated with the number of interval differences of successive NN intervals greater than 50 ms (pNN), and additional features which described the spectral power statistics for very-low, low, and high frequency bands, for which individual sets of statistical features were computed. Six time domain statistical features for each of the three other signals were also computed. Additionally, four features that described the combined statistics for BVP and Respiration taken together were also computed to provide a total of 73 features per feature vector.

There were a total of 406 recordings for the physiological modality, with 205 alert recordings and 201 drowsy recordings, as well as 82 undistracted and 324 distracted recordings, respectively.

4.4 Segmentation

It was observed that the class distribution for the Drowsiness label was balanced, but for the Distraction label was imbalanced by a factor of approximately 1:4, which was the result of how the scenarios were set up. In particular, while there were two sessions resulting in a balanced class distribution for the Drowsiness label, recordings for distracted driving consisted of four sub-recordings, namely physical, cognitive, emotional and frustration, as discussed earlier. Hence, for every one freedriving recording, there were four distracted recordings, causing the imbalance. However, the use of segments meant that the total duration of a recording had more importance than the number of recordings themselves. Here, freedriving recordings would average around 15 minutes per recording, where all the four distractors combined resulted in a cumulative average of around 12 minutes. The exact ratio of the class imbalance would vary in a range of 1.4:1 to 2.2:1 depending on the modality and window size in question.

After the features were extracted for each modality, the data was normalized by dividing each feature vector in a subject's recording by the corresponding afternoon baseline recording, which was selected experimentally. Post-normalization, each recording was split into segments, where a segment size was based on a window of a fixed duration for all modalities. All the samples in each segment were averaged over each feature to generate one feature vector per segment per modality. We chose five window sizes, starting from the smallest window size of 2 seconds, and going upwards in powers of two, for 4, 8, 16, and 32-second windows in total. Testing a wide range of window sizes allowed us to analyze the effect of the window size on the performance. Given that the frame rate across modalities differed, the number of samples used in a given segment would change, but the duration was constant. Accordingly, the samples in each segment were averaged to produce one feature vector per segment.

5 EXPERIMENTAL SETUP

For classification, we used the Extreme Gradient Boosted classifier (XGB) as the sole classifier for all experiments due to its known performance in previous literature [14] and based on preliminary experiments that we performed when preparing the setup. All experimental results were evaluated using Leave-One-Subject-Out Cross Validation (LOSO CV), where each test fold consisted of all recordings of one subject to ensure that the classifier had no prior information about the test subject during training. Two binary classification problems were modeled for our experiments. The first was the Drowsiness label, with 'alertness' and 'drowsiness', and the second was the Distraction label, with 'undistracted' and 'distracted' as their binary classes. The performance metrics used for evaluation were the average accuracy, the average F1-score, the recall of class 0 (alertness/undistracted) and the recall of class 1 (drowsiness/distracted).

For further analysis of the performance of the distraction-based models, we also built separate classification models for each of the four distractors. Accordingly, class 0 referred to the undistracted class, while class 1 referred to a specific distractor. It should be noted that, in this case, there was a strong imbalance in favor of the undistracted class as each individual distractor had only about three minutes of recording time compared to 12 minutes in freedriving.

In addition, early fusion was applied to the different ROIs of the thermal modality as well as to the four channels of the physiological modality. In particular, we used the five thermal regions including the face, forehead, eyes, cheeks, and nose as individual modalities and also combined to form a singular thermal modality. For the physiological modality, the same approach was used, where the BVP, skin temperature, skin conductance and respiration were used as four distinct channels as well as fused to be used as a single modality. The results of early fusion for these modalities are discussed in Section 6.1 in further detail.

Finally, cascaded late fusion was used to allow for a multimodal analysis of the dataset. In general, this approach utilized separate classifiers for each modality, the predictions of which were then used in a specified voting scheme to generate a final prediction, as discussed below. We first carried out LOSO CV to generate predictions for each segment in each modality for the first stage of the fusion. For the second stage, we implemented three voting strategies to analyze their performance. For all experiments in cascaded late fusion, 4-second segment windows were used, owing to their stability in terms of recall for both classes as observed in unimodal classification. The three voting strategies were as follows:

Modality vote with confidence scoring: In this approach, the mean prediction for a given modality averaged the predictions of all the segments pertaining to that recording, giving us a value $\in [0, 1]$ which we call the confidence score. Next, for a given recording, the confidence scores were summed up across the modalities used for classification, and then divided by the number of modalities used to give a final prediction for a recording. Binarizing this final score provided the prediction that is then used against the true label to evaluate performance. Using confidence scores, instead of specific decisions, would prevent a modality that predicted segments without a clear preference to one class from biasing the final vote.

