Driver's Condition Detection System Using Multimodal Imaging and Machine Learning Algorithms

Paulina Leszczełowska[#1], Maria Bollin[#2], Mateusz Żak[#3], Karol Lempkowski[#4]

[#]*Faculty of Electronics Telecommunications and Informatics, Gdansk University of Technology*
Gdansk Siedlicka street 5a Poland

[1]s175902@student.pg.edu.pl

[2]s175642@student.pg.edu.pl

[3]s176086@student.pg.edu.pl

[4]s175691@student.pg.edu.pl

***Abstract*: To this day, driver fatigue remains one of the most significant causes of road accidents. Despite technological advancements, the automotive industry still lacks a clear solution to mitigate or even detect drowsiness. Many approaches have been tried and tested, but none have been widely adapted. In this paper, a novel way of detecting and monitoring a driver's physical state has been proposed. The goal of the system was to make use of multimodal imaging from RGB and thermal cameras working simultaneously to monitor the driver's current condition in real time. Based on chosen and further implemented machine learning algorithms, the system would actively notify the driver in cases of fatigue detection. In order to create such a system, the necessary video data was required. Multiple open-source datasets were taken into consideration, but in order to have full control over the type and nature of data, a simulated driving environment was prepared and used for the simultaneous acquisition of thermal and RGB video samples. Acquired data was further processed and used for the extraction of necessary metrics pertaining to the state of the eyes and mouth, such as the eye aspect ratio (EAR) and mouth aspect ratio (MAR), respectively. Breath characteristics were also measured. The created data vectors were used in later stages of the project for model training and testing. A customized residual neural network was chosen as the final prediction model for the entire system. The results achieved by the proposed model validate the chosen approach to fatigue detection by achieving an average accuracy of 85% on evaluation data.**

***Keywords*— Thermal, RGB, Machine Learning, Neural Networks, Simulator, Fatigue, Drowsiness, Driver**

I. Introduction

The National Highway Traffic Safety Administration reports that approximately 91,000 accidents are caused by drowsy driving every year. Approximately 50,000 people are injured and 800 are killed as a result of these accidents [1]. Since many of these accidents are not reported, and even if they are, it is sometimes difficult to determine whether a driver was drowsy at the time, the number of car accidents caused by drowsy driving is likely grossly underestimated. According to the European Commission, 10 to 25 percent of all accidents were due to driver fatigue [2].

As a result, a growing number of companies are implementing driver-state monitoring systems as a safety measure. This number will increase annually as a result of laws and regulations mandating the installation of this system in every new automobile. Due to this, driver fatigue detection has become a topic of great interest in the scientific community.

The existing solutions can be broken down into three categories: monitoring the status of the equipment in the controlled vehicle; measuring the physiological parameters of the driver; and monitoring the behavior of the driver. Systems belonging to the first category are those that utilize vehicle-specific data. This includes the movements of the steering wheel, the measurements of lane departure, and the patterns of braking. This system is non-intrusive, but its effectiveness is contingent on road conditions and driver skill. Systems that measure physiological responses are the most invasive and distracting for the driver because they require the driver to wear measuring devices like an electroencephalogram or an electrocardiograph. The mere act of wearing such devices can skew the system's results. The final category is concerned with observing and measuring the driver's behavior using computer vision methods. This type of system is frequently more dependable than its predecessors and is as non-intrusive as the system that monitors the vehicle's equipment. [3]

In this paper, we present a system that simultaneously uses video data captured by thermal and RGB cameras to determine the state of the driver. This multimodal approach enables the extraction of driver-specific features that would otherwise be lost when monitoring the vehicle's state. We have conducted a data acquisition process via the simulator that we designed, resulting in our own unique dataset.

II. State of knowledge

To initiate implementation of a solution for driver fatigue detection, a systematic literature review was conducted. A person's drowsy state is a fairly abstract concept, as each individual experiences it differently, so we wanted to

determine how other researchers have approached it and what the state-of-the-art solutions were. Each of the four authors contributed to the annotation and analysis of this paper. The relevant papers were gathered using three literature databases: IEEExplore, Scopus, and Springer. As a result, we collected 296 papers published between 2017 and 2022, which were then reduced by removing duplicates and excluding papers without a DOI identifier.

