Gaze Zone and Drowsiness Identification in Unconstrained Scenarios for Advanced Driver Assistance Systems

先進的運転支援システムのための運転シーンを限定しない視線及び眠気の識別に関する研究

February, 2022

Catherine Elena LOLLETT PARAPONIARIS
ロレット　パラポニアリス　カテリン エレナ

Gaze Zone and Drowsiness Identification in Unconstrained Scenarios for Advanced Driver Assistance Systems

先進的運転支援システムのための運転シーンを限定しない視線及び眠気の識別に関する研究

February, 2022

Waseda University Graduate School of Creative Science and Engineering

Department of Modern Mechanical Engineering, Research on Human-Robot Interface

Catherine Elena LOLLETT PARAPONIARIS
ロレット　パラポニアリス　カテリン エレナ

Abstract

Many fatal accidents are the result of a driver's inattention or fatigue. Several studies have been undertaken on driver gaze zone and drowsiness classifiers to solve this issue. However, making this classification under unconstrained situations remains extremely challenging. Examples include the driver's face partially covered (e.g., masks, scarves, eyeglass sticks), the driver's face being in a profile pose, the environment having considerable changes in light conditions, and the driver's eyeglasses having reflections. By merging computer vision techniques and several deep-learning models, the proposed system intends to distinguish the driver's gaze zone and drowsiness under highly unconstrained conditions.

This study proposes a robust pipeline for analyzing drivers' behavior. The first step in the pipeline is to adjust the frame's brightness using a technique called Contrast Limited Adaptive Histogram Equalization (CLAHE) on the frames' Lab Color Space lightness channel to mitigate the effects of strong lighting changes. Next, the pipeline uses dense landmark detection, optical flow estimation, and constructs pose flows by associating cross-frames poses to accurately identify the face, eyes, pupils, and shoulder joints. These methods also allow the tracking of pupil and eyelid movement even in situations where the face is partially occluded or in a profile pose. The pipeline proposes two modules: Module 1 focuses on driver gaze detection, while Module 2 focuses on driver drowsiness detection.

In Module 1, the driver's gaze direction is detected by considering two different face postures for the gaze classifier: frontal and profile. A separate DNN model is trained for each face pose, and the feature vector parameters for these models are based on the relationships between pupil and eye landmarks in proportion to the driver's facial structure. This takes into account the fact that everyone's facial structure is different. The model will then retrieve a predefined standard driver's gaze area based on the direction of the eyes.

In Module 2, a GRU model is used with a novel input feature vector that considers the driver's eyelid closure, lower-face contour, and chest movement landmarks to detect drowsiness. Unlike previous studies, mouth closure is not included in the feature vector due to the possibility of the mouth being covered by a mask. One of the key contributions of this approach is the inclusion of chest movement and lower face contour as possible parameters of the feature vector, which helps to address the issue of mask-wearing. Drowsiness is detected by looking for yawning and eyelid closure. To recognize a yawning state, the lower facial contour and chest movement landmarks are used, and the current location of these landmarks is subtracted from their original position obtained in the first frame. Eyelid closure is measured by tracking the closure of the eyes in each frame. The driver's drowsiness is then determined by combining these parameters and using them as the feature vector of a GRU-based model.

The proposed method has been shown to be more effective than existing methods at detecting the driver's gaze direction and drowsiness in challenging situations, as demonstrated by the results obtained from a dataset featuring highly challenging driving conditions. These results suggest that the method is able to accurately identify the driver's gaze zone and drowsiness in a variety of challenging circumstances.

To everyone who helped make my dream come true

Acknowledgements

First, I am sincerely thankful to my supervisor, Prof. Sugano, who gave me the chance to study in this amazing laboratory.

To Prof. Kamezaki, my whole gratitude because his guidance made me go beyond my boundaries, challenge myself, and make me love more research every day.

To the Ministry of Education, Culture, Sports, Science, and Technology of Japan for making my dream come true by providing me with a scholarship to study in Japan.

Forever indebted to my part-time job friends in Sola K.K., especially to Mr. Yoshimura, because their endless support. Thank you for giving me the opportunity to work while studying. Without them, everything would have been much harder.

To Simón Bolívar University, Waseda University, Nagaoka University of Technology, and The University of Tokyo, for not only giving me the tools to become the professional that I am but also for the incredible experiences that shaped me into the person I am.

My heartfelt gratitude to my professors for constantly reminding me that we are not alone in this path. Thank you for being my role model.

To all my friends, my students, and their families from Venezuela, Japan, and all around the world, because inside my weakness, they never let me down.

To Watanabe, Motoi, Mizukami, Futagawa, and Maeda families because knowing that I was far away from my family, they decided to make me their own daughter.

Finally, nothing in this world has any worth without the endless love of my family and Seurin-Mena's family. Thank you for working so hard for me and for your unconditional support. Nothing can ever repay all you have done.

To everyone that in any way supported me, thank you. You are my heroes.

Table of Contents

List of Tables

List of Figures

Abbreviations

ADAS Advanced Driver Assistance System.

AFW Annotated Faces in the Wild.

CAM Class Activation Map.

CLAHE Contrast Limited Adaptive Histogram Equalization.

CNN Convolution Neural Network.

CPU Central Processing Unit.

DCNN Deep Convolutional Neural Network.

DNN Deep Neural Network.

EDA Electro-dermal Activity.

EEG Electroencephalograms.

EKG Electrocardiograms.

FAN 3D-Facial Alignment Network.

FDDB Face Detection Dataset and Benchmark.

GAN Generative Adversarial Network.

GPU Graphics Processing Unit.

GRU Gated Recurrent Units.

HE Histogram Equalization.

HOG Histograms of Oriented Gradients.

ICP Iterative Closest Point.

LSTM Long Short-Term Memory Networks.

NIR Near-Infrared.

NMS Non-Maximum Suppression.

PERCLOS Percentage of Eyelid Closure over the pupil over time.

PF-NMS Pose Flow Non-Maximum Suppression.

POSIT Projection-based Pose Iteration.

PRC Precision-Recall Curve.

RGB Red, Green and Blue.

RGB-D Red, Green, Blue, Depth.

S3FD Single Shot Scale-invariant Face Detector.

SDM Supervised Descent Method.

SVM Linear Support Vector Machines.

Glossary of Terms

Class Activation Map Visualization technique used to show which regions of an image were important for a specific classification.

Convolutional Neural Network Neural network that uses convolutional layers to analyze visual data.

Deep Convolutional Neural Network Neural network that uses convolutional layers and multiple layers to analyze visual data.

Deep Neural Network Neural network with multiple layers that can learn complex representations of data.

Electrocardiograms Test that records the electrical activity of the heart.

Electrodermal Activity Measurement of the electrical activity of the sweat glands in the skin.

Electroencephalograms Test that records the electrical activity of the brain.

Equalization Process of adjusting the distribution of brightness values in an image to enhance its overall contrast.

Facial landmarks Specific points on a face such as the corners of the eyes mouth and jawline that can be used as reference points for facial analysis.

Facial Alignment Process of detecting and adjusting the position and orientation of facial features in an image to a standardized reference frame.

Frame Single image of a video sequence.

Gated Recurrent Units Recurrent neural network that uses gates to control the flow of information.

Generative Adversarial Network Neural network used for generative tasks where it learns to generate new data from a given dataset.

Histogram Graphical representation of the distribution of numerical data.

Histograms of Oriented Gradients Feature descriptor used for object detection in computer vision.

In the wild Describes images or videos captured in uncontrolled natural environments.

Iterative Closest Point Method for aligning point clouds in 3D space.

Linear Support Vector Machines Machine learning algorithm used for classification and regression tasks.

Long Short-Term Memory Networks Recurrent neural network that can learn long-term dependencies in sequential data.

Near-Infrared Light that has a longer wavelength than visible light and is often used in imaging and sensing applications.

Neural Network Machine learning model composed of interconnected nodes that process and transmit information inspired by the human brain.

Non-Maximum Suppression Method for reducing multiple detections of the same object in object detection.

Pixel Basic unit of a digital image.

Pose Flow Technique for estimating the movement of an object in an image.

Recurrent neural network Neural network that uses feedback connections to process sequential data over time.

Scale-invariant Feature or model that can accurately detect or classify objects regardless of their size in an image.

x Position of a point along the horizontal axis in a two-dimensional space.

y Position of a point along the vertical axis in a two-dimensional space.

Chapter 1

Introduction

1.1 Motivation

There is strong evidence that driver inattention and drowsiness are major causes of fatal and injury crashes. Advanced Driving Assistance Systems (ADAS) that can detect these states in advance have the potential to improve safety by warning drivers and even taking control of the situation in autonomous vehicles. Tracking the driver's head and eyes can provide an accurate assessment of their state. However, current studies on this topic have mostly been conducted under ideal conditions and often use expensive sensors or complex systems, leading to poor performance in unrestricted conditions. This research aims to develop a high-performance driver gaze and drowsiness classifier that can handle challenging situations such as mask-wearing faces, face occlusions, eyeglasses reflections, strong light changes, profile face poses, and faces and eyes facing different directions. By accurately identifying distractions with fewer mistakes, this system has the potential to significantly reduce the number of automobile accidents.

1.2 Contributions

This work aims to develop a single-camera, robust classifier for driver gaze zone and drowsiness that can handle various challenging situations effectively.

These challenging situations may involve one or more of the following scenarios:

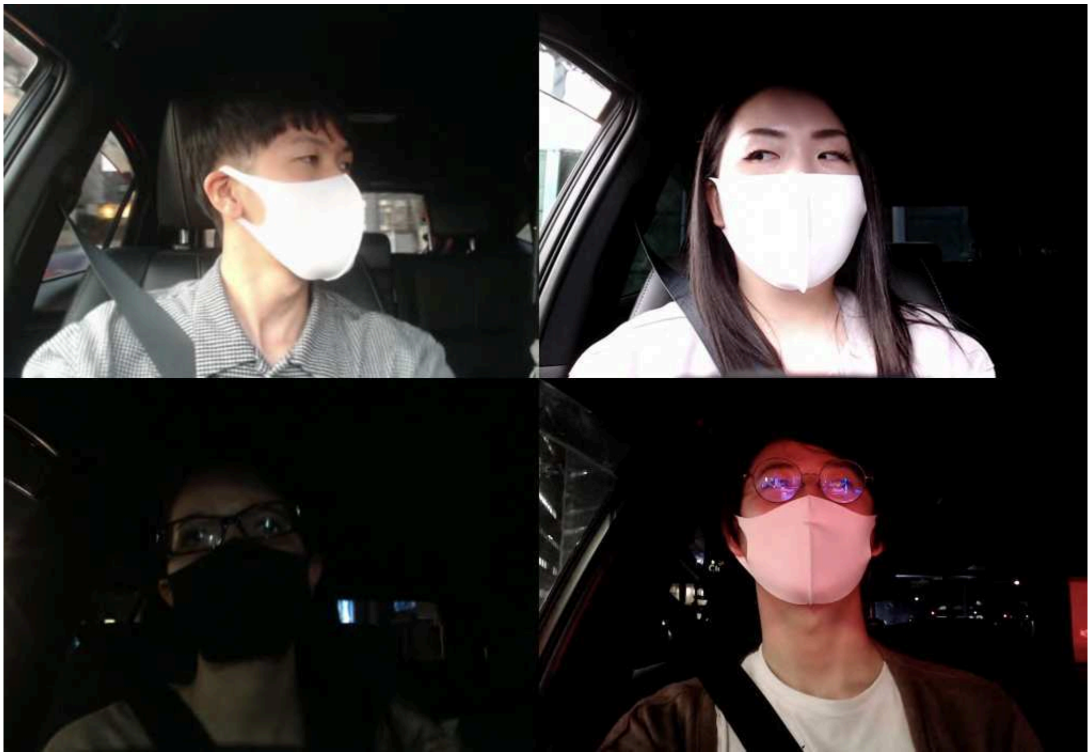


Figure 1.1: Left: Driver with profile face, eyes facing the window, wearing white mask in a daylight condition (top) and driver with frontal face with eyeglasses reflection, pupil noise, wearing a black mask in a poor-light night environment (bottom). Right: Driver with frontal face, eyes facing right window direction, wearing white mask in a daylight condition (top) and driver with frontal face, eyes facing towards back-mirror, wearing white mask, eyeglasses reflection, pupil noise, wearing a white mask in a night environment (bottom).

- Mask-wearing faces
- Face partial occlusions
- Eyeglasses reflection
- Strong daylight variations
- Pupil noise

Some of the edge conditions considered in this study are shown in Figure 1.1.

By using a single camera, the system can be more cost-effective and easier to implement. The goal is to create a classifier that can accurately detect driver gaze zone and drowsiness in a wide range of challenging conditions, improving safety and reducing the risk of accidents.

To achieve this goal, two modules have been implemented: Module 1, which recognizes the driver's gaze direction, and Module 2, which recognizes the driver's drowsiness. By addressing these two aspects of driver attention and alertness, the system has the potential to improve safety and reduce the risk of accidents.

To reach a robust and portable system, the key seven steps that this study involves are as follows:

1. Frames' Lab's color space manipulation: To address the strong light variation issue, equalize the brightness of the frames by manipulating its Lab's color space's luminance channel using a Contrast Constrained Adaptive Histogram Equalization is one of the keys of this study.

2. Robust recognition of face, eyes, and body-joints landmarks: Combining an anchor-based real-time face detector with a normalized dense alignment for landmark identification that incorporates 3D eyelid and facial expression movement tracking for face and eye landmarks detection makes the data less noisy. In addition, an online optimization framework for the shoulder joints recognition is used. It builds an association of cross-frame poses and form pose flows robust to unconstrained situations body pose. This gives robustness to strong light condition variations and various facial occlusions such as masks, scarves, eyeglasses reflections, eyeglass sticks, small eyes, partial occluded pupils, and profile face poses.

3. DNN models structure and introduction of novel feature vectors parameters: Module 1 of the framework consists of two main DNN models: the Face Frontal model and the Face Profile model. The feature vector parameters used by the models include information about the relative positions of different facial landmarks, such as the pupils and eyes, in relation to the overall geometry

of the face. Since the geometric facial structure varies from person to person, these feature vectors allow the models to adapt to individual differences in facial structure. Using just these relations helps to make the training dataset much smaller, as the pattern is clear and specific. Each model classifies different standard areas of the face that are known to be essential for drivers to check. It also considers when the driver's face and eyes are facing in different directions. This classification is done on a per-frame basis.

4. GRU model structure, introduction of lower face contour and chest movement landmarks: Module 2 of the framework involves using a Gated Recurrent Units (GRU) model to classify driver drowsiness. In this module, the feature vector parameters consider the landmark information about the eyes, lower face contour, and chest movement. Unlike some previous studies, this model does not include mouth closure as a feature vector parameter, as masks may cover this area. Including chest movement and lower face contour as possible parameters is a key contribution to address the issue of mask-wearing in the drowsiness classifier. This classification is based on video, rather than individual frames. Using the eyes landmarks, the system measures the closure of the eyes. To measure the yawing state, was use the movement of the lower face contour and chest. For these landmarks, the current position of each landmark is compared to its original position in the first frame. Finally, the driver's drowsiness is determined by combining spatiotemporal features based on the previously mentioned subtractions, which are used as the feature vector for the GRU model.

5. High Normalization: The feature vectors are highly normalized, making the patterns easy to discern. This helps the system to reduce the required dataset.

6. Portability and extensivity: One of the advantages of this system is its portability and extensiveness, as it requires only a single camera and a computer to operate.

7. Generalization: This system has good generalization capabilities, as it can classify correctly regardless of the subject being analyzed.

In addition, the performance of the proposed system was compared to the general approach, and also was evaluated the importance of each stage of the proposed pipeline. When tested on a dataset featuring highly unconstrained driving conditions, the system outperformed the general approach in accurately classifying the driver's gaze zone and drowsiness in challenging situations. These results highlight the effectiveness of the proposed system in handling a wide range of conditions.

1.3 Thesis Outline

The thesis is organized into seven chapters. In Chapter 1, the reader is introduced to the main goals and motivation for implementing robust driver gaze and drowsiness classifiers. Chapter 2 goes through the current literature and makes a rough analysis of their robustness. Chapter 3 describes how to manipulate the Lab's color space of the frames to mitigate strong lighting variations. Chapter 4 introduces the robust libraries for strong face, facial landmarks, pupil, and body posture detection. The application of dense landmark detection and optical flow estimation methods for accurately identifying the face, eyes, and pupils, as well as tracking pupil and eyelid movement, are described. Also examines the performance of different face detectors with and without masks to demonstrate the importance of conducting the research under challenging conditions. Chapter 5 outlines the implementation of the Gaze Classifier Module (Module 1). The module considers two face poses: frontal and profile, and uses a separate DNN model for each pose, with the output being the driver's gaze zone. The feature vector parameters for these DNN models are based on the relative positions of the pupil and eye landmarks in relation to the overall geometry of the face. Chapter 6 covers the Drowsiness Classifier Module (Module 2). This module explores using a combination of lower-face contour, eyes, and chest landmarks to compare the effectiveness of three different feature vectors and explains how to integrate spatiotemporal features using a GRU model. The modules' performance is evaluated and compared to other methods in Chapters 5 and 6. Finally, Chapter 7 discusses the results obtained, the strengths and weaknesses of the system, and potential areas for improvement in the future.

Chapter 2

Background

2.1 Previous Research

During typical driving situations, tracking the driver's face, eyes, and body position can be an effective way to detect signs of distraction or drowsiness early on to prevent car accidents. However, occlusions can occur in certain driving scenarios, such as when the driver wears a mask or scarf when there are reflections on eyeglasses, or intense light changes in RGB frames. These occlusions can pose a challenge for classifiers if they are not robust enough to handle them.

To date, no available gaze and drowsiness classifiers have been evaluated using a dataset with a wide range of highly unconstrained conditions. However, a long list of classifiers has been implemented for more constrained conditions. This chapter will review various gaze and drowsiness classifiers that use a single camera (or sensor) to give some context to the reader and a preliminary assessment of these classifiers' robustness.

2.1.1 Previous Research for Gaze Zone Classifiers

Recognizing the driver's face and other facial landmarks is often the first step in most gaze classifiers, as it provides the foundational data that will be used in subsequent process steps. If this step is unsuccessful, the rest of the classification will also be highly impacted due to poor performance. For this step, some studies, such as [1]–[7], rely on Cascade Classifiers operating on Haar Feature Descriptors, Histograms of Oriented Gradients (HOG), or Linear Support Vector Machines (SVM). However, as [8] notes, these methods can struggle to localize faces accurately or predict landmarks under unconstrained conditions. Under unconstrained conditions, face detection also tends to have lower recognition rates and slower processing times than other methods, as discussed in Chapter 3.

One study that tries to develop a robust gaze classifier is [9]. They use Faceness for face detection, but this method has the weakest Precision and Recall Curve performance of all face detection models when tested on the Wider-Face dataset [10], a benchmark for face detection. This may lead to misclassifications under unconstrained conditions. They use estimations derived from cascaded regression models for landmark detection, as described in [11]. However, as pointed out by [12], this type of regression can be ineffective in certain situations, such as when the face is in a profile pose, as it can only regress visible spots on the face and cannot describe the occluded parts.

Several studies, including [13], [14], and [15], use pre-trained CNN models for their classifiers. However, these models may struggle to classify patterns accurately that do not resemble those in the training dataset, leading to incorrect classifications in unconstrained conditions, such as when the driver is wearing eyeglasses or other occlusions like masks or scarves. A good advantage of [13]'s work is that is able to run on low-capacity devices.

[16] proposes an interesting approach that observes the driver's gaze zone and the surrounding driving scenario environment to understand how the driver is processing information from the outside world, referred to as a vision-in/vision-out strategy. However, this method requires a head-eye tracker to measure the gaze direction, which can be invasive to drivers in actual driving situations. Despite this limitation, the approach has great potential, and mapping the objects with the driver's gaze could improve the accuracy of gaze classifiers. [17] also uses a head-eye tracker, while

[18] uses a stand-alone eye tracking device. However, eyes trackers, as [18] states, may not always be able to accurately capture the gaze due to the constant changing light conditions that can cause reflections that will turn into noisy data taken from the device.

Using an RGB-D camera, [19] implements a multi-zone Iterative Closest Point (ICP)-based head location tracking and gaze estimation system based on appearance. The Viola-Jones face detector is used for face recognition, but as noted in [9], this detector may not be robust enough for challenging scenarios. [20] trains a random forest classifier using head vectors and eye image features. The POSIT algorithm is used to compute the head vector by combining facial landmark identification with a 3D face model. The SDM facial landmark detector is used to locate eye corner points and other facial points. However, the SDM method relies on cascaded regression for face alignment, which has limitations for large-head poses.

[21], [22], and [23] only use face pose information for their classification, which can be problematic as it only estimates the head pose without considering the pupils. [21] creates intervals based on continuous gaze angles and treats the grid of quantized gaze angles as an image for dense prediction using a headband, which can be intrusive. [22] builds on pre-existing CNN structures with minor adjustments and estimates the pose using the POSIT algorithm with a 3D generic face and selected rigid landmarks. While the landmark detection method is more robust than other research, it relies on a method that may incorrectly classify occluded landmarks for profile faces. [23] uses point clouds, which have the potential to be robust for large head poses.

[24] proposes an interesting approach that attempts to solve the problem of eyeglass reflection in challenging driving situations. However, their dataset does not include cases where one eye is covered, does not show results per gaze class, and does not evaluate how well their method works with other types of face occlusions.

[25] uses Fine-Grained Head Pose Estimation Without Keypoints [26] to determine the Euler angles for the head posture. This method estimates the intrinsic Euler

angles directly from the intensities of the images using joint binned pose classification, and regression [26]. However, regression methods are not suitable for scenarios with occlusions as they are not robust enough to handle them.

After analyzing various gaze classifiers and considering the strengths and weaknesses of my previous work, this study proposes an improved pipeline that takes another step toward an accurate classification in situations with unconstrained conditions.

2.1.2 Previous Research for Drowsiness Classifiers

Drivers who are too exhausted to handle their vehicles safely and effectively are a potential cause of serious accidents. Drowsiness causes a delay in reaction time, a reduction in awareness, and impaired judgment. This increases the chances of an accident. According to the research that has been done thus far, two key indicators of tiredness are yawning and eyelids' closure. It should not be difficult to spot these behaviors in a constrained driving environment. Unfortunately, these drivers' features are frequently threatened by occlusions in a typical driving situation. The previously mentioned occlusions can be, for example, masks, eyeglass reflections, or strong daylight changes, making it difficult to achieve the previously discussed recognition. Same as for the gaze zone research, work done in the past has yet to fully investigate and evaluate how tiredness while driving should be classified in an unrestricted environment. However, drowsiness classifiers for drivers have a long history of use in restricted scenarios.

This subsection provides an overview of various drowsiness classifiers to provide context for the reader and a preliminary assessment of their robustness.

In the same line as the gaze zone classifier, the goal is to develop a system that can be applied in real-world settings using a single sensor that is not invasive to the user, so the list of the studies that were reviewed involves these aspects. In light of this, is not included in the analysis any work that consists of the use of invasive sensors such as electrocardiograms (EKG), electroencephalograms (EEG), head-eye trackers, electro-dermal activity (EDA), or head-bands; similarly, is not considered any studies that were carried out in simulated environments. The analysis is made on [27]'s lists of the most recent research on drowsiness detection based on behavioral factors, in addition to other more recent works that aren't featured inside it.

Behavioral parameters are a way to detect drowsiness without using intrusive procedures. These methods assess a driver's level of fatigue based on observable behavioral factors such as head position, the ratio of times their eyes are closed, facial expressions, eyes blink's frequency, and yawning. Cameras and computer vision techniques are often used in behaviorally-based approaches to extract these behavioral features [27].

To evaluate previous studies' robustness, there are two important aspects to be focused on: the reliability (robustness) of the libraries or approaches used to extract features and the robustness of the characteristics used to assess whether or not the driver is drowsy. Ensuring that the feature extraction methods and the characteristics used to identify drowsy drivers are robust is essential to produce reliable results.