Modality vote with binary scoring: This approach is very similar to the confidence scoring approach, with one exception. The mean prediction using all segments for a modality would not be left as a confidence score, but instead would be binarized at this stage to provide a definite prediction from the modality itself. Binary scoring on a modality level would test a converse hypothesis than that of confidence, where uncertainty from a modality would be eliminated to give a stronger final decision. The idea behind using this strategy is the case where many modalities had a weaker confidence for the correct prediction. Accordingly, they would be able to override a stronger wrong prediction by another modality to give a correct prediction overall.

Segment voting: This approach is significantly different from the previous modality-based voting strategies, as the averaging of segment predictions per modality was not implemented. Here a final prediction for a recording was generated by taking a mean score using all segments present for the recording irrespective of the modality it came from. Binarization of this mean score gave the final prediction to be used for evaluation. Segment voting allows for modalities with more segments to have a stronger influence on the final decision. Given a modality with a greater number of segments, and thus having more information, we would be able to test if this modality would further benefit the final prediction; in such a case, the modality should be prioritized.

It should be noted here that we are testing the above strategies for their performance, not claiming either is the best one for our experiments. The results from each voting strategy are discussed in further detail in Section 6.4 below.

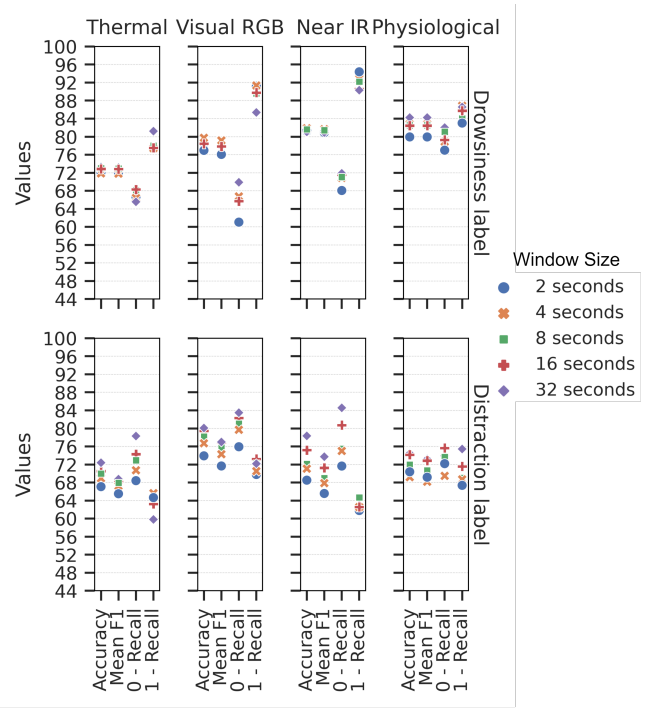

Figure 4: Performance of Modalities in Unimodal Classification

6 EXPERIMENTAL RESULTS

6.1 Unimodal Classification

Starting with unimodal classification across the four modalities as seen in Fig. 4, it can be noticed that all modalities perform better on the Drowsiness label compared to the Distraction label. One possible reason for this could be that using the four different types of distractors combined might not provide consistent patterns for the classifier which leads to a lower overall accuracy.

The physiological modality is the best performer with an F1-score of just above 84% when using a 32-second window for the Drowsiness label. Smaller window sizes, such as four seconds and eight seconds performed similarly at around 83%, indicating that a confident and correct prediction about the driver's state could be provided efficiently using small window sizes, hence increasing the responsiveness of the driver's monitoring system. For the Distraction label, larger window sizes of 32 seconds and 16 seconds are seen to be the better performers throughout the four modalities. The visual modality has the best accuracy, being slightly over 80% when using the 32-second window. However, the Physiological modality has the highest recall for the distraction class, with the 32-second window achieving a recall of approximately 75%.

Even though there is a very slight degradation of F1-scores by approximately one to two percent observed for a 4-second window prediction compared to a 32-second window prediction, their overall performance is quite similar. However, using a smaller window results in a more stable recall performance for both classes.

