Towards Detecting Levels of Alertness in Drivers Using Multiple Modalities

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

Michalis Papakostas
Electrical Engineering
& Computer Science
University of Michigan
mpapakos@umich.edu

Hussein Kokash
Mechanical Engineering
University of Michigan-Flint
hkokash@umich.edu

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

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

Rada Mihalcea
Electrical Engineering
& Computer Science
University of Michigan
mihalcea@umich.edu

ABSTRACT

Distracted and drowsy driving are two very common causes of car accidents as they contribute to 2.3% of all the fatalities caused on the US roads. Therefore, in the era of smart driving there is an increased need of technologies able to monitor driver's alertness and provide timely alerts to the driver. In this paper, we conduct as pilot study and we present a preliminary, yet novel multimodal dataset, collected from 10 subjects using three different modalities. Our modalities include a thermal camera, an RGB camera, and four physiological indicators. The dataset consists of two recording sessions for each subject, thus, offering in total 20 multimodal driving sessions. We propose a machine learning framework aiming to investigate the hypothesis that multimodal features have higher potential towards driver alertness detection. Our dataset and analysis focus on exploring the differences between alertness and drowsiness as they intersect with the presence of different distractions. The results highlight the validity of our hypothesis and introduce interesting future directions for research.

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, distracted driving, action units, classification, machine learning , thermal imaging

PETRA '20, June 30-July 3, 2020, Corfu, Greece

ACM ISBN 978-1-4503-7773-7/20/06…$15.00
https://doi.org/10.1145/3389189.3389192

ACM Reference Format:
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 *The 13th PErvasive Technologies Related to Assistive Environments Conference (PETRA '20), June 30-July 3, 2020, Corfu, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3389189.3389192

1 INTRODUCTION

Sleeping at the wheel is one of the main causes of car accidents worldwide. According to a AAA Traffic Safety Foundation study, 37% of drivers reported having slept behind the wheel at least once at a certain point in their lives. Furthermore, an estimated 21% of fatal accidents, 13% of accidents resulting in severe injury, and 6% of all accidents involve a drowsy driver [36].

As reported by the National Highway Traffic Safety Administration (NHTSA), in 2015 there was an increase of 7% of fatal crashes (2346 fatal crashes) in the US compared to the year before [13]. In addition, there were 824 fatalities (2.3 percent of all fatalities) attributed to driving while being drowsy or sleepy. NHTSA found that this types of accidents often involve a single vehicle, with no passengers apart from the driver, speeding off the road with no braking evidence. That can be attributed to the fact that drowsiness slows reaction time and significantly impacts and delays the decision of braking. It was also estimated that about 100,000 reported accidents involve a low level of alertness of the driver. These crashes, do not only cause financial loss, but also severe physical damages, including approximately 1,550 deaths, 71,000 injuries and $12.5 billion in financial damage.

Driving while being drowsy not only affects the driver and passengers, but also all road users such as pedestrians, cyclists and motorists. Exhausted driving has become a normality due to increased workload, constant pressure, over-exertion, and lack of sleep. Despite the presence of research on the factors that affect the alertness of drivers, there is still no mean to measure or classify them precisely and reliably. In order to reduce the risks associated with drowsy driving, more extensive research is required to detect and understand the various driver states.

Nonetheless, a report by the National Highway Traffic Safety Administration in 2012 indicated a reduction in the number of fatal

accidents between 1995 and 2012 as automotive technology and safety features were gradually increased [19]. The aforementioned findings, observations and needs have motivated an increasingly evolving field of research that focuses on monitoring and preventing drowsy and distracted driving.

For the purposes of this work, we have collected a multimodal dataset, targeting driver's alertness detection and we provide an analysis of three modalities regarding their ability to distinguish between alert, drowsy and distracted driving. The dataset is composed of physiological, visual and thermal modalities collected using a driving simulator for 10 subjects at different times of the day. The main purpose of this research is to determine which modalities have higher discriminative capability in measuring the alertness level of the drivers and whether the integration of multimodal features can induce further improvements. Moreover, we research how different facial features and action units [12] may be associated with alert, drowsy and distracted driving behaviors. Finally, unlike previous studies, this work expands the binary classification of drivers from "Alert" vs "Not Alert" into three and four different categories in an effort to understand how various alertness levels may correlate to distracted driving.

This paper is organized as follows. Section 2 overviews previous work that was proposed to detect driver's alertness. Section 3 details the preliminary dataset used in our experiments. Section 4 introduces our proposed approach to create a system that can discriminate between different behavioral states of drivers. Our experimental results are presented in Section 5. Section 6 discusses how the proposed framework could be utilised in a real-life scenario and finally, conclusions and guidelines for future work are provided in Section 7.

2 RELATED WORK

Several approaches have been developed to track driver's alertness using visual, physiological, sensorial, behavioral, and environmental information [1, 17, 24].

