An Artificial Intelligence Approach to attention evaluation in assisted driving systems

Alessia Ciarlo[1,†], Alessandro Garbetta[1,†]

[1]Student of Artificial Intelligence and Robotics, DIAG, Sapienza University of Rome

Abstract

Monitoring driver's attention is an important task in order to maintain the driver safe. The estimation of driver's gaze can help us to evaluate if the driver is not focusing his attention on driving. For an evaluation of this type, we need to know what happens inside and outside the vehicle, so it was necessary to create a specific dataset for the task. In this work, in fact, we realize a machine learning oriented approach to evaluate driver's attention by means of a coupled visual perception system. It is developed, by analyzing the road and the driver's gaze simultaneously in order to understand if the driver looks at the traffic signs detected along the roads, while he is driving. Evaluating if a determined ROI contains a traffic sign or not, is performed through a Convolutional Neural Network (CNN), in particular YOLOv8. After checking which ROI contains a traffic sign, the head and gaze position of the driver is evaluated to know if he is looking at that specific ROI which contains the signal.

HIGHLIGHTS

- innovative approach to filming the driver while driving, use of a single internal camera and a single external camera;
- construction of a new dataset for head-gaze position consisting of people images during the driving;
- developed of a composed head-pose and eye-gaze-position classification model with 10 classes that represents portions of street (HEGClass);
- training of YOLOv8 to perform the detection and recognition of traffic signs;
- creation of an application that can be used directly by the final user providing only a pair of internal-external videos.

1. Introduction

The field of Artificial Intelligence (AI) applied to attention evaluation in assisted driving is rapidly expanding, driven by the development of autonomous vehicles and existing hybrid systems that support drivers, such as cruise control, lane-keeping assistance, automatic parking, and other features integrated into the latest generation of vehicles. It is well known that driver inattention is a major cause of road accidents [1, 2]. According to the statistics provided from ISTAT [3], in Italy, the top three causes of fatal accidents are: distraction while driving (15.4%), failure to observe precedence (14.3%), excessive speed (10.0%). Monitoring driver attention is emerging as an essential requirement for automotive safety systems, aiming to identify potential risks and directly prevent accidents, as in [4, 5]. Systems that allow for a complete attention monitoring should necessarily include a precise analysis of the driver's posture, head position and rotation angles, and gaze direction. Each of this information allows the observation of driver behavior and the identification of factors that characterize a person's reactions to specific conditions and scenarios, contributing to preventing future incidents related to distractions and drowsiness. Existing literature, represented by state-of-the-art studies described in section 2, predominantly focuses on the problem of driver attention by separating the internal component from the external component. Often the analysis of the vehicle cabin and of the driver gaze, in fact, is not followed by the evaluation of the environment, road situation and driver's corresponding response to an event. Some studies utilize cameras to observe the internal part of the vehicle without considering the surrounding environment [6, 7, 2, 4], while others employ external cameras and sensors to analyze driver responses solely from an external perspective [8]. The lack of comprehensive work addressing both perspectives without the use of complex equipment that is difficult-to-access for end-users, prompted us to develop our research in this direction. Our approach involves analyzing internal information through the observation of the driver inside the vehicle and simultaneously recording external information regarding the road and points of interest (signs, pedestrians, etc.) during driving. In fact, this work aims to integrate different internal and external methods for gaze recognition and correspondence with external Regions Of Interest (ROIs) into a readily usable application, addressing the issue

EAI - AI for Visual Perception in HCI & HRI, Final Project

[†] These authors contributed equally.

ciarlo.1796690@studenti.uniroma1.it (A. Ciarlo);
garbetta.1785139@studenti.uniroma1.it (A. Garbetta)

CEUR Workshop Proceedings (CEUR-WS.org)

of driver attention comprehensively. This would have immediate practical applications in everyday life, such as:

Car crashes: Having information about driver attention during a road accident could facilitate the execution of investigations, checks, and insurance procedures [9]. By utilizing an affordable camera system, video data on the driver involved in the accident could be collected and provided to an application, like the one proposed here, enabling an initial evaluation of the attention during the incident. Regarding the insurance premium, it could be adjusted based on the insured individual's overall behavior, assessing whether they are more or less conscientious while driving. In fact, in the past 20 to 30 years, more and more insurance companies around the world have launched vehicle usage-based insurance (UBI) products based on driving style analysis as shown in [9].

Driving schools: Another practical example of potential applications is to deploy the new system on vehicles used in driving schools. It could serve as a valuable tool for instructors to understand and potentially correct the driving behavior of novice drivers, enhancing road safety in the future.

Autonomous driving cars: Understanding where a driver's attention is focused during specific driving moments, how long they pay attention to certain aspects and what they consider irrelevant can be crucial for autonomous driving projects. As seen in [10], the attention behavior of the driver is used in the ADAS (Advanced Driver Assistance System) solutions.

To develop our work, we focused on computer vision-based approaches, which are becoming increasingly popular compared to physiology-based methods that require the use of biological sensors attached to the driver. Vision-based methodologies, in fact, only utilize one or more cameras (reducing the cost of experiments and fostering the growth of approaches for this problem) to monitor and analyze driver behaviors without disturbing or inconveniencing the individual by wearing devices such as eye-tracking glasses or brainwave recognition devices. The works that have most inspired our approach are [4, 11, 12], from which we took the ideas of developing a grid of 9 cells for predicting the ROI of the driver's gaze, of using a VGG16 network for feature extraction from facial video frames and of adding the head-pose (roll, pitch, yaw) information to the gaze-position. The most in-depth analysis in our work focused on finding the best method and features to extract from images that would be useful in accurately determining the driver's gaze-position, thus understanding what they were looking at while driving. The difference between tracking gaze-position when a person is looking at a monitor, as in [11, 12], and tracking it while they are driving a car is substantial. When looking at a monitor, head movements are imperceptible, so the only discriminant is the position of the pupil. During driving, however, the driver tends to rotate their head to look at vehicles and pedestrians on the right and left sides of the road or tilt it to see street names, signs, or higher traffic lights. They also shift their gaze to look at mirrors or to initiate a reverse maneuver. For this reasons, analyzing only pupil movement was insufficient for our task and it was necessary to have additional information about head pose (rotation angles) and characteristics of eyes or facial images, Figure 2. In addition to methodological research, another challenge in this work was finding a suitable dataset for driver attention monitoring. The final decision to develop our own dataset arose from the fact that all the datasets we found and analyzed had proper documentation of driver

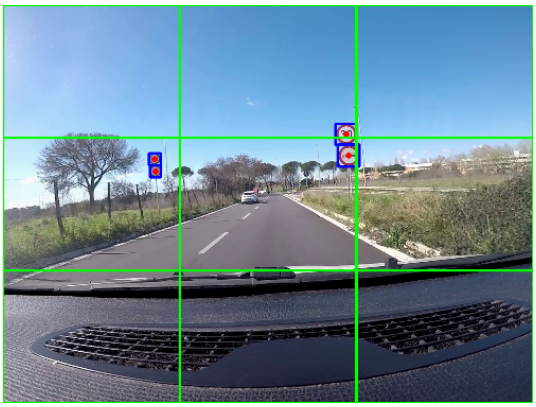

Figure 1: Example of external image at final step of our work. This is an external frame with different regions of interest. In this case the cells 4, 5 and 6 contain at least one traffic sign.