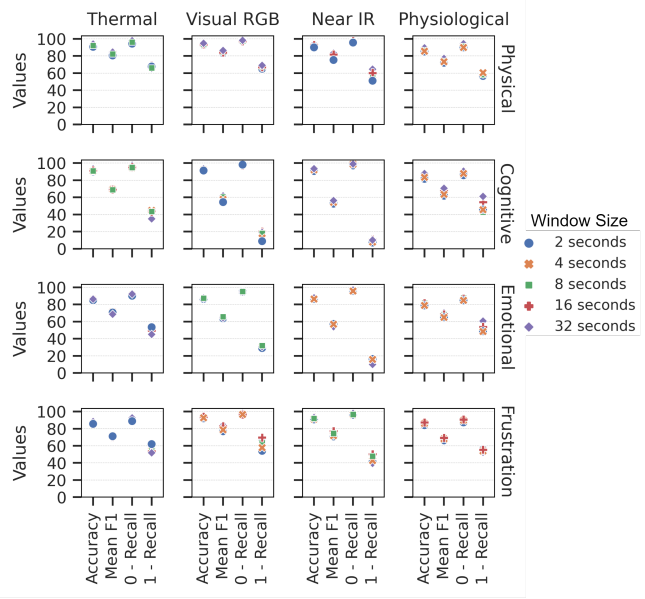

Figure 5: Performance of Modalities in Unimodal Classification against Individual Distractor Classification

For example, for the Physiological modality, the difference in recall between the two classes using a 4-second window is less than

one percent. However, using a 32-second window, the recall has a difference of over four percent.

To further analyze the behavior of different distractors, we evaluated the performance when classifying each distractor separately, as shown in Fig. 5. The lower distractor recall performance is due to the imbalance discussed in Section 5. Compared to merging all the distractors with a single label in Fig. 4, it can be seen that some modalities, such as the visual and NIR perform very poorly on the cognitive and emotional distractors while performing better for the physical and frustration distractors. This seems to be an effect of facial cues and expressions being more prominent for the latter two distractors compared to the former. Furthermore, the thermal and physiological modalities are more consistent for all four distractors, but this does not translate into a better F1-score when combined. These results show that it is easier to detect distraction in general, while recognizing the specific kind of distractor might be more challenging.

6.2 Thermal Sub-modalities

In Fig. 6 it can be noticed that the best two performing regions of the thermal modality are the face and the eyes, with the eyes reaching an F1-score of just under 76% for the Drowsiness label. This compares to the combined performance of 74% when using all thermal ROIs in unimodal classification. This indicates that using all regions might be harming the performance of the thermal modality, with fewer regions resulting in more optimal performance. However, the converse is seen with reference to the Distraction label, where no ROI achieves an accuracy greater than 68%, but when combined, an improvement of up to 72% is seen in Fig. 4 for the thermal modality as a whole. These results suggest that the fusion of the ROIs for the Distraction label allows the classifier to generalize better to the different distractors included during training by learning more information from multiple channels.

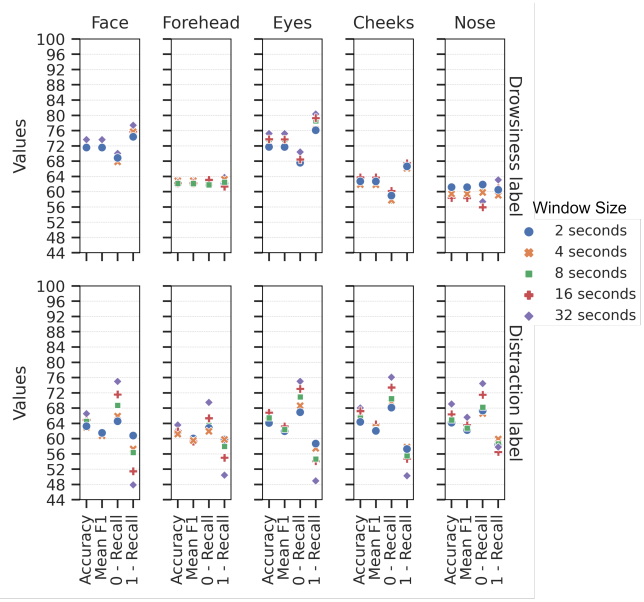

Figure 6: Performance of Regions in the Thermal Modality

Another observation that could be made is regarding the performance optimization of the fusion itself. It was noticed that using three regions, the forehead, eyes and nose outperformed the combined thermal modality (all five ROIs) using the Drowsiness label with an F1-score of just under 76%, and is behind using the Distraction label by less than a percent, with an F1-score of 71%, which is also better than the F1-score of each of the ROIs when used individually. Such results demonstrate the viability of requiring fewer regions to make a correct prediction while requiring lower computational power.