In earlier studies, Palvidis et al.[31] collected data from 6 subjects to detect facial patterns of anxiety, alertness, and/or fearfulness on different scenarios using a computer vision approach. The same number of subjects was used by Lin et al. [25] to predict driver's drowsiness based on EEG signals. For their experiments, the authors used a VR-based driving simulator where subjects had to perform long-term monotonous driving on a straight highway. More recently, using a similar setup on a different task, Papakostas et al., exploited facial features and EEG to detect signs of cognitive fatigue on a group of 19 participants [30]. Their results highlighted the advantage of multimodal machine learning towards detecting the short-term deterioration of cognitive performance.

Craye et al. [10] attempted to detect drowsiness by collecting data from 12 participants. In addition to physiological and car input signals, they used an RGB camera and an infrared camera to monitor the driver in a driving simulation framework.

Kiashari et al. [21] showed that the respiratory state of a person can be tracked without interference using thermal imaging and that the observed fluctuations may be highly correlated to wakefulness and drowsiness. Thermal imaging-based respiration monitoring was accomplished by observing the difference in temperature between the air entering and leaving the respiration system. A respiration signal was constructed by localizing the nostrils' zone, which was followed by putting together the mean temperature of the nostril region in all of the frames.

By utilizing the visual channel carrying facial expressions and the auditory channel carrying voice intonations, Cowie et al. [9] developed a hybrid framework towards recognizing the affective state of individuals. Kolli et al. [22] were able to detect driver's emotions using an infrared thermal camera, while Lopez et al. [26] used a thermal camera to identify fatigued individuals by applying an SVM classifier on the feature vectors extracted from two convolutional neural networks .

Other studies have investigated automatic drowsiness and distraction recognition during driving using different types of visual features. Jie et al., [20] investigated yawning behavior in simulated driving scenarios as a sign of sleepiness, by monitoring spontaneous gestures of drowsy drivers. Furthermore, Du et al. [11] extracted facial features using the OpenFace library [5] and combined them with speech features and car driving measures to detect distracted behavior in drivers. In most related work that was based on multimodal feature analysis, concatenating the different features in an early stage has proven to be one of the most prominent fusion methods [1, 10, 23]. Overall, the majority of the proposed work used environmental, behavioral, physiological and visual modalities while our multimodal approach is solely based on diverse and implicit, human-generated signals.

3 DATASET

For our experiments, we gathered a multimodal dataset consisting of thermal, audiovisual and physiological recordings from 10 subjects, 7 males and 3 females, during two recording sessions per subject while using a driving simulator. The first recording session took place in the morning assuming that the subjects were more alert compared to a later point in the day, after a long day at work or after attending a series of classes at the University. After a typical working day, the second session was recorded during the evening to simulate drivers' drowsiness.

There were three different types of recordings for each session. The first recording was a normalization baseline, where the subjects were seated without any activity.

The second recording consisted of a total of twenty minutes of free driving. For this step, only the last three minutes of driving were recorded to make sure the driver was deeply involved in driving and had familiarized him/herself with the simulator.

During a ten-minute interval, the last recording was captured for four minutes where three different types of distractors were introduced to the subjects. First, a cognitive distractor was introduced called the N-Back task [29], where the subject had to listen to a sequence of letters and loudly pronounce any letters repeated after a sequence of two other letters (ie. N=2). Second, an emotional distractor was applied, where the subject shared his personal experience on an emotional topic of his choice. Finally, a physical distractor, where the subject was asked to search for different addresses using the GPS on his phone while driving.

The same set of recordings were repeated for the evening session. For the thermal recordings, the FLIR SC6700 high-resolution thermal camera was used to detect and track the driver's thermal patterns in five different areas of the face including the entire face, forehead, eyes nose and cheeks. The camera is a state-of-the-art scientific-grade thermal camera. The resolution of the FLIR SC6700 is 640x512 and 7.2 M electrons, achieving a frame rate of approximately 100 frames per second.

The visual recordings consist of the subject's frontal view focusing on the face area, which is recorded using a Raspberry Pi camera.

The physiological recordings contain measurements of the subject's skin conductance, blood volume pulse, respiration rate and skin temperature from four physiological biosensors attached to the subject's hands and thoracic area. For the physiological recordings we used the "ProComp infinity" biosensor device which has been used in several studies in the past [3]

4 METHODOLOGY

4.1 Feature Extraction

We extracted features from the individual modalities to create our multimodal model using two approaches, early fusion and late fusion. In the early fusion approach, we concatenated the features before classification is performed. For the late fusion approach, the final prediction was performed using majority voting of the individual decisions from the three modalities.

4.1.1 ***RGB video:*** The OpenFace library was used to extract visual features describing the facial behavior of the subjects [5] from videos captured using the Raspberry Pi device. For the detection of facial landmarks, we deployed a Constrained Local Neural Field algorithm (CLNF) [5]. CLNF was preferred over a Constrained Local Model (CLM) [6] approach for its improved results. CLM struggled to perform in poor lighting and was significantly affected by occlusions thus, introducing a lot of noise in the facial-landmark detection task.