Then, each paper was given a score based on how pertinent the title appeared to be to our study, and only the articles that received the highest score from all of the annotators were included in the abstract relevance tagging. At this stage, we have disregarded any papers that did not use RGB or thermal video data to classify the driver's state. Following abstract relevance tagging in the same manner as title tagging, only 32 articles with a perfect score that were not literature reviews were considered for full text analysis.

A. Environment

The first piece of information essential for further development was the context in which the data were collected. We were able to extract information about the type of environment from 28 of the 32 articles. Three types of environments were identified: real, simulated, and regular.

Real-life settings are videos in which a driver is captured driving a vehicle. This type of environment is challenging because it can cause a driver to become distracted and cause dangerous situations on the road; as a result, they were only used in five of the discovered papers. However, models trained using such data samples may be the most trustworthy, as they directly correspond to the system's target environment.

A second type of environment was a simulated one in which data was collected using a car driving simulator. This method eliminates the potential dangers associated with actual driver video recording. A simulated environment can replicate real-world scenarios quite accurately, but requires additional devices and software. This configuration was utilized in eight of the 32 articles.

Researchers employed the final category the most frequently. They utilized videos depicting people in everyday settings that were unrelated to driving a car. Typically, they portrayed a person in front of the camera who was either drowsy or awake. This strategy appeared in 19 of 32 papers. This enables the simplest data acquisition, but it may lead to inconsistency in the final system's decisions.

B. Datasets

There were twelve distinct datasets found among the collected articles, and fifteen articles implemented their solutions using their own unique datasets.

Only three of the discovered datasets were directly associated with driver drowsiness detection. The Driver Drowsiness Detection Dataset collected by the NTHU Computer Vision Lab in a simulated environment was the most popular dataset [4]. The dataset was utilized eight times. Another dataset that was utilized twice was INVEDRIFAC, which contained driver data recorded while driving a vehicle. The final dataset (UTA-RDD) was only utilized once. This dataset includes videos of participants in their everyday environments.

Other datasets were utilized for the system's intermediate components, such as face detection, eye state detection, and yawn detection. There were four separate face detection dataset: WIDER FACE (used 3 times), MTFL (used 1 time), FER2013 (used 1 time), and Celeba (used 1 time). For eye state detection, five datasets were utilized: CEW (used 3 times), MRL (used 2 times), ZJU (used 1 time), and two datasets downloaded from Kaggle. The final 2 datasets used for yawn detection were YawnDD and one of the previously mentioned Kaggle datasets.

C. Features

There were only six distinct characteristics identified across the 32 articles. The majority of the discovered articles took a similar approach to selecting features for drowsiness detection, focusing on eye and mouth state detection (27 and 19 times, respectively). Additionally, six articles identified head bending or nodding as a sign of fatigue. Three other features were used only once each: eyebrow furrowing, measuring face temperature, and facial muscle movement speed.

D. Machine learning algorithms

Fourteen algorithms used in the implementation of driver drowsiness detection systems have been identified among the articles analyzed. In the vast majority of solutions (23 articles), Convolutional Neural Networks were utilized for either feature extraction or decision-making. Seven papers utilized VGG architectures. Most prominent were VGG16 for the final classification of the driver’s state and VGG-FaceNet for facial feature extraction. Support Vector Machines (SVM) were the third most frequently used model for both tasks and were implemented five times. Three times out of four, SVM was combined with the histogram of oriented gradients (HOG). The Long Short-Term Memory model was utilized four times to estimate driver drowsiness. General Deep Neural Networks (DNN) were the last algorithm to be used multiple times; they were employed twice. This algorithm was implemented once for feature extraction and classification and once solely for driver state classification.

Each of the remaining algorithms was only utilized once. Recurrent Neural Network, K-Nearest Neighbors, Bayesian classifier, Fisher's Linear Discriminant Analysis, Random Forest, and Long-term Recurrent Convolutional Network were applied for driver state classification. The Viola Jones method was used for face detection and extraction of the eye region. Adaboost was implemented for both the improvement of the face detection model and the final classifier.

III. Methodology

A. Simulator

Purpose of reliable acquisition of needed video data required a controlled driving environment. Due to safety and practicality reasons, the simulator was set up in a laboratory environment. The main components of the prepared system included a PC running the simulation as well as thermal and RGB cameras on separate stands, both of which were connected to the Google Coral computer.

