Keeping Eyes on the Road: Understanding Driver Attention and Its Role in Safe Driving

Francesca Fiani[1], Valerio Ponzi[1,2] and Samuele Russo[3]

[1]Department of Computer, Control and Management Engineering, Sapienza University of Rome, 00185 Roma, Italy
[2]Institute for Systems Analysis and Computer Science, Italian National Research Council, 00185 Roma, Italy
[3]Department of Psychology, Sapienza University of Rome, 00185 Roma, Italy

Abstract

Monitoring the driver's attention is an important task to maintain the driver's safety. The estimation of the driver's gaze direction can help us to evaluate if the drivers are not focusing their attention on the street. For an evaluation of this type, comparing the inside view and outside scenery of the vehicle is essential, therefore we decided to create a specific dataset for this task. In this work, we realize a machine-learning-oriented approach to driver's attention evaluation using a coupled visual perception system. By analyzing the road and the driver's gaze simultaneously it is possible to understand if the driver is looking at the traffic signs detected. We evaluate if a determined Region Of Interest (ROI) contains a road sign or not through YOLOv8.

Keywords

Visual Attention Estimation, Machine Learning, Artificial Intelligence, ADAS (Autonomous Driver Assistance Systems), YOLO

1. Introduction

Artificial Intelligence (AI) employed in assessing driver attention within assisted driving scenarios is swiftly advancing, propelled by the evolution of autonomous vehicles and the integration of hybrid systems designed to assist drivers. These systems encompass a range of functionalities, including cruise control, lane-keeping assistance, automatic parking, and various other features integrated into modern vehicles. It is well known that driver inattention is a major cause of road accidents [1, 2, 3], with violations of the expected driver behavior being a fundamental factor [4]. Due to its significant contribution to accidents, monitoring driver attention has become a critical necessity for automotive safety systems, aiming to detect potential risks and proactively prevent accidents. To achieve comprehensive attention monitoring, it is imperative to conduct precise analyses of various factors, including the driver's posture, head position, rotation angles, and gaze direction. These insights into driver behavior enable the identification of factors influencing reactions to different conditions and scenarios, thereby mitigating distractions and drowsiness-related incidents in the future [5].

Literature primarily addresses driver attention by dividing the internal and external components. Typically, the analysis of the vehicle cabin and the driver's gaze is conducted independently, without considering the evaluation of the surrounding environment, road conditions, and the driver's reaction to specific events.

Several studies focus either on observing the driver's behavior through internal vehicle cameras or analyzing external road conditions using external cameras and sensors [6, 7, 8, 9, 10]. However, a gap exists in comprehensive research that integrates both internal and external perspectives without relying on complex and inaccessible equipment. To address this gap, our research adopts a novel approach. We simultaneously analyze internal driver information, such as posture and gaze, and external data about road conditions and points of interest, like signs and pedestrians, during driving. This integrated approach allows for a more holistic understanding of driver attention and behavior.

Machine learning is playing a pivotal role in creating a safer society. In the realm of energy [11], machine learning algorithms are optimizing data systems [12, 13], improving supply-demand forecasting, and enhancing the efficiency of renewable energy sources. This not only ensures a stable energy supply but also reduces the risk of blackouts. When it comes to fostering a green environment, machine learning is at the forefront of monitoring and predicting environmental changes, enabling us to take timely action against potential threats [14, 15]. Social benefits are manifold, including improved healthcare through predictive diagnostics, personalized education, and effective public services, all contributing to an improved quality of life [16, 17, 18]. In the context of urban driving, machine learning is the driving force behind autonomous vehicles [19]. These vehicles promise to sig-

SYSYEM 2023: 9th Scholar's Yearly Symposium of Technology, Engineering and Mathematics, Rome, December 3-6, 2023

fiani@diag.uniroma1.it (F. Fiani); ponzi@diag.uniroma1.it (V. Ponzi); samuele.russo@uniroma1.it (S. Russo)
0009-0005-0396-7019 (F. Fiani); 0009-0000-2910-0273 (V. Ponzi); 0000-0002-9421-8566 (S. Russo)

nificantly reduce traffic accidents, improve traffic flow, and reduce carbon emissions, making our cities safer and more sustainable. Thus, machine learning is a key enabler in our pursuit of a safer society.