The CLNF algorithm involves a Local Neural Field (LNF) patch expert, which learns about both adjacent and long distance pixels by gaining information on similarity and sparsity constraints over long distances. This provides the local variation of each landmark's appearance. The second main component for facial landmark detection is Point Distribution Model, which captures variation in the shape of facial landmarks.

The CLNF model is initialised using the facial landmarks detected in past frames, while processing the videos. This allows 68 facial landmarks to be detected at every frame [5]. The 3D facial landmarks detected at each step, are then projected using an orthographic camera projection, to detect the pose of the head[15].

For eye gaze detection, the system first detects landmarks in the image region associated with the eyes and then estimates the position of the pupil based on the intersection of a ray passing through the pupil and the eyeball sphere. The outcome of the eye gaze detection is a feature vector for each eye describing the position of the pupil. Figure 1 presents gaze and head pose estimations and facial landmark detection on video frames from our dataset; green lines represent the estimated eye gaze vectors, 3D representation shows the head pose estimation.

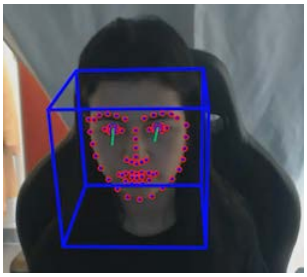

Figure 1: Gaze and head pose estimations and facial landmark detection on video sequences

For extracting facial appearance features, OpenFace uses similarity measures transformed from the presently noticed facial landmarks to a neutral expression frontal landmark representation. The Histogram of Oriented Gradients (HOG) is extracted from the aligned face producing a high-dimensional vector of 4464 features. In order to reduce the dimensionality, Principal Component Analysis (PCA) is applied [38]. CLNF's facial shape features and reduced dimensionality HOG features are then used for predicting Action Units (AUs) and measuring the AU intensity. An AU describes facial deformation due to each facial muscle movement [12]. OpenFace is able to identify 18 different AUs and provides a metric indicating the intensity or the presence of each of these AUs. The presence is recorded as 0 (absent) or 1 (present) respectively and the intensity ranges from 0 to 5, where 0 means the AU is not present, 1 is the lowest intensity level and 5 represents the maximum intensity. Table 1 summarizes the details of each of the 18 AUs.

Table 1: List of AUs in OpenFace. I corresponds to "Intensity" and P to "Presence".

AU	Full name	Prediction
AU1	Inner brow raiser	I
AU2	Outer brow raiser	I
AU4	Brow lowerer	I
AU5	Upper lid raiser	I
AU6	Cheek raiser	I
AU7	Lid tightener	P
AU9	Nose wrinkler	I
AU10	Upper lip raiser	I
AU12	Lip corner puller	I
AU14	Dimpler	I
AU15	Lip corner depressor	I
AU17	Chin raiser	I
AU20	Lip stretched	I
AU23	Lip tightener	P
AU25	Lips part	I
AU26	Jaw drop	I
AU28	Lip suck	P
AU45	Blink	P

OpenFace framework is implemented using state-of-the-art methods for AU recognition [4, 37], and it is tailored to be easily applicable on natural videos sequences from unseen datasets [6]. Similar approaches that deployed more complex algorithms such as deep learning or physical geometry-based features [16, 37] were outperformed by the OpenFace approach when tested on the SEMAINE dataset [27], which shares some similarities to our data. Therefore, we chose OpenFace for providing a more reliable and suitable solution to our problem.

4.1.2 ***Thermal Imaging:*** For analysing the thermal video, we followed three main processing steps. The first step was to segment the faces of the participants into five different regions, including the entire face, forehead, eyes, cheeks, and nose. Thereafter, our tracking algorithm proposed in [7], was applied to track these regions throughout the recording. Finally, by creating thermal maps for all regions of interest, we generated the final thermal feature vectors.

In more details, the first step was to manually segment the Regions Of Interest (ROIs) from the first frame of each of the video recordings, as automatic methods of face detection did not show acceptable performance on thermal images. Then, to capture points of interest in the detected ROIs, a variation of the Shi-Tomasi corner detection algorithm [34] was applied by computing the weighted square difference between two successive frames.

As the method compares an image patch $I_1(x_i)$ with a shifted version of the image, $I_1(x_i + \Delta u)$, an auto-correlation function S was used.

$$S(\Delta u) = \sum_i w(x_i)(I_0(x_i + \Delta u) - I_0(x_i))^2 \quad (1)$$

where u is the displacement vector and $w(x_i)$ is a window function. The function is approximated using Taylor Series expansion into

$$S(\Delta u) \approx \sum_i w(x_i)(\nabla I_0(x_i).\Delta u)^2 \quad (2)$$

where,

$$\nabla I_0(x_i) = (\frac{\partial I_0}{\partial x}, \frac{\partial I_0}{\partial y})(x_i) \quad (3)$$

We used a fixed-size Gaussian filter to smooth the calculated gradient. Thus, S can be rewritten as:

$$S(\Delta u) = \Delta u^T V \Delta u \quad (4)$$

where V denotes the auto-correlation matrix. The interesting corner points to be tracked were located using the variation in S by computing the minimum eigenvalues from V.