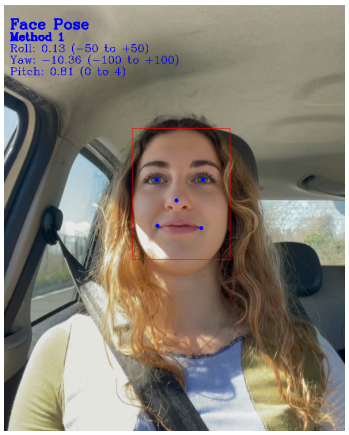

Figure 2: Example of internal image extrapolated from one of the videos in correspondence with an external one to classify the gaze cell. In this case the driver sees cell number 5.

behavior but lacked corresponding external real-world observations. In other cases, datasets focused exclusively on gaze analysis consist of collections of images of people looking at points on a computer screen.

Therefore, we collected our own dataset to work with both external and internal videos during driving sessions and so to allow the final application to simultaneously process and correlate information from different perspectives. For training the two parts that compose the final application, we used two additional datasets. For the internal part, which involves the model prediction of the driver's gaze-position, we created the HEAD-POSE dataset with four subjects (rather than just one, as in many other datasets). For the external part, which involves to the model prediction of the position of road signs, we used a customized datasets of signs available on the internet. We chose to use images from MAPILLARY and GTSDB datasets and, from the first one, to remove foreign signs that did not align well with the signs encountered during our recordings, retaining only those conforming to European traffic regulations governed by the Vienna Convention of 1968.

In section 2, we investigate the relevant works from recent years.

In section 3, we explain the type of data we collected and the models and approaches tested in our work. The first phase involved extracting facial features. After identifying the face and the landmarks through MTCNN model, we extracted the eyes pupils and the head angles. Subsequently, we tested differents types of classification models, including SVM, CNN and VGG. The VGG-Net 16 model inspired approach proved to be the most effective for gaze ROI classification. We passed to it the facial images and the gaze features, such as the coordinates of the pupils and the roll, pitch, and yaw values of the head, to have the final prediction. Next, we trained YOLOv8 model for road sign detection, which allowed us to analyze the extracted external frames from the driving sessions' videos. Finally, we developed an application that unify and cooperatively work with both components. It analyzes the external frames and, when a road sign was detected, it examines the corresponding internal frame to check if the driver's gaze was towards one of the objects of interest in our investigation.

In section 4, we present the results and discuss the critical aspects leading to our final conclusions.

2. Related Works

As previously discussed, there has been a growing interest in analyzing driver attention during driving in recent years. This includes understanding whether a person is observing the road, being distracted, remaining vigilant, or experiencing drowsiness.

Most state-of-the-art, as in [6, 7, 2, 4], approaches are based on the unique observation of the driver's interior cabin to understand their behaviors. One or more internal cameras are used to observe the driver and determine if they are looking at the infotainment system, the road, the mirrors, or, for example, other passengers.

For instance, in [2], the position of the hands and arms is studied to assess whether the driver keeps their hands on the steering wheel or in other positions, such as holding a phone.

On the other hand, other approaches solely focus on external factors by studying the surrounding environment and collecting information about the vehicle's movement (speed, position) to study the driver's reactivity in specific circumstances. For example, in [8], various sensors such as cameras and lidar, applied to the external part of the vehicle, allow the observation of the driver's reaction in certain situations.

However, there are few projects that simultaneously analyze both internal and external images of the vehicle while assessing driver attention to the road from the driver's perspective. In cases where interior cabin images are associated with external frames, the driver's viewpoint is often recorded using glasses or equipment that track eye movements, as demonstrated in [10], which directly indicates what is being observed.

Regarding the specific gaze detection, various approaches are used in simulated or real environments, both indoors and outdoors [13]. In most literature works and datasets, recordings are made using a personal computer's webcam while the subject looks at specific points on the screen for certain moments [12]. With this regression problem, the aim is to recognize the precise gaze position on the monitor by studying the direction of the pupils and gaze triangulation [14, 12, 5, 15, 16, 17].

These types of problems can also be approached through classification. For example, in [11], images are taken of a stationary subject in front of a personal computer screen, ideally divided into a 9-cell grid, and the gaze position is returned not as precise point coordinates but as the number of the observed cell (classification). In the mentioned paper, pupil characteristics are extracted and then classified using an SVM model. This latter work has inspired our choice to implement a classification approach, given the problems and requirements already described and specific to the field of driving.

Another approach for studying gaze position [12] is based on a dataset of tens of thousands of photos, collected using a personal computer's webcam, from which facial information are extracted and eyes images are cropped and passed to a VGG network.

In [4], in addition, information related to head position (rotation angles - roll, pitch, yaw) is utilized. Specifically, images collected from multiple subjects, in multiple vehicles, and under different weather conditions were used

to regress the viewpoint position.
Finally, additional approaches make use of recordings in simulated environments using various technologies, from simulators to simple computer-played videos. For example, in [13], the user's gaze position is recorded while watching driving videos shortly before certain incidents, in order to understand which objects the driver (simulated in this case) would have focused on.

3. Methods

This work investigates a novel approach for gaze recognition during driving to assess driver attention. The first step involved collecting new data aligned with our specific task. Subsequently, for the internal part of the vehicle, we have proceeded training and testing neural networks for gaze classification, experimenting different models: SVM, ClassNET, VGG16-based Net and HEG-Class Net. Additionally, for the external part, a training phase was conducted using a custom dataset of traffic sign objects. Once the various results were obtained, the parts were combined, allowing for a comprehensive study on video recordings captured during real-world driving scenarios. This integrated approach provided a more holistic understanding of gaze behavior and its relationship to driver attention in everyday driving situations.

3.1. Dataset

A series of images and videos were collected and used in various stages of development. They can be divided into four distinct collections described below in details.
HEAD-GAZE Dataset[1]: A collection of images, taken by us, of 4 individuals in a driving environment to classify their gaze position in a grid of ten cells.
DRIVERS-on-DRIVE Dataset[2]: A collection of external and internal videos, recorded by us, of 4 individuals while driving.
TRAFFIC OBJECTS: A modified version of Mapillary Dataset[3] containing images of traffic signs.
TRAFFIC SIGNS DATASET in YOLO format (TSDY): a collection of images from GTSDB (German Traffic Sign Detection Benchmark) that can be downloaded from Kaggle[4].