6.3 Physiological Sub-modalities

For different data channels in the physiological modality as seen in Fig 7, we can observe that respiration is by far the strongest channel for modelling the Drowsiness label with an F1-score of 79%. However, using all physiological data channels as a single modality is far more beneficial as shown in Fig. 4, where the physiological modality had a F1-score of 84%, higher than any of the individual channels. Similarly, this behaviour is also seen using the Distraction label, with BVP achieving the best performing accuracy of 69% against the combined modality accuracy of 75% using the 32-second window.

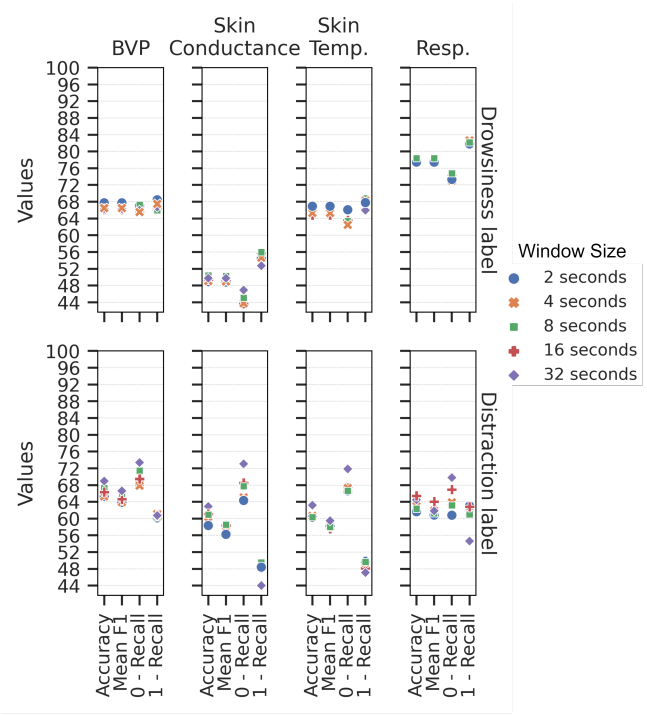

Figure 7: Performance of the physiological Channels.

6.4 Cascaded Late Fusion

Fig. 8 shows the performance of late fusion when applied to two different modality selections using a 4-second window when using different voting strategies, as discussed in Section 5. Here, a modality selection is of two types, the first being '4 Modalities' which

refers to the usage of the thermal, visual, NIR, and physiological modalities, as seen earlier with the unimodal classification. The second selection is '11 Modalities' which refers to the usage of the five individual regions for the thermal modality and the four channels for the physiological modality, separately, along with the visual and NIR modalities for a fusion of 11 modalities in total. A general improvement in performance can be noticed using late fusion, with '4 Modalities' achieving an F1-score of 90% using the 'confidence scoring' fusion approach for the Drowsiness label. This outperforms the 32-second window performance seen in unimodal classification, which was the best performing window size.