The interesting points were usually located where sharper changes in the colors existed indicating the possibility of the presence of a blood vein controlling the temperature of the surrounding region. Figure 2 shows the points of interest detected in the face region with a relatively lower threshold, allowing more points to be detected.

To stabilize the ROI bounding box for the duration of the videos we tracked them by applying a fast version of the tracking method KLT [35], which is a tracking method that provides very accurate results when the tracked objects maintain their shape over time.

The algorithm estimates the relocation of points of interest between two successive frames by assuming a slight displacement between the pixels in a frame at times t and t + τ, which was very

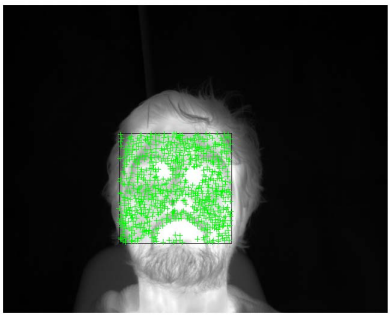

Figure 2: Points of interest detected in the face region

suitable to our tracking needs. Following the tracking process and the displacement estimation, a geometric transformation was applied to map the interesting points between the frames. This latter transformation globally estimated the transformation of interesting points based on similarity . We set a threshold of 95% of successfully mapped points between two successive frames as a precaution to account for potential occlusion such as not having the subject face in the frame or getting it partially. In case of occlusion, the current frame will be skipped, and the tracking will resume.

Lastly, in order to extract potentially indicative thermal features of drowsy and distracted behavior residing in our five regions of interest, we created a thermal map outlining the thermal patterns in the regions of interest. For this purpose the following steps had to be implemented in that order: a) ROI segmentation, b)segment binarization, c)image masking and d) thermal map cropping for each ROI. This process is demonstrated in Figure 3.

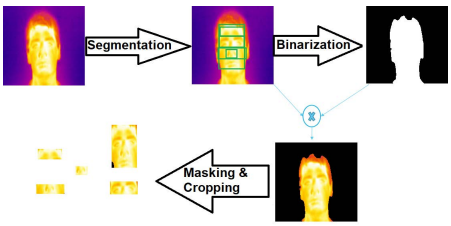

Figure 3: The process of segmenting, binarizing, masking and cropping the thermal faces

The pixel values represented temperatures in Fahrenheit. The map was formed by extracting the mean of the pixel values within a ROI, the minimum temperature, the maximum pixel value representing the highest temperature, the difference between the maximum and minimum temperatures, the mean of the 10% highest pixel values representing the mean of the 10% highest temperatures, and a histogram over the values of the pixels in a ROI, which corresponds to the temperature distribution in that region.

4.1.3 ***Physiological Sensors:*** Four physiological sensors were used to obtain the raw physiological measurements as well as statistical features derived from them over time. The extracted features

include the maximum and minimum values, means, power means, standard deviations and mean amplitudes, among others.

The first sensor is a Blood Volume Pulse (BVP) sensor or photoplethysmograph that used infrared light against the subject's skin surface to detect blood variation in the skin by measuring the amount of reflected red light. That is due to the fact that blood tends to absorb most colors other than those in the spectrum of red.

The second sensor is a Skin Conductance (SC) measuring module. Skin Conductance is an index of the Sympathetic Nervous System (SNS)'s activation and emotional arousal. A small electrical potential is introduced in order to assess SC between two electrodes that are strapped or attached to the fingers and measure the amount of current between the electrodes.

The third module captured Skin Temperature (ST) and is a temperature measuring sensor attached to the small finger.

Finally a Respiration Rate (RR) sensor was used to extract abdominal respiration features. RR is a very motion sensitive device, capturing the stretches that occur during respiration as the system is wrapped around the thoracic region.

The final feature set consisted of a total of 77 physiological features including 50 BVP features, 7 SC features, 9 RR features, 7 ST features, and 4 features extracted from the BVP and the RR sensors combined, such as the mean and heart rate max-min difference, which is a measure of breath to heart rate variability. After calculating the maximum, mean, minimum and standard deviations to obtain four different vectors, we concatenated these new measurements into a new vector of 308 measurements.

4.2 Classification

Motivated by previous work that highlighted the potentials of Decision Tree classifiers (DT) on producing promising results with similar types of data [2, 32, 33], we used the same algorithm to estimate a benchmark performance on our dataset.

We performed a 10 leave-one-out cross validation scheme on the features extracted from the 39 recordings for the three modalities since we discarded one recording for low data quality. The final result is the average of those ten runs.

We did explore the use of additional classifiers such as SVM, Nearest Neighbor, and Naive Bayes, however, DT showed steady improvement in the classification performance over all other classifiers, as well as better model interpretability.