To build our own collections we consistently used the same equipment setup, described in Figure 3 : a city car for the HEAD-GAZE and DRIVERS-on-DRIVE Datasets and the camera of an iPhone 12 to collect the internal images and to record the interna videos. The iPhone was positioned behind the steering wheel to capture the driver prominently, while minimizing irrelevant information. Additionally, a GoPro Hero5 camera was placed centrally on the car's dashboard to record external footage during the drive.

[1]Release forms signed by the participants
[2]Release forms signed by the participants
[3]https://www.mapillary.com/dataset/trafficsign
[4]https://www.kaggle.com/datasets/valentynsichkar/traffic-signs-dataset-in-yolo-format

HEAD-GAZE Dataset specification

Classes	Male	Male_glasses	Female	tot
0	39	27	34	100
1	41	22	37	100
2	41	24	36	101
3	41	22	37	100
4	39	25	35	99
5	38	30	33	101
6	40	29	32	101
7	44	22	39	105
8	41	22	36	99
9	42	21	37	100
tot	406	244	356	1006

Table 1
The HEAD-GAZE Dataset collect the gaze positions of four drivers. The collection contains a total of 1006 images, 406 for male individual without glasses, 244 for man with glasses and 356 images for female.

Regarding the HEAD-GAZE Dataset, we collected images from four different subjects (two males and two females) and, for two of them, both with and without glasses. The photos were taken at various hours of the day and under different lighting conditions. A total of 1006 images were collected and distributed as shown in Table 1. The subjects were seated inside a car and, once their driving position was established by adjusting the

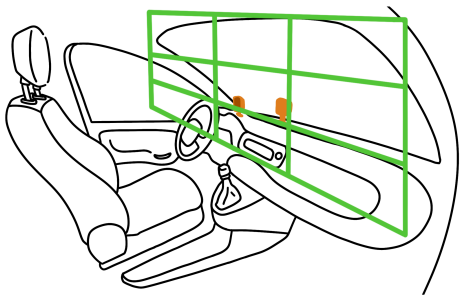

Figure 3: The setup of city-car environment during the collection of images and videos. In orange there are the two cameras, GoPro Hero5 to record the external street and iPhone 12 to record the the driver. In green the virtual grid that represent how the gaze position area is divided.

seat distance, photos were taken with different gaze and head positions. To achieve the desired classification, an imaginary grid was created to divide the external view and the driver's gaze into a 9-cell grid. This approach establishes a correspondence between the head and eye positions and specific portions of the external images, allowing for the identification of ROIs matching during the experiments. The classes in the dataset are the 9 cells of the grid, plus an additional tenth position indicating lack of attention towards the road (e.g., face turned sideways, gaze directed downwards, upwards, etc.). In this initial image collection the subjects were not driving, the photos were taken while they were changing only facial positions looking at specific points in front of them.

The DRIVERS-on-DRIVE Dataset consists of videos recorded during driving. Using the same setup as the previous dataset, the subjects were filmed during various driving sessions, capturing both the driver and the road ahead simultaneously. Approximately one hour of recordings was collected, featuring different subjects and varying lighting and weather conditions. The videos, both external and internal, underwent a pre-processing stage where synchronization was performed using voice cues to ensure they were completely matched. Third-part software, specifically DaVinci Resolve, was used for this purpose. Additionally, the videos were divided into subclips of 30 seconds each, optimizing the manipulation of them during subsequent steps. Each of the 191 sub-clips extracted was labeled with "CAREFUL" or "NOT CAREFUL" based on the driver's driving style and the attention given to the road; there is also the information if the driver wear glasses or not. Following the pre-processing stage and frame extraction, the resolution of the external images is (1440x1080), while the resolution of the internal images is (1080x1920).

Figure 4: Some examples of images by the HEAD-GAZE Dataset after the extrapolation of head pose with bounding boxes. The pictures are cropped (224,224).

The first dataset used for the traffic signs detection and recognition is the TRAFFIC OBJECT from the Mapillary Traffic Sign Dataset (It is provided under the Creative Commons Attribution NonCommercial Share Alike (CC BY-NC-SA license)). This dataset contains tens of thousands of images collected from roads worldwide and includes information about the type of sign and its pixel coordinates within the image. To ensure that the images were suitable for our purpose, we decided to focus solely on Italian/European traffic signs, filtering out images with signs that significantly differed in shape or content. Approximately 20,000 images were reviewed, and out of these, around 3,000 images were selected for our dataset. The chosen images exhibit a wide diversity in terms of brightness (captured at different times of the day, exposure to sunlight), positioning within the frame, and context (various degrees of obstruction by foliage or other objects present, and background). From these 3,000 images, the corresponding JSON files were extracted and converted into TXT files, formatted in a way that could be interpreted by the YOLO training model. The conversion process involved extracting relevant information from the files, such as the bounding box coordinates and class labels of the traffic signs, and organizing them into a format compatible with YOLO's training requirements. This step was essential to prepare the dataset for training the model to recognize and localize traffic signs. Furthermore, it was decided to modify the labels of the dataset. Mapillary contains a wide range of classes for traffic signs (over 300). To simplify the classification task, the decision was made to reduce these classes to their super categories, which were extracted from the JSON file. Three groups were created, namely ['PROHIBITORY', 'DANGER', 'MANDATORY'], which encompass the majority of traffic signs that are relevant during driving. This categorization allowed for a more streamlined and focused training process, focusing on the specific types of signs that are crucial for driver attention and safety.

The last dataset used for this work (for the traffic signs detection part) is the TRAFFIC SIGNS DATASET in YOLO format (TSDY), a refinement of the larger GTSDB (from Institut für Neuroinformatik RUB). The collection contains 741 images with labels already express in Yolo format. Each image resolution is 1360x800. The number of classes, modified from the GTSDB version (about 40), are four ['PROHIBITORY', 'DANGER', 'MANDATIRY', 'OTHER'].