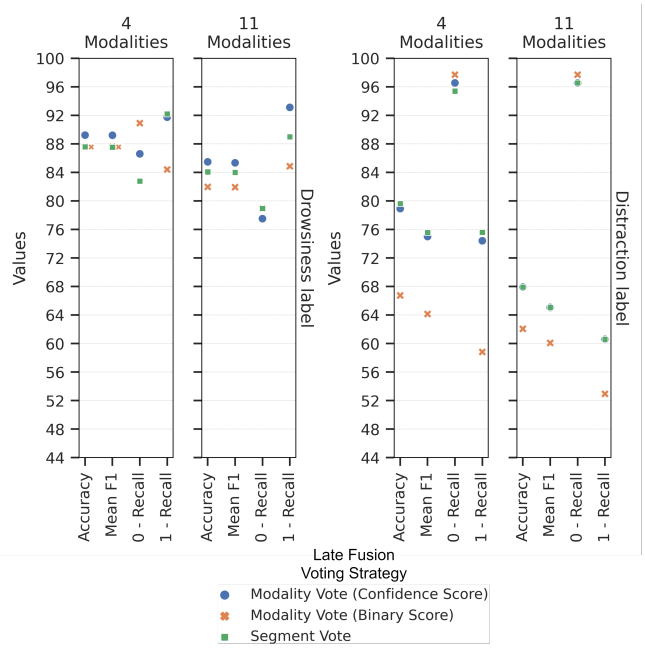

Figure 8: Performance of Cascaded Late Fusion

The confidence score approach is the best performing late fusion strategy amongst the three used for both drowsiness and distraction detection. On the other hand, the hypothesis we proposed for using the binary score voting strategy does not perform as well, as it seems that a strong correct prediction could be overridden by several weaker wrong predictions, which hence skews the final prediction in a detrimental manner. Segment voting suffers from a similar bias, where a modality with a correct prediction could be outvoted by a modality with more segments that provides a wrong prediction. However, it can still be noted that the segment voting scheme achieves similar performance to the confidence score approach using the Distraction label. This is interesting as it could suggest that while the Drowsiness label classification benefits from a confidence-based approach with no need for segments as an additional weight, more segments generated from different types of distractors is supporting distraction detection.

It is also observed that, in general, using individual regions or channels ('11 modalities') did not improve the performance as much as using the combined modality selection ('4 modalities'). While '11 Modalities' achieves accuracy of 86% and 62% for the Drowsiness and Distraction labels, respectively, '4 modalities' provides better performance in both labels with accuracies of 90% and 80%, respectively. The '4 modalities' results outperform any unimodal classification for the Drowsiness label and also provide a matching performance for the Distraction label compared to the 32-second window visual modality, which was the best performer, as seen earlier in Fig. 4. In addition, we still observe that late fusion greatly benefits the recall of the undistracted class, with a highest value of 98% seen for this metric, far higher than any unimodal classification. Furthermore, there is an improvement to the distracted class recall as well, outperforming all other modalities in this regard except physiological, where it matches the performance at 76% recall.

7 CONCLUSION

In this paper we analyzed driver's drowsiness and distraction using a multimodal dataset and a segment-based approach. Each of these behaviors distinctly affect the alertness level of drivers and may lead to severe consequences. We evaluated unimodal performance across all modalities, where the physiological modality provided the best performance of 84% F1-score when using a 32-second window for the Drowsiness label and the visual modality provided the best performance of a 78% F1-score for the Distraction label. We then observed the effects of early fusion in the thermal and physiological modalities using different ROIs and signals. For the thermal modality, we observed that the face and eyes regions were better indicators of drowsiness, with the nose being the best indicator of distractedness. Moreover, we showed that the fusion of certain ROIs, such as the forehead, eyes, and nose allowed for the most optimal balance between fusion and performance.

We also observed that certain physiological signals, such as skin conductance, did not perform well in detecting either labels. The respiration rate was the best physiological indicator for drowsy driving while heart rate was a better indicator of distracted driving, while taking into consideration the imbalanced nature of the data. When observing the performance of the individual and merged distractors, it appeared that the detection of distraction as a general incident during driving was more feasible than detecting a specific distractor at a time. Furthermore, the visual modality was the best modality at detecting physical and frustration distractors, while the physiological modality was better at cognitive and emotional distractor detection. Finally, we presented the results using a cascaded late fusion approach across multiple modalities, testing three voting strategies to see which provided the best performance. The modality-based confidence scoring strategy had the best results, with a 90% F1-score for the Drowsiness label, and a 76% F1-score for the Distraction label when using 4-second windows with four modalities. Overall, we observed that detecting drowsiness was easier than detecting distraction.

ACKNOWLEDGMENTS

This content is based in part on research funded by the Toyota Research Institute ("TRI"). Any opinions, observations, assumptions, or recommendations shared in this content are the authors' only and do not necessarily represent the views of TRI or any other Toyota entity.

REFERENCES

[1] [n.d.]. 100-Car Naturalistic Study Fact Sheet. https://www.csg.org/sslfiles/dockets/2011cycle/31B/31Bbills/100-Car_Fact-Sheet.pdf. Accessed: 2021-05-12.

[2] [n.d.]. Distracted Driving 2018. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812926. Accessed: 2021-05-22.