As already discussed, for the purposes of this paper we experimented with 3 different classification approaches:

- **Binary**:
  (1) "Drowsy": drowsiness without distraction + drowsiness with distraction
  (2) "Alert": alertness without distraction + alertness with distraction
- **3 Classes**:
  (1) "Drowsy": drowsiness without distraction
  (2) "Alert": alertness without distraction
  (3) "Distracted": drowsiness with distraction + alertness with distraction
- **4 Classes**:
  (1) "Drowsy": drowsiness without distraction
  (2) "Alert": alertness without distraction
  (3) "Drowsy Distracted": drowsiness with distraction
  (4) "Alert Distracted": alertness with distraction

4.2.1 ***Multimodal Classification:*** Multimodal classification was performed by integrating the features from the different modalities using two approaches. Firstly we performed an early modality fusion by concatenating the features obtained from the thermal, visual and physiological streams, to create a single feature vector, which was then used for classification.

Secondly a late fusion approach was followed, where the final overall accuracy was determined by using the majority vote of the three decisions derived from each modalities. The fused decision is computed as:

$$F(x) = arg\max_{y} \sum_{i=1}^{N} f_i(x) \tag{5}$$

where N is the number of modalities; N=3 in our experiments.

5 RESULTS

For our evaluation we report the average overall accuracy, and per class recall and f1 score. We argue that monitoring recall will provide a more trustworthy evaluation of our system compared to just accuracy. That is due to the fact that the goal of this task is to maximize the detection of positive samples. Observing f1 helps to keep track of false-positive rate so that the model is not overfitted. As it can be observed in the results discussed below, in most cases recall and f1 values were close indicating the stable behavior of the final models. For our evaluation, we run a 10 leave-one-out cross validation scheme and we show the averaged results on all three metrics.

Thermal features were created using different histogram sizes: 20, 60, 120, 180 and 255 to detect driver's drowsiness. The features were extracted as detailed in Section-4.1.2 from five different regions of the face. We normalized our features vectors by dividing each vector with the mean of the normalization vector calculated from the baseline recording, in which the subjects were recorded without activity. The aim of this kind of normalization is to produce features that represent the personalized behavioral changes given the subject's baseline recordings and avoid biases introduced by the variations across different subjects.

Based on our preliminary experiments, the 20-bin histogram showed the best performance compared to the other evaluated histogram sizes. Each region had a set of 25 features resulting in a total of 125 combined thermal features from the five regions.

Table 2 illustrates the overall accuracy, recall and f1 score on the different classification schemes using only the thermal modality. As it can be observed from the table, as the number of classes increases the discrimination ability of the thermal features significantly drops. However, it is worth noting that in the binary classification task, the thermal modality seems to be quite effective on distinguishing between alertness and drowsiness giving an improvement of approximately ~20% over the random choice baseline in terms of accuracy. An observation that is also reflected by the improved performance in terms of recall and f1. However in both multicalss problems, the model's accuracy significantly drops and is comparable or slightly worse than the baselines. An explanation for this

behavior can be attributed to the small number of available samples, especially for multiclass classification, which makes it more difficult to generalize well.

Table 2: Thermal Features Classification Results

4 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	25.64	23.32	16.00	16.08
	Alert Distracted	25.64		32.00	27.94
	Drowsy	25.64		24.00	23.79
	Drowsy Distracted	23.07		21.10	24.33
3 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	25.64	30.25	18.00	18.82
	Drowsy	25.64		31.00	31.78
	Distracted	48.71		36.31	34.70
Binary					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	51.29	71.02	68.50	70.72
	Drowsy	48.71		73.68	71.19

Table 3 shows the classification results using the RGB video from the raspberry pi camera. The results indicate that RGB features can consistently offer improved performance over the baseline in all classification scenarios. As expected, similar to the case of the thermal features, a larger number of classes results in lower performance and imbalanced results. However, in contrast to the thermal modality, the improvements observed here over the baselines are more profound in all cases. Interestingly, recall shows a general improvement when measured on distraction classes ("Alert Distracted", "Drowsy Distracted" and "Distracted"), which may indicate that distractions can be detected more effectively by the RGB visual features than features extracted from the Thermal domain.

Table 3: Visual Features Classification Results

4 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	25.64	31.53	27.00	26.32
	Alert Distracted	25.64		29.00	32.19
	Drowsy	25.64		23.00	22.39
	Drowsy Distracted	23.07		48.88	45.33
3 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	25.64	55.63	29.00	28.81
	Drowsy	25.64		28.00	26.96
	Distracted	48.71		84.21	84.90
Binary					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	51.29	83.84	86.00	84.49
	Drowsy	48.71		81.57	83.09

Table 4 shows the classification results using only the physiological modalities. The results highlight the superior performance of the physiological-based model to recognise states related to the "Alert" class, in both multiclass problems.

It is interesting to note that for binary classification, compared to other modalities and in contrast to the multiclass schemes, the physiological features exhibit lower performance. This observation comes to confirm our hypothesis that a multimodal approach is highly suited for the purposes of our task.