3.2. Gaze Classification

Various approaches, represented in Figure 5, were tested to find the one that had the right balance between high accuracy in predicting gaze position and generalization ca-

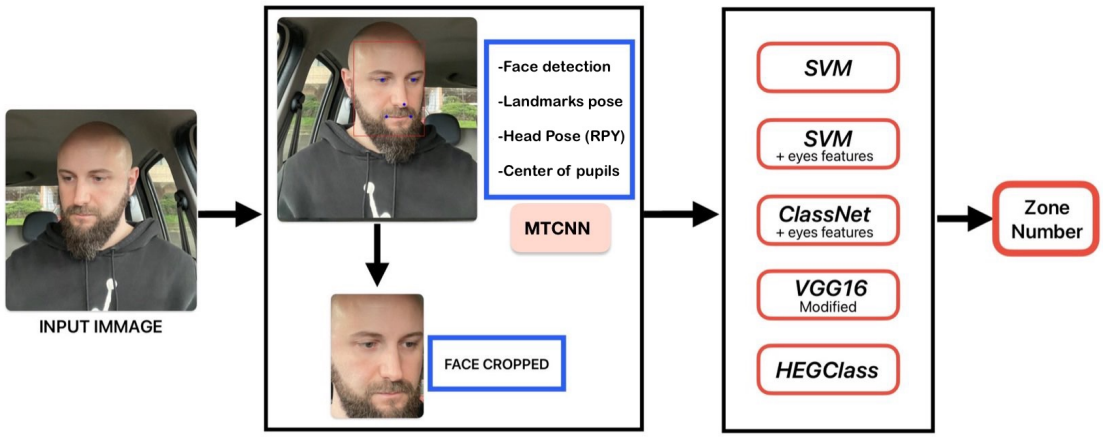

Figure 5: Gaze Classification scheme, composed by Head-Pose Estimation Part and Classification Part (with all the tested approaches). The input image is mirrored by the MTCNN to have a more user friendly representation of the right gaze position (cell 3-3 for the image in figure). The result of the model is the zone number referred to the gaze position.

pability for images with different contrast and/or brightness characteristics and different subjects.
The general method takes an input image (or video frame) and analyzes it through two steps or macro-parts: the *head pose estimation part* and the *classification part* for the final gaze position prediction. The first part is common to all the approaches tested, while the second part has been modified several times in terms of structure, model type, and input received until reaching the final solution.

3.2.1. Head-Pose Estimation Part

To detect the face in the image, we used a pre-trained MTCNN model [18], which, after various tests with other methods, proved to be the most accurate in finding the face region (bounding boxes) and accurately positioning the 5 landmarks (left eye, right eye, nose, and 2 mouth corners).
Multi-task Cascaded Convolutional Networks is a face detection method proposed by Zhang et al. in 2016 [18]. It is developed as a solution for both face detection and face alignment. The process consists of a cascaded structure with three stages (P-Net, R-Net, and O-Net) of convolutional networks that can recognize faces and landmark locations such as eyes, nose, and mouth. In the P-Net stage, it uses a shallow CNN to produce candidate windows. In the R-Net stage, it refines the proposed windows through a more complex CNN. Lastly, in the O-Net stage, it uses a third more complex CNN to output facial landmark positions. This method performs very well with any type of face, even in the presence of long beards and glasses and it has much better results compared to the Haar Cascade Classifier method [19], especially regarding landmark positioning in profile photos or photos with partially closed eyes.
Then, using the landmarks found, it geometrically calculate the head rotation angles of the driver, roll, pitch, and yaw, and the position of the pupils as features to be passed to the final classifier. This information proves to be useful and necessary for the classification model to determine the gaze position accurately, regardless of the physical characteristics of the drivers.

3.2.2. Classification Part

To determine the driver's attention in presence of road signs, our approach is based on classification. Therefore, the field of view is divided into 9 parts, to which a tenth zone was added to classify head and gaze positions that do not fall into the 9 frontal zones, i.e., gaze towards the mirrors or to the passenger seat.
Classification methods based on SVM, CNN, VGG16-net structure, both trained or fine-tuned, were developed and tested. The first method described is the novelty of this paper. Then an overview of the other ones and comparison between them all is done.

HEGClass The classification model that achieved the best results was the HEGClass (Head-Eyes-Gaze Classifier), the hybrid approach introduced in this paper. It takes as input the cropped face images resulting from the bounding boxes found in the head-pose estimation step, as well as the rotation angles of the head and the pupil

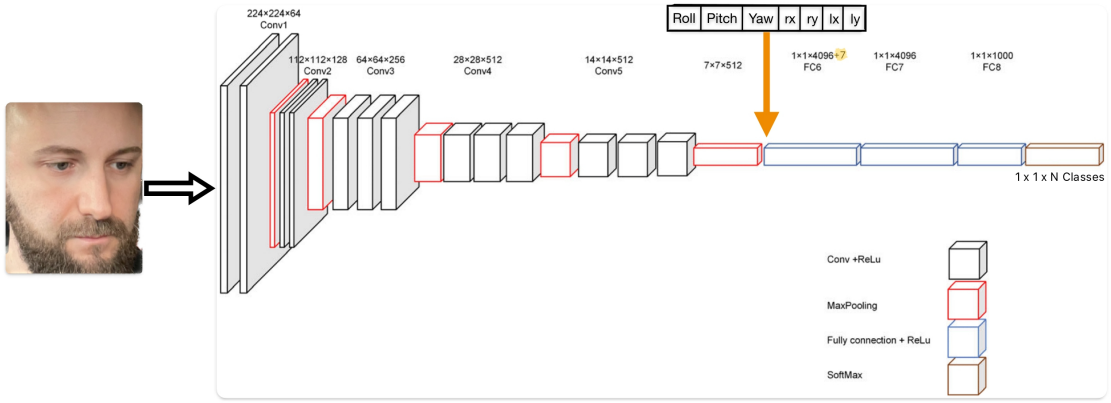

Figure 6: HEGClass - The hybrid approach with a pretrained VGG16 part to extract image features and a fully connected part in which the other 7 features [Roll, Pitch, Yaw, rx, ry, lx, ly] are added to arrive to the final prediction

centers. The use of all these characteristics combined has allowed this approach to achieve high levels of precision in the classification of the Region of Interest towards which the gaze is directed.
Inside the HEGClass network, as seen in Figure 6, first useful features are extracted from the RGB images of the face using a pre-trained VGG16 network, returning the output of its penultimate layer. Then, these features are flattened and concatenated with the normalized array, of size 7, containing roll-pitch-yaw of the head and the 4 coordinates of the pupil centers. The combination of these features pass through two fully connected linear layers followed by two relu activation functions and finally through a last fully connected linear layer, which, aided by the presence of the softmax activation function, determines the class membership among the 10 possibilities (9 ROIs for the frontal regions and 1 for others).
The training of this model was performed using our own HEAD-GAZE dataset, with 10 epochs (to avoid overfitting), 32 samples per batch, Cross-entropy loss function and Adam Optimizer.
We further evaluated the following methods to make a comparison of the results:

SVM Support Vector Machine is a supervised classification algorithm that constructs a hyperplane or a set of hyperplanes in a high-dimensional space to separate elements from different classes. The best hyperplane is the one with the maximum distance from the nearest point of each class. In our case, for the classification of 10 classes, we used an SVM with a polynomial kernel of degree equal to 4, a regularization parameter of 100, and a coefficient of 10. The training was performed always using images from the HEAD-GAZE dataset.
From these images, the roll, pitch, yaw, and pupil centers were extracted by passing through the previous head-pose-estimation part. Subsequently, using the Haar Cascade Classifier [19], eye patches were extracted from each grayscale image and passed to a pre-trained ResNet to obtain the 2048 features for each eye. This two sets of features are then averaged together to have a single array of 2048 elements containing the useful information of the 2 eyes.
The final samples, composed by the features of the eyes, roll, pitch, yaw and pupil centers for each image in the dataset, are l2 normalized and then are passed to the SVM for training and testing. The results are shown in the Results Section Table 2 and compared with the other approaches.