[3] [n.d.]. NHTSA Drowsy Driving Research and Program Plan. https://www.nhtsa.gov/sites/nhtsa.gov/files/drowsydriving_strategicplan_030316.pdf. Accessed: 2021-04-28.

[4] [n.d.]. Teens and Distracted Driving 2018. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812931. Accessed: 2021-05-22.

[5] [n.d.]. Undercounted is Underinvested: How incomplete crash reports impact efforts to save lives. https://www.nsc.org/getmedia/88c97198-b7f3-4acd-a294-6391e3b8b56c/undercounted-is-underinvested.pdf. Accessed: 2021-05-22.

[6] Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2015. Cascaded Multimodal Analysis of Alertness Related Features for Drivers Safety Applications. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '15)*. Association for Computing Machinery, New York, NY, USA, Article 59, 8 pages. https://doi.org/10.1145/2769493.2769505

[7] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2017. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security* 12, 5 (2017), 1042–1055.

[8] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. 2017. Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498* (2017).

[9] J. Todd Arnedt, Gerald Wilde, Peter Munt, and Alistair Maclean. 2001. How do prolonged wakefulness and alcohol compare in the decrements they produce on a simulated driving task? *Accident; analysis and prevention* 33 (06 2001), 337–44. https://doi.org/10.1016/S0001-4575(00)00047-6

[10] Yusuf Artan, Orhan Bulan, Robert P Loce, and Peter Paul. 2014. Driver cell phone usage detection from HOV/HOT NIR images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 225–230.

[11] Muhammad Awais, Nasreen Badruddin, and Micheal Drieberg. 2017. A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability. *Sensors* 17, 9 (2017), 1991.

[12] Karel A Brookhuis and Dick De Waard. 2010. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention* 42, 3 (2010), 898–903.

[13] Simiao Chen, Michael Kuhn, Klaus Prettner, and David E Bloom. 2019. The global macroeconomic burden of road injuries: estimates and projections for 166 countries. *The Lancet Planetary Health* 3, 9 (2019), e390–e398.

[14] Tianqi Chen and Carlos Guestrin. 2016. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug 2016). https://doi.org/10.1145/2939672.2939785

[15] Anuva Chowdhury, Rajan Shankaran, Manolya Kavakli, and Md Mokammel Haque. 2018. Sensor applications and physiological features in drivers' drowsiness detection: A review. *IEEE Sensors Journal* 18, 8 (2018), 3055–3067.

[16] Yulun Du, Chirag Raman, Alan W. Black, Louis-Philippe Morency, and Maxine Eskénazi. 2018. Multimodal Polynomial Fusion for Detecting Driver Distraction. *CoRR* abs/1810.10565 (2018). arXiv:1810.10565 http://arxiv.org/abs/1810.10565

[17] Yulun Du, Chirag Raman, Alan W Black, Louis-Philippe Morency, and Maxine Eskenazi. 2018. Multimodal Polynomial Fusion for Detecting Driver Distraction. *arXiv preprint arXiv:1810.10565* (2018).

[18] Tiziana D'Orazio, Marco Leo, Cataldo Guaragnella, and Arcangelo Distante. 2007. A visual approach for driver inattention detection. *Pattern recognition* 40, 8 (2007), 2341–2355.

[19] Paul Ekman and Wallace V Friesen. 1978. *Manual for the facial action coding system*. Consulting Psychologists Press.

[20] Centers for Disease Control and Prevention. 2020. *Cost of Injury Data*. https://www.cdc.gov/injury/wisqars/cost/index.html

[21] Ishani Janveja, Akshay Nambi, Shruthi Bannur, Sanchit Gupta, and Venkat Padmanabhan. 2020. InSight: Monitoring the State of the Driver in Low-Light Using Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[22] Michael J Kane, Andrew RA Conway, Timothy K Miura, and Gregory JH Colflesh. 2007. Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, 3 (2007), 615.

[23] Serajeddin Ebrahimian Hadi Kiashari, Ali Nahvi, Hamidreza Bakhoda, Amirhossein Homayounfard, and Masoumeh Tashakori. 2020. Evaluation of driver drowsiness using respiration analysis by thermal imaging on a driving simulator. *Multimedia Tools and Applications* (2020), 1–23.

[24] Neslihan Kose, Okan Kopuklu, Alexander Unnervik, and Gerhard Rigoll. 2019. Real-Time Driver State Monitoring Using a CNN Based Spatio-Temporal Approach. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 3236–3242.