Table 4: Physiological Features Classification Results

4 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	25.64	37.94	61.00	64.07
	Alert Distracted	25.64		28.00	29.11
	Drowsy	25.64		31.00	27.88
	Drowsy Distracted	23.07		31.10	30.52
3 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	25.64	38.97	54.00	55.07
	Drowsy	25.64		20.00	17.91
	Distracted	48.71		41.05	43.02
Binary					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
DT	Alert	51.29	61.02	60.00	61.19
	Drowsy	48.71		62.10	60.76

Table 5: Multimodal Classification Results from combining thermal, physiological and raspberry pi camera features

4 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
Early Fusion	Alert	25.64	55.12	73.00	78.72
	Alert Distracted	25.64		32.00	31.34
	Drowsy	25.64		72.00	66.45
	Drowsy Distracted	23.07		42.22	42.27
Late fusion	Alert	25.64	35.12	60.00	41.69
	Alert Distracted	25.64		33.00	33.95
	Drowsy	25.64		16.00	17.99
	Drowsy Distracted	23.07		31.10	44.33
3 Classes					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
Early Fusion	Alert	25.64	79.73	69.00	76.06
	Drowsy	25.64		77.00	69.47
	Distracted	48.71		86.84	87.30
Late fusion	Alert	25.64	46.14	53.00	43.52
	Drowsy	25.64		16.00	21.67
	Distracted	48.71		58.41	57.04
Binary					
Classifier	Class	Baseline	Accuracy	Recall	f1 score
Early Fusion	Alert	51.29	64.09	59.5	62.72
	Drowsy	48.71		68.94	64.83
Late fusion	Alert	51.29	82.04	79.92	81.72
	Drowsy	48.71		84.36	82.32

5.1 Integrated Modality

Table 5 summarizes the performance of the two multimodal approaches with early and late fusion, as discussed in Section 4.2.1. For our multimodal experiments, we combined all thermal, physiological and visual features and evaluated using the same three classification schemes.

Compared to using the features of individual modalities, the multimodal representations show a steady improvement for multiclass classification. For the three-class classification scheme when using early fusion, the accuracy exceeds 79% and for the four-class exceeds 55%. Moreover, the per-class recall exceeds the corresponding baseline in both multiclass approaches when using early fusion.

Nevertheless, the accuracy of the late fusion method did not show comparable results to early fusion.

What is most interesting is that despite the fact that in both multiclass problems early modality fusion outperforms all other approaches, in the binary class problem visual models dominated the task. This might indicate a strong correlation of visual features with the states described by the "Alert" and "Drowsy" classes but also their inability to perform equally well when distractions are targeted as explicit states.

Lastly, there are few cases where low recall values are observed. Fluctuations in recall across the different experiments are even more noticeable when the thermal or the RGB data were exclusively used. Physiological data seem to provide slightly more stable results and probably have a significant positive impact on our multimodal experiments as well. A possible explanation about this observation could be the limited amount of available data, which fail to represent adequately the targeted classes, especially in the 3-class and 4-class problems. We plan to further investigate these observations in the future by expanding our dataset with more participants.

Table 6: Summary of the results

	Modality				
Classification	Thermal	Visual	Physiological	Early Fusion	Late Fusion
Binary	71.02	**83.84**	61.02	64.09	82.04
3-Class	30.25	55.63	38.97	**79.73**	46.14
4-Class	23.32	31.53	37.94	**55.12**	35.12

A summary of the classification results is presented in Table 6. The visual modality outperformed all other modalities with an accuracy of 83.84% using a DT classifier in the binary case of "Alert" and "Drowsy" classes. The visual output in the multiclass classification, however, was outperformed by the early fusion method which achieved an accuracy of 79.73% in the classification of the three classes "Alert", "Distracted" and "Drowsy." In addition, using early fusion, the accuracy increased to 55.12% compared to 37.94% obtained by the best individual modality in the classification of the four classes.

5.2 Action Unit Analysis

Figure 4 provides a more in depth analysis on how the different AUs are associated with drowsiness and alertness. To find patterns correlated with alertness vs. drowsiness, the graph bars are calculated by subtracting from the alertness values the average of the drowsiness AU feature values.

A positive result therefore specifies an association between an AU and alertness, while a negative result indicates an association between an AU and drowsiness. The resulting figure gives insightful observations. Alertness is highly associated with the Lip Stretched and Lips Part. Furthermore, for the majority of subjects, Inner brow raiser and Outer brow raiser are strongly associated with alertness. In addition, it is interesting how higher blinking rates have a higher alertness association rather than drowsiness. All the above are important observations that we plan to investigate deeper in the near future.

6 DISCUSSION

Applying such a multimodal system in real life could be proven very beneficial towards increasing road safety. However, a future utilization of our approach would demand several modifications in terms of hardware so that the different sensors will not interfere with the driving process.

First and foremost,the wired, body-placed physiological sensors should be replaced with alternative devices that would be attached either on different parts of the car or would be wireless, wearable devices worn by the drivers. Sensors like the blood volume pressure, body temperature and skin conductance could be potentially placed on the steering wheel, while breathing rate sensors could be located on the safety belt of the driver. Works like the ones presented by Choi et al. [8] and Muhlbacher et al. [28] are excellent paradigms of such methods. Other off-the-shelf devices like Empatica, which was designed by the team of Rosalind Picard at MIT could also play an important role towards that end [14] . Another alternative would be to extract the physiological features from the thermal data as discussed by Hessler et al. [18].