In this research, we merge various internal and external techniques for gaze recognition and correlate them with external Regions Of Interest (ROIs) to develop an easily applicable solution that comprehensively tackles the issue of driver attention. This approach holds significant practical implications for everyday scenarios, including:

- Autonomous vehicle development: Understanding the driver's focus during critical driving situations, including the duration of their attention to specific elements and their perception of irrelevant factors, plays a pivotal role in the advancement of Advanced Driver Assistance System (ADAS) solutions.
- Car crashes: Having information about driver attention during a road accident could facilitate the execution of investigations, checks, and insurance procedures. By utilizing an affordable camera system, video data on the driver involved in the accident could be collected and provided to an application.
- Emergency services: Emergency response vehicles, including ambulances and fire trucks, often need to navigate through traffic quickly and safely. Driver attention monitoring systems can help emergency service providers ensure their drivers remain vigilant while responding to emergencies, minimizing the likelihood of accidents and delays.
- Public transportation infrastructure: Driver attention monitoring systems can also be integrated into public transportation infrastructure, such as traffic lights and pedestrian crossings. By detecting instances of driver distraction or inattention, these systems can improve traffic flow and pedestrian safety, reducing the risk of accidents and congestion in urban areas.

To advance driver attention monitoring, we have directed our efforts towards computer vision-based methodologies, which are gaining traction over physiology-based approaches. Unlike physiological methods, in fact, vision-based techniques rely solely on cameras to observe and analyze driver behaviors, eliminating the need for intrusive devices such as eye-tracking glasses or brainwave recognition gadgets and consequently reducing the cost associated with experiments.

The most in-depth analysis in our work focused on finding the best method and features to extract from images to accurately determine the driver's gaze direction and point of focus. Our novel approach involves the use of a grid of nine cells to predict the Regions Of Interest (ROIs) of the driver's gaze, as illustrated in Figure 1. To achieve this, we employ a VGG16 network to extract features from facial video frames, augmenting this information with head-pose data (i.e. roll, pitch, and yaw angles) to enhance gaze-position prediction [20, 21, 22].

The difference between tracking gaze-position when a person is looking at a monitor and while they are driving is, in fact, substantial [21, 22]. 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 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 these 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, see Figure 2.

In addition to methodological research, another significant challenge we faced was sourcing an appropriate dataset for driver attention monitoring. We encountered existing datasets with comprehensive documentation of driver behavior, but they lacked corresponding real-world external observations. Furthermore, datasets focused solely on gaze analysis typically consisted of images of individuals looking at points on a computer screen, which did not align with our real-world driving scenario. To address this gap, we decided to create our own dataset, encompassing both internal and external videos captured during driving sessions. This approach enabled our final application to process and correlate

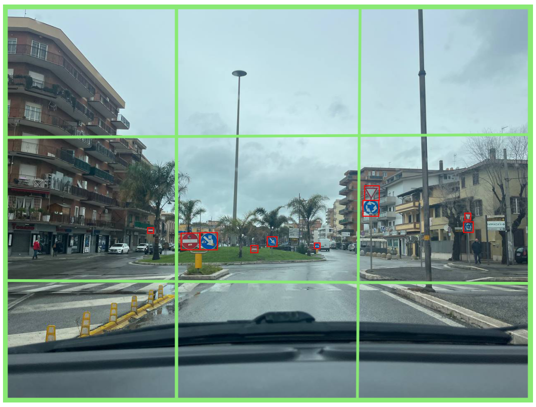

Figure 1: Example of an external image after road signs detection, with the ROI grid in green. Several regions of interest which contain one or more road signs have been identified in the image (specifically, cells 4, 5 and 6). The red rectangles represent the traffic signs bounding boxes.

information from multiple perspectives simultaneously.

To train the two components of our application, we utilized two additional datasets. For the internal component, which involves predicting the driver's gaze position, we curated the HEAD-POSE dataset, featuring data from different subjects. Unlike many existing datasets that often focus on a single subject, our dataset offers a broader and more diverse range of observations. For the external component, which entails predicting the position of road signs, we leveraged a customized dataset of road signs sourced from the internet. We carefully selected images from datasets such as MAPILLARY and GTSDB, ensuring that they adhered to European traffic regulations governed by the Vienna Convention of 1968. This meticulous curation process ensured the relevance and accuracy of the data used in our research and development efforts.

2. Related Works

As previously discussed, in recent years a growing interest in analyzing driver attention during driving has been noticed. This includes understanding whether a person is observing the road, being distracted, remaining vigilant, or experiencing drowsiness. Most state-of-the-art approaches are based on the unique observation of the driver's interior cabin to understand their behaviors [23, 24, 2, 20]. 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. Several methods can be used to determine the attention level of the driver, with the most classical metric being the gaze direction, generally assessed by analyzing facial features such as the face mesh. Other approaches are however available, for instance, the position of the hands and arms, which can be used to assess whether the driver keeps their hands on the steering wheel or in other positions, such as holding a phone [2].