Volunteers made use of commercially available steering wheels and gas and brake pedals to control simulated vehicles. The simulation was placed inside a city environment and did not include any particular goal. Participants were instructed to drive aimlessly through the city streets. Settings for the simulations were pre-set by our research team. Specific crowd density and weather conditions were chosen to provide a stress-free scenario that could induce drowsy behavior.

Recordings of participants were taken with aforementioned RGB and Thermal cameras with the use of Google Coral. Both cameras and the Coral device were placed inside 3D-printed cases tailored specifically to those devices. Separate cases were printed for Coral, which had a directly connected RGB camera, and for the Lepton Purethermal Mini Module, which was connected using a USB interface. Using openCV, our team managed to take recordings from both cameras simultaneously in a specified loop for a given session. Recordings were offloaded after recording to the researcher's laptop, which was connected to the Google Coral throughout the given session for control and monitoring purposes.

In order to get a clear view of the drivers' faces, a whiteboard was placed behind them in a position that completely covered the background of the person's head and upper torso. The RGB camera was recording from a top-down view, which ensured proper capture of facial features, while the thermal camera was recording from the level of the steering wheel. Placing the thermal camera in such a way allowed for the capture of the nostrils of participants, which made it possible to analyze breath metrics.

The use of commercially available hardware and open-source software allowed for easy setup of the environment and provided a safe and fully controllable source of data. Recording sessions with the use of the simulator were conducted under the guidance of at least one member of our research team.

Fig. 1 Simulator

B. Acquisition procedure

It was essential to optimize the process of acquiring thermal and RGB recordings. As was already discussed in the previous section, it was decided that at least one member of the team be present during recording sessions, so the procedure also needed to be adjusted in such a way that a single person had complete control over every part of the procedure. A custom bash and python script were prepared in order to automate recording with our cameras. Through the use of these scripts, we were able to create a setup where recordings would be taken one after the other without any additional input. A team member only needed to specify the ID assigned to a specific person, and recordings would be taken one after the other. Having tested the validity of such a solution, it was decided that recording sessions could begin.

The procedure for each volunteer was the same. Participants would first arrive at the laboratory with a simulator installed. A team member would explain the process and ask them to sign the required agreement before taking part in the project. Following the signing, a team member would conduct a short survey to determine if the person was currently feeling any kind of fatigue or if they had drank any caffeine earlier in the day.

After that, a given person had up to 10 minutes to get familiar with the setup and steering of the simulator while the present team member prepared to start recording. When the

participant confirmed that they were ready to start, a team member would fasten the Respiration Monitor Belt, which was used to gather data about breath frequency that would be used as reference for later stages of the project. Once it was done, the participant would start driving in a preset environment, and a team member would start the recordings.

An exit survey was conducted after recordings were completed, and a team member confirmed that the system worked properly through the file inspection. In this form, participants were asked if they felt any fatigue during the procedure. Besides that, the form included questions based on the user's sentiment towards the proposed system. Such a procedure would be repeated for every volunteer. Recordings were taken at different times of the day in an indoor setting with mainly artificial lighting.

C. Dataset

As a result of conducting recording sessions, our team managed to acquire over 200 RGB and thermal recordings from nearly 20 different volunteers. The volunteers' age group varied, but most of them were students aged between 20 and 25 years. Each recording is placed in a directory named after a specific participant's ID number. Recordings are further divided into two subdirectories. One for RGB videos and one for thermal. Each recording in the series includes a specific ID of a given participant with an additional number including the recording's order in the series. Having structured our dataset in such a way allows for easy access to the necessary data.

Every single recording was annotated based on participant responses to the surveys mentioned in the previous section. A CSV file with labels for every recording is placed in the main directory. Label 1 means that the person on a given recording is fatigued, while label 0 denotes the opposite. Other than labels for entire recordings, our team also annotated RGB videos frame by frame. These annotations included information on when a given person was yawning, had an open mouth, tilted their head, or blinked. Having both types of annotations allowed for much more detailed analysis in the next stages of the project and was a solid reference for further data processing and analysis.

IV. Proposed system

A. System architecture

Our system is based around the simultaneous recording of the driver with two camera modules, then processing the acquired recordings in the context of specific metrics and their further use, after being appropriately processed, in a recurrent neural network to arrive at the verdict.