Lastly, despite the high quality offered by the FLIR SC6700 thermal camera, the expense of such a device makes it impossible to place it in the a real vehicle. Hence, in our future work we plan to use a low-resolution thermal camera to compare the trade-of between performance and resolution when using the different thermal sensors.

7 CONCLUSIONS & FUTURE WORK

In this study, we introduced a preliminary novel multimodal dataset for driver's alertness detection and we preformed a pilot study on discriminating between drowsiness, alertness, and distraction. To our knowledge, this is the first approach to monitor the three driver's states combined using physiological, visual, and thermal modalities. In addition to the dataset, this research addressed three main objectives. First, we investigated the advantages of thermal features towards monitoring alertness levels as it is a highly underresearched modality. Second, our work extended the standard binary classification problem into a three- and four-classes problem to detect different levels of driver's alertness. Last, we studied which modalities have higher discrimination ability towards classifying alertness in drivers.

Our experimental results highlighted the advantages offered by multimodal feature learning, showing significant improvement over all individual modalities for both multiclass classification schemes. Early modality fusion lead to improved performance compared to individual modalities for multiclass classification with an overall accuracy of 79.73% for the three-class scheme and 55.12% for the four-class approach. Similar behaviors were observed by the recall per class as well.

On the other hand, this did not hold true in the binary classification approach where visual features showed better discrimination ability between alert and drowsy states. Thermal modalities showed promising performance on the binary task as well but failed to represent either of the multiclass problems. Physiological data at last, showed high correlation to the state of alertness when evaluated in the multiclass classification schemes which was contradictory to the performance of the other individual modalities under the same

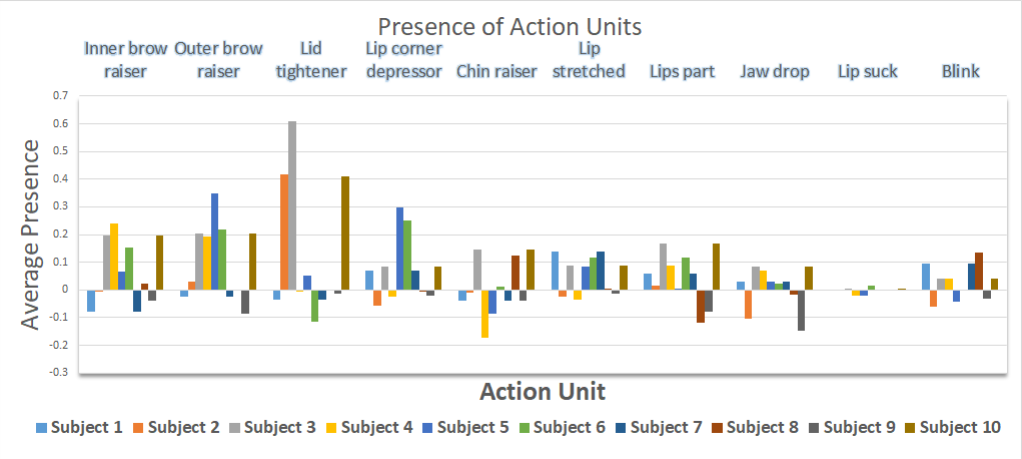

Figure 4: Action Units

evaluation conditions. In addition, we provided an analysis of facial action units to discover interesting behavioral correlations between facial expressions and the states of drowsiness and alertness.

The promising results and the insightful findings extracted from this analysis will be valuable towards the future directions of our research in the topic. In the next steps, we plan to significantly expand our dataset and investigate in further depth the scalability of our current findings. Moreover, we plan to explore more sophisticated methods towards beating the benchmark set in this paper. Lastly our research will be targeted not only towards identifying universal patterns of the different driver's states but also towards detecting personalized features that are associated with alertness, drowsiness and distraction.

Overall, this work indicates that developing a multimodal driver's alertness system can aid in improving the quality of driving and road safety thus, being potentially very useful towards reducing the number of related accidents.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the Toyota Research Institute ("TRI"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of TRI or any other Toyota entity.

REFERENCES

[1] 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*. ACM, 59.

[2] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2016. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security* 12, 5 (2016), 1042–1055.

[3] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2017. Multimodal gender detection. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 302–311.

[4] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–6.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[6] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[7] Mihai Burzo, Mohamed Abouelenien, David Van Alstine, and Kristen Rusinek. 2017. Thermal discomfort detection using thermal imaging. In *ASME 2017 International Mechanical Engineering Congress and Exposition*. American Society of Mechanical Engineers Digital Collection.

[8] YouJun Choi, HeeSung Shin, and JaeYeol Lee. 2014. Smart steering wheel system for driver's emergency situation using physiological sensors and smart phone. In *2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings*. IEEE, 281–286.

[9] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* 18, 1 (2001), 32–80.