[25] Miguel Bordallo Lopez, Carlos R del Blanco, and Narciso Garcia. 2017. Detecting exercise-induced fatigue using thermal imaging and deep learning. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 1–6.

[26] Anthony D. McDonald, Thomas K. Ferris, and Tyler A. Wiener. 2020. Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures. *Human Factors* 62, 6 (2020), 1019–1035. https://doi.org/10.1177/0018720819856454 PMID: 31237788.

[27] Rizwan Ali Naqvi, Muhammad Arsalan, Ganbayar Batchuluun, Hyo Sik Yoon, and Kang Ryoung Park. 2018. Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor. *Sensors* 18, 2 (2018), 456.

[28] World Health Organization. 2020. *Road traffic injuries*. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[29] Michalis Papakostas, Kapotaksha Das, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2021. Distracted and Drowsy Driving Modeling Using Deep Physiological Representations and Multitask Learning. *Applied Sciences* 11, 1 (2021), 88.

[30] Anna Persson, Hanna Jonasson, Ingemar Fredriksson, Urban Wiklund, and Christer Ahlström. 2019. Heart Rate Variability for Driver Sleepiness Classification in Real Road Driving Conditions. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6537–6540.

[31] Xuli Rao, Feng Lin, Zhide Chen, and Jiaxu Zhao. 2019. Distracted driving recognition method based on deep convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing* (2019), 1–8.

[32] Aashreen Raorane, Hitanshu Rami, and Pratik Kanani. 2020. Driver Alertness System using Deep Learning, MQ3 and Computer Vision. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 406–411.

[33] Bryan Reimer and Bruce Mehler. 2011. The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics* 54, 10 (2011), 932–942.

[34] Kais Riani, Michalis Papakostas, Hussein Kokash, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2020. Towards detecting levels of alertness in drivers using multiple modalities. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 1–9.

[35] Anwesha Sengupta, Anirban Dasgupta, Aritra Chaudhuri, Anjith George, Aurobinda Routray, and Rajlakshmi Guha. 2017. A multimodal system for assessing alertness levels due to cognitive loading. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 7 (2017), 1037–1046.

[36] Jianbo Shi et al. 1994. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 593–600.

[37] Heung-Sub Shin, Sang-Joong Jung, Jong-Jin Kim, and Wan-Young Chung. 2010. Real time car driver's condition monitoring system. In *SENSORS, 2010 IEEE*. IEEE, 951–954.

[38] Sudipta N Sinha, Jan-Michael Frahm, Marc Pollefeys, and Yakup Genc. 2006. GPU-based video feature tracking and matching. In *EDGE, workshop on edge computing using new commodity architectures*, Vol. 278. 4321.

[39] Oxford University Press USA. 2018. *Sleep deprived people more likely to have car crashes*. https://www.sciencedaily.com/releases/2018/09/180918082041.htm

[40] Michel F. Valstar, Timur Almaev, Jeffrey M. Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F. Cohn. 2015. FERA 2015 - second Facial Expression Recognition and Analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 06. 1–8. https://doi.org/10.1109/FG.2015.7284874

[41] R. Verma, B. Mitra, and Sandip Chakraborty. 2019. Avoiding Stress Driving: Online Trip Recommendation from Driving Behavior Prediction. *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom* (2019), 1–10.

[42] K. Wang, Y. L. Murphey, Y. Zhou, X. Hu, and X. Zhang. 2019. Detection of driver stress in real-world driving environment using physiological signals. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, Vol. 1. 1807–1814.

[43] A M Williamson and Anne-Marie Feyer. 2000. Moderate sleep deprivation produces impairments in cognitive and motor performance equivalent to legally prescribed levels of alcohol intoxication. *Occupational and Environmental Medicine* 57, 10 (2000), 649–655. https://doi.org/10.1136/oem.57.10.649 arXiv:https://oem.bmj.com/content/57/10/649.full.pdf

[44] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2519–2528.

[45] Sebastian Zepf, Neska El Haouij, Jinmo Lee, Asma Ghandeharioun, Javier Hernandez, and Rosalind W. Picard. 2020. Studying Personalized Just-in-Time Auditory Breathing Guides and Potential Safety Implications during Simulated Driving *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 275–283. https://doi.org/10.1145/3340631.3394854