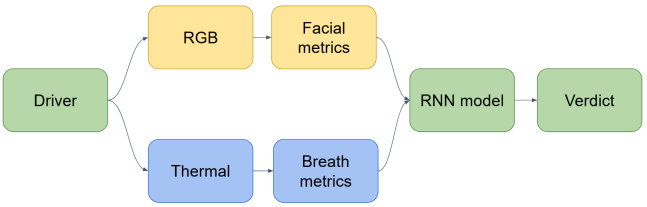

Fig. 2 Proposed system architecture

The presented architecture is the representation of workflow in the final version system, with the goal of being able to operate and arrive at a verdict in real time. For the purposes of the research, each stage of the system was implemented and tested separately. Choosing such an approach allowed for better control of the development phase and made it possible to conduct specific experiments to validate the approach chosen for a specific part of the system. The presented architecture includes everything from the driver and their vehicle to the cameras connected to the computer, which processes the data and notifies the driver through chosen means if fatigue is detected. Specific parts of the architecture, their details, and how they contribute to the whole system have been described in the further sections of this article.

B. RGB pipeline

The eye state and mouth state, which are two of the most crucial features for assessing a driver's condition, are determined using RGB videos.

The first step in RGB video processing is the detection of the face and its characteristic points in every frame of recording. It was done with the usage of dlib library which provides facial landmarks detector with pre-trained models and is capable of estimating the location of 68 coordinates (x, y) that map the facial points on a person's face.

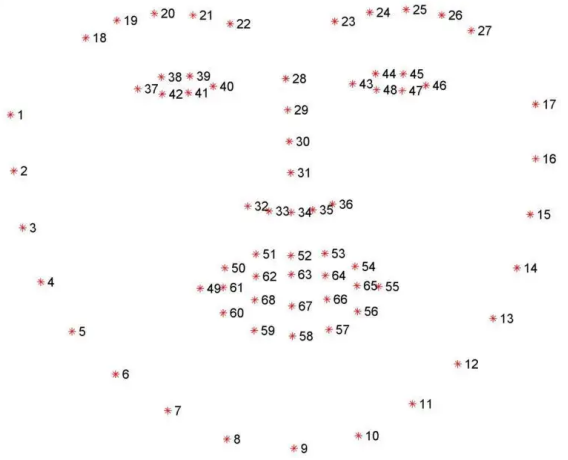

Fig. 3 Facial landmarks

$$MAR = \frac{||p_{62} - p_{68}|| + ||p_{63} - p_{67}|| + ||p_{64} - p_{66}||}{3} \quad (1)$$

In order to determine the mouth state of the driver, six of the detected points are used to calculate a measure called the Mouth Aspect Ratio (MAR). It is defined by the mean of three distances between points from the upper and lower lips to identify how widely the mouth is opened.

For the purpose of eye state evaluation, two approaches are used. The first of them is based on the same concept as mouth state determination. Therefore, in order to assess the degree of eye closure, twelve facial points are used (six for each eye) to calculate Eye Aspect Ratio (EAR).

$$EAR = \frac{||p_{38} - p_{42}|| + ||p_{39} - p_{41}||}{2 \times ||p_{37} - p_{42}||} \quad (2)$$

A second approach is used to classify eyes as open or closed. The classifier is based on a convolutional neural network with VGG16 architecture. The model is initialized with the weights of the VGGFace model and further trained with the Closed Eyes in the Wild (CEW) dataset.

C. Thermal pipeline

The respiratory rate plays an important role in determining the driver's state. It is determined using data extracted from thermal imaging.

The first step of data processing is histogram equalization. It equalizes the distribution of intensities for a given range of values. In the case of images, it essentially improves the contrast. Due to the low resolution of thermal imaging, the face features may not be clearly visible. This step helps highlight those features. To obtain the desired facial region, a simple face detection method with Haar cascades was employed. Once the region of interest was obtained, the mean value of pixels for each frame was calculated. The resulting values can then be treated as a signal representing the average pixel value over time. During respiration, the pixels around the nose and mouth area should change values due to the difference in temperature.