ClassNET Using the same set of features for each image - roll, pitch, yaw, pupil centers and 2048 features extracted from both eyes using pre-trained ResNet - the ClassNet network was trained. This Classifier Network is composed by two convolutional layers with relu activation functions, a maxpooling layer and two fully connected layers with relu and softmax activation functions. ClassNet training was performed in 300 epochs, with MSEloss function (Mean Squared Error), Adam Optimizer and it predicts, for each photo, which of the 10 zones the driver is looking at. The results are shown in the Results Section Table 2 and compared with the other approaches.

VGG16-based Net Before reaching the final model (HEGClass), another test was conducted by creating a network based on the architecture of a VGG16 network, consisting of 13 convolutional layers and 3 final fully connected layers. This network was trained with 1006 samples from the HEAD-GAZE dataset, each composed by the face image tensor, the head rotation angles (roll,

pitch, yaw) and the pupil centers. Here, the features of each image, are extracted directly within the classification network from the RGB image of the entire face, not just from the eyes. The other features (roll, pitch, yaw and pupil centers), collected previously through the Head-Pose Estimation Part, were added to the 512 features of the image in the first fully connected layer. The training was performed in 100 epochs, with Cross Entropy Loss function and Adam Optimizer. The results are shown in the Results Section Table 2 and compared with the other approaches.

3.3. YOLO training for traffic signs

For the detection and recognition of traffic signs, the pre-trained YOLOv8 model was used and the YOLOv5 model was tested before transitioning to the v8 version. We performed fine-tuning of YOLOv8 using two different datasets, TRAFFIC OBJECTS and TSDY. The final set of weights chosen for the ultimate application of this work comes from the dual fine-tuning of YOLOv8 with both previously mentioned datasets.
The first fine-tuning with the TRAFFIC OBJECTS dataset involved 1802 images for the training set and 919 for the validation set. Starting from a collection of 3000 images, various tests were conducted by modifying the training and validation sets. The original dataset did not have a balanced set, and furthermore, by adjusting the labels based on sign categories (as described in the Dataset section), these imbalances remains.
Then, using the weights resulting from the first training, it was retrained with images from the TSDY dataset using 600 images for training set and 141 for validation set.
The dual fine-tuning approach yielded improved performance in terms of final accuracy (as shown in the Results section Figure 11) and enhanced generalization capabilities in detecting road signs, even in images (frames from the Drivers-on-drive dataset) under various lighting conditions.
During all the training process, in fact, various image augmentation techniques were employed to enhance the diversity and generalization capability of the training data. The techniques used, from the Albumentations library, are: Blur, MedianBlur, and CLAHE (Contrast Limited Adaptive Histogram Equalization). These techniques introduce controlled variations to the images, such as blurring or enhancing contrast, to simulate different real-world conditions and improve the robustness of the model. To optimize the training process, the Stochastic Gradient Descent (SGD) optimizer was employed. SGD is a popular optimization algorithm in deep learning, which updates the model's parameters based on the gradients computed on small batches of data. A learning rate of 0.01 was chosen as the initial value for controlling the step size of parameter updates during training.
After the training of the network, for each image passed to the model, it returns a text file with one line per detected sign and its position. From this file, by processing the pixel coordinates, the information was extracted to reconstruct the position of the sign's center and edges in order to be located within the grid cells. In many cases, the signs spanned across multiple cells, which is why it was necessary to identify the sign using multiple coordinates.

3.4. Application part: fusion of branch

The final application, composed of two main parts as shown in Figure 7, allows to have an output report in CSV format of the overall behavior of a person driving. It requires as input an internal video that films the driver and an external one that films the street. To facilitate the analysis, we require 30-second video synchronized between the two parts and we provide functions for video processing. After segmenting the video into frames, the application initializes the various data structures, files and models. The first step involves the analysis of external frames using the YOLOv8 model trained on road signs. The result is a text file containing information and specifications of the detected signals, including the Regions of Interest in which these signals are found. For each image with at least one sign detected, the program takes the corresponding internal image from which information are extracted through the Head-Gaze part. Image and info pass through the network to classify the position of the driver's gaze. Once the prediction of the cell observed in that particular frame is obtained, it is compared with the position of the corresponding sign. In some external frames, multiple marks are detected in a single image and therefore multiple regions of interest, allowing the driver to be "signal alert" even when looking at one of them. Therefore, it was necessary to provide more acceptable positions without favoring a particular type of signal among those extracted. Since a single road sign can extend over several cells, in the design phase, it was decided to keep 5 points which characterize the position of the object, the four corners and the centre. Observing only a part of the sign is also considered as attention to the sign. Therefore, if the driver looks at the cell that contains a partial view of the sign, that sign is considered observed, also taking into account the peripheral vision of the human eyes. Furthermore, even when there are no signals present and/or the driver is looking at cells 4/5/6 (which identify the entire road surface, including the side areas from which cars or pedestrians can arrive), he is still considered "attentive to street". If gaze is rated a 7 or 8, the person is considered to be engaged in looking at

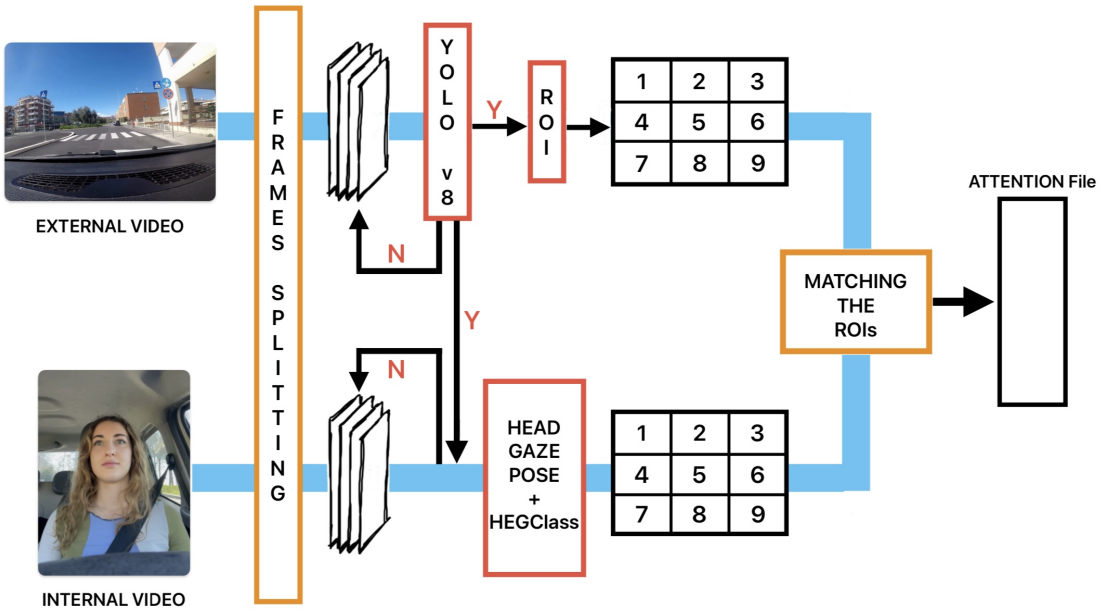

Figure 7: Scheme of the final application that unify Gaze Classification part and YOLO for traffic signs part