Analysis of the libraries' robustness used to extract facial features

Firstly, the reliability (robustness) of the libraries or approaches used to extract the key landmark features will be analyzed. Many previous studies have used face and facial landmark identification as a first step in the process, as it provides the foundation for subsequent processing and classification. Is important to note that if this step does not succeed, the entire classification process will be unsuccessful.

Cascade Classifiers that are based on Histograms of Oriented Gradients (HOG), Haar Feature Descriptors, Linear Support Vector Machines or Ada-Boost, are what [28] - [34] use for this step. These approaches either fail to accurately locate the faces or produce predictions of the landmarks that are not properly aligned when attempting to identify face attributes under unconstrained conditions, as [8] explains straightforwardly and concisely. [35] - [36] use Viola-Jones face detection. Nevertheless, as [9] describes, the Viola-Jones face detection method does not provide sufficient robust performance, particularly when the face is significantly rotated or occluded. Therefore, it is important to carefully consider the robustness of the feature extraction methods used in drowsy driver detection systems to ensure reliable results.

Near-infrared (NIR) cameras are used in [30] [32] [35] implementation. NIR cameras are sensitive to wavelengths of light in the near-infrared range. These cameras have several advantages, including capturing images in low-light conditions.

However, NIR cameras also have some drawbacks. One major drawback is the lack of color. When sunglasses, masks, or other occlusions cover the face, the lack of color makes it harder to recognize faces and landmarks. This is because the difference between the skin and surroundings' color, plays a major role in detecting facial features and other landmarks while wearing masks or sunglasses.

Another drawback of NIR cameras is the problem of specular reflections on eyeglasses, which can cause a continual occlusion of the pupils and degrade the performance of a face detector system. This is because the reflections from the eyeglasses can create multiple light blobs in the image, which can confuse the tracking algorithm. Some researchers have attempted to address this issue by developing algorithms specifically designed to optimize tracking with eyeglasses, but these approaches have not yet been widely adopted. As [30] also describes, when users wear glasses, the performance of the tracker suffers because different light blobs appear in the image as a result of the reflections caused by the NIR camera in the glasses. Additionally, they have not used any specific algorithms to enhance the tracking with the glasses [30]. Moreover, the pattern used to detect faces and other landmarks is the same as [28].

Regarding the libraries used, [37]'s work utilizes the most robust libraries inside the current research. Their work uses the same face detector used in this work, which is called the Single Shot Scale-invariant Face Detector (S^{3}FD) [38]. As for the landmark detector, is used 3D-Facial Alignment Network (FAN) which is both accurate and reliable.

Analysis of the features considered to determine drowsiness

The second robustness analysis will be based on the features that different studies employ. Two characteristics are crucial for determining whether or not a driver is drowsy: eye closure and yawing frequency. The ratio of eye closure, or PERCLOS, can be used to classify the eyes as open or closed and can indicate drowsiness or fatigue. Similarly, detection systems based on yawning frequency can identify changes in the geometric shape of the mouth, such as the size of the mouth opening or the placement of the lips, as potential signs of drowsiness.

However, it is important to note that these features may not always be sufficient for detecting drowsiness in all circumstances. For example, if a driver is wearing sunglasses or a mask that occludes the eyes or mouth, these features may not be visible or may be difficult to detect accurately. In these cases, other features or methods may need to be used to detect drowsiness, or the system may need to be designed to be more robust to these types of occlusions.

PERCLOS was the only feature used in [36], [39], [30], [29], [40], [32], [41], [42] work. Eye closure can indeed be an important feature for detecting driver drowsiness. Eye closure is often considered a strong indicator of fatigue and drowsiness, and many studies have found that drivers that have a smaller eye closure are more likely to be drowsy or fatigued. However, it is also important to note that relying on any single feature, such as PERCLOS may not be sufficient for accurately detecting drowsiness in all cases. Citing [37]'s explanation, yawning is a key parameter in a drowsiness detection system. As a result, drowsiness detection systems need to consider yawning frequency as a potential feature.

[31],[43],[44], [45], [34],[33],[28], [37] incorporates frame sections of the mouth or other locations into their feature vector. Using features such as frame sections of the mouth or other locations as part of the feature vector can be useful for detecting drowsiness. Still, these features may not be reliable in all circumstances. For example, if a driver wears a mask that covers the mouth, the mouth feature may become very

noisy data for the model, as it is hidden and does not provide any useful information. Similarly, using the full set of face landmarks as features without filtering or processing them can also result in noisy data, as each person has their own facial configuration, leading to unique landmark spaces for each person. This can make it challenging for the model to identify a consistent pattern.

After reviewing the related work, it is clear that both classifiers lack robustness to daily and basic conditions such as masks or eyeglass reflections. This lack of robustness is a major concern, as it can lead to accidents due to the high risk of misclassification. This study aims to address this issue by developing classifiers that are more robust to these conditions.

Chapter 3

Frame Environment Manipulation

3.1 Near-Infrared vs. RGB Cameras

Using Near-Infrared Cameras (NIR) is a common approach to address the issue of strong light variations in a large number of studies. NIR cameras, however, come with their own set of disadvantages. To begin, specular reflections can occur when using active lighting to assist face imaging using NIR cameras. This results in inaccurate localization, alignment, and recognition of the eyes [46]. Specular reflections on eyeglasses can create big occlusions in the area where pupils are located. Besides this, frames captured by NIR cameras lack color information, a crucial feature for landmark and face detection libraries to work accurately and efficiently during occlusions. Color information is valuable for distinguishing between skin color and the surrounding environment, which helps the model understand the pattern and improves the reliability and speed of processing time during occlusions. This explanation is not intended to dissuade the reader from using NIR cameras; rather, it is intended to point out the limitations of these cameras. Since RGB and NIR both have their benefits and drawbacks, an interchangeable RGB and NIR camera system can be considered to be implemented in future works. For this project, only an RGB camera frame was utilized, and an algorithm-related approach was used to deal with the light variations.

In terms of the characteristics of the frame itself, it is important to note that size and compression can have a significant impact on the performance of machine-learning models. First, larger image sizes will require more memory and computational resources to process, which can slow down training and inference times. Additionally, large images may contain more noise and less relevant information, which can negatively impact model performance. Compression, on the other hand, can affect the quality of the image and the information it contains. Lossy compression, in particular, can remove important details and features from the image, which can negatively impact model performance. Lossless compression, on the other hand, preserves all of the information in the image but still increases the image size, which affects the memory and computational resources. It is essential to strike a balance between image size and compression to avoid a negative impact on the performance of machine-learning models. The current project aims to develop models that can handle variations in image size and compression by normalizing the feature vector parameter values.

As previously stated, in gray-scale images, time processing increases and decreases performance and precision while detecting faces and landmarks inside highly challenging frames. As a result, it was necessary to alter the image Lab's space to normalize the brightness. The following section will detail how it is normalized in further depth.

3.2 Lab Color Space

The Lab* color space is a method for representing and manipulating colors using three parameters or axes: L, a, and b. These axes calculate a numerical value for each color, ensuring that no two hues have the same number.

The L axis represents lightness and ranges from 0 to 100. The a-axis represents the red-to-green spectrum, with negative values representing green and positive values representing red. The b-axis represents the yellow-to-blue spectrum, with negative values representing blue and positive values representing yellow.

The Lab* color space is characterized by three color attributes: hue, saturation, and brightness. Hue refers to the color itself and changes as it moves around the Lab* diagram. Saturation refers to the vividness or dullness of a color, with more vivid colors located further from the center of the color wheel. Brightness refers to the lightness or darkness of a color.

The a* and b* axes define a color's hue and saturation by forming a Cartesian plane, while the L axis represents the color's brightness and is perpendicular to the ab* axes.

There are several advantages on using the Lab* color space. First, it has a larger color space than other color spaces, meaning it has a broader range of colors. Second, it is an independent color space, making it a useful resource for color management and conversions. Finally, it allows for the brightness of each pixel to be manipulated without affecting the color value, making it particularly useful in this study.

3.3 CLAHE

Histogram equalization (HE) is a common technique for improving image contrast, according to [47]. However, it can result in overexposed highlights, large contrast differences, and unusual pixel distribution in the image. To address these issues, this work used CLAHE, which operates on small sections of the image called tiles and expands partial histograms to enhance pixels without causing large contrast differences. Bilinear interpolation is used to blend the surrounding tiles and remove false boundaries. CLAHE can also adjust the histogram range to eliminate any artificiality in the enhanced image [48].

While CLAHE is typically used on grayscale images, this work applied a different approach to improve performance and precision in detecting faces and landmarks in challenging frames, as explained previously. To adjust the image's brightness, the Lab color space was used, and the CLAHE algorithm was applied to the "L" channel of the image in the Lab space.

In a standard HE process, the image's contrast is enhanced by using a mapping function that transforms the original intensity values into enhanced intensity values. The slope of this function, which determines the amount of contrast enhancement, can be controlled by adjusting the height of the histogram at a particular gray level [49], as shown in (1). To prevent the enhancement of pixel intensity transitions caused by noise, the slope of the mapping function can be limited in the process. The contrast-enhanced intensity S_k is directly proportional to the cumulative probability density, and the transformation of the slope at any gray level is related to the histogram height at that gray level. By clipping the height of the histogram to a threshold, the slope of the mapping function can be restricted [49].

$$S_K = \frac{I-1}{M \times N} \sum_{j=0}^{k} n_j \quad k = 0, 1, 2, ..., I-1 \, (1)$$

where:

- $I - 1$ is the highest intensity attainable level.
- $M \times N$ corresponds to the pixels' total number.
- n_j corresponds to the occurrences number of the jth intensity.

The image is separated into non-overlapping contextual sections or tiles in CLAHE. Each tile's histogram is trimmed to a user-specified clip limit. The clip limit is a multiple of the average height of the contextual region's histogram. The histogram's average height is the ratio of the total number of pixels in the contextual area to the number of gray levels. The local histogram is then computed. In this proposal, was used the following clip limit and tile grid size:

- Module 1's *Face Frontal Model* and Module 2: Tile grid size of 5x5 pixels and clip limit of 100.0.
- Module 1's *Face Profile Model*: Tile grid size of 1x1 pixels and clip limit of 50.0.

After CLAHE was applied to the image light ("L") channel in the Lab color space, the enhanced channel was combined with the "a" and "b" channels, and the frame was converted back to the RGB space. This process results in more evenly and uniformly distributed intensity values, creating a more balanced luminance across the image and making the details of the driver's landmarks clearer. The use of CLAHE can be seen in Figures 3.1 and 3.2.

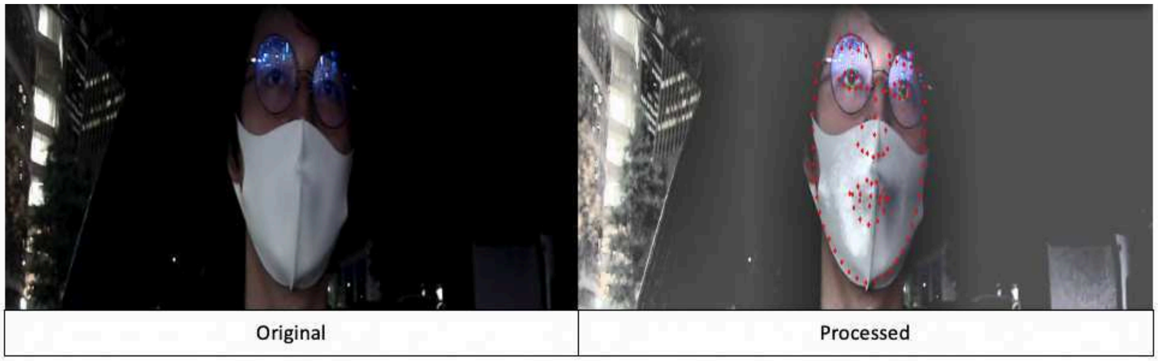


Figure 3.1: Frames before and after applying CLAHE, landmark detection (red), and pupil detection (green) for Frontal Face.

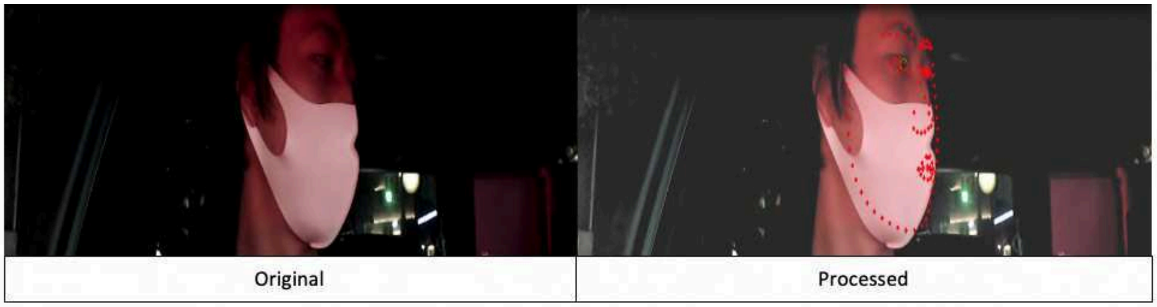


Figure 3.2: Frames before and after applying CLAHE, landmark detection (red), and pupil detection (green) for Profile Face.

Chapter 4

Robust Libraries for Data Extraction

This chapter will introduce the libraries that were chosen to be utilized for extracting the data that would serve as a base for making the different feature vectors for the models that have been implemented and introduce the features that become useful to the present work.

This appears to be a minor step, but it is the key difference between a model that will outperform and one that will not. The reduction of noisy data is critical not only for creating a better performance model but also for reducing on a great scale the model's training data.

4.1 Face Recognition

Face detection is the first and most important phase in many applications that deal with faces in some way, including face alignment, face recognition, face tracking, and many more.

This study requires robust face recognition as it aims to develop classifiers that can accurately classify in difficult-to-recognize situations, such as when they are wearing masks. [10] list the most robust face detection classifiers, all of which are based on RGB images. From this list, the Single Shot Scale-invariant Face Detector (S^3FD) face classifier was selected for use in this study. S^3FD is an anchor-based real-time face detector that detects several pre-set anchors created by tiling a series of boxes with different sizes and aspect ratios on the image at regular intervals and classifying and regressing them. These anchors are associated with one or more convolutional layers, whose spatial size and stride size determine the position and interval of the anchors, respectively. The anchor-associated layers are convolved to classify and align the corresponding anchors. Anchor-based detection methods are more robust in complex scenes, and their speed is invariant to the number of objects compared to other methods.

One important feature of this classifier, which is very useful in this study, is its special dedication to ease the outer faces' miss-classification. They employ a technique that they refer to as a *anchor matching strategy.* In anchor-based detection frameworks, anchor scales are often discrete; nevertheless, the face scale is continuous. So, those faces whose scales distribute in a direction different from the anchor scales are unable to match sufficient anchors, which results in a low recall rate for those faces. As an example, tiny and outer faces. Outer faces often occur when occlusions occur. To solve this, they offer a scale compensating anchor matching approach with two stages to increase the recall rate of these faces that were overlooked. The first step uses the current anchor matching algorithm but modifies the threshold to make it more flexible. Through scale compensation, the second stage ensures that each scale of faces matches adequate anchors.

Since this research must be conducted in real-time, another crucial aspect of the library that must be considered is speed processing. S^{3}FD outperforms other benchmark face detectors by a large margin across the different face detection benchmark datasets as Annotated Faces in the Wild (AFW), PASCAL face, Face Detection Dataset, and Benchmark (FDDB), and WIDER FACE running at 36 FPS on an Nvidia Titan X for VGA-resolution images [38], [10].

4.1.1 Face Recognition Libraries Comparison

It was important to analyze the performance of various face detectors commonly used in current literature, both under normal conditions and when faces are occluded. This serves as the foundation for the rest of the data extraction process. This analysis allows an understanding of how well the face detectors can recognize and detect faces under different conditions. Furthermore, this analysis gives an insight into how the performance of a classifier can greatly vary when tested under constrained and unconstrained conditions. This information is essential for understanding the limitations and potential of these face-detection methods and can aid in developing more robust and reliable algorithms.

To better understand how to face detectors perform under different occlusions, was conducted an analysis of various face detectors commonly used in current literature. As a starting point, the detailed research conducted by Wilber [50] explains some important details on this topic since makes a classification occluding various key facial features such as eyes, ears, nose, and mouth. This research showed that face recognition performance was lowest when the nose was covered, a common occlusion when using masks or scarves. This is corroborated after experiments with different face detectors used in current literature, concluding that they struggle to detect faces when the nose is occluded by strong lights or masks/scarves. With this in mind, an experiment was designed to explore this issue further.

Introduction

This experiment aims to demonstrate the significant difference in performance between face detectors when dealing with scenarios with mask-wearing drivers and those without. The ultimate goal is to show that while face detectors commonly used in current literature perform well in scenarios with no mask-wearing drivers, they perform poorly when faced with a dataset containing mask-wearing drivers.

To test this, are used five face detectors on two different datasets: one containing

Table 4.1: Results of the Running Time (RT) in seconds, and Accuracy (Acc) in percentage in decimal form of diverse Driver Face Detectors under No Mask-wearing and Mask-wearing datasets.

Face Detector	No Mask-wearing Drivers		Mask-wearing Drivers	
	RT	Acc	RT	Acc
Dlib HOG	0.29	0.83	0.031	0.01
Dlib CNN	3.03	0.81	0.024	0.48
Viola Jones	0.13	0.13	0.042	0.00
S^3FD	0.07	1.00	0.049	0.99
Retina	0.03	1.00	0.050	0.98

no mask-wearing drivers and one containing mask-wearing drivers. These five face detectors are Dlib HOG (HOG) for Frontal Faces, Dlib CNN (CNN), Viola Jones, S^3FD, and RetinaFace [51] (Retina). The first three face detectors are commonly found in current literature, while the latter two are newly proposed for use in this study.

Dataset

Different RGB camera frames of different videos taken in a real car with and without movement were part of the dataset.

- No mask-wearing dataset: consists of 1138 image frames. Video frames contained in this dataset are from 2 females and 8 males.

- Mask-wearing dataset: consists of 1120 image frames. Video frames contained in this dataset are from 3 females and 7 males.

The images inside the testing dataset does not include no face showing and there are few fully profile faces. Images can have noise on the pupil area as small eyes and pupils partially occluded (e.g., reflections in the eyeglasses or eyeglass sticks). There can be intense light enviroments.

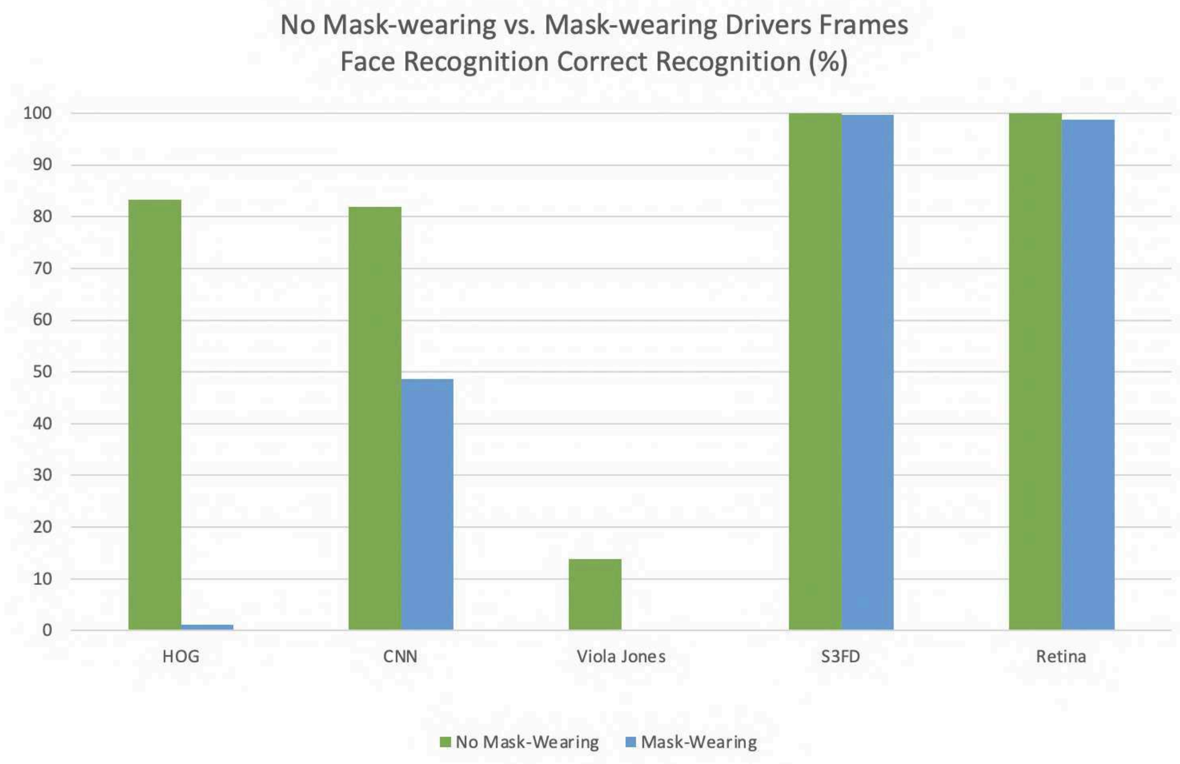


Figure 4.1: No Mask-wearing vs. Mask-wearing Drivers Frames Face Detection Correct Classification measured in percentage.

Metrics

The following metrics were used to evaluate the face detectors' performance:

Accuracy: Model's performance measurement based on how many instances it correctly recognizes. It is calculated by dividing the number of true positives by the total number of frames.

Average CPU Running Time: CPU processing time measured in seconds. As the proposed system aims to run in real-time for use in advanced driver assistance systems (ADAS) development, the CPU processing time of each experiment was evaluated.

Results for the Running Time (RT) and Accuracy (Acc) are shown in Table 4.1.

Figure 4.1 illustrates the accuracy of various face detectors when applied to datasets of driver images with and without masks. Some possible conclusions that can be drawn from this information:

- The performance of HOG, CNN, and Viola Jones varies significantly when detecting faces in datasets with and without occlusions (e.g., masks), while S^{3}FD and Retina have more stable performance.

- S^{3}FD and Retina both have good performance in detecting the overall shape of the face, with S^{3}FD having particularly constant boundary (i.e., the boxes around the detected faces are consistently sized) detection (Figure 4.2).

- S^{3}FD has the highest correct classification rate in the dataset of drivers wearing masks, at 99%.

- All the face detector models are already pre-trained models. The poor performance of Viola-Jones may be due to the fact that the acquired pre-trained model was not trained on a sufficient number of images. Viola Jones method should have a similar performance to HOG. It uses an Integral Image and a Haar-like feature with an AdaBoost process to create a cascade classifier, while the HOG method uses a sliding window to extract HOG descriptors and apply a classifier to each image at various scales. If the classifier detects an object that looks like a face with sufficient probability, it records the bounding box of the window and applies non-maximum suppression [52].

- Retina has slightly better performance than S^{3}FD in detecting heavily occluded key points (e.g., face partially profile plus both eyes with a great noise in the pupils plus black shadow in part of the profile). Figure 4.3 shows an example of this.

- CNN performs poorly in low light conditions (Figure 4.4) and when the driver's face is rotated away from a frontal position when wearing a mask.

- When wearing mask CNN can understand that there is a face when the eyes are seeing frontal. However, if the eyes are seeing in another direction it fails the face detection (Figure 4.5). This does not happen when there is a situation when the nose and mouth is not occluded.

- Mask color may impact the performance of Retina, with brighter colors leading to better performance due to a closer match to skin color (Figure 4.6).

- S^{3}FD and Retina has a better performance of the geometrical face shape. We can see that S^{3}FD and Retina makes a more rectangle shape that shows better the boundaries of the actual face. Both S^{3}FD and Retina are very precise, however until the frames that have been analyzed until now S^{3}FD has a very meticulous precision of the limits of the face (Figure 4.7).

- S^{3}FD seems to be more robust when there are masks and the driver is closing the eyes.

- Even though CNN correctly detects the face in different cases, when the driver closes the eyes cuts half the bounding box of the face detection which can lead to a mistake for the next steps as face landmark detection (Figure 4.8).