The final steps aim at reducing the noise and filtering the signal to obtain a waveform used for respiration rate prediction. First, the signal is smoothed out using the asymmetric least squares smoothing algorithm with $\lambda = 10$ and $p = 0.1$. Finally, the resulting signal is filtered using a digital Butterworth filter with critical frequencies of 0.046 and 0.23.

D. Final discriminator

The final decision model was based on RNN to which video segments were fed. Each video was separated into 10-second-long windows. This window was then shifted by one second's worth of frames, resulting in twenty windows per video. Then, for each window extracted from a video, five features were calculated using the metrics described in earlier sections. These characteristics included mean EAR, percentage of frames with closed eyes, mean MAR, maximum MAR, and mean breath length.

The final discriminator comprises a single GRU unit and two linear layers. The model outputs a single value in the range (0, 1), which indicates the confidence that the driver is in a drowsy state. A decision was made based on a chosen threshold of 0.5. If the confidence was above the threshold, that would indicate that the driver was drowsy.

V. Achieved results

A. Eye state detection

From the calculated EAR metrics for both eyes separately, the mean was taken in order to estimate the average state of both eyes at a time. In Figure 5 we can see exemplary frames from our dataset with mean EAR calculated.

The trained VGG16 model achieved very good results while tested on our dataset. In Figure 4 we can see the confusion matrix for the classification calculated on the basis of 30 videos with about 800 frames each. The classification results were taken in order to calculate the number of eye blinks per video. The results were evaluated using two error measures: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The Table 1 shows the corresponding error values with the comparison of classification by simple EAR threshold. As we can deduce from the table, CNN model gives much better results with lower error rates compared to simple EAR.

		True class: Open	True class: Closed
Predicted class	Open	20443	365
	Closed	2019	1479

Fig. 4 Confusion matrix for classification by VGG16

Table 1 Confusion matrix for classification by VGG16

Metric	VGG16	EAR threshold
MAE	**3.55**	25.38
RMSE	**5.41**	32.7

B. Mouth state detection

Given that the data was annotated as either open or closed mouth, it is difficult to estimate the accuracy of the calculated metric in terms of mouth state. For the purposes of quality control, we applied a threshold of 5.5 to the MAR value, and any value above this threshold is deemed an open mouth.

Based on the binary classification of a mouth state to evaluate the correctness of detection, we computed the average accuracy, Area Under Receiver Operating Charateristic Curve (AUROC), and F1 score for sample 5 recordings. The mouth state classification achieved great results, with an average accuracy score of approximately 0.915, an average AUROC of 0.895, and an average F1 score of 0.787. This ensured that the MAR values calculated for specific frames accurately reflected the mouth's actual condition.

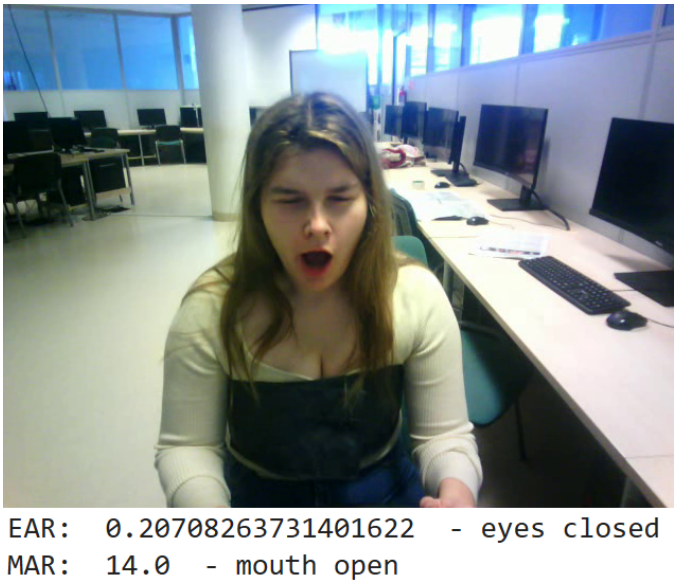

Fig. 5 Exemplary frame wit EAR and MAR values calculated

C. Respiratory rate detection

For the purposes of the problem, a mean respiration length was calculated. A duration of a single breath was obtained by taking the number of frames between valleys in the signal resulting from the processing described in Section IV-C and dividing it by the frame rate of the Lepton camera. The mean lengths varied between recordings, however, the overall average came to about 3 seconds per breath. Medical studies show that the average respiratory rate for adults is between 12 and 20 breaths, which means that the results are consistent with medical knowledge. An example signal is presented in Fig. 6.