the car's dashboard and infotainment system. Finally, as already mentioned, a CSV file was created to store the information obtained during the analysis of the 30-second videos. Each row contains the following data: the number of frames (same for internal and external video), the number of road signs present in that frame, the number of the cell(s) in which the detected signals are located, the predicted value from the network of the cell watched by the driver, the number of signals observed after matching the two internal and external regions of interest and a value that indicates whether the driver was attentive or not.

4. Results

The goal of this work is to provide a compound system that can be directly used to analyze a posteriori a person's attention to driving by taking only two synchronized videos as input. Since this is the first application that works simultaneously with both sides (internal and external) we do not have a direct references or comparisons to do with our final results and the ones of other similar works in the scientific literature. All the evaluations and comparisons that will follow, therefore, will concern the individual parts that make up the final work. However the final application, from the numerous tests carried out using the DRIVERS-on-DRIVE dataset consisting of about 194 videos of 30 seconds each, shows good performance and gives as output a file with a percentage of attention of the driver that is consistent with the label assigned, at the time of data acquisition, to the respective video evaluating the driving behavior of the person concerned.

4.1. Results Gaze Classification

Also the internal part, like the final application, is composed of multiple sub-models, the results of which need to be described individually in order to provide an optimal view of the final predictions.
Regarding the face detection and the landmarks extraction to calculate the angles of facial rotation, the MTCNN model [18] has immediately shown much better results compared to using the Haar Cascade Classifier [19]. This because the Haar Cascade performs well when the person's face is clearly visible. However, our dataset for training the internal part (HEAD-GAZE Dataset) includes images with rotated or profiled faces, and the frames extracted from the testing videos may be blurred or have lowered eyelids. The MTCNN has met our specifications especially because one of the predicted landmarks corresponds to the center of the pupil, which was a crucial data for training the final classification network. Nevertheless, in some rare cases, the network fails to detect the face or slightly misplaces

Gaze Classifiers Results

Method	Accuracy	F1-Score
HEGClass	**96**	**94,3**
SVM	74,5	74
ClassNet	56	45,6
VGG16-based Net	81	79

Table 2
Accuracy and F1-Score of all the tested and compared methods for Gaze Classification.

the landmarks, introducing a small error in this initial part.

Regarding the prediction of gaze direction, due to the inability to precisely determine the exact point on the road that the person is looking at and considering the human eye's ability to have a holistic view of an area in front of oneself, an approach of "zone" classification was chosen instead of regression (predicting coordinates of a point in space). Compared the tested methods, the one based on SVM was inable to surpass an accuracy level of 70%, probably due to the fact that many of the features of the 1006 samples have very similar values (especially those resulting from passing the eye images to the pre-trained ResNet network). Moving to a neural network-based approach, the ClassNet network has lower accuracy even compared to SVM, despite tests with different features: only head rotation angles and pupils or adding those extracted from ResNet directly from the face or eye images. Trying a network based on the structure of VGG-16 and retraining it from scratch with our data, yielded better results arriving to an accuracy level of 81%. The issue in this case lies in the number of samples in our dataset, which is too low to train such a network and the computing capabilities at our disposal, which, even with data augmentation on our dataset, did not allow us to complete sufficiently long training sessions. For these reasons, we opted for the hybrid approach, HEGClass, which achieved 96% accuracy and 94,3% F1-score with training without additional data. In Figure 8 and Figure 9 the Confusion matrix and the loss plot of HEGClass method. All the accuracy and F1-score results are shown in Table 2.

4.2. Results YOLOv8 Classification and Detection

With the dual fine-tuning of YOLOv8 network, the final F1-score achieved is around 95%, as shown in Figure **??**. In the same image the results obtained from training the network with only the first dataset (TRAFFIC OBJECTS) are also compared. It can be observed that the F1-score value, in the case of training with a single dataset, is significantly lower, even considering the two confusion matrices in the Figure 12.
About the single training of YOLOv8 with TRAFFIC OBJECTS dataset only, the overall accuracy is around 65%-70%. With the YOLOv5 version the accuracy is higher, but extraneous elements different from road signs are recognized. For example, empty spaces between bare branches are often identified and classified as a mandatory or danger sign, as shown in Figure 10. The values of boxloss, precision, recall and F1-score of YOLOv8 and YOLOv5 trained only with the first dataset, are similar and do not undergo significant variations. One noticeable aspect is that the number of signs detected by the models varies slightly. The problem of the unbalanced dataset mainly concerns the class of signs, with the result that some classes could be not recognized due to the "subsampling" of the class.

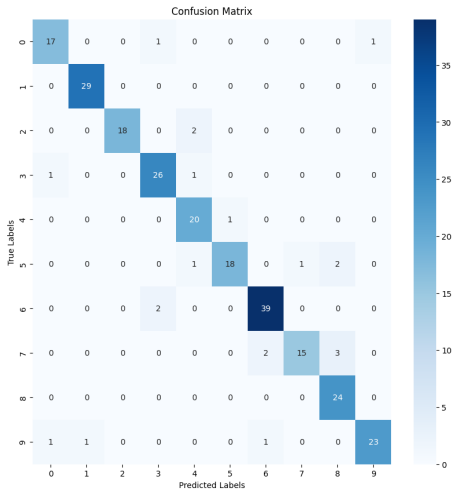

Figure 8: Confusion Matrix resulting from the training of HEGClass on the HEAD-GAZE Dataset (10 classes) over 250 validation samples.

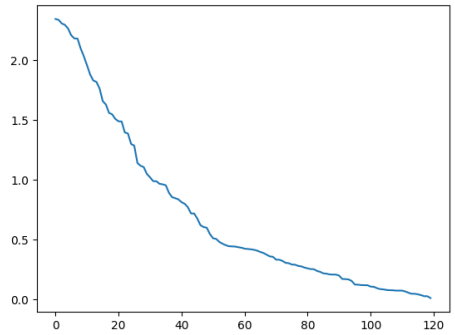

Figure 9: Plot of the change of loss during the training process of HEGClass on the HEAD-GAZE Dataset (10 classes).