[10] Céline Craye, Abdullah Rashwan, Mohamed S Kamel, and Fakhri Karray. 2016. A multi-modal driver fatigue and distraction assessment system. *International Journal of Intelligent Transportation Systems Research* 14, 3 (2016), 173–194.

[11] 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).

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

[13] National Center for Statistics and Analysis. (2017, October). Drowsy Driving 2015 (Crash•Stats Brief Statistical Summary. Report No. DOT HS 812 446). Washington, DC: National Highway Traffic Safety Administration.

[14] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. 2014. Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. IEEE, 39–42.

[15] Christopher Geyer and Konstantinos Daniilidis. 1999. Catadioptric camera calibration. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 1. IEEE, 398–404.

[16] Amogh Gudi, H Emrah Tasli, Tim M Den Uyl, and Andreas Maroulis. 2015. Deep learning based facs action unit occurrence and intensity estimation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–5.

[17] Amjad Hashemi, Valiallah Saba, and Seyed Navid Resalat. 2014. Real time driver's drowsiness detection by processing the EEG signals stimulated with external flickering light. *Basic and clinical neuroscience* 5, 1 (2014), 22.

[18] Christian Hessler, Mohamed Abouelenien, and Mihai Burzo. 2018. A Survey on Extracting Physiological Measurements from Thermal Images. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. 229–236.

[19] National highway traffic safety administration. 2012. national statistics Available from. http://www-fars.nhtsa.dot.gov/Main/index.aspx

[20] Zhuoni Jie, Marwa Mahmoud, Quentin Stafford-Fraser, Peter Robinson, Eduardo Dias, and Lee Skrypchuk. 2018. Analysis of yawning behaviour in spontaneous expressions of drowsy drivers. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 571–576.

[21] Serajeddin Ebrahimian Hadi Kiashari, Ali Nahvi, Amirhossein Homayounfard, and Hamidreza Bakhoda. 2018. Monitoring the Variation in Driver Respiration Rate from Wakefulness to Drowsiness: A Non-Intrusive Method for Drowsiness Detection Using Thermal Imaging. *Journal of Sleep Sciences* 3, 1-2 (2018), 1–9.

[22] Abhiram Kolli, Alireza Fasih, Fadi Al Machot, and Kyandoghere Kyamakya. 2011. Non-intrusive car driver's emotion recognition using thermal camera. In *Proceedings of the Joint INDS'11 & ISTET'11*. IEEE, 1–5.

[23] Nanxiang Li and Carlos Busso. 2014. Predicting perceived visual and cognitive distractions of drivers with multimodal features. *IEEE Transactions on Intelligent Transportation Systems* 16, 1 (2014), 51–65.

[24] Chin-Teng Lin, Yu-Chieh Chen, Teng-Yi Huang, Tien-Ting Chiu, Li-Wei Ko, Sheng-Fu Liang, Hung-Yi Hsieh, Shang-Hwa Hsu, and Jeng-Ren Duann. 2008. Development of wireless brain computer interface with embedded multitask scheduling and its application on real-time driver's drowsiness detection and warning. *IEEE Transactions on Biomedical Engineering* 55, 5 (2008), 1582–1591.

[25] Fu-Chang Lin, Li-Wei Ko, Chun-Hsiang Chuang, Tung-Ping Su, and Chin-Teng Lin. 2012. Generalized EEG-based drowsiness prediction system by using a self-organizing neural fuzzy system. *IEEE Transactions on Circuits and Systems I: Regular Papers* 59, 9 (2012), 2044–2055.

[26] 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.

[27] Gary McKeown, Michel F Valstar, Roderick Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 1079–1084.

[28] Stephan Mühlbacher-Karrer, Ahmad Haj Mosa, Lisa-Marie Faller, Mouhannad Ali, Raiyan Hamid, Hubert Zangl, and Kyandoghere Kyamakya. 2017. A driver state detection system—Combining a capacitive hand detection sensor with physiological sensors. *IEEE Transactions on Instrumentation and Measurement* 66, 4 (2017), 624–636.

[29] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. 2005. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping* 25, 1 (2005), 46–59.

[30] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. 2019. CogBeacon: A Multi-Modal Dataset and Data-Collection Platform for Modeling Cognitive Fatigue. *Technologies* 7, 2 (2019), 46.

[31] Ioannis Pavlidis, James Levine, and Paulette Baukol. 2000. Thermal imaging for anxiety detection. In *Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No. PR00640)*. IEEE, 104–109.

[32] Tiantian Qin, Judee Burgoon, and Jay F Nunamaker. 2004. An exploratory study on promising cues in deception detection and application of decision tree. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. IEEE, 23–32.

[33] Tiantian Qin, Judee K Burgoon, J Pete Blair, and Jay F Nunamaker. 2005. Modality effects in deception detection and applications in automatic-deception-detection. In *Proceedings of the 38th annual Hawaii international conference on system sciences*. IEEE, 23b–23b.

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

[35] 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.

[36] Brian C Tefft et al. 2014. *Prevalence of motor vehicle crashes involving drowsy drivers, United States, 2009-2013*. Citeseer.

[37] 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. 6. IEEE, 1–8.

[38] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.