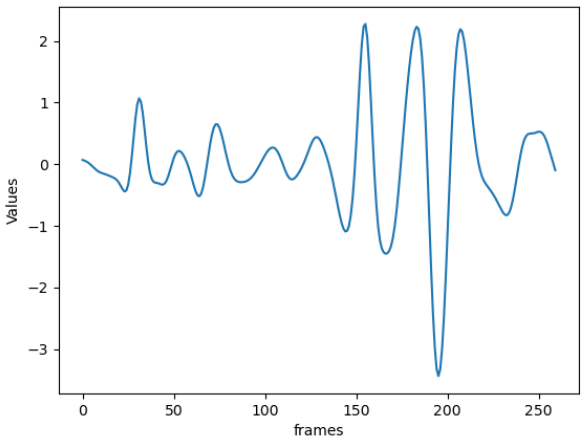

Fig. 6 Respiration signal resulting from processing thermal imaging data

D. Drowsiness detection

The model was trained using 92 videos, with 59 demonstrating alert drivers and 33 demonstrating drowsy drivers. It was trained with the Adam optimizer with early stopping based on the number of epochs in which the balanced accuracy score on the validation dataset did not decrease. The balanced accuracy score, presented in Fig. 7, is defined as the mean of recall for each predicted class.

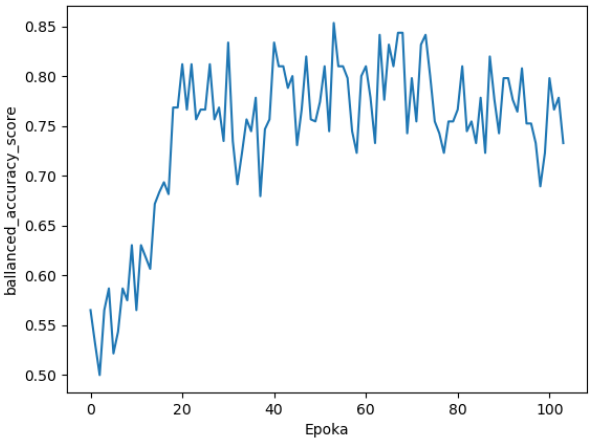

Fig. 7 Balanced accuracy score on the validation dataset

With such a limited dataset, the model was able to achieve balanced accuracy of approximately 85%, which is an extremely encouraging score.

VI. Summary

In the course of this research, a driver drowsiness detection system was developed. It is based on a multimodal approach and makes use of both RGB and thermal cameras. We were therefore able to obtain the driver's physiological characteristics in a non-invasive manner. With information regarding a participant's eyes, mouth, and breathing rate, we were able to achieve a validation accuracy of approximately 0.85. This result demonstrates that this particular type of system is reliable, and with additional data samples, it is possible that it could be improved even further. It is possible that the addition of vehicle state variables that are commonly used in the automotive industry could further improve the prediction and increase overall road safety.

Acknowledgment

We wanted to issue a special thank you to Gdańsk University of Technology for supporting us in the process of developing this system and giving us access to the university's resources.

References

[1] *Drowsy driving*. NHTSA. (n.d.). Retrieved January 15, 2023, from https://www.nhtsa.gov/risky-driving/drowsy-driving

[2] *Fatigue and crash risk*. Mobility & Transport - Road Safety. (n.d.). Retrieved January 15, 2023, from https://road-safety.transport.ec.europa.eu/statistics-and-analysis/statistics-and-analysis-archive/fatigue/fatigue-and-crash-risk_en

[3] Huang, R., Wang, Y., & Guo, L. (2018). P-FDCN based eye state analysis for Fatigue Detection. *2018 IEEE 18th International Conference on Communication Technology (ICCT)*. https://doi.org/10.1109/icct.2018.8599947

[4] Weng, C.-H., Lai, Y.-H., & Lai, S.-H. (2017). Driver drowsiness detection via a hierarchical temporal deep belief network. *Computer Vision – ACCV 2016 Workshops*, 117–133. https://doi.org/10.1007/978-3-319-54526-4_9