Regarding the final dual training of YOLOv8, despite the substantial increase in the network's generalization capabilities, the presence of errors in sign recognition has not been completely eliminated. In some cases, for instance, there are objects along the road that can easily be mistaken for road signs, such as advertisements containing elements that, with low resolution, could be confused. This issue is present, but not in a quantity that would significantly affect the quality of the obtained results. With a wider variation of images, this problem could likely be mitigated. Another problem when classifying signals arises from the fact that, for convenience, signs of different shapes and colors are grouped in the same class (e.g. STOP signs were different from other signs in their class). This created a bias in their classification. Also, in some cases, there are signs that contain other signs within them (e.g. parking sign reserved for disabled people, which often includes a prohibition sign inside it). The problem is that, while driving, this type of signal is not of interest until the moment when you need to park. To address this problem, some signs have been eliminated from the training phase. Since the accuracy values in detection are quite high, the confidence threshold can be adjusted to mitigate the issue of misclassification and eliminate recognized elements that are not of interest. It sometimes happens that signs are recognized (due to their shape) even when they are rotated or facing the opposite direction of the lane, or when they are part of a road section that is not relevant to us (e.g., left signs while we need to turn right). In such cases, they will be counted as points of inattention.
The problem of imbalance in the presence of various classes leads to the correct detection of the sign in the image, but not always to the accurate classification of the type of road sign. Since, for the purpose of our work, the specific type of recognized sign is not of particular interest, it was decided not to extract the label from the YOLOv8 network results, but only the information related to the bounding box that defines its position.
However, a final attempt was made by modifying the dataset to recognize only one class, "TRAF_SIGN". Yet, this approach also encountered difficulties between detecting signs and identifying environmental areas unrelated to road signs. After several attempts, the decision was made to revert to using the dataset with the original labels.

4.3. Final results

The final results of the application, regarding the prediction of the driver's average attention while watching a video, are very good in almost all cases with a few specific exceptions. In Figure 13 we plotted the distribution of attention results. Each value of attention (from 0% to 100%) represents the ratio of matching between the driver's gaze and traffic signs' position. The tests were done on 191 videos of 30 seconds each present in the DRIVERS-on-DRIVE dataset. After some tests, videos were eliminated from the ratings because they were compromised due to low light (video in an almost night environment) or excessive blurring of frames that did not allow prediction. The final number of videos used is 158. In cases where the images are not well-defined, blurry or shaky, the most predicted class is 0, indicating that the model fails to recognize the correct gaze position. Furthermore, in nighttime or low-light videos, the brightness significantly affects the accurate evaluation of the gaze. Additionally, YOLO also struggles to correctly detect the relevant signs, often confusing them.
Moreover, classes 5 and 6 are frequently detected as the gaze position during driving, but this is consistent with the fact that these two areas correspond to the central ones.In some cases, such as when the vehicle is stopped at a traffic light or in traffic, the system recognizes the same signs in multiple frames. However, the driver may not pay attention to them for the entire time because they have already seen them and they are not of significant relevance in that particular moment.

Figure 10: Frame extracted from the DRIVERS-on-DRIVE dataset predicted by the fine-tuned models YOLOv5 and YOLOv8 (used in this work), the upper and lower images respectively.

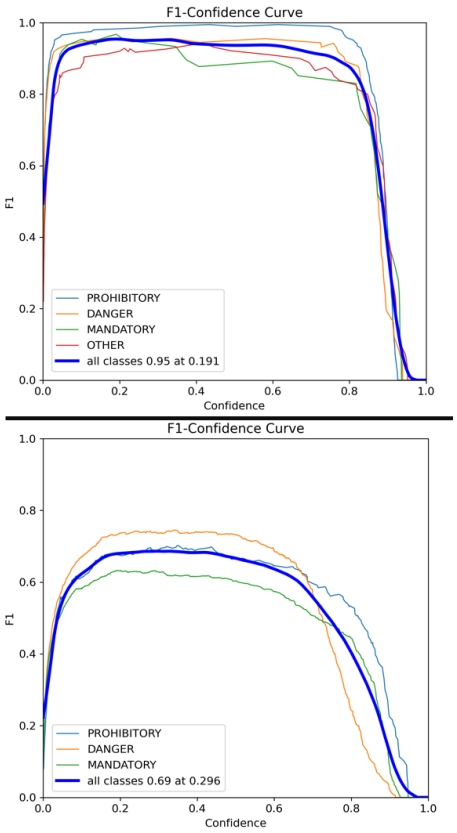

Figure 11: F1-Confidence Curve of the second fine-tunining of YOLOv8 with TRAFFIC OBJECTS and TSDY dataset vs the single fine-tuning of YOLOv8 with TRAFFIC OBJECTS only, the upper and lower images respectively.

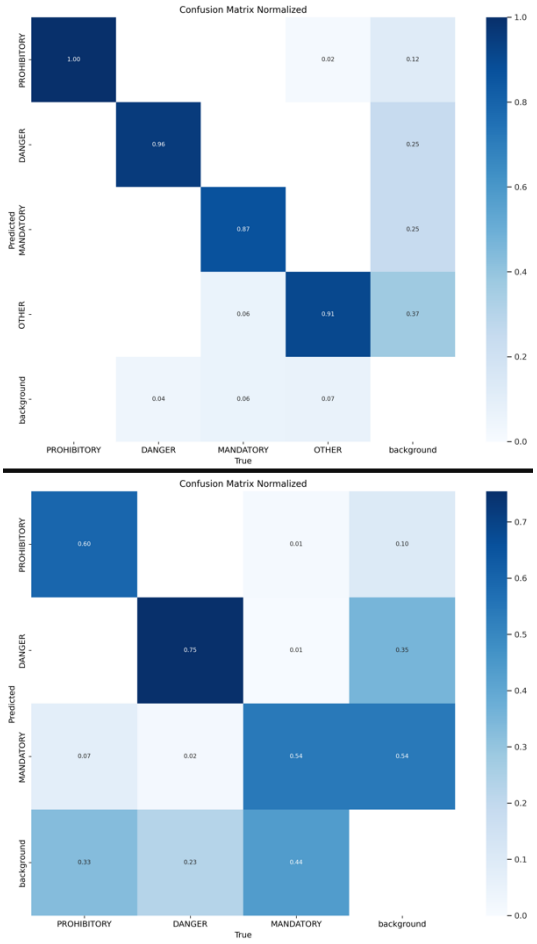

Figure 12: Confusion Matrix Normalized of the second fine-tunining of YOLOv8 with TRAFFIC OBJECTS and TSDY dataset vs the single fine-tuning of YOLOv8 with TRAFFIC OBJECTS only, the upper and lower images respectively.