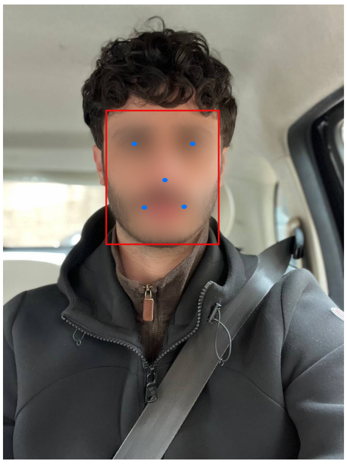

Figure 2: Example of internal image with facial features extraction. The red rectangle is the measured face bounding box and the blue dots represent the found facial landmarks. The face of the subject has been blurred according to privacy regulations.

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, various sensors such as cameras and lidar, applied to the external part of the vehicle, can allow the observation of the driver's reaction in certain situations [10]. Another classical study when analyzing the external environment surrounding the car is the analysis of road elements present in the scene via neural networks such as YOLO [25, 26].

While poor in number compared to decoupled approaches, some studies 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, which directly indicates what is being observed [27]. There are also some recent datasets created in a controlled setting that simulate common driving situations, such as the DGAZE dataset with its corresponding algorithm I-DGAZE [28].

Regarding specifically the gaze detection task, various approaches are used in simulated or real environments, both indoors and outdoors [29]. 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. 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 [30, 22, 31, 32, 33, 34]. These types of problems can, however, also be approached through classification. For example, images can be taken of a stationary subject in front of a personal computer screen, ideally divided into a 9-cell grid, and the gaze position can then be returned not as precise point coordinates, but instead as the ID number of the observed cell (classification) [21]. In such cases, pupil characteristics are generally extracted and then classified using classic machine learning methods such as Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), or Deep Neural Networks (DNNs). This last approach in particular has inspired our choice to implement a classification algorithm, given the problems and requirements already described and specific to the field of driving.

Other existing algorithms for studying gaze position start, as previously mentioned, from datasets of tens of thousands of photos collected using a personal computer's webcam, and then extract facial information from the given images to crop eye images and pass them to networks such as VGG-16 [22]. In addition, information related to head position (rotation angles - roll, pitch, yaw) can also be considered [20]. It is particularly of note that to improve regression on the viewpoint position it is fundamental to collect images from multiple subjects, in multiple vehicles, and under different weather conditions. Finally, additional approaches make use of recordings in simulated environments using various technologies, from simulators to simple computer-played videos. For example, the user's gaze position can be 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 [29].

3. Methods

This research explores an innovative method for recognizing gaze patterns while driving to evaluate driver attention. Subsequently, we focused on the internal aspect of the vehicle, where we trained and tested neural networks for gaze classification. Our experimentation involved various models, including SVM, ClassNET, VGG16-based Net, and HEGClass Net. Additionally, we conducted a training phase for the external aspect using a custom dataset comprising traffic sign objects. Once we obtained results for both components, we merged the two modules to conduct a comprehensive analysis of video recordings obtained during real-world driving scenarios. This integrated approach facilitated a more holistic comprehension of gaze behavior and its correlation with driver attention in typical driving situations.

3.1. Dataset

A variety of images and videos were gathered and utilized at different stages of development to construct custom datasets tailored to our research objectives. These datasets can be categorized into four distinct collections:

- Gaze Directions and Head Posture Dataset (GDHPD): This dataset comprises images captured by us, featuring individuals in a driving environment. The images are utilized to categorize the gaze position of individuals within a grid consisting of nine cells, including the exterior of the grid.
- People Driving Dataset (PDD): This dataset consists of both external and internal videos, recorded by our team, showcasing the driving activities.
- Traffic Objects Dataset: This dataset is a modified version of the Mapillary Dataset, containing images depicting various traffic signs.
- Traffic Signs Dataset in YOLO format (TSDY): This dataset comprises images sourced from the German Traffic Sign Detection Benchmark (GTSDB), available for download from Kaggle.

To create our dataset, we maintained a consistent equipment setup as depicted in Figure 3. For both the Gaze Directions and Head Posture Dataset and People Driving Dataset we utilized a city car, while an iPhone 15 camera was employed to capture internal images and record internal videos. The iPhone was strategically positioned behind the steering wheel to ensure clear visibility of the driver while minimizing extraneous details. Furthermore, we positioned a GoPro Hero10 camera at the center of the car's dashboard to capture external footage throughout the drive.

For the GDHPD, we compiled images from distinct subjects, consisting of two males and two females. In some cases, subjects wore glasses, while in the other they did not. The image collection process encompassed various times of the day and diverse lighting conditions, resulting in a total of 1012 images. Table 1 provides a breakdown of the distribution of these images.

Subjects were positioned inside a car, with their seating adjusted to achieve a standard driving posture. Subsequently, images were captured while subjects varied their gaze and head positions. To facilitate classification, we devised a virtual grid dividing the external view and the driver's gaze into a 9-cell configuration. This