- All the face detectors performed better on average under the dataset while drivers are not using masks (Figure 4.9).

- HOG, Viola-Jones, and CNN can have a significant change in performance when it comes to detecting faces under difficult situations, such as those with occlusions. This is because it can be more challenging for the models to identify the face pattern under these conditions. On the other hand, S^{3}FD and Retina tend to be more stable and consistent in their performance regardless of the context in which the face is being detected. This suggests that these detectors may be better suited for handling difficult or unconstrained conditions.

Figure 4.10 shows the running time performance of the different face detectors' performance under no mask-wearing and mask-wearing drivers frames datasets. Different statements can be concluded:

- Dlib CNN face detector is the slowest face detector while Retina is the fastest one.

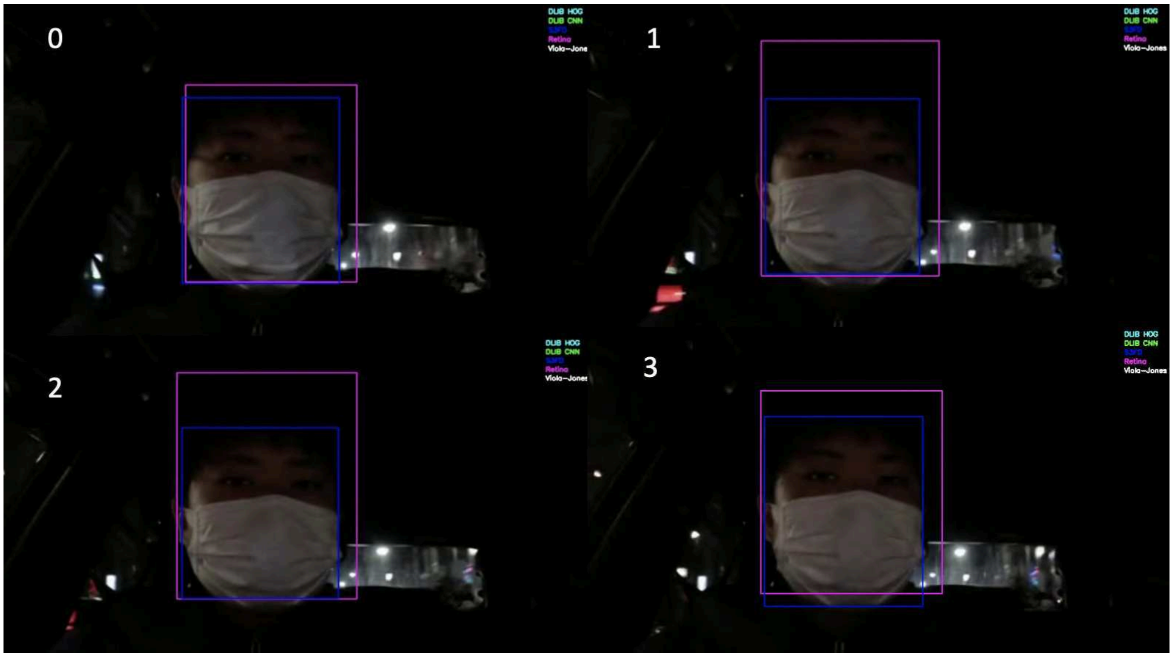


Figure 4.2: In datasets with occlusions, S^{3}FD has a better bounding box stability (i.e., the boxes around the detected faces are consistently sized) than Retina. White numbers are the correspondent frame number.

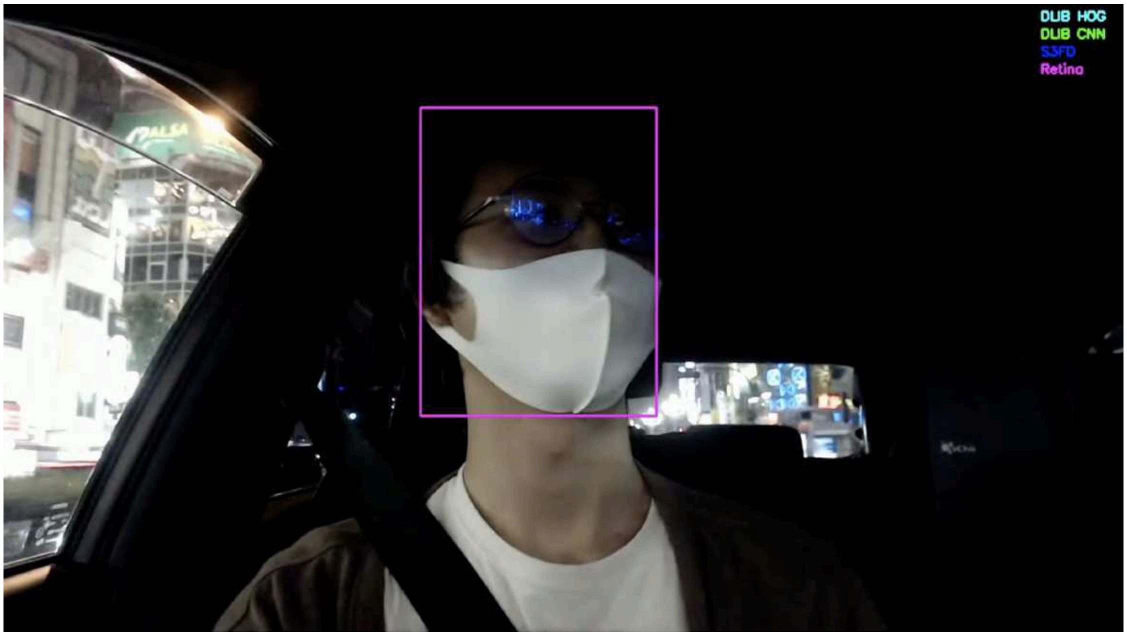


Figure 4.3: Retina has slightly better performance than S^{3}FD in detecting heavily occluded keypoints (e.g., partially occluded eyes and nose).

- HOG and CNN time performance change greatly when come to detecting faces under difficult situations. This is because searching the face pattern under

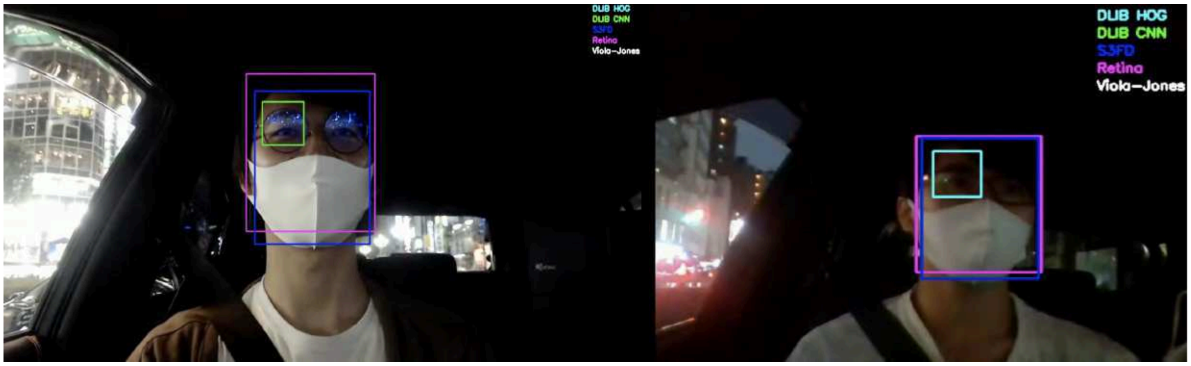


Figure 4.4: CNN performs poorly in very low light conditions.

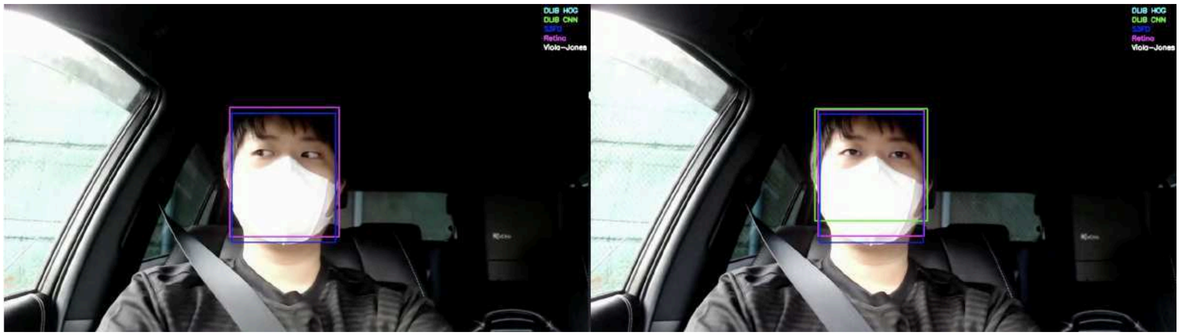


Figure 4.5: CNN while detecting under a dataset of mask-wearing drivers may turn into misdetection if the eyes are seeing in another direction. This does not happen when there is a situation where the nose and mouth are not occluded.

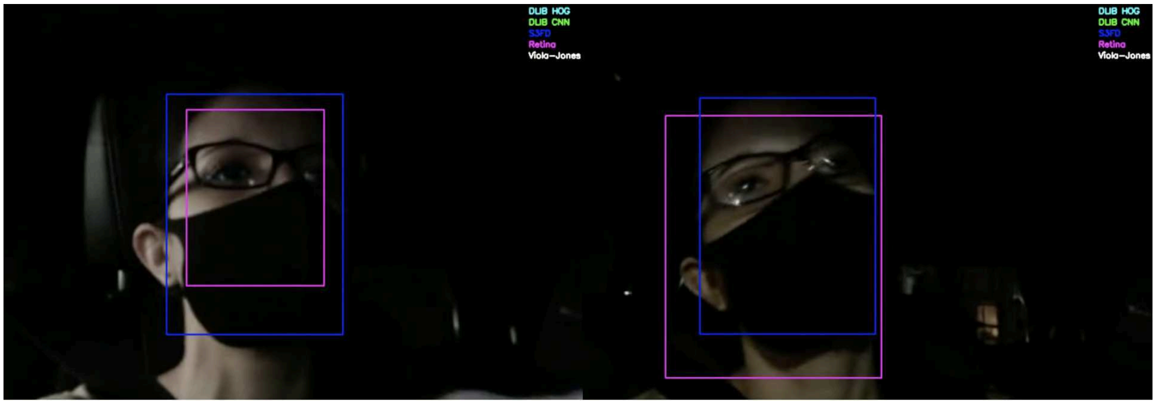


Figure 4.6: Mask color may impact the performance of Retina, with brighter colors leading to better performance due to a closer match to skin color.

occluded faces is much harder for the models. Viola-Jones, S^{3}FD, and Retina are quite stable regardless of the context in which the face is being detected.

- Even though Retina is the fastest, S^{3}FD's processing time is nearly the same

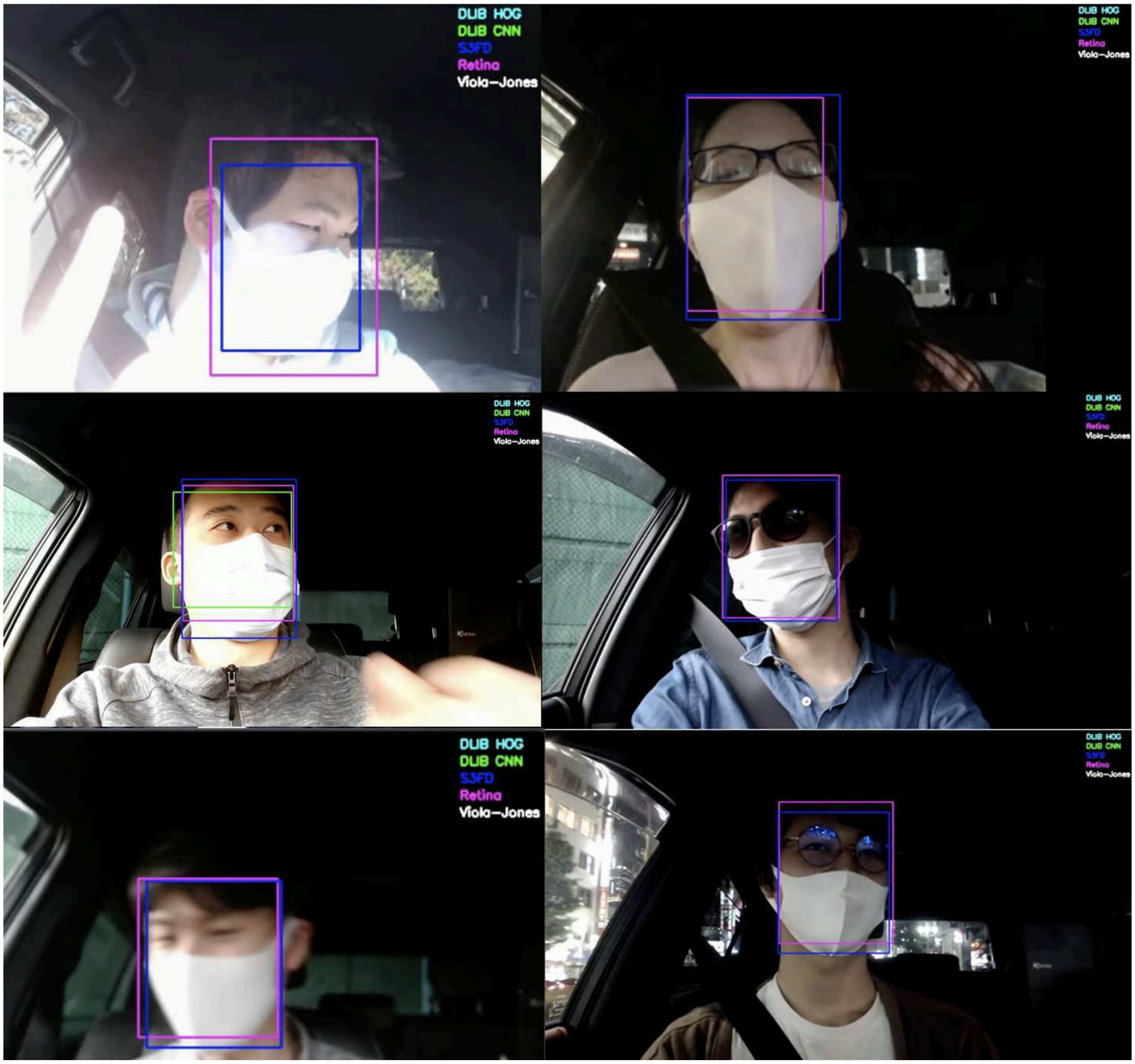


Figure 4.7: Both S^{3}FD and Retina are very precise, however until the frames that have been analyzed until now S^{3}FD has a very meticulous precision of the limits of the face.

as Retina.

As a general resume, to have a fair comparison it is important to evaluate these models on datasets that include a diverse range of images, such as those with occlusions, light differences, and other highly unconstrained situations. These situations can greatly affect the performance of a model, and therefore should be considered when comparing the effectiveness of different face detection models.

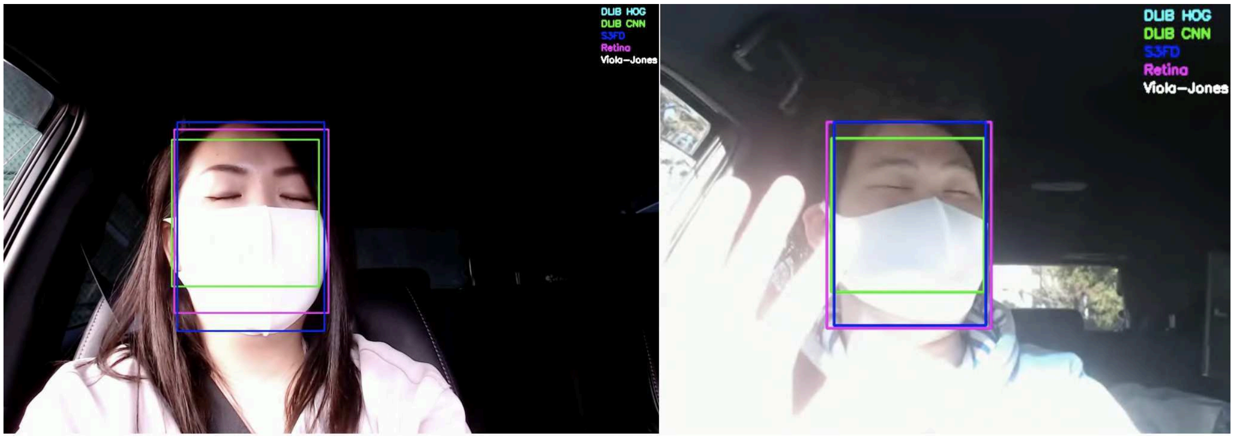


Figure 4.8: Even though CNN correctly detects the face in different cases, when the driver closes the eyes cuts half the bounding box of the face detection, which can lead to a mistake for the next steps as face landmark detection.

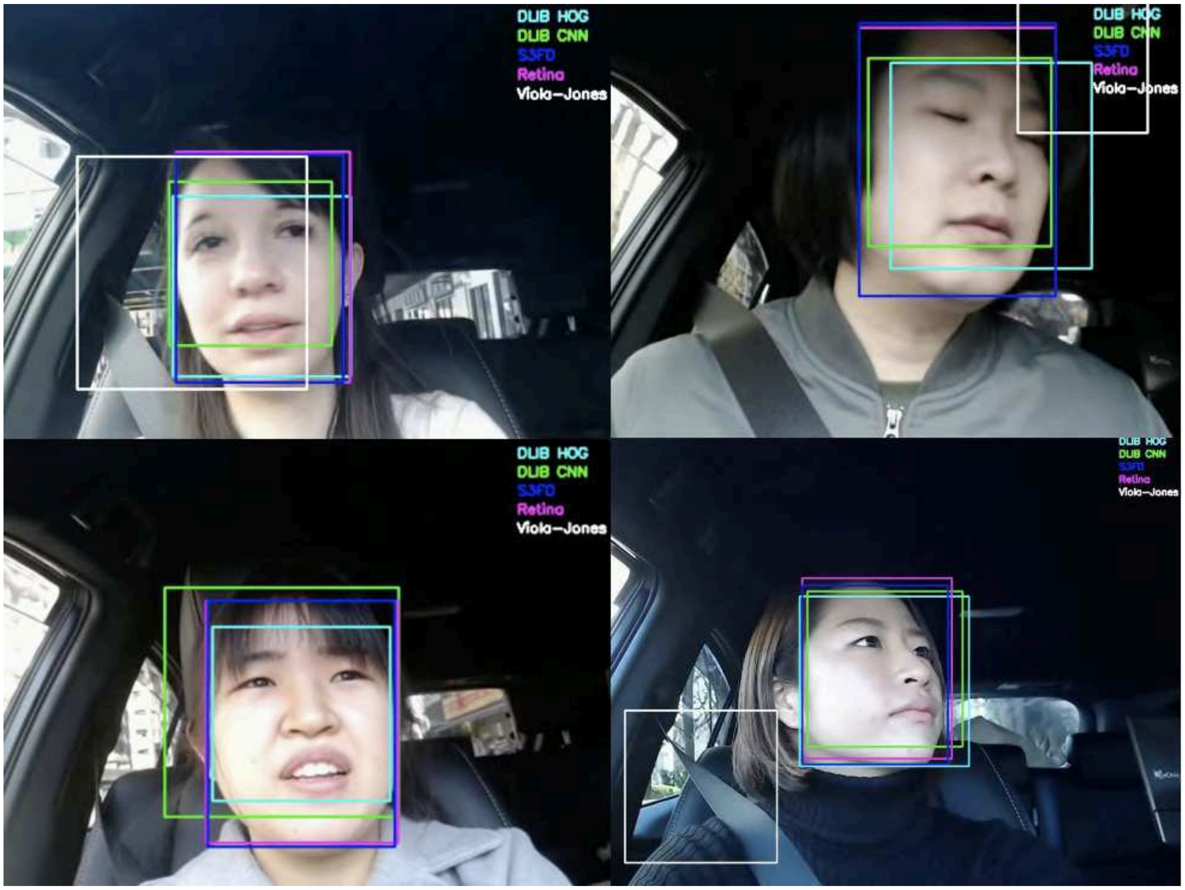


Figure 4.9: Face detectors performance in a dataset that is not under a mask-wearing situation.

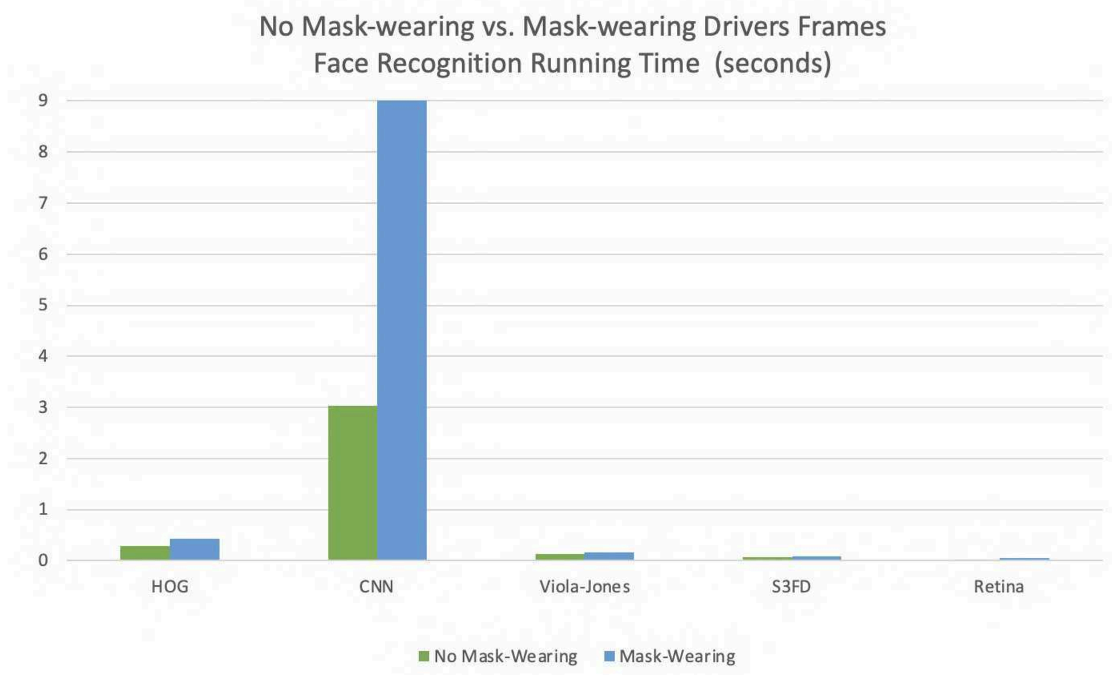


Figure 4.10: No Mask-wearing vs. Mask-wearing Drivers Frames Face Recognition Running Time Measured in seconds.

4.2 Face Landmark Recognition and Pupil Extraction

It is essential to have a robust and reliable system for detecting landmarks and pupils as serves as the foundation of many eye gaze tracking systems. [53] presents a real-time, accurate method for 3D eye gaze capture that overcomes the limitations of using random forest classifiers for this purpose. To achieve this, they propose using deep convolutional neural networks (DCNNs) to automatically extract iris and pupil pixels from input frames. These DCNNs are constructed using the capabilities of Unet and Squeezenet, and are used for pixel classification.

The process begins by automatically identifying relevant 2D facial features and optical flow constraints for each frame. These features and constraints are then used to recreate 3D head positions and large-scale facial deformations through the use of multilinear expression deformation models. The fast optical flow estimate approach is applied to the surrounding frames of the input video within the face region to extract the motion flow, which is a crucial element in this reconstruction. The landmarks are aligned with facial expressions in this case, indicating that it is a dynamic recognition system that adapts to the circumstances. This is particularly useful for tracking the movement of eyelids or the lower-face contour, for example. The facial landmark detectors that are currently utilized in academic literature lack this specific feature, and in general, this feature is considered to be quite uncommon among facial landmark detectors.

Other studies have often approached the problem of pupil detection as a regression problem in the context of iris and pupil recognition. However, eye movement can be complex, with features such as fixation, saccades, and smooth pursuit. A temporal tracking approach may be prone to error accumulation when an eye saccade occurs. To address this issue, [53]'s approach employs a DCNN-based segmentation method to extract iris and pupil pixels on a per-frame basis and tracks the eye gaze to update the eye state. This distinguishes their approach from other state-of-the-art methods, which may not consider the complexity of eye movement.

4.3 Body Posture

The pose tracker used to obtain the shoulder joints was implemented by [54][55][56][57] in their work on AlphaPose. They have developed an online optimization framework called PF-Builder (pose flow builder) that associates cross-frame poses and constructs pose flows, which represent a sequence of poses of the same person instance in different frames. Pose flow is a valuable feature for future research in this field. To improve the robustness of the pose flow, [54] also proposes a novel pose flow non-maximum suppression (PF-NMS) method to reduce redundant pose flows and re-link temporal disjoint ones.

PF-Builder iteratively constructs pose flows from pose proposals within a short video clip selected using a temporal sliding window. Instead of using a greedy matching approach, it uses an efficient objective function to find the pose flow with the highest overall confidence among the feasible flows. This optimization strategy helps to stabilize the pose flows and associate discontinuous ones (due to missing detections). On the other hand, PF-NMS uses pose flow as the unit of processing in the non-maximum suppression (NMS) process, rather than using NMS at the frame level like traditional techniques. As a result, temporal information is fully considered in the NMS process, leading to improved stability. These features significantly contribute to the robustness of the categorization in this study.

Chapter 5

Module 1: Gaze Classifier

5.1 Overview

Driver's inattention and distraction are major contributing factors in fatal, and injury crashes [58]. Gaze zone classifiers in ADAS have been key research for distraction detection support [59]. If an ADAS can detect a driver's inattention, the car's system can warn the driver or, in the near future, take control of the situation, reducing the number of fatal accidents. There are vast implementations of driver's gaze classifiers; however, these studies employ methodologies that may highly fail when challenging situations occur. Some of these challenging situations are face occlusions, eyeglasses reflection, strong daylight variations, profile face poses, and face and eyes facing different directions, as shown in Figure 5.1. Moreover, many use expensive sensors, equipment sensitive to light, or complex systems.

Furthermore, no former research explicitly does experiments with highly unconstrained datasets. These edge conditions recurrently appear in driving scenarios, the reason why driver's gaze classifiers urge a more robust implementation. To overcome part of this problem, this module presents a single-camera driver's gaze zone classifier approach that can robustly make a correct classification during various beforementioned complex situations.

This study involves three key steps to achieve a robust system:

1. Frames' Lab's color space manipulation: To address the strong light variation

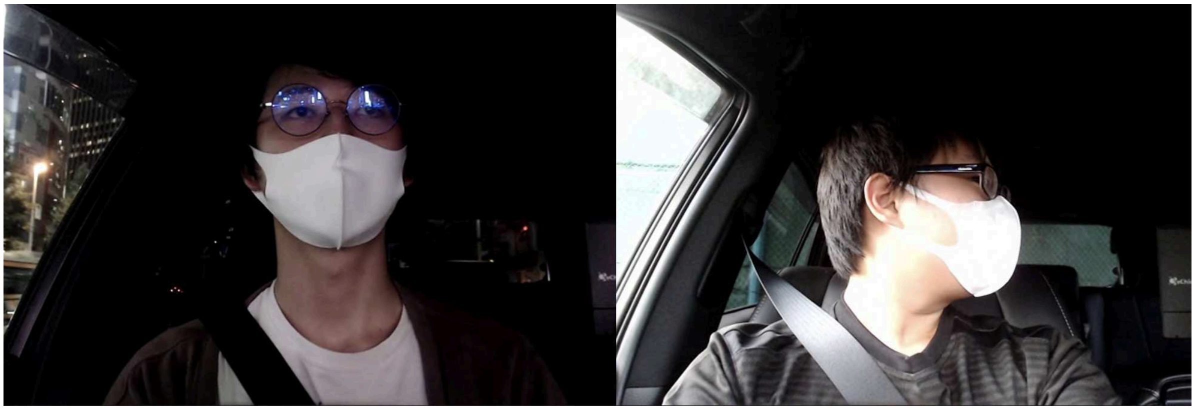


Figure 5.1: Left: Driver's face with a mask, eyeglass reflection, and low brightness environment. Right: Driver profile face with glasses, eyes, and face facing different directions.

issue, the brightness of the frames are equalized by manipulating its Lab's color space's luminance channel using a CLAHE. This process was covered in Chapter 3.

2. Usage of robust libraries for face, facial landmarks, pupil, and eyelid movement detection: After having frames with balanced lighting, robust recognition of the face, eyes, and pupil landmarks is achieved by combining an anchor-based real-time face detector with a dense landmark alignment that includes optical flow estimation methods for pupil and eyelid movement tracking. This process was covered in Chapter 4.

3. DNN models structure and feature vectors parameters: This framework involves two main models, *Face Frontal* and *Face Profile* DNN models. Since the geometric facial structure varies per person, the feature vector parameters consist of different relations between pupil and eye landmarks in proportion to the driver's geometrical face configuration. Each model's possible output is summarized in Table 5.1. Face and eyes facing different directions are considered. *Face Profile* Model can discriminate between the right and left sides. The occlusions caused by profile faces may be harmful data for the model, so *Face Frontal* and *Face*

Table 5.1: Gaze Group Labels evaluated in the DNN models: Face Front DNN Model and Face Profile (Window Direction) DNN Model. Face Profile DNN Model fits for both sides, right and left window.

Face Front Model			Face Profile Model		
Face Direction	Eyes Direction	Label	Face Direction	Eyes Direction	Label
Front	Front	FF	*Profile (Window)*	Window	PW
	Right Window	FRW		Front	PF
	Back Mirrow	FBM			
	Speed Meter	FS			
	Navigator	FN			

Profile models were separated. This chapter will go into extensive detail on the steps involved in this procedure.

A flowchart of the system is shown in Figure 5.2.

The main contribution of this module is a gaze classifier that includes:

1. Robustness to strong light condition variations and various facial occlusions as could be: masks, scarves, eyeglasses reflections, eyeglasses sticks, small eyes, partial occluded pupils, and profile face poses.

2. Portability and extensivity, as it needs only one camera and a computer.

3. Generalization, as it can classify correctly regardless of the subject.

Moreover, this study will compare the performance of the proposed system in contrast to the general approach and show the importance of each stage of the proposed pipeline to achieve good results over a dataset involving highly unconstrained driving conditions.

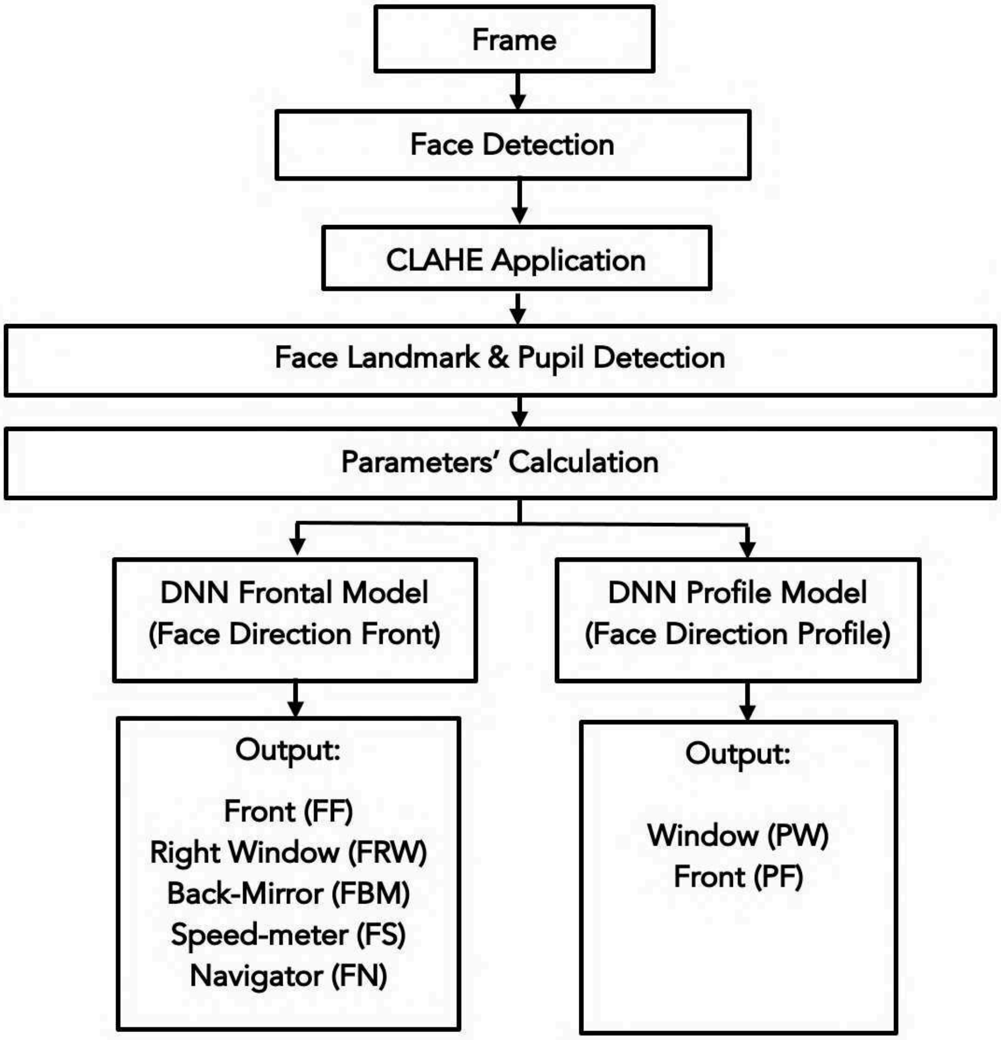


Figure 5.2: System's Diagram. The system is built on a single frame. The system's inputs a frame, detects the face and use CLAHE to equalize its brightness. Then a facial landmark and pupil are detected. After extracting this data, are calculated the parameters, which will be passed to either the frontal or profile face model depending on the face position. There are five possible outputs for the frontal model and a binary output for the profile face.

5.2 Feature Vector

The quality of the input data is a crucial factor in the success of a DNN model, particularly when the model must handle data with strong variations. Many studies underestimate the importance of data quality. Focusing on improving the quality

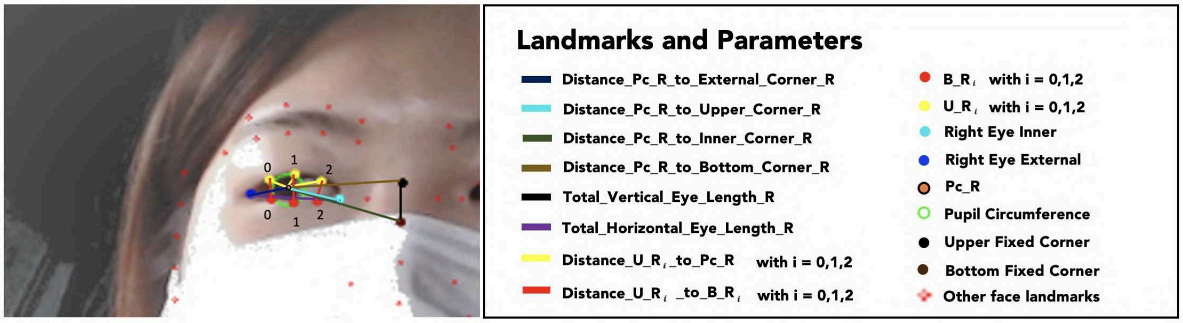

0
1
2
0
1
2
Landmarks and Parameters
Distance_Pc_R_to_External_Corner_R
Distance_Pc_R_to_Upper_Corner_R
Distance_Pc_R_to_Inner_Corner_R
Distance_Pc_R_to_Bottom_Corner_R
Total_Vertical_Eye_Length_R
Total_Horizontal_Eye_Length_R
Distance_U_R$_i$_to_Pc_R with i = 0,1,2
Distance_U_R$_i$ _to_B_R$_i$ with i = 0,1,2
B_R$_i$ with i = 0,1,2
U_R$_i$ with i = 0,1,2
Right Eye Inner
Right Eye External
Pc_R
Pupil Circumference
Upper Fixed Corner
Bottom Fixed Corner
Other face landmarks

Figure 5.3: Different key landmarks and parameters used in the proposed feature vector.

of the input data can lead to better performance from a fixed model, including the reduction of the training data. With this in mind, great effort was put into ensuring that the input data was as clean as possible. Since the face is symmetrical, only one side of the face landmarks was used, with the system deciding which side to use. The center of the left and right eye pupils are referred to as Pc_L and Pc_R, respectively. If the pupil detector fails to fully recognize a pupil, its radius becomes smaller. The system compares the radii of both pupils and analyzes the landmarks of the side with the larger pupil radius. The input feature vector has eleven parameters. The different key landmarks and parameters used in the proposed method can be seen in Figure 5.3. The details of each parameter are:

Right Side Face Landmarks Case:

1. Pc_R radius length: Length of the radius r_r. Depending on the head rotation and proximity to the camera, the pupil radius varies.

2. Pc_R distance to the right eye external corner proportional with the total horizontal distance of the right eye:

$$ecp\_r = \frac{Distance\_Pc\_R\_to\_External\_Corner\_R}{Total\_Horizontal\_Eye\_Length\_R}$$

3. Pc_R distance to the right eye inner corner proportional with the total horizontal distance of the right eye:

$$icp\_r = \frac{Distance\_Pc\_R\_to\_Inner\_Corner\_R}{Total\_Horizontal\_Eye\_Length\_R}$$

4. Pc_R distance to the highest point of the fixed eyelid proportional with the total fixed vertical distance of the right eye:

$$ucp\_r = \frac{Distance\_Pc\_R\_to\_Upper\_Corner\_R}{Total\_Vertical\_Eye\_Length\_R}$$

   Note that the word fixed is used. As an eyelid movement tracker was used, the eye's total vertical size is always in movement. So was framed an approximation of the vertical size of the eye using its relation with the nose landmarks that are permanently fixed. This step is crucial because all faces are different. Using data relative to the face (proportions) reduces the data noise that causes having different faces with different measures and transforms the data as equally as possible.

5. Pc_R distance to the lowest point of the fixed eyelid proportional with the total fixed vertical distance of the right eye:

$$bcp\_r = \frac{Distance\_Pc\_R\_to\_Bottom\_Corner\_R}{Total\_Vertical\_Eye\_Length\_R}$$

6. Closure of the upperlid with respect bottomlid (3 parameters): By analyzing the data, was noticed that the eyelid closure gives a better clue to the algorithm to understand eyes' up/down movement than the pupil's position. The upperlid landmarks U_R_i are extracted, 3 for each eye. Consequentially, is measured the distance between the upperlid landmark and its proximal vertical equivalent bottomlid landmark B_R_i. The following equation represents the distance d:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (2)$$

   Then was calculated the absolute value for each $distance_i$, and was divided by the vertical eye distance. With each landmark l_i_r was made a Closure Parameter respect the bottomlid cb_i_r following the formula:

$$cb_i\_r = \frac{Distance\_U\_R_i\_to\_B\_R_i}{Total\_Vertical\_Eye\_Length\_R}$$

with $i = 0, 1, 2$

7. Closure of the upperlid with respect to the pupil (3 parameters): It is similar to the closure of the upperlid to the bottomlid calculation with the discrepancy that was measured the distance between the upperlid landmarks and the pupil position instead of the bottomlid. Equation (2) was used to make the distance calculation and is not used the absolute value of the difference for making the calculation. Having each landmark l_i, a Closure Parameter is made with respect to the pupil cp_i following this formula:

$$cp_i\_r = \frac{Distance\_U\_R_i\_to\_Pc\_R}{Total\_Vertical\_Eye\_Length\_R}$$

with $i = 0, 1, 2$

Finally for the feature vectors for each model are:

- Frontal Face Model: The Frontal Face model's feature vector is annotated as ff_fv.

  Right Side Face Landmarks Case:

  $$ff\_fv = \{r\_r, ecp\_r, icp\_r, ucp\_r, bcp\_r, cb_0\_r, cb_1\_r, cb_2\_r, cp_0\_r, cp_1\_r, cp_2\_r\}$$

  Left Side Face Landmarks Case:

  Since face landmarks are symmetric, the feature vector when using the landmarks of the left side of the face will be almost the same. The difference is that the results of positions two and three of the feature vector will be swapped, so the relations are kept the same (inner and external corners for right and left eyes are opposites). Therefore, the representation of the left side of the face landmarks is:

$$ff\_fv = \{r\_l, icp\_l, ecp\_l, ucp\_l, bcp\_l, cb_0\_l, cb_1\_l, cb_2\_l, cp_0\_l, cp_1\_l, cp_2\_l\}$$

- Profile Face Model:

  The Profile Face model's feature vector is annotated as pf_fv and consists in the exact same parameters as the *Frontal Face Model.*

  Right Side Face Landmarks Case:

  $$pf\_fv = \{r\_r, ecp\_r, icp\_r, ucp\_r, bcp\_r, cb_0\_r, cb_1\_r, cb_2\_r, cp_0\_r, cp_1\_r, cp_2\_r\}$$

  Left Side Face Landmarks Case:

  $$pf\_fv = \{r\_l, icp\_l, ecp\_l, ucp\_l, bcp\_l, cb_0\_l, cb_1\_l, cb_2\_l, cp_0\_l, cp_1\_l, cp_2\_l\}$$

5.3 DNN Models

5.3.1 Frontal Face

The model can have five outputs: eyes are facing front, eyes are facing right window, eyes facing speed meter, eyes facing back-mirror, and eyes facing navigator. The input is a feature vector whose parameters are relations between pupil and eyelid landmarks in proportion to the driver's geometrical face configuration. The details of the model are:

- Input: The input is one of the feature vectors fv described in Section 5.2.

- Network Topography and Hyperparameters: The topography and hyperparameters are illustrated in Figure 5.4 (a).

- Output: Probabilistic prediction of the direction of the eyes that can be front, right window, back-mirror, speed-meter, and navigator.

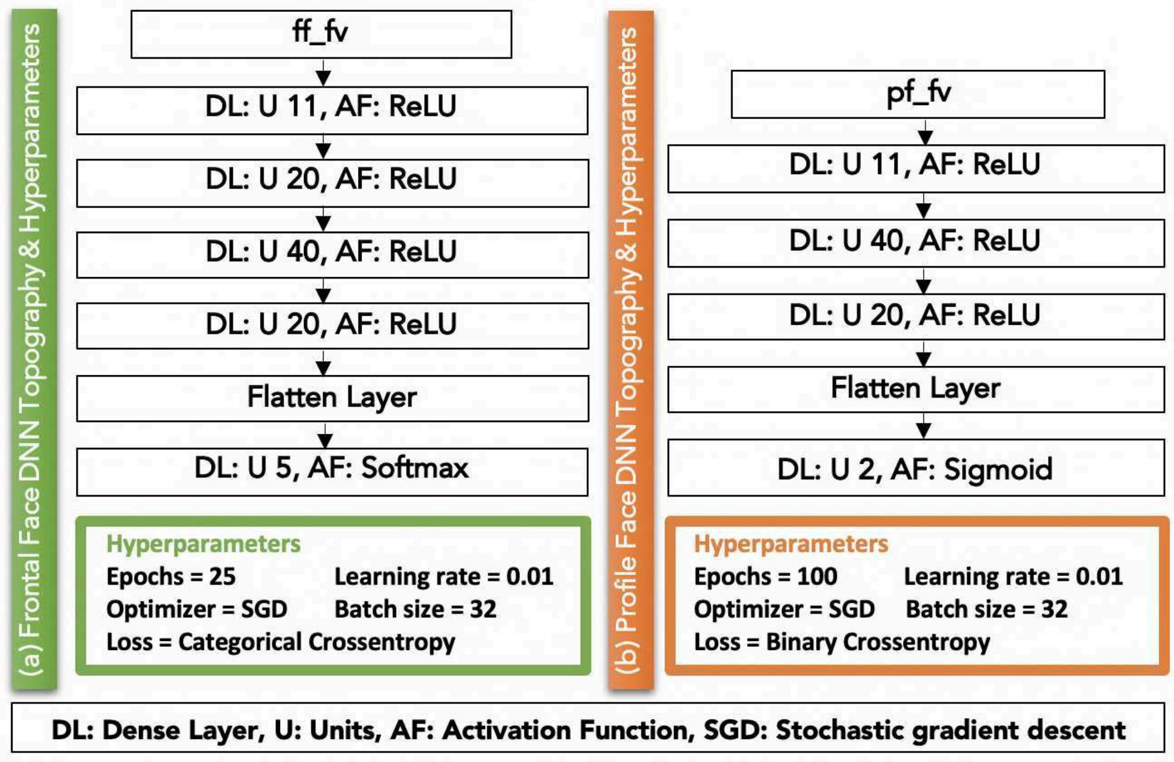


Figure 5.4: Proposed DNN Topographies and Hyperparameters. Left: Frontal Face DNN Model. Right: Profile Face DNN Model.

5.3.2 Profile Face

The model has a binary output: eyes are facing front, eyes facing window. The inputted feature vector parameters are different relations between the pupil and the eye landmarks in proportion to the driver's geometrical face configuration of the side that is not occluded. The data landmarks that was use to evaluate an instance are from the fully visible profile side of the driver. The system can decide which side is visible by checking the pupil radius size. The pupil radius size gets smaller when it gets more occluded or less visible. So, if the radius of the right eye pupil $r\_R < r\_L$, the driver sees at the right side, so was used the left side face landmarks to make the process and vice versa. The details are as follows:

- Input: The input is one of the feature vectors fv described in Section 5.2.

- Network Topography and Hyperparameters: The topography and hyperparameters are illustrated in Figure 5.4 (b).

- Output: Probabilistic prediction of the direction of the eyes (binary output) that can be *Window* or *Front*.

5.4 Experimental Evaluation and Results

5.4.1 Overview

This study aims to show the positive effect of each step of the proposed algorithm to correctly classify the driver's gaze zone under challenging settings, as can be partial sun reflection, face occlusions, eyeglasses reflection, and drastically daylight variations. For this, the system was evaluated in a challenging dataset with the before-mentioned conditions. The experimental evaluation is divided into two parts to demonstrate two key points: first, how the conventional methods fail in unconstrained situations, and second, how each feature of the suggested system contributes to the system's robustness.

Experiment 1 The goal of this experiment is to highlight the weaknesses of existing approaches in unconstrained scenarios. To demonstrate this, two different model structures of a widely used approach are tested, as there is no current state-of-the-art approach. Experiment 1 is divided into two sub-experiments for this purpose:

Experiment 1.a: A similar 2D CNN model architecture proposed in [60] where the input of the model will be the frame's face and eyes image data was used, with the difference that the section where each occluded pixel data is replaced with a same 8-bit integer value 255 (white) was not implemented as this step is not covered in the conventional method. This approach is one of the most often utilized in the literature. The testing dataset was evaluated without applying the frame luminance equalization step.

Experiment 1.b: An open-source code done by [61] that contains a Class Activation Map (CAM) to visualize how the model learns the patterns used to make the classification is used. Since is a broadly used model, different from the one already proposed, this model was chosen. Also since we can visualize the pattern the model observes, is very useful. The input is the full face of the driver image data. It was assessed without the frame luminance equalization step.

Experiment 2 The goal of this experiment is to highlight the significance of each proposed feature in the proposed methodology. Experiment 2 is divided into three sub-experiments to demonstrate this:

Experiment 2.a: Testing dataset was assessed without using CLAHE.

Experiment 2.b: Testing dataset was assessed without the parameters relative to the face proportion with the following adjustments:

- $ecp\_r = Distance\_Pc\_R\_to\_External\_Corner\_R$
- $icp\_r = Distance\_Pc\_R\_to\_Inner\_Corner\_R$
- $ucp\_r = Distance\_Pc\_R\_to\_Upper\_Corner\_R$
- $bcp\_r = Distance\_Pc\_R\_to\_Bottom\_Corner\_R$
- $cb_i\_r = Distance\_U\_R_i\_to\_B\_R_i$ with $i = 0, 1, 2$
- $cp_i\_r = Distance\_U\_R_i\_to\_Pc\_R$ with $i = 0, 1, 2$

  For the case when is used the right side of the face landmarks. The same modifications are made when is used the left side.

Experiment 2.c: Testing dataset was assessed using the full procedures proposed in this work.

5.4.2 Dataset

A single RGB camera frame of different videos taken in a real car with and without movement was part of the dataset. Twenty distinct participants - 5 females, 15 males

- participate in the experiments. For the training dataset, 3 females and 9 males data were used and for the testing dataset, 2 females and 6 males data was used. The training dataset contains around 700 images for each class but does not include images taken under challenging conditions. The training dataset was supplied by using [62]'s dataset to add extra data. As for the testing dataset, the subjects are different from the training dataset, each class containing around 100 images all of them from the taken own dataset. The majority of the images in the testing dataset have been taken under at least one of the following conditions: occlusions (such as masks or scarfs), reflections from eyeglasses, or intense lighting environments (excluding extremely dark or bright conditions where the face is not visible).

The images inside the testing dataset do not have strong variations of the driver's distance to the camera, nor strong variations in the *pitch, roll, yaw* of the driver's face, except for the profile face. Images can have noise on the pupil area as small eyes and pupils partially occluded (e.g., reflections in the eyeglasses or eyeglass sticks). Eyes' direction was recognized with profile faces as well.

5.4.3 Metrics

Predicted Results Confusion Matrix: In statistical classification, a Confusion Matrix is a layout that displays the performance of a model by showing how many instances were correctly classified by the model compared to the actual class labels. It helps to identify which specific instances were misclassified and how. The columns of the matrix represent the predicted labels, while the rows represent the ground truth labels.

Macro-average and Micro-average Accuracy: In general, accuracy is the percentage of correct predictions made by a model. A macro-average computes the metric individually for each class and then takes the average, therefore all classes equally contribute to the final averaged metric. Micro-average aggregates all classes' contributions to compute the average metric, therefore, all samples equally contribute to

Table 5.2: Results of Running Time (RT) in seconds, Macro-Average Accuracy (Mac-Avg) and Micro-Average Accuracy (Mic-Avg) in percentage in decimal form, for each Experiment (Exp) for Face Front Model and Face Profile Model.

Exp	Face Front Model			Face Profile Model		
	RT	Mac-Avg	Mic-Avg	RT	Mac-Avg	Mic-Avg
1.a	0.033	0.75	0.76	0.031	0.54	0.52
1.b	0.024	0.18	0.18	0.024	0.46	0.48
2.a	0.043	0.56	0.57	0.042	0.33	0.34
2.b	0.051	0.77	0.77	0.049	0.37	0.38
2.c	0.051	0.95	0.95	0.050	0.91	0.92

the final averaged metric.

The following equations correspond to micro-average and macro-average:

$$\text{Macro-average accuracy} = \frac{1}{N}\sum_{j=1}^{N}\frac{(\text{True positive})_j}{(\text{Total Population})_j} \tag{3}$$

$$\text{Micro-average accuracy} = \frac{\sum_{j=1}^{N}(\text{True positive})_j}{\sum_{j=1}^{N}(\text{Total Population})_j} \tag{4}$$

with N = Classes' number

Average CPU Running Time: The proposed system aims to run in real-time to be useful for on-current development ADAS, so the CPU processing time of each experiment was evaluated. It is measured in seconds.

5.4.4 Results

The running time, macro-average accuracy, and micro-average accuracy of each experiment are shown in Table 5.2.

Experiment 1:

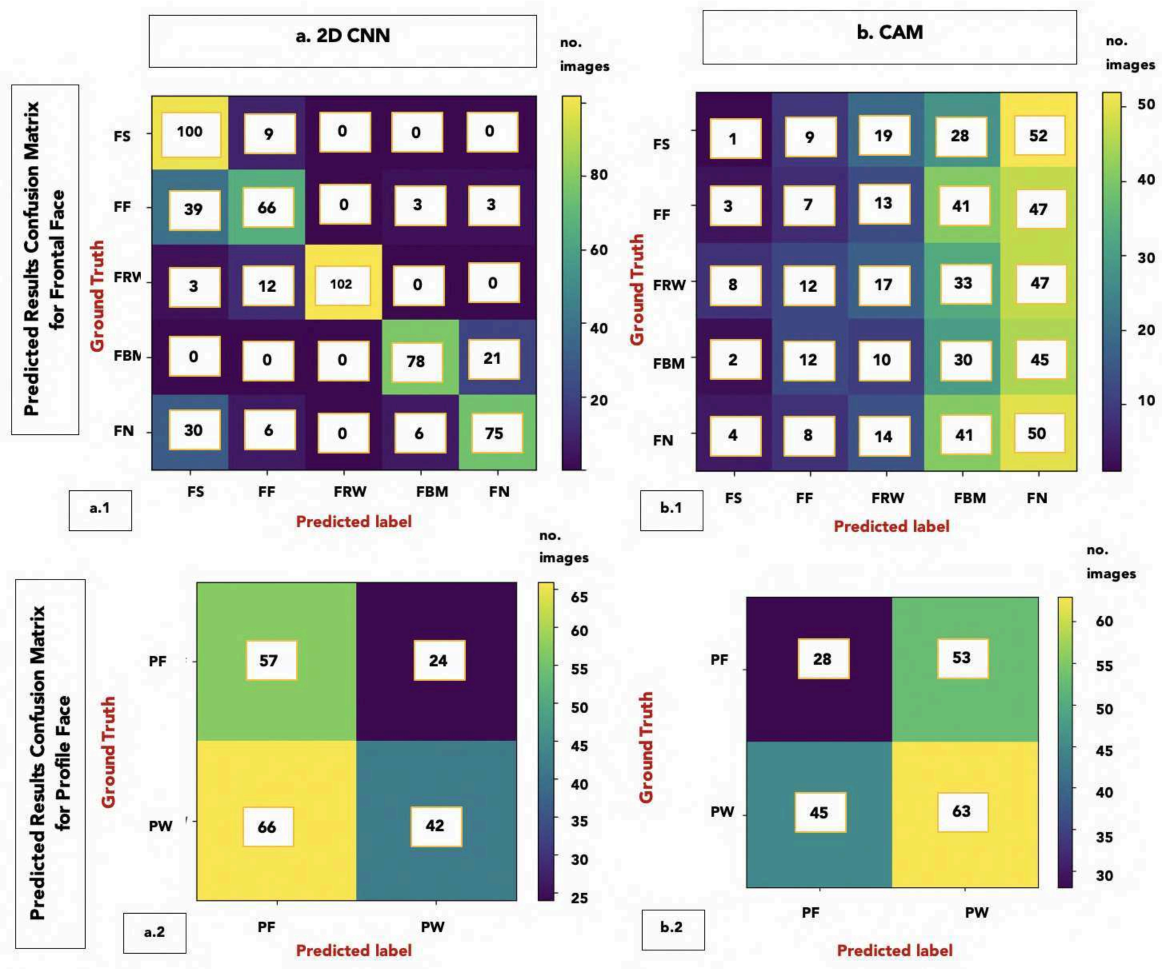


Figure 5.5: First row: Predicted Results Confusion Matrix for each experiment for the Frontal Face on Experiment 1. Second row: Predicted Results Confusion Matrix for each experiment for the Profile Face Model. From left to right: Experiment 1.a and Experiment 1.b.

Experiment 1.a: Figure 5.5 (a.1) shows that Frontal Face Model can distinguish correctly between classes. However, for the Profile Face Model, it still does not make a full correct classification.

Experiment 1.b: Figure 5.5 (b.1) and Figure 5.5 (b.2), show that for this model is hard to properly learn the underlying patterns in this challenging dataset. One probable explanation is that the only small nuance between each image is the eyes' direction. Also, some images have pupils partially occluded, adding difficulty in understanding the pattern. Only the face of the driver is used as an input. However,

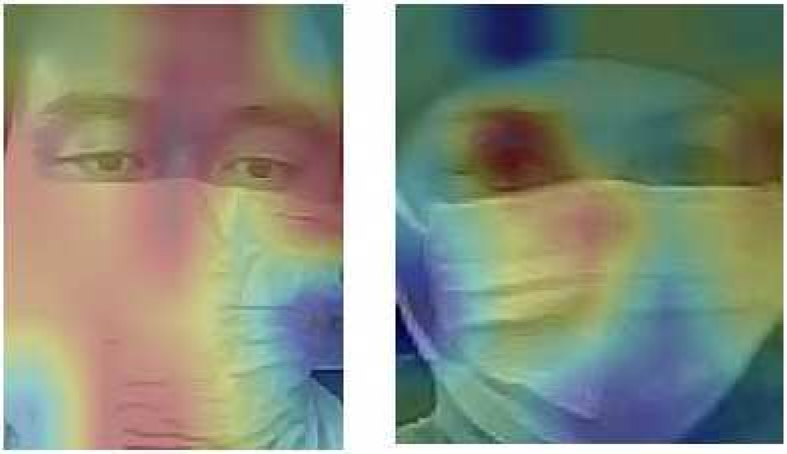


Figure 5.6: Example of CAM's performance. As can be observed, the network has not learned the underlying patterns in the unconstrained dataset effectively.

this model has the best timing comparing the rest because there is almost no pre-processing. Results of this model's CAM performance can be seen in Figure 5.6. This model performed worst among all models in terms of Macro-average accuracy and Micro-average accuracy for Frontal Faces, as shown in Table 5.2.

Experiment 2:

Experiment 2.a: Figure 5.7 (a.1), shows that the model has the lowest performance for the Frontal Face Model among the experiments made in Experiment 2. Figure 5.7 (a.2) shows poor performance when having flipped results. The landmark recognition accuracy was low because for the model is very hard to understand the pattern when having extreme intensities in one frame. So the pupil center context may fall into the wrong place. For Example, in FF, the pupil is typically just in the center. However, if the eye landmarks are slightly shifted to the left, the inner corner of the right eye will be in the ground truth position of the center eye of the right eye, and the model will make a miss-classification with FBM or FN. Figure 5.10 illustrates one example of this. The same pattern is observed across different classes. This result shows the importance of manipulating the luminance of the images. During driving scenarios, we have highly drastic light changes that should be filtered. This model performed worst

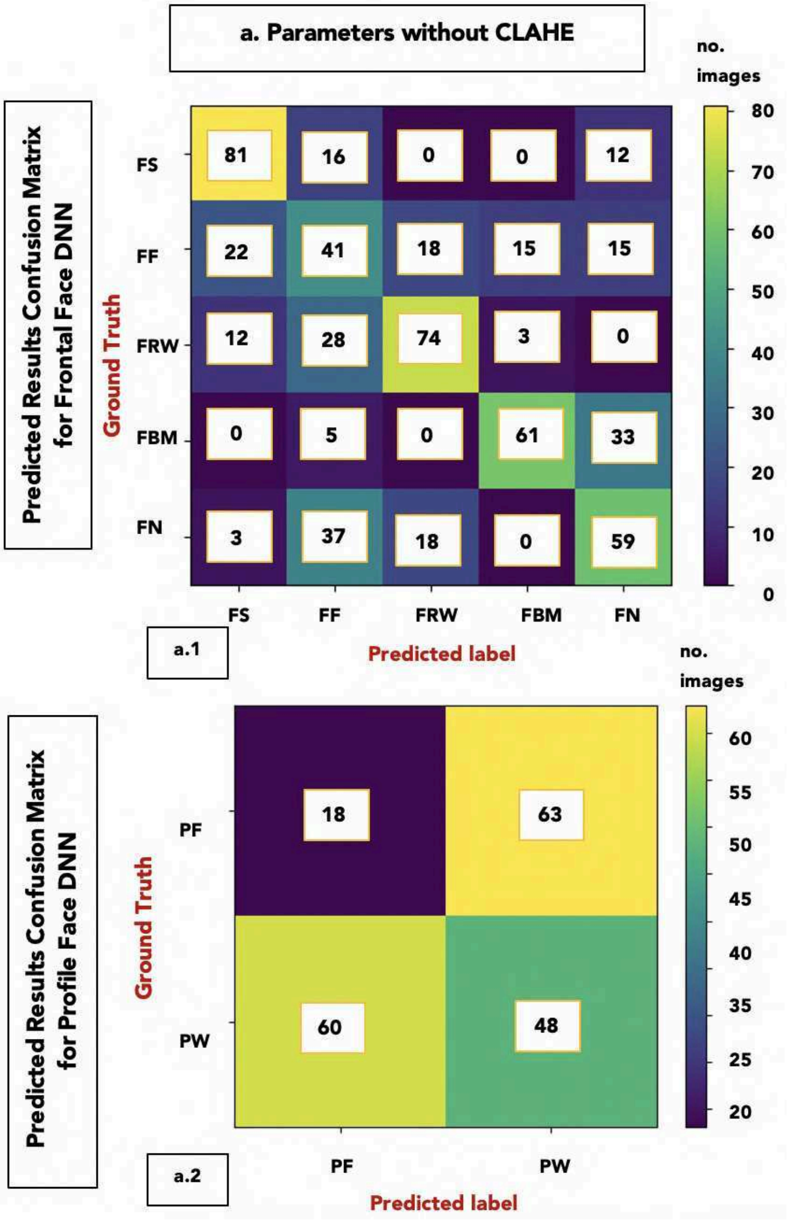


Figure 5.7: Experiment 2.a. First row: Predicted Results Confusion Matrix for the Frontal Face DNN Model. Second row: Predicted Results Confusion Matrix for Profile Face DNN Model.

among all models in terms of Macro-average accuracy and Micro-average accuracy for Profile Faces as shown in Table 5.2.

Experiment 2.b: In Figure 5.8 (b.1), we can observe that the model reached a much better performance than the previous experiment. However, Figure 5.8 (b.2) shows flipped results. Each person's face has differences, so the data is unequal when the

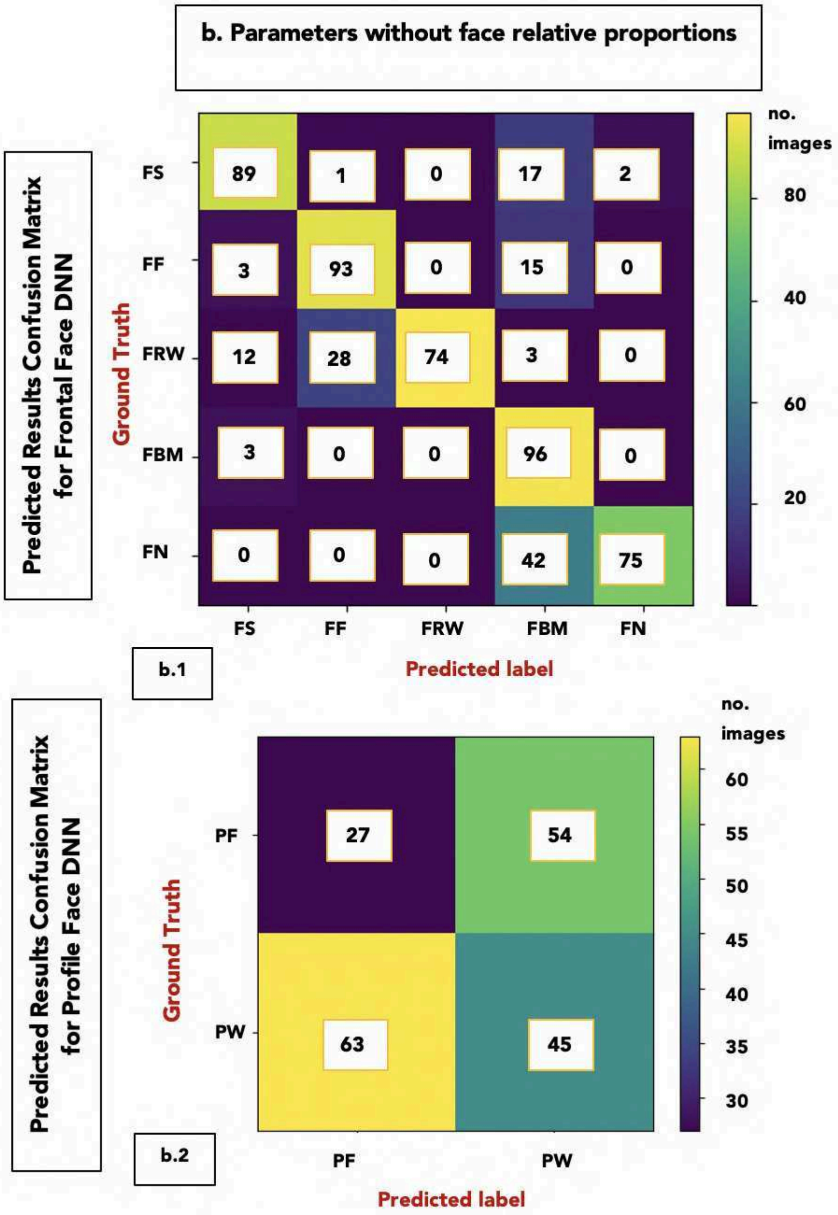
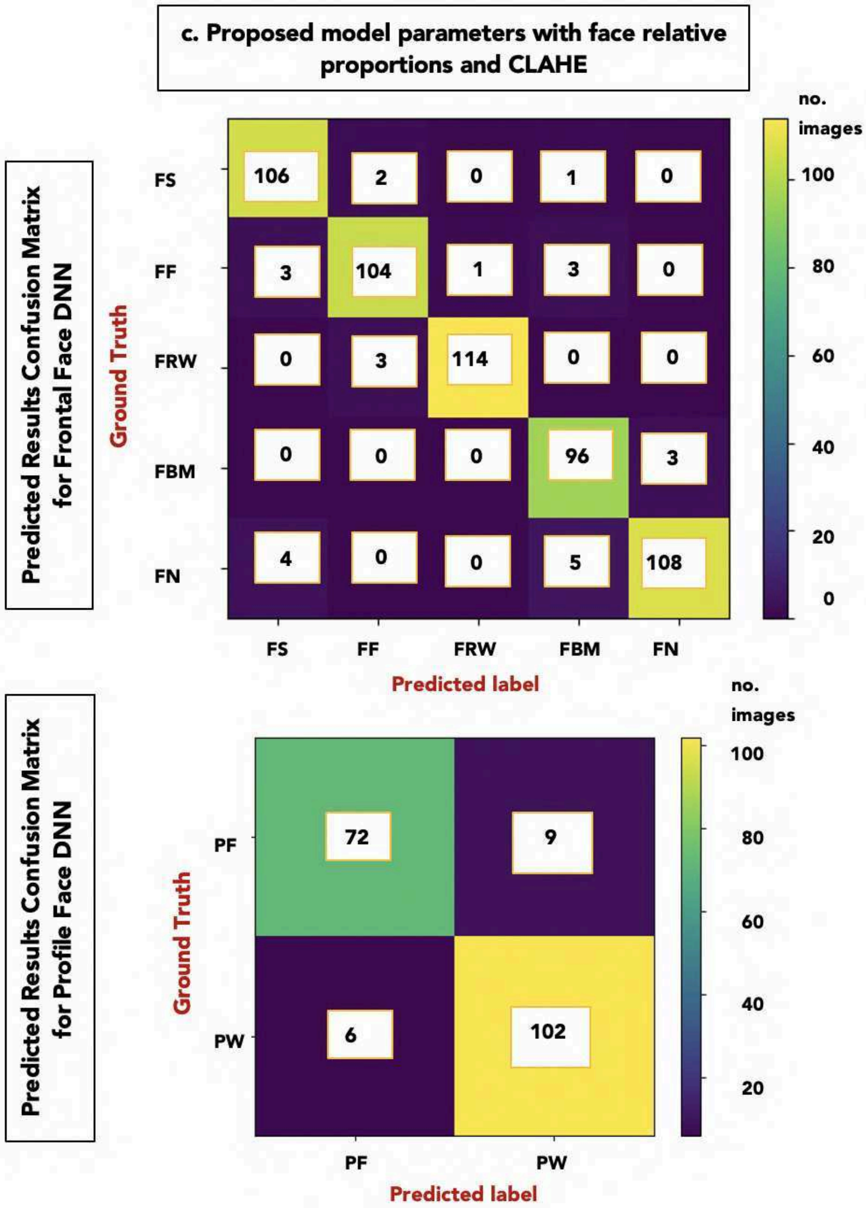


Figure 5.8: Experiment 2.b. First row: Predicted Results Confusion Matrix for the Frontal Face DNN Model. Second row: Predicted Results Confusion Matrix for Profile Face DNN Model.

face proportion relation is not considered. The model has difficulty understanding the data pattern because it does not have a generalization of the data. This can also be reflected since the unconstrained training dataset is small, so if the model's data is not general enough, the model will fall into miss-classifications. Figure 5.11 shows



	FS	FF	FRW	FBM	FN
FS	106	2	0	1	0
FF	3	104	1	3	0
FRW	0	3	114	0	0
FBM	0	0	0	96	3
FN	4	0	0	5	108

	PF	PW
PF	72	9
PW	6	102

Figure 5.9: Experiment 2.c. First row: Predicted Results Confusion Matrix for the Frontal Face DNN Model. Second row: Predicted Results Confusion Matrix for Profile Face DNN Model.

some of the photos that were miss-classified in Experiment 2.b but correctly classified in Experiment 2.c.

Experiment 2.c: From all models, this has a significant out-performance. For Figure 5.9 (c.1), the minimum correct classification percentage was 92%. For Figure 5.9 (c.2), the minimum correct classification percentage was 88%. Also, there is a clear improvement compared to the best performance on conventional approaches shown in Figure 5.5 (a.1). This is considered as a great achievement considering the highly challenging testing dataset. From images with much noise, the system could transform and deliver the data as equally as possible so the model can understand what is happening regardless of the noise. For Figure 5.9 (c.2), the images under PF label were hard to classify when the driver had a thick black eyeglass stick. The thick black eyeglass stick and pupil are both around the same color, so it is hard for the system to detect accurately and precisely. Also, itself the landmark alignment on profile face frames is hard and less accurate, especially when there are occlusions. Figure 5.12 shows an example of this. Future works should consider either making a semantic segmentation of the eyeglass to make a distinction or incorporating context (video recognition) to track the pupil better. However, not all of the frames with these characteristics were misclassified. In terms of time, even though it is double expensive as the general approach, the difference in accuracy is very significant. Future works may explore other techniques for reducing time processing. In terms of accuracy, this model outperformed all models in terms of macro-average accuracy and micro-average accuracy for both, frontal and profile faces, as shown in Table 5.2. The resulting running time in all the experiments represents a significant achievement for us, considering that one goal is a system that runs in real time.

Some overall comments on the performance of the system are:

1. The experiments using CLAHE had a higher processing time as shown in Figure 5.13, although they were still running in real-time. Reducing the image size could potentially improve performance, but the impact on the overall performance of the results should be carefully evaluated.

2. The profile face model performed poorly overall. This is because profile faces are generally difficult to classify, even with robust landmark detectors, due to the number of occluded regions on the face.

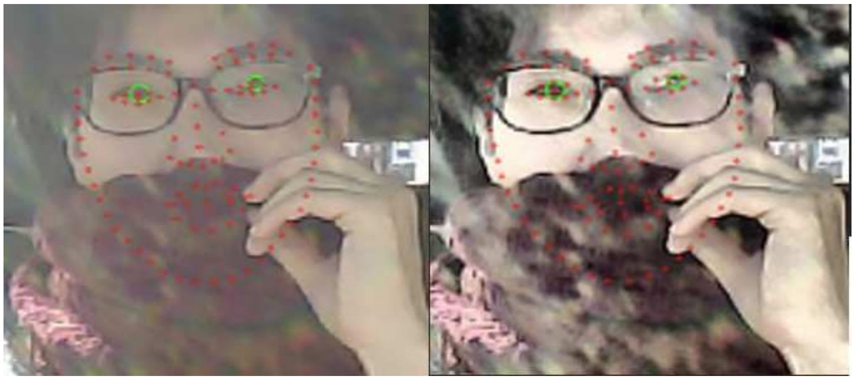


Figure 5.10: Comparison between a frame without applying CLAHE (left) and with CLAHE (right). We can observe that the left image has a slight shift in eye landmark recognition. This will lead to misclassification. Instead, on the right frame, the landmark detection is very clean, which helps the model make a correct classification.

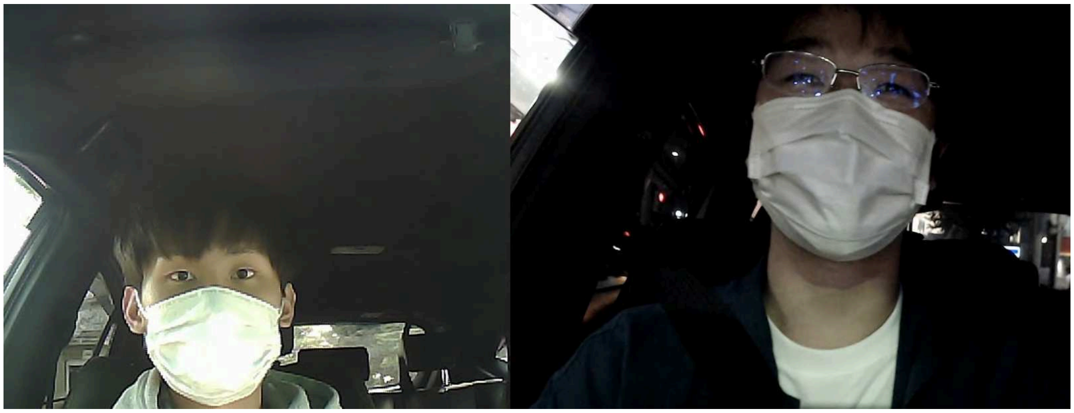


Figure 5.11: Drivers can have different distances to the camera and different face configurations, e.g., eyes size. The values of the parameters might substantially fluctuate, so the model may not comprehend the pattern, especially with a small dataset. Thus, these discrepancies can be harmless information to the system. These images are examples where Experiment 2.b failed (Left: FN, Right: FS), and Experiment 2.c succeed (both FF). Using relations proportional to the face position and configuration helps the system normalize the data and deal with the fluctuation.

3. To improve the performance, it is necessary to explore and consider additional information, such as context, to increase robustness.

Figure 5.13, 5.14, and 5.15, shows visually the results of Table 5.2.

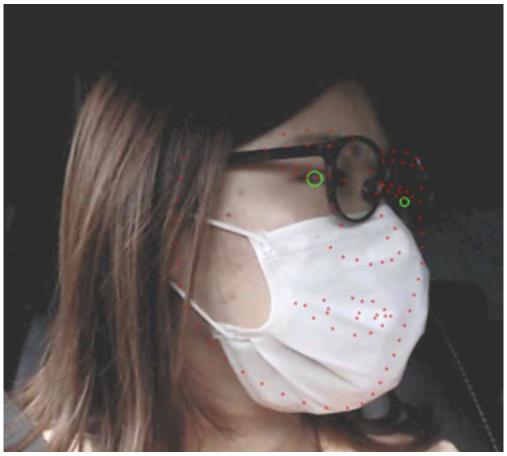


Figure 5.12: Example of miss-classification on Experiment 2.c with profile faces. The pupil tracker may fail when the driver has very thick eyeglasses frames.

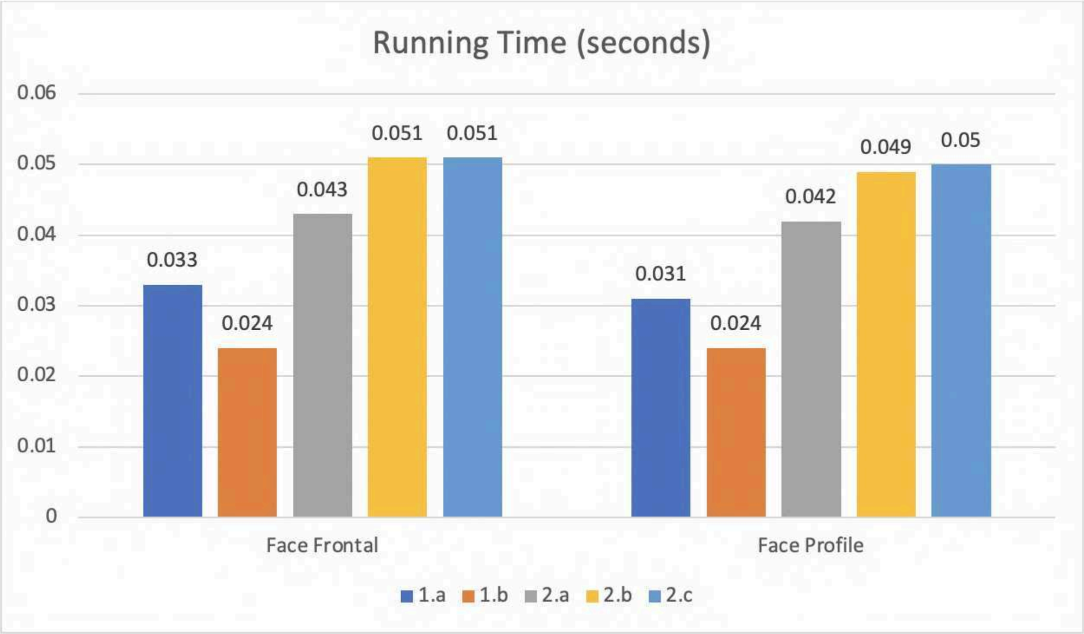


Figure 5.13: Comparative Graph of the Running Time.

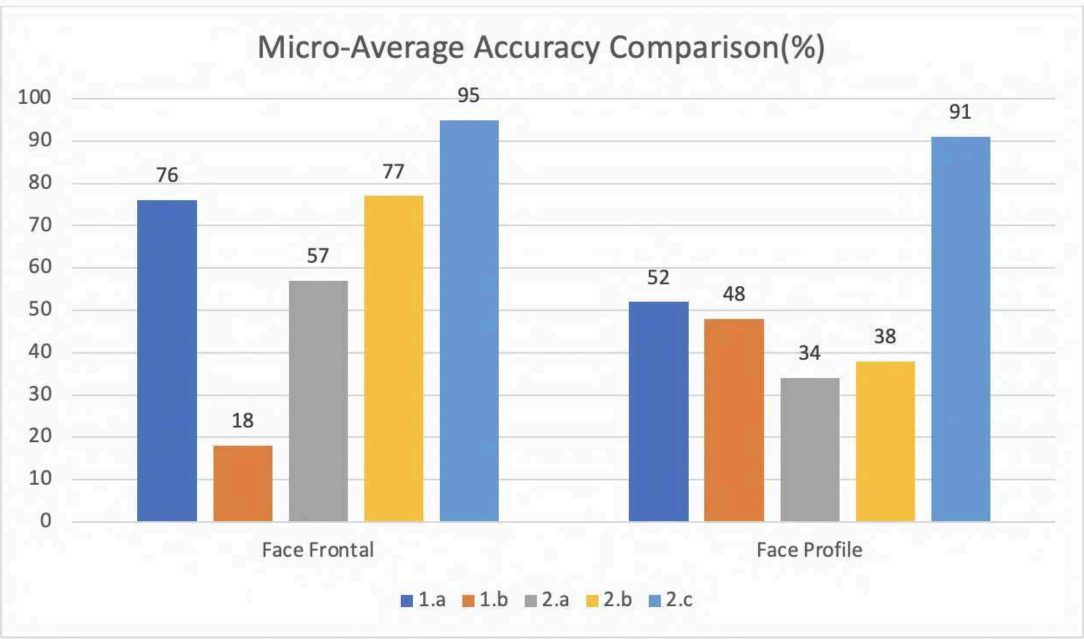


Figure 5.14: Comparative Graph of the Micro-Average Accuracy.

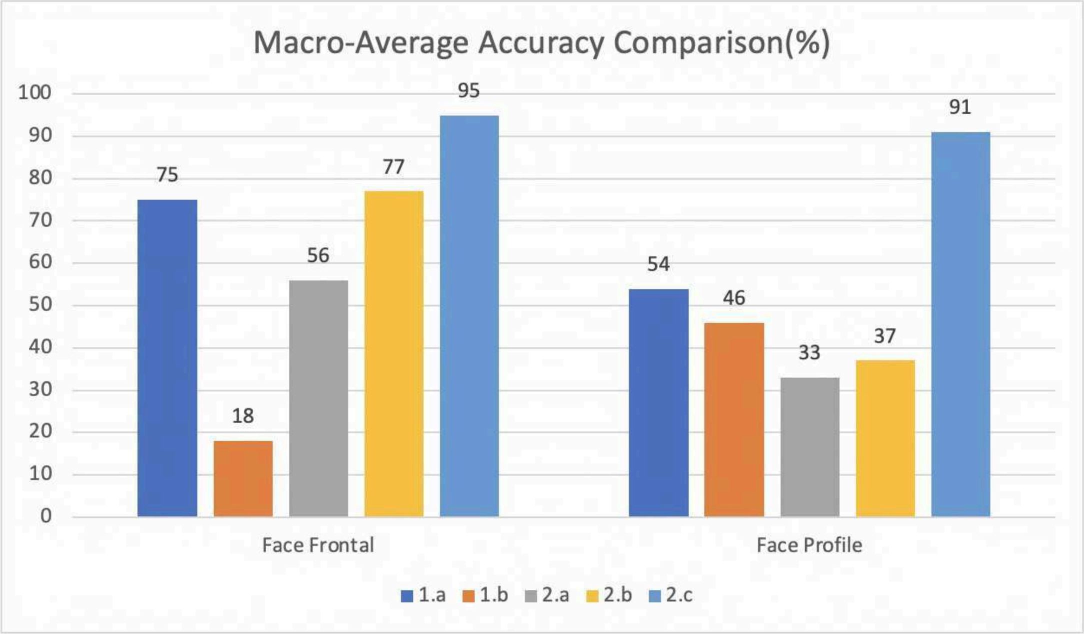


Figure 5.15: Comparative Graph of the Macro-Average Accuracy.

Chapter 6

Module 2: Drowsiness Classifier

6.1 Overview

According to different studies on driver monitoring, detecting yawning and eye closure is essential for identifying driver drowsiness, which is a major contributor to traffic accidents according to various sources [63].

However, in situations with challenging driving conditions, such as the use of masks by drivers, reflections on eyeglasses, or significant changes in lighting, current research tends to rely on unstable techniques as it can be difficult for computers to recognize patterns under these circumstances and the current research has not made an effort to work under this kind of conditions. Additionally, many of these systems rely on expensive sensors that may be sensitive to light or intrusive to the driver. Figure 6.1 shows frame samples from this study's dataset that represents the aforementioned difficult scenarios. To address the limitations of existing drowsiness detection systems, a novel single-camera approach was developed that is able to accurately classify drowsy drivers, even when they are wearing masks, experiencing eyeglass reflections, or undergoing significant changes in daylight conditions. It entails a pipeline with three novel steps:

1. To solve the issue of strong light variance, the frames' brightness were equalized by modifying their Lab's color space using CLAHE. This process was covered in Chapter 3.

2. The face, eye, and body joint landmark recognition in this system is robust. The face and eye landmark detection use a combination of an anchor-based real-time face detector with a normalized dense alignment that considers 3D eyelid and facial expression movement tracking. For shoulder joint recognition, an online optimization framework is used. It links cross-frame poses and generates pose flows that are robust to unconstrained body pose in different situations. This process was covered in Chapter 4.

3. This framework uses three key landmarks: lower-face contour, eyes, and chest movement to classify drowsiness. Unlike other studies, mouth closure is not included in the feature vector because masks may cover it. Instead, the inclusion of the lower-face contour and chest movement as potential parameters in the feature vector is a key contribution of this paper in addressing the issue of mask-wearing. This classification is video-based. The closure of the eyes is measured in each frame. For the lower-face contour and chest movement, the current position of each landmark is subtracted from its position in the first frame (original). The driver's drowsiness is then determined by combining spatio-temporal features based on these subtractions, which are used as the feature vector in a GRU-based model. This chapter will provide a detailed step-description of this process.

The system's flowchart is shown in Figure 6.2.

The method outperforms in correctly classifying driver drowsiness in challenging situations, as evidenced by the performance evaluation results with a dataset containing intense light differences, eyeglasses reflection, and mask-wearing situations.

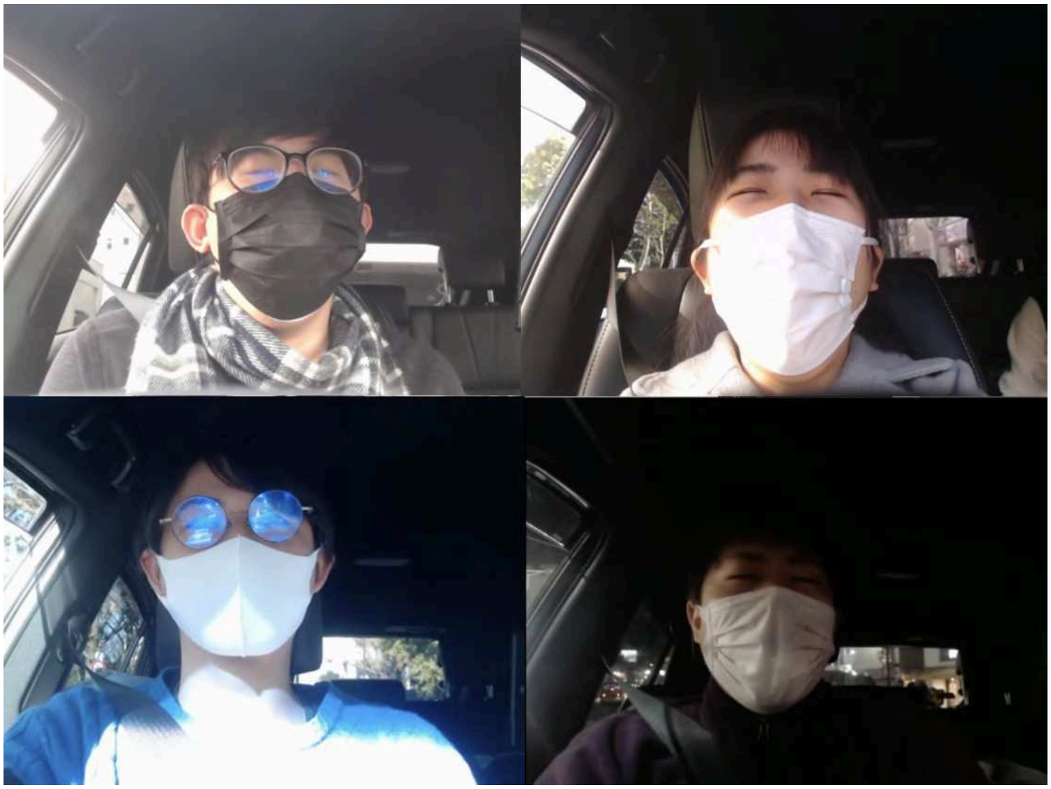


Figure 6.1: Examples of frames from the dataset in which the drivers are wearing masks. Left-top: Driver yawning with a black mask. Left-bottom: Driver not yawning with eyeglass reflections and a white mask. Right-top: Driver yawning with a white mask. Right-bottom: Driver yawning with low brightness and a white mask.

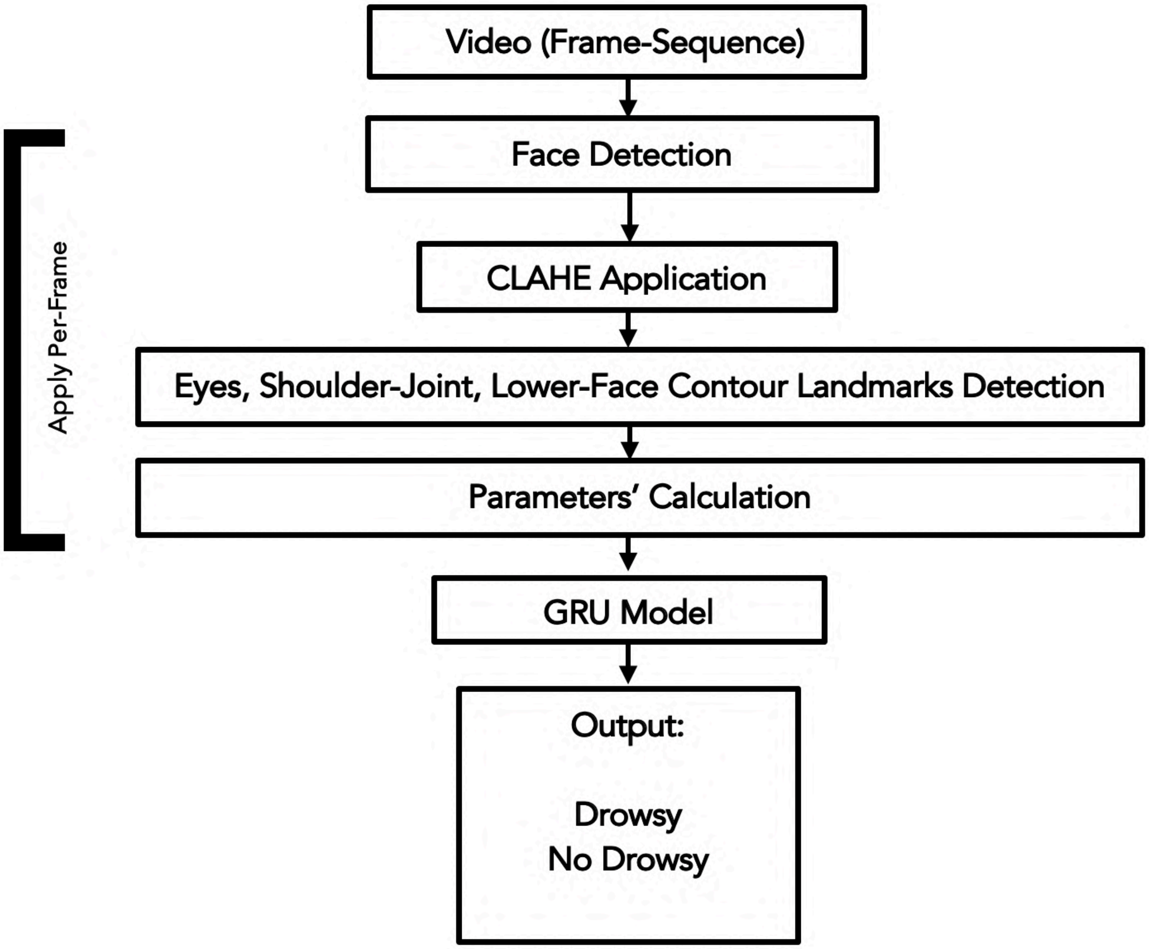


Figure 6.2: System's Diagram. The input is a video frame-sequence. For each frame inside the video, first, it detects the face, and then CLAHE is applied to equalize its brightness. Then eyes, shoulder-joint, and lower-face contour landmarks detection are applied. After extracting this data, the parameters are calculated, and then the feature vector will be constructed and after processing all the frames will be passed to the GRU model. There are two possible outputs: Drowsy or No Drowsy.

6.2 Feature Vector

Three proposed feature vectors are one of the key novelties of this work. To determine which parameters produce the best results, different combinations of these parameters are compared. The quality of the data that is input into a model is one of the most important factors in determining how successful the model will be, and this is especially true when the model requires dealing with data that contains large differences. In general, many studies overlook the importance of data quality. If more attention is paid to the quality of the input data, it is possible to achieve the best performance from a static model while also reducing the training data amount required. With this in mind, this study focuses on providing the input data to the model in the clearest form possible. Each frame of data, denoted by the letter F, is a two-dimensional array of pixel values denoted by $F(x, y)$, where the first index (row) represents the x-coordinate and the second index (column) represents the y-coordinate. A pixel within the frame F with the following structure is considered to be an extracted landmark L in the frame F being evaluated.

$$L = F(x_L, y_L)$$

x_L represents the x coordinate and y_L the y coordinate of landmark L within the two-dimensional pixels array that compose the frame data F.

To construct the feature vectors, three key area landmarks are used: the eyes, the contour of the lower face, and the shoulder landmarks. There were made different relations between these landmarks and carried out pre-processing on these parameters to cut down on noise and obtain distinct patterns for the model. The following is a description of the pre-processing steps of each parameter:

1. Eye closure (3 parameters): The eyelid closure distance over the pupil over time is the result of measuring the vertical distance between the eye center and each upper lid landmark. Since the face has a symmetrical structure, the eye landmarks on only one side of the face were used, and the system automatically decides which side to use. The eye pupil center landmark is named Pc_L for the left eye and Pc_R for the right eye. If the pupil detector is unable to detect

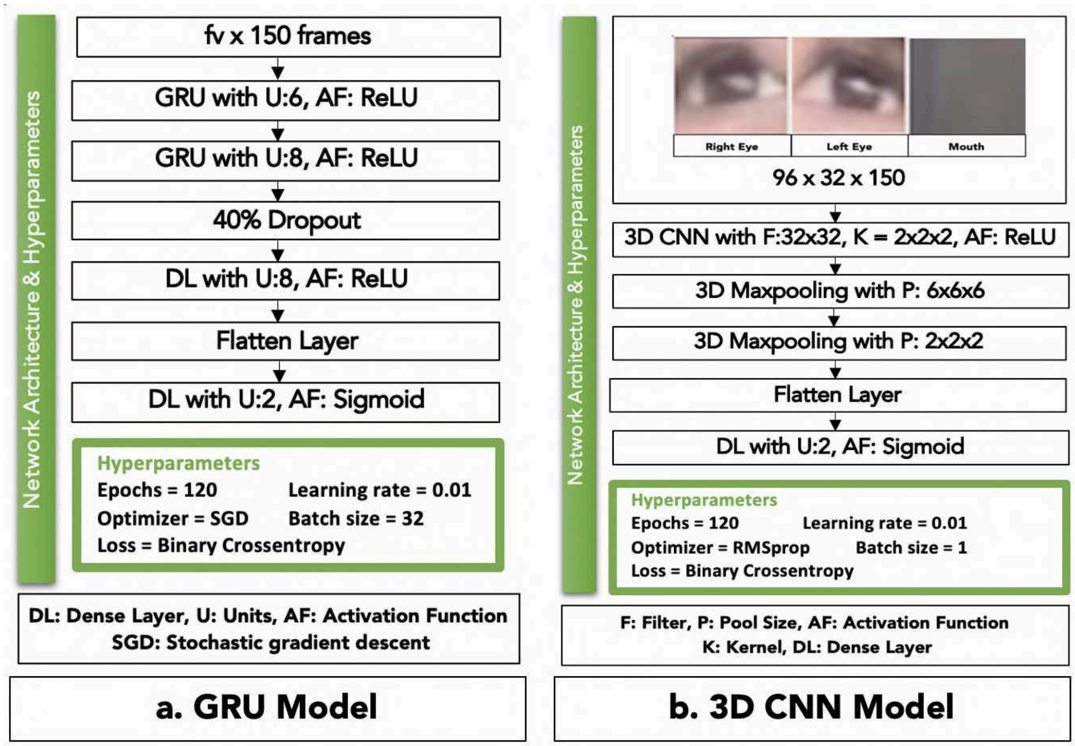


Figure 6.3: Left: Proposed Network Architecture for GRU's Model. Right: Proposed Network Architecture for 3D CNN model.

a pupil accurately, the size of the pupil is reduced. The system then compares the sizes of the pupils in both eyes and uses the landmarks of the eye with the larger pupil size (referred to as Pc) to filter out noise in the data. The upperlid landmarks Uj, with $j = 0, 1, 2$ was extracted. The vertical distance from the eye center landmark Ec was then measured (y-coordinate). It is important to note that the eye center is not the same as the pupil, identified through segmentation. The eye center is determined based on the positions of the eyelid landmarks. The eye closure e_c_j is represented by:

$$e\_c_j = y_{U_j} - y_{Ec} \quad with \ j = 0, 1, 2$$

2. Lower-Face Contour (15 parameters): Another important metric used to detect

drowsiness is the distance of each lower-face contour landmark from its initial position to its position over time. This distance is significantly larger when drivers are yawning than when they are in a normal or talking state. There are 15 lower face contour landmarks, denoted by n_fc . To determine the size of this difference is stored each lower face contour landmark as o_fc_j in the first frame. Then, in each subsequent frame, the following equation is used to determine the vertical difference between each current lower-face contour landmark c_fc_j with o_fc_j:

$$f\_d_j = y_{c\_fc_j} - y_{o\_fc_j} \quad with\ j = 0, ..., (n\_fc - 1)$$

3. Shoulder Joint (1 parameter): Yawning involves a deep intake and outlet of air [64]. Therefore, the movement of the shoulders as we inhale and exhale was included as a criterion in this study. To measure the difference in position, the original position of the shoulder joint landmark was saved in the first frame, and the vertical difference between the current shoulder joint landmark and the original position was calculated in each frame. This difference was used as a feature in the classification model to distinguish between normal and yawning states. Only the left shoulder joint landmark was used to reduce noise in the data, as shoulder joints are symmetrical, and the position of the left shoulder joint can be used to represent the position of both shoulders. The first shoulder joint in the first frame was saved as o_{sc}, and the vertical difference between each current shoulder joint landmark c_{sc} and o_{sc} was calculated in each frame using the following equation:

$$s\_d = y_{c\_sc} - y_{o\_sc}$$

However, an extra step was required. For face landmarks, the recognition zone is small enough to achieve a high level of precision. However, for shoulder joints, the joint point may be valid over a wider range of motion. As a result, to validate that the origin possesses the joint point with the lowest elevation, the value of the current shoulder joint landmark is first determined, then compared to the value of the origin shoulder joint landmark, and finally, the current shoulder joint landmark is updated if its value is lower than the value of the origin shoulder joint landmark. This step reduces noise in the data by ensuring that the origin shoulder joint is located at the lowest point.

After describing the different parameters, will be introduced three feature vectors fv, and conduct a comparative analysis of them in the next Section 6.4.4. Each feature vector is composed of the following combination of parameters:

- Eyes Closure $\cup$ Shoulder-Joint
- Eyes Closure $\cup$ Lower-Face Contour
- Eyes Closure $\cup$ Shoulder-Joint $\cup$ Lower-Face Contour

6.3 GRU model

Given that space-temporal features are important for detecting a drowsiness state, a GRU-based model would fit to this study. Figure 6.3 represents the model's topography and hyperparameters.

Model:

In Recurrent Neural Networks (RNNs), connections between nodes can create a cycle, allowing the output of some nodes to influence subsequent input to the same nodes, enabling the system to exhibit temporally dynamic behavior. Variants of RNNs include GRUs (Gated Recurrent Units) and LSTMs (Long Short-Term Memory Networks). A GRU model was adopted because it has a simpler structure and fewer

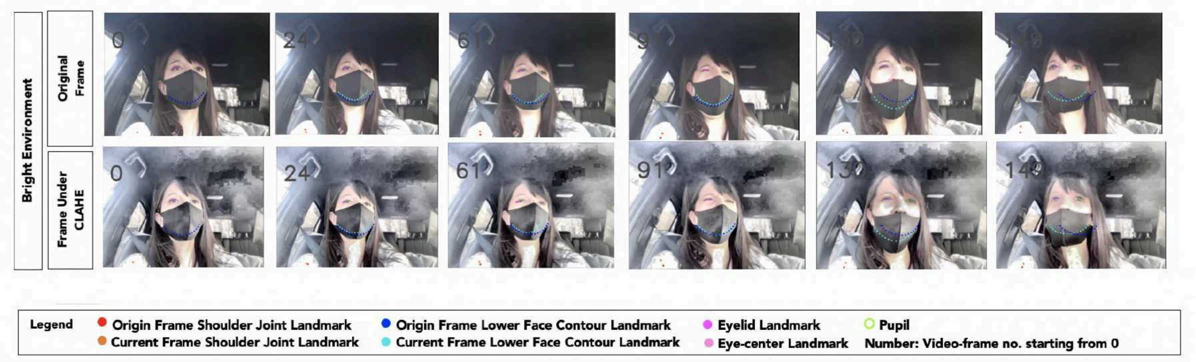

Bright Environment
Original Frame
Frame Under CLAHE
0
24
61
91
0
24
61
91
Legend
● Origin Frame Shoulder Joint Landmark
● Current Frame Shoulder Joint Landmark
● Origin Frame Lower Face Contour Landmark
● Current Frame Lower Face Contour Landmark
● Eyelid Landmark
● Eye-center Landmark
○ Pupil
Number: Video-frame no. starting from 0

Figure 6.4: Comparison of the detection of the key landmarks used in this module under a day-light environment.

matrix multiplications, making it more efficient, especially when the dataset is small compared to LSTM, as shown in the literature [65]. The details are as follows:

Input: One of the feature vectors fv described in Section 6.2.

Network Topography and Hyperparameters: GRU's model topography and hyperparameters are shown in Figure 6.3 (a).

Output: Prediction probability between the binary output, *No Drowsy* or *Drowsy.*

It is important to understand how detecting these key landmarks can vary based on different conditions, such as lighting conditions. Figure 6.4 compares the detection of key landmarks in a day-light environment, while Figure 6.5 compares the detection of key landmarks in a night-light environment. Additionally, the performance of key landmark detection may also be affected by using CLAHE, which is the method explained in 3.3. Therefore, it is useful to compare the detection of key landmarks with and without the use of CLAHE to understand how it may impact the accuracy and reliability of the results.

6.4 Experimental Evaluation and Results

6.4.1 Overview

This study's purpose is to show that the different feature vectors proposed are capable of accurately detecting driver drowsiness in challenging situations such as mask-

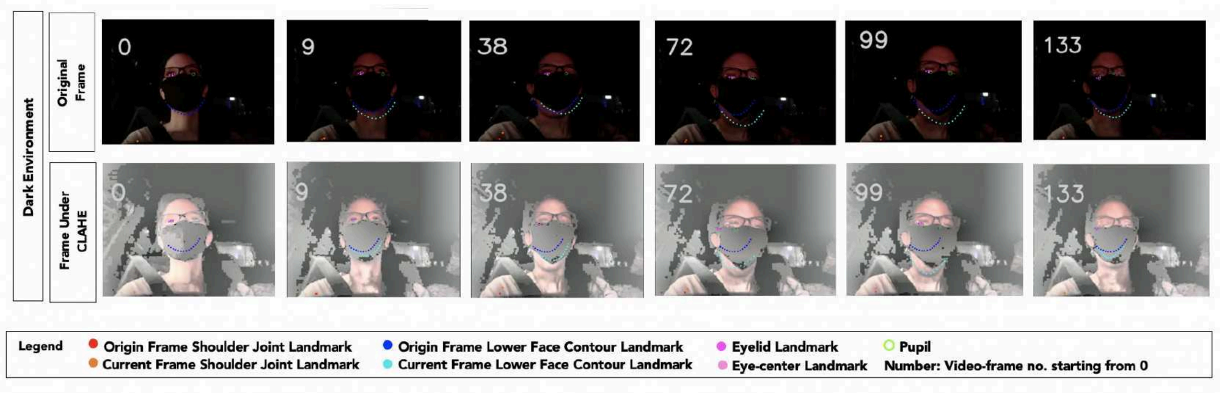

Dark Environment
Original Frame
Frame Under CLAHE
0
9
38
72
99
133
Legend
● Origin Frame Shoulder Joint Landmark
● Current Frame Shoulder Joint Landmark
● Origin Frame Lower Face Contour Landmark
● Current Frame Lower Face Contour Landmark
● Eyelid Landmark
● Eye-center Landmark
○ Pupil
Number: Video-frame no. starting from 0

Figure 6.5: Comparison of the detection of the key landmarks used in this module under a night-light environment.

wearing, sun reflection, eyeglass reflections, and changes in lighting. To do this, experiments were conducted using a challenging dataset with the aforementioned conditions. The experimental evaluation was divided into two parts to illustrate two important points: first, how the general methodology fails when applied to unrestricted situations, and second, compare each proposed feature vector's performance.

Experiment 1:

The goal of Experiment 1 is to demonstrate where the current frameworks presented in the literature have their flaws in unconstrained settings. No approach is considered to be state-of-the-art, so the two selected methods are the ones utilized the most frequently in the literature. For this demonstration, Experiment 1 was divided into two separate sub-experiments:

- Experiment 1.1: The evaluation of the testing dataset is made without the luminance equalization step. The model's architecture is the same as what is described in Section 6.3 model, with the difference that the face detector that is used will be dlib's face detector and that the model's input will be the complete output of dlib's face landmark detection.

- Experiment 1.2: The evaluation of the testing dataset is made without the luminance equalization. The eyes and mouth image data from the video will

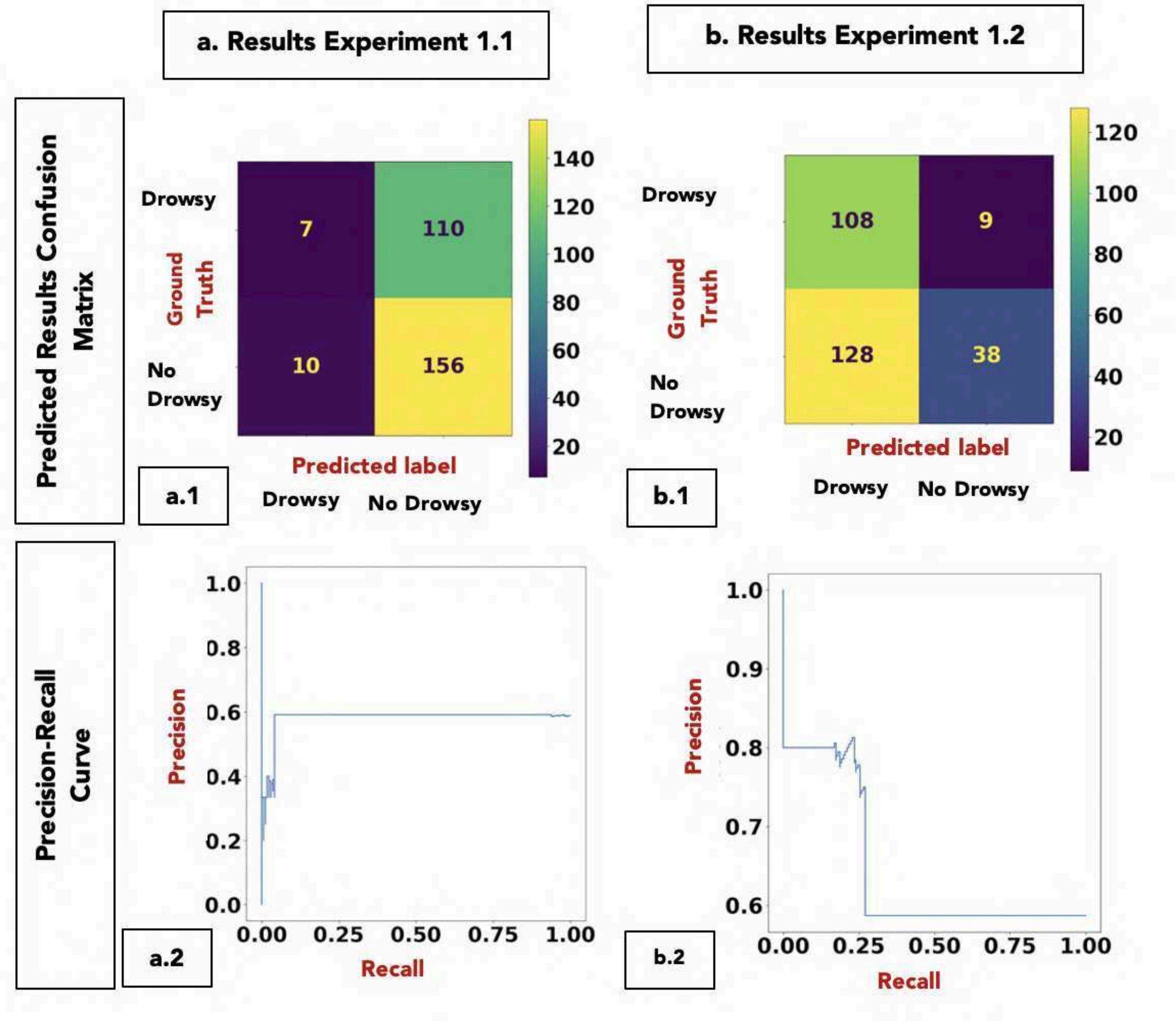


Figure 6.6: Results obtained for Experiment 1. From left to right: Experiment 1.1 (a) and Experiment 1.2 (b). Top: Each experiment Confusion Matrix. Second Bottom: Each experiment Precision-Recall curve.

be used as the model's input. The model's architecture is a three-dimensional convolutional neural network, represented in Figure 6.3 (b).

Experiment 2:

Experiment 2 compares each proposed feature vector's performance explained in Section 6.2. For this demonstration, Experiment 2 was broken down into the following four sub-experiments:

- Experiment 2.1: The evaluation of the testing dataset was done with the proposed GRU model using as an input Eyes Closure $\cup$ Shoulder-Joint 's feature vector. The frame luminance equalization step (CLAHE) is applied.

- Experiment 2.2: The evaluation of the testing dataset was done with the proposed GRU model using as an input Eyes Closure $\cup$ Lower-Face Contour's feature vector. The frame luminance equalization step (CLAHE) is applied.

- Experiment 2.3: The evaluation of the testing dataset was done with the proposed GRU model using as an input Eyes Closure $\cup$ Shoulder-Joint $\cup$ Lower-Face Contour's feature vector. The frame luminance equalization step (CLAHE) is applied.

- Experiment 2.4: The evaluation of the testing dataset was done with the proposed GRU model using as an input the feature vector that got overall better results among Experiments 2.1, 2.2, and 2.3 with the difference that in this experiment was not applied the frame luminance equalization step (CLAHE).

6.4.2 Dataset

The training dataset consists of videos from [66] and contains around 400 videos that do not include challenging conditions. Data augmentation was applied, and the dataset was divided into two classes: *No Drowsy* and *Drowsy*.

As for the testing dataset, all the videos were taken especially for conducting this study's experiments, and there are 283 videos. These videos were captured in a real car, sometimes when the car was moving and when it was not. Experiments were carried out with a total of thirteen participants, four of whom were female, and nine of whom were male. The ages of those involved range from 19 to 30 years old. Only frontal faces are considered in the testing dataset. There are no faces that are strongly rotated within the same video, faces that are too close to the camera, or no faces at all. All the drivers wore a mask and the videos were also subjected to at least one of the following conditions: environments with intense light (excluding completely dark

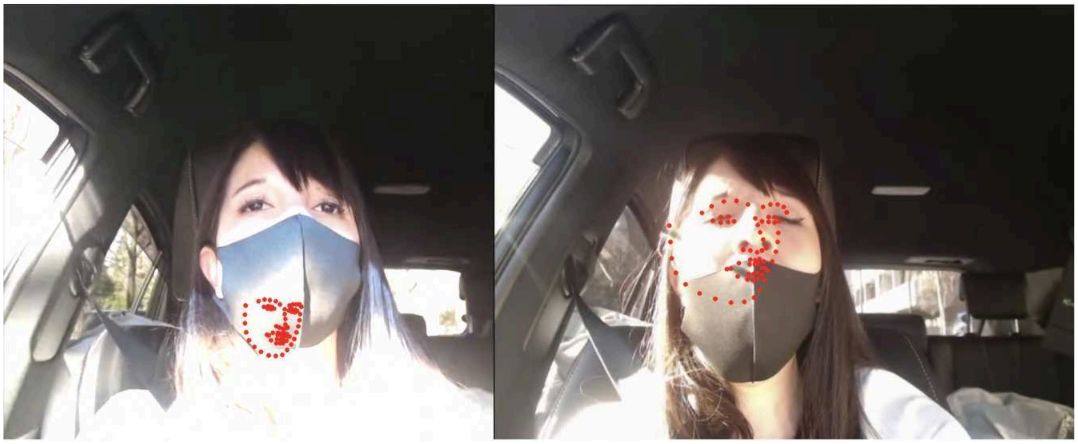


Figure 6.7: In Experiment 1.1, the method demonstrated poor performance in face detection, and in the few cases where it was able to recognize the face, it also struggled to identify landmarks.

or bright videos where the face could not be seen even after applying CLAHE) or eyeglass reflections. Each video contains 150 frames and was recorded in a variety of daylight conditions, including nighttime. The dataset includes as well same videos but with data augmentation to create darker or brighter versions to demonstrate that the classifier is robust to variations in lighting conditions.

6.4.3 Metrics

- Predicted Results Confusion Matrix: In statistical classification, it is the presentation of the performance of the model by illustrating the number of instances that are correctly classified. The rows contain the ground truth labels, while the columns contain the predicted labels.

- Precision-Recall Curve (PRC): In statistics, this diagram illustrates how the balance between recall and precision changes at various thresholds.

  - A precision-recall curve is a graph that plots the precision (along the y-axis) and the recall (along the x-axis) for various thresholds. High recall and high precision are both represented by a large area under the curve, while a point on the curve denotes a perfect skill model (1,1).

- Precision is a measure of a model's accuracy in predicting the positive class. It is calculated by dividing the number of true positives by the total number of true and false positives. Precision is also referred to as specificity, which refers to the ability of a model to predict the presence of the positive class accurately.
- Recall is a measure of a model's ability to correctly identify positive instances and is calculated as the number of true positives divided by the total number of true positives and false negatives. Sensitivity is another term used to refer to the ability of a model to correctly make predictions and is often used synonymously with recall.

6.4.4 Results

Experiment 1:

Experiment 1.1: Seeing this results, the general approach, using Dlib's face recognition and landmark recognition, has a poor performance. In Figure 6.6 (a.1) is reflected that the model is not capable of making an accurate distinction between *Drowsy* and *No Drowsy*. The poor performance can be seen in Figure 6.6 (a.2). The curve area does not have the expected shape and the precision value is less than 0.6. The fact that it was unable to recognize landmarks on multiple occasions even the few times that could make a correct facial recognition was clearly reflected in the results, as can be seen in Figure 6.7. The results of this experiment suggest that the model is unable to understand the pattern formed by the features used for the classification.

Experiment 1.2: The results of this experiment were not significantly improved compared to those of Experiment 1.1. The model cannot detect changes in the mouth because this feature is covered. It is clear by seeing Figure 6.6 (b.1) that the model is unable to differentiate accurately between the two states denoted by the labels *No Drowsy* and *Drowsy*. Compared to Figure 6.6 (a.2), Figure 6.6 (b.2) shows a better PRC result. However, even though a slight regular PRC shape is visible, the

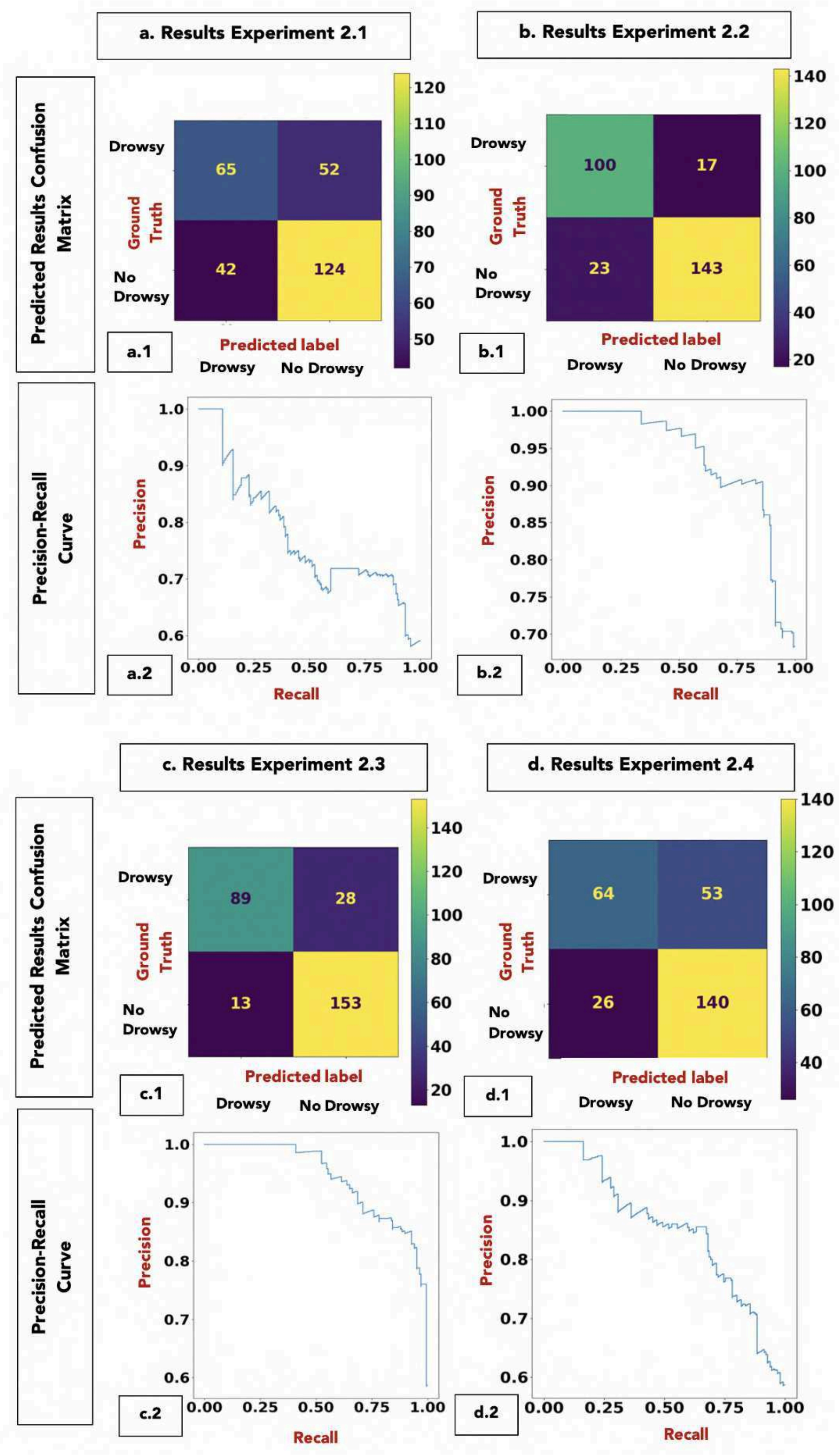


Figure 6.8: Results obtained for Experiment 2. Odd rows: Each experiment Confusion Matrix. Even rows: Each experiment Precision-Recall curve.

performance is still considered a failure because the precision drops significantly at low recall thresholds. Figure 6.6 (a.2) displays a worse PRC result. This is likely due to the model's difficulty in interpreting the data pattern because the methodology does not provide a generalization. If the data is not refined, the model will try to search for a pattern under various conditions, such as the influence of mask wearing, changes in the environment, and eyeglass reflections, which is a very challenging task.

Experiment 2:

Experiment 2.1: Figure 6.8 (a.1) illustrates that the model is unable to correctly differentiate between the *No Drowsy* and *Drowsy* state. This is reflected in the *Drowsy* class that contains approximately half of the tested videos. Poor performance can be seen in Figure 6.8 (a.2) because the precision is less than 0.8. According to the findings of this study, only the pure shoulder joint feature may not be effective in predicting driver drowsiness. There are several potential reasons for this, including the fact that the marker for the shoulder joint is not in the same exact location in each frame. The potential correct landmark area for the shoulder joint is larger, which means that the landmark may appear in each frame with a varied position within a region. This is in contrast to face landmark detection, where each landmark has a nearly precise and consistent location. Even though the shoulder joint is roughly in the same place as the previous frame, it may still have an error because the location does not remain constant from frame to frame. In the future, the development will implement a neighborhood-repositioning capability, which will allow for the enhancement of the benefits offered by this feature.

Classification Percentage *Drowsy Class: 55.55% No Drowsy Class: 74.69%.*

Experiment 2.2: In terms of overall performance, this model was superior to all others. Both classes can be recognized without any problems, as can be seen in Figure 6.8 (b.1). According to Figure 6.8 (b.2), this experiment has the highest area under the curve among the four experiments; indicating that it is the model with higher accuracy. This has a significant meaning: from noisy frame data, the relevant features were provided in an even and consistent manner so that the model could recognize the pattern within the frame. The use of robust landmark detection libraries, as well as comparing the landmark position in each frame to its initial position, are critical steps inside the pipeline in achieving these results. However, can be seen that the curve contains a spike in the middle of it. The fact that it was difficult to identify landmarks in some of the frames is one possible explanation.

Classification Percentage *Drowsy Class: 85.47% No Drowsy Class: 86.14%.*

Experiment 2.3: This model came in second place compared to others in terms of its overall performance. Figure 6.8 (c.1) shows that the model did a better job of classifying the *No Drowsy* category than the model used in Experiment 2.2. This result shows that the shoulder-joint characteristic, combined with other features, can provide the model crucial information to comprehend the pattern. In addition, the model's balance and smoothness can be seen in Figure 6.8 (c.2), which displays a fairly large area under the curve. As the shoulder-joint potential correct landmark area is bigger, for the same reason as Experiment 2.1, Experiment 2.3 did not perform better than Experiment 2.2. Therefore, if a previous frame neighborhood-repositioning capability for shoulder joint landmark detection can be included if the shoulder joint landmark position of the current frame is significantly different from the previous one, then it is possible that this model may perform better than the model used in Experiment 2.2.

Classification Percentage *Drowsy Class: 76.06% No Drowsy Class: 92.18%.*

Experiment 2.4: Experiment 2.2 yielded the best results overall, so this experiment was based on it. The difference is that was not applied CLAHE. It is clear from looking at Figure 6.8 (d.1) that the model is unable to differentiate clearly between the two states denoted by the labels *Drowsy* and *No Drowsy.* For a medium recall threshold, the precision was lower than 0.8 and Figure 6.8 (d.2) displays unsatisfactory performance. The importance of adjusting the brightness of the frames is emphasized by the results of this experiment. Significant shifts in the lighting conditions occur while driving, and these should be filtered. Due to variations in lighting, the model may struggle to distinguish between different faces or misidentify landmarks if the brightness is not balanced. On the other hand, CLAHE can be of great assistance when operating in low-light conditions. Since the face features are being significantly emphasized in CLAHE, yawning detection can achieve a higher level of precision.

Classification Percentage *Drowsy Class: 57.65% No Drowsy Class: 84.33%.*

Chapter 7

Conclusions & Future Works

7.1 Conclusions

This study presented a novel driver's gaze zone and drowsiness identification pipeline using a single camera robust to highly challenging situations as could be:

- Mask-wearing faces
- Face partial occlusions
- Eyeglasses reflection
- Strong daylight variations
- Pupil noise

The robustness of the driver's gaze zone and drowsiness identification pipeline was achieved through the following key points:

1. Equalizing the brightness of the frames through the use of Contrast Constrained Adaptive Histogram Equalization on the luminance channel of the Lab color space. This helps to improve the visibility of features in the frames, making it easier for the algorithm to accurately identify key landmarks such as the eyes, face, and body joints.

2. Recognizing the key landmarks as face, eyes, pupils, and body-joints with libraries that are highly robust under unconstrained situations. This reduces on a great scale the possible noise that the data may have thanks to various facial occlusions as could be masks, scarves, eyeglasses reflections, eyeglasses sticks, small eyes, partially occluded pupils, and profile face poses.

3. The construction of DNN models using novel feature vector parameters for the Gaze Zone Classifier was introduced. Since the geometric facial structure varies per person, the feature vector parameters consist on different relations between pupil and eye landmarks in proportion to the driver's geometrical face configuration. It considers when the driver's face and eyes are facing different directions. This is a per-frame classification. Related works general approaches got a best Micro-Average and Macro-Average Accuracy for the Frontal Model of 76% and 75%, respectively, while the proposed method got 95% and 95%. As for the Profile Face Model, the related works approaches got a best Micro-Average and Macro-Average Accuracy of 52% and 54%, respectively, while the proposed method got 92% and 91%. In both models, the proposed method outperformed the related works approaches. However, the related works method had the best processing time of 0.024 seconds, whereas the proposed method had a processing time of 0.051 seconds.

4. The introduction of a GRU model structure, where the input feature vector considers eyes closure, lower-face contour and chest movement landmarks for the Drowsiness Classifier. In contrast to prior studies, mouth closure was not considered as part of the feature vector because mouth occlusions like masks could cover it. One of the most significant contributions to addressing the mask-wearing situation issue for the drowsiness classifier is the addition of chest movement and the lower-face contour as possible feature vector parameters to detect a yawing state. For the face contour and chest motion, each current

landmark location is subtracted from its original position. Also, it is measured the eye closure in each frame. Finally, this information is the feature vector of a GRU-based model that, based on spatial-temporal features, derives the driver's drowsiness state. This is a video-based classification. Both related works' methodologies got very unbalanced models. Inside the proposed feature vectors, the feature vector that uses Eyes Closure ∪ Lower-Face Contour as parameters got the best classification results with a classification percentage in the *Drowsy Class* of 85.47% and for the *No Drowsy Class* of 86.14%.

Inside the study, both classifiers were compared with the general approach. Moreover, was also demonstrated the significance of each step in the suggested pipeline. After evaluating the current implementation from a dataset with highly unconstrained driving conditions, was observed that this approach can correctly handle the classification of the driver's gaze zone and drowsiness in a wide variety of difficult situations.

Finally, the current work and its extensions have the potential to serve as a base for establishing more stable and robust ADAS systems towards challenging scenarios.

7.2 Contribution to Intelligent Transportation Systems

According to findings from different studies [67], tiredness and distracted driving are two of the primary contributors to automobile accidents. If Advanced Driving Assistance Systems (ADAS) can detect drowsy or distracted drivers, this can greatly improve traffic safety. These systems will not only alert the driver but also take control of the vehicle in semi-autonomous cars, preventing severe accidents and promoting safe driving.

Classifiers for driver gaze and drowsiness identification are crucial in determining if the driver is alert and aware of their surroundings, allowing an ADAS (Advanced Driver Assistance System) to make a warning or, if it is inside an autonomous car, to take control of the vehicle if necessary. When combined, tracking a driver's head, eyes, and body features can provide a reasonably accurate estimate of where the driver is looking or the driver's alertness. However, the studies in the prior research were carried out under ideal circumstances. It is highly challenging to make this categorization under unconstrained settings, which is why the performance of the most recent investigations may lead to very poor performance under these difficult conditions. The development of a high-performance driver gaze and drowsiness classifier that can operate in unconstrained environments using a single camera is the primary focus of this study. When operating in unrestricted conditions, having a high-performance system means recognizing distractions and alertness with fewer mistakes, leading to a considerable reduction in automotive accidents.

7.3 Future Work

Making a robust classifier requires concentrating efforts on very specific details. So to improve the performance of the current gaze zone and drowsiness detection, future works should explore and extend the proposed method in the following topics:

1. Combining both modules (gaze zone and drowsiness detection) into one system to operate simultaneously.

2. Extending Module 1 to consider spatial and temporal information to provide a more refined result based on context.

3. Proposing a system that utilizes both RGB and NIR cameras to handle totally dark frames.

4. Implementing an autonomous parameter tuning algorithm for CLAHE (Contrast Limited Adaptive Histogram Equalization) that is able to detect the frame's luminance and choose the optimal settings for equalization.

5. Exploring how the algorithm performs when the driver is at different distances from the camera and has significant differences in pitch, yaw, and roll.

6. Mapping the driver's gaze to the object they are looking at, to provide more information to the system.

7. Extending the implementation to consider the case of multiple passengers in the vehicle.

8. Incorporating a module to detect daydreaming as a component in a driver assistance or autonomous vehicle system can aid in identifying when the driver's attention is not fully on the road, and they may be concentrated on their own thoughts.

9. Implementing a continuous zone classification system for determining where a driver is looking, instead of a discrete zone classification system, could provide more accurate information to the system. This is because a continuous zone classification system allows for a more precise representation of the eye's direction rather than being limited to a set of predetermined regions.

Bibliography

[1] L. Fridman, P. Langhans, J. Lee, and B. Reimer, “Driver gaze region estimation without use of eye movement,” in IEEE Intell. Sys., vo. 31, no. 3, pp.49–56, 2016.

[2] L. Fridman, J. Lee, B. Reimer, and T. Victor, “‘Owl’and ‘Lizard’: Patterns of head pose and eye pose in driver gaze classification,” IET Computer Vision, vo. 10, no. 4, pp. 308–314, 2016.

[3] L. Fridman, H. Toyoda, S. Seaman, B. Seppelt, L. Angell, J. Lee, B. Mehler, and B. Reimer, “What can be predicted from six seconds of driver glances?,” in Proc. Conf. on Human Factors in Computing Sys., pp. 2805–2813, 2017.

[4] M.C. Chuang, R. Bala, E.A. Bernal, P. Paul, and A. Burry, “Estimating gaze direction of vehicle drivers using a smartphone camera,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis., pp. 165–170, 2014.

[5] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, “Face recognition using histograms of oriented gradients,” Pattern Recognit. Lett., vo. 32, no. 12, pp.1598–1603, 2011.

[6] A. Naqvi, M. Arsalan, G. Batchuluun, S. Yoon, and R. Park, “Deep Learning-Based Gaze Detection System for Automobile Drivers Using a NIR Camera Sensor,” Sensors, vo. 18, no. 2, 2018.

[7] L. Fridman, J. Lee , B. Reimer, and B. Mehler, “A framework for robust driver gaze classification,” SAE Technical Paper, 2016.

[8] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proc. IEEE Int. Conf. Comput. Vis., pp. 3730–3738, 2009.

[9] S. C. Martin, “Vision based, Multi-cue Driver Models for Intelligent Vehicles,” PhD diss., 2016.

[10] S. Minaee, P. Luo, Z. Lin and K. Bowyer, “Going Deeper Into Face Detection: A Survey,” arXiv preprint arXiv:2103.14983, 2021

[11] X. Burgos-Artizzu, P. Perona, and P. Doll´ar, “Robust face landmark estimation under occlusion,” in IEEE Intl. Conf. Comput. Vis., 2013.

[12] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, “Joint 3d face reconstruction and dense alignment with position map regression network,” in Proc. Europ. Conf. Comput. Vis., pp. 534–551, 2018.

[13] I.R. Tayibnapis, M.K. Choi, and S. Kwon, “Driver’s gaze zone estimation by transfer learning,” in IEEE Int. Conf.Consumer Electronics, pp. 1–5, 2018.

[14] X. Shan et al., “Driver Gaze Region Estimation Based on Computer Vision,” in Int. Conf. Measuring Technology and Mechatronics Automation (ICMTMA), pp. 357-360, 2020.

[15] S. Vora et al., “Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis,” in IEEE Trans. Intell. Transp., vol. 3, no. 3, pp. 254–265, 2018.

[16] J. Schwehr and V. Willert, “Driver's gaze prediction in dynamic automotive scenes,” in Proc. IEEE Int. Conf. Intell. Transp, pp. 1-8, 2017.

[17] A. Tawari et al., “Where is the driver looking: analysis of head, eye and iris for robust gaze zone estimation,” in Proc. IEEE Int. Conf. Intell. Transp, pp. 988–994, 2014.

[18] S. Guasconi, M. Porta, C. Resta and C. Rottenbacher, “A low-cost implementation of an eye tracking system for driver's gaze analysis,” in Int. Conf. Human System Interactions, pp. 264-269, 2017.

[19] Y. Wang, G. Yuan, Z. Mi, J. Peng, X. Ding, Z. Liang and X. Fu,“Continuous driver's gaze zone estimation using rgb-d camera,” Sensors, vol. 19, no. 6, p.1287, 2019.

[20] Y. Wang, T. Zhao, X. Ding, J. Bian and X. Fu, “Head pose-free eye gaze prediction for driver attention study,” in IEEE Int. Conf. Big Data and Smart Computing, pp. 42-46, 2017.

[21] S. Jha and C. Busso, “Probabilistic Estimation of the Gaze Region of the Driver using Dense Classification,” in IEEE Int. Conf. Intell. Transp, pp. 697-702, 2018.

[22] K. Yuen, S. Martin and M. M. Trivedi, “Looking at faces in a vehicle: A deep CNN based approach and evaluation,” in IEEE Int. Conf. Intell. Transp, pp. 649-654, 2016.

[23] T. Hu, S. Jha and C. Busso, “Robust Driver Head Pose Estimation in Naturalistic Conditions from Point-Cloud Data,” in Proc. IEEE Intell. Veh. Symp., pp. 1176-1182, 2020.

[24] A. Rangesh, B. Zhang and M. Trivedi, “Driver Gaze Estimation in the Real World: Overcoming the Eyeglass Challenge,” in Proc. IEEE Intell. Veh. Symp., pp. 1054-1059, 2020.

[25] S. Dari, N. Kadrileev and E. Hüllermeier, “A Neural Network-Based Driver Gaze Classification System with Vehicle Signals,” in Int. Conf. Neural Networks, pp. 1-7, 2020.

[26] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” in Proc. IEEE Comput. Soc. Conf. Comput Vis. Pattern Recognit., pp. 2074-2083, 2018

[27] M. Ramzan, H. U. Khan, S. M. Awan, A. Ismail, M. Ilyas and A. Mahmood, “A Survey on State-of-the-Art Drowsiness Detection Techniques,” in IEEE Access, vol. 7, pp. 61904-61919, 2019, doi: 10.1109/ACCESS.2019.2914373.

[28] Ji, Q., Zhu, Z., Lan, P. “Real-time nonintrusive monitoring and prediction of driver fatigue,” in IEEE Trans. Veh. Technol. 53(4), 1052–1068 (2004)

[29] Brandt, T., Stemmer, R., Rakotonirainy, A. “Affordable visual driver monitoring system for fatigue and monotony,” in IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), vol. 7, pp. 6451–6456 (2004)

[30] Bergasa, L.M., Nuevo, J., Sotelo, M.A., Barea, R., Lopez, M.E., “Real-time system for monitoring driver vigilance,” IEEE Trans. Intell. Transp. Syst. 7, no. 1, pp. 63–77, 2006.

[31] Lew, M., Sebe, N., Huang, T., Bakker, E., Vural, E., Cetin, M., Ercil, A., Littlewort, G., Bartlett, M., Movellan, J. “Drowsy driver detection through facial movement analysis,” in Human-Computer Interaction, vol. 4796, pp. 6–18, 2007.

[32] Zhang, Z., Zhang, J. “A new real-time eye tracking based on nonlinear unscented kalman filter for monitoring driver fatigue,” J. Contr. Theor. Appl. 8, pp. 181–188, 2008

[33] R. Jabbar, M. Shinoy, M. Kharbeche, K. Al-Khalifa, M. Krichen and K. Barkaoui, “Driver Drowsiness Detection Model Using Convolutional Neural Networks Techniques for Android Application,” in IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), pp. 237-242, 2020.

[34] F. You, X. Li, Y. Gong, H. Wang and H. Li, “A Real-time Driving Drowsiness Detection Algorithm With Individual Differences Consideration,” in IEEE Access, vol. 7, pp. 179396-179408, 2019, doi: 10.1109/ACCESS.2019.2958667.

[35] Flores, M.J., Armingol, J.M., Escalera, A.D.L. “Driver drowsiness detection system under infrared illumination for an intelligent vehicle,” in IET Intell. Transp. Syst. 5, no. 4, pp. 241–251, 2011.

[36] Chirra, V.R.R., ReddyUyyala, S. and Kolli, V.K.K., 2019. “Deep CNN: A Machine Learning Approach for Driver Drowsiness Detection Based on Eye State,” Rev. d’Intelligence Artif., 33, no. 6, pp.461-466.

[37] B. Dong and H. Lin, “An On-board Monitoring System for Driving Fatigue and Distraction Detection,” 2021 22nd IEEE International Conference on Industrial Technology (ICIT), pp. 850-855,2021.

[38] Zhang, Shifeng, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. “S^{3}fd: Single shot scale-invariant face detector,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 192-201, 2017.

[39] Shen, W., Sun, H., Cheng, E., Zhu, Q., Li, Q., “Effective driver fatigue monitoring through pupil detection and yawing analysis in low light level environments,” Int. J. Digit. Technol. Appl 6, pp. 372–383, in 2012.

[40] Dorazio, T., Leo, M., Guaragnella, C., Distante, A. “A visual approach for driver inattention detection,” in Pattern Recog. 40, pp. 2341–2355, 2007.

[41] Liu, D., Sun, P., Xiao, Y., Yin, Y., “Drowsiness detection based on eyelid movement,” in Second International Workshop on Education Technology and Computer Science, vol. 2, pp 49–52, 2010.

[42] Varona, L., Ortega, J.D., Leškovský, P. and Nieto, M., “Robust Real-time Driver Drowsiness Detection System for Heterogeneous Lightning Conditions.”

[43] Fan, X., Yin, B.C. and Sun, Y.F., “Yawning detection based on gabor wavelets and LDA,” in Journal of Beijing university of technology 35, no. 3, pp.409-413, 2019.

[44] J. Yu, S. Park, S. Lee and M. Jeon, “Driver Drowsiness Detection Using Condition-Adaptive Representation Learning Framework,” in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 11, pp. 4206-4218, Nov. 2019.

[45] Abtahi, S., Hariri, B. and Shirmohammadi, S. “Driver drowsiness monitoring based on yawning detection,” in IEEE International Instrumentation and Measurement Technology Conference, pp. 1–4, 2011.

[46] S.Z. Li, R. Chu, S. Liao and L. Zhang, “Illumination invariant face recognition using near-infrared images,” in IEEE Trans. Pattern Anal. Mach. Intell., 29(4), pp.627-639, 2007.

[47] S. J. Sangwine and R. E. N. Horne, “The Colour image processing handbook,” in New York: Chapman and Hall. 1998.

[48] W.Yi and S. Dongbin, “Joint exact histogram specification and image enhancement through the wavelet transform,” IEEE Transactions on Image Processing, 16, pp. 2245–2250, 2007.

[49] Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B. and Zuiderveld, K. “Adaptive histogram equalization and its variations,” in Computer vision, graphics, and image processing, vol.39, pp.355-368. 1987.

[50] M. J. Wilber, V. Shmatikov and S. Belongie, “Can we still avoid automatic face detection?,” 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-9, 2016.

[51] Deng, Jiankang, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. “Retinaface: Single-shot multi-level face localisation in the wild,” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5203-5212. 2020.

[52] Rahmad, C., R. An Asmara, D. R. H. Putra, I. Dharma, H. Darmono, and I. Muhiqqin. “Comparison of Viola-Jones Haar Cascade classifier and histogram of oriented gradients (HOG) for face detection,” In IOP conference series: materials science and engineering, vol. 732, no. 1, p. 012038. IOP Publishing, 2020.

[53] Z. Wang, J. Chai and S. Xia, “ Realtime and Accurate 3D Eye Gaze Capture with DCNN-based Iris and Pupil Segmentation,” in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 1, pp. 190-203, 2021.

[54] Xiu, Y., Li, J., Wang, H., Fang, Y. and Lu, C. “Pose Flow: Efficient online pose tracking,” in arXiv preprint arXiv:1802.00977, 2018

[55] Fang, Hao-Shu and Xie, Shuqin and Tai, Yu-Wing and Lu, Cewu. “RMPE: Regional Multi-person Pose Estimation,” in ICCV, 2017.

[56] Li, Jiefeng and Wang, Can and Zhu, Hao and Mao, Yihuan and Fang, Hao-Shu and Lu, Cewu. “Crowdpose: Efficient crowded scenes pose estimation and a new benchmark,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.

[57] Li, Jiefeng and Wang, Can and Zhu, Hao and Mao, Yihuan and Fang, Hao-Shu and Lu, Cewu. “Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021.

[58] V. Beanland, M. Fitzharris, K.L. Young, and M.G. Lenné, "Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study," in Accident Analysis Prevention, 54, pp.99-107, 2013.

[59] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," in IEEE Trans. Intell. Transp., vol. 17, no. 4, pp. 980–992, 2015.

[60] C. Lollett, H. Hayashi, M. Kamezaki and S. Sugano, "A Robust Driver's Gaze Zone Classification using a Single Camera for Self-occlusions and Non-aligned Head and Eyes Direction Driving Situations," in Proc. IEEE Int. Conf. Syst. Man Cybern., pp. 4302-4308, 2020.

[61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in IEEE Proc. of Computer Vision and Pattern Recognition, pp. 2921-2929, 2016.

[62] I. Martinikorena, R. Cabeza, A. Villanueva, and S. Porta, "Introducing i2head database," in Proceedings of the 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction, pp. 1-7, 2018.

[63] Forsman, P.M., Vila, B.J., Short, R.A., Mott, C.G. and Van Dongen, H.P. "Efficient driver drowsiness detection at moderate levels of drowsiness," In Accident Analysis and Prevention, vol. 50, pp. 341-350. 2013.

[64] Script, B. "Their diaphragm when yawning or hiccupping. The sensations when engaging. Applied Cognitive Behavioral Therapy in Schools," pp. 181, 2021.

[65] Su, Rui, Wenjing Huang, Haoyu Ma, Xiaowei Song, and Jinglu Hu. "SGE net: Video object detection with squeezed GRU and information entropy map," arXiv preprint arXiv:2106.07224 (2021).

[66] Abtahi, Shabnam, Mona Omidyeganeh, Shervin Shirmohammadi, and Behnoosh Hariri. "YawDD: A yawning detection dataset," in Proceedings of the 5th ACM multimedia systems conference, pp. 24-28, 2014.

[67] Khan, M.Q. and Lee, S., "A comprehensive survey of driving monitoring and assistance systems," Sensors, vol. 19, no.11, pp. 2574, 2019.

List of Research Achievements

Journals (Peer-reviewed)

○ [1] Catherine Lollett, Mitsuhiro Kamezaki, and Shigeki Sugano. “Single Camera Face Position-Invariant Driver’s Gaze Zone Classifier based on Frame-Sequence Recognition using 3D Convolutional Neural Networks,” Sensors, vol. 22, no. 15, 2022.

Proceedings of International Conferences (Peer-reviewed)

○ [2] Catherine Lollett, Mitsuhiro Kamezaki, and Shigeki Sugano. “Driver’s Drowsiness Classifier using a Single-Camera Robust to Mask-wearing Situations using an Eyelid, Lower-Face Contour, and Chest Movement Feature Vector GRU-based Model,” In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 4–9 June 2022.

○ [3] Catherine Lollett, Mitsuhiro Kamezaki, and Shigeki Sugano. “Towards a Driver’s Gaze Zone Classifier using a Single Camera Robust to Temporal and Permanent Face Occlusions,” In Proceedings of the 2021 IEEE Intelligent Vehicles Symposium (IV), Nagoya, Japan, 11–17 July 2021.

○ [4] Catherine Lollett, Hiroaki Hayashi, Mitsuhiro Kamezaki, and Shigeki Sugano. “A Robust Driver’s Gaze Zone Classification using a Single Camera for Self-occlusions and Non-aligned Head and Eyes Direction Driving Situations,” In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 4302–4308.

[5] Hiroaki Hayashi, Mitsuhiro Kamezaki, Udara E. Manawadu, Takahiro Kawano, Takaaki Ema, Tomoya Tomita, Lollett Catherine, and Shigeki Sugano, “A Driver

Situational Awareness Estimation System Based on Standard Glance Model for Unscheduled Takeover Situations," In Proceedings of the IEEE Intelligent Vehicles Symposium, Paris, France, 9–12 June 2019; pp. 718–723.

Domestic Conferences

[6] Catherine Lollett, Mitsuhiro Kamezaki, and Shigeki Sugano. "DNN-based Driver's Eyelid Closure Classifier using a Single Camera Robust to Challenging Driving Scenarios," 39th Annual Conference of the Robotics Society of Japan (RSJ), 2021.

[7] Lollett Catherine, Hayashi Hiroaki, Kamezaki Mitsuhiro, Sugano Shigeki. "2D CNN Driver's Gaze Direction Classification in Non-Aligned Head and Eyes Direction Situations," 37th Annual Conference of the Robotics Society of Japan (RSJ), 2019.