5. Conclusions

With this work we wanted to develop and analyze a complete system concerning the evaluation of attention to the traffic signs in the driving environment. In order to do this we have created two new datasets (HEAD-GAZE, DRIVERS-on-DRIVE) and we have used others two (TRAFFIC OBJECTS, TSDY) with some modification to make them more suitable for our task. We divided the final application into two parts and used YOLOv8 for sign prediction and MTCNN + HEGClass for gaze position classification. The overall accuracy of the final system is very good, despite having accumulated the partial errors of the two parts that compose it.
The problems encountered during the various training and testing phases of the two parts, described in the Results, are the starting points for future works.
In fact, the work lends itself to great margins for improvement, including: keep track of the signals seen and those still to be seen, in this way, a signal already recognized as seen by the driver is no longer taken into account in subsequent frames; improve both internal and external prediction in different light and atmospheric conditions by adding more samples to datasets or working with computer vision methods on all images before passing them to networks; improve our datasets by adding more elements to make them more complete.

References

[1] G. Fitch, S. Soccolich, F. Guo, J. McClafferty, Y. Fang, R. Olson, M. Pérez-Toledano, R. Hanowski, J. Hankey, T. Dingus, The impact of hand-held and hands-

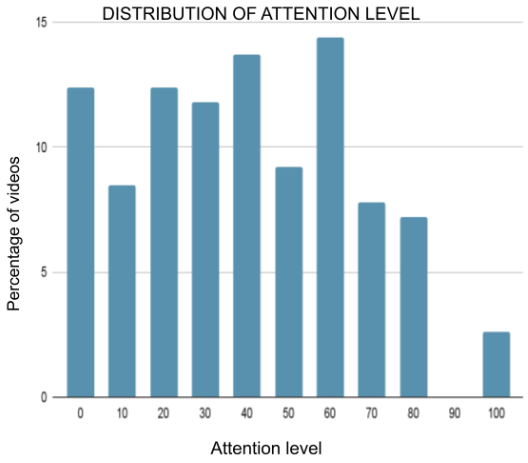

Figure 13: Attention results distribution - distribution of the attention values resulting from the test on the 158 videos passed to the final application. On the x-axis the attention values are shown (from 0% = TOTALLY INATTENTIVE to 100% = TOTALLY CAREFUL), while on the y-axis there are the respective percentages of their occurrences, so the number of videos in which we have this kind of attention

free cell phone use on driving performance and safety-critical event risk, 2013.

[2] W. Wang, X. Lu, P. Zhang, H. Xie, W. Zeng, Driver action recognition based on attention mechanism, in: 2019 6th International Conference on Systems and Informatics (ICSAI), 2019, pp. 1255–1259. doi:10.1109/ICSAI48974.2019.9010589.

[3] ISTAT, ACI, Incidenti stradali in italia. anno 2021 [comunicato stampa], 2022. URL: https://www.istat.it/it/files//2022/07/REPORT_INCIDENTI_STRADALI_2021.pdf.

[4] D. Yang, X. Li, X. Dai, R. Zhang, L. Qi, W. Zhang, Z. Jiang, All in one network for driver attention monitoring, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 2258–2262. doi:10.1109/ICASSP40776.2020.9053659.

[5] K. Guo, G. Yu, Z. Li, An new algorithm for analyzing driver's attention state, in: 2009 IEEE Intelligent Vehicles Symposium, 2009, pp. 21–23. doi:10.1109/IVS.2009.5164246.

[6] S. Vora, A. Rangesh, M. M. Trivedi, Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis, 2018. URL: https://arxiv.org/abs/1802.02690.

[7] N. Mizuno, A. Yoshizawa, A. Hayashi, T. Ishikawa, Detecting driver's visual attention area by using vehicle-mounted device, in: 2017 IEEE 16th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC), 2017, pp. 346–352. doi:10.1109/ICCI-CC.2017.8109772.

[8] E. Yüksel, T. Acarman, Experimental study on driver's authority and attention monitoring, in: Proceedings of 2011 IEEE International Conference on Vehicular Electronics and Safety, 2011, pp. 252–257. doi:10.1109/ICVES.2011.5983824.

[9] W. Nai, Z. Yang, Y. Wei, J. Sang, J. Wang, Z. Wang, P. Mo, A comprehensive review of driving style evaluation approaches and product designs applied to vehicle usage-based insurance, Sustainability 14 (2022) 7705. doi:10.3390/su14137705.

[10] A. Palazzi, D. Abati, S. Calderara, F. Solera, R. CUcchiara, Predicting the driver's focus of attention: The dr(eye)ve project abs/1807.02588 (2018). URL: https://arxiv.org/abs/1705.03854. arXiv:1705.03854.

[11] D. Melesse, M. Khalil, E. Kagabo, T. Ning, K. Huang, Appearance-based gaze tracking through supervised machine learning, in: 2020 15th IEEE International Conference on Signal Processing (ICSP), volume 1, 2020, pp. 467–471. doi:10.1109/ICSP48669.2020.9321075.

[12] X. Zhang, Y. Sugano, M. Fritz, A. Bulling, MPIIgaze: Real-world dataset and deep appearance-based gaze estimation, 2017. URL: https://arxiv.org/abs/1711.09017.

[13] A. Yoshizawa, H. Iwasaki, Analysis of driver's visual attention using near-miss incidents, in: 2017 IEEE 16th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC), 2017, pp. 353–360. doi:10.1109/ICCI-CC.2017.8109773.

[14] H. M. Peixoto, A. M. G. Guerreiro, A. D. D. Neto, Image processing for eye detection and classification of the gaze direction, in: 2009 International Joint Conference on Neural Networks, 2009, pp. 2475–2480. doi:10.1109/IJCNN.2009.5178924.

[15] H. Lee, J. Seo, H. Jo, Gaze tracking system using structure sensor zoom camera, in: 2015 International Conference on Information and Communication Technology Convergence (ICTC), 2015, pp. 830–832. doi:10.1109/ICTC.2015.7354677.

[16] A. G. Mavely, J. E. Judith, P. A. Sahal, S. A. Kuruvilla, Eye gaze tracking based driver monitoring system, in: 2017 IEEE International Conference on Circuits and Systems (ICCS), 2017, pp. 364–367. doi:10.1109/ICCS1.2017.8326022.

[17] H. Mohsin, S. H. Abdullah, Pupil detection algorithm based on feature extraction for eye gaze, in: 2017 6th International Conference on Information and Communication Technology and Accessibility (ICTA), 2017, pp. 1–4. doi:10.1109/ICTA.2017.8336048.

[18] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multi-task cascaded

convolutional networks, 2022. URL: https://arxiv.org/abs/1604.02878. doi:10.48550/ARXIV.2210.07548.

[19] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 1, 2001, pp. I–I. doi:10.1109/CVPR.2001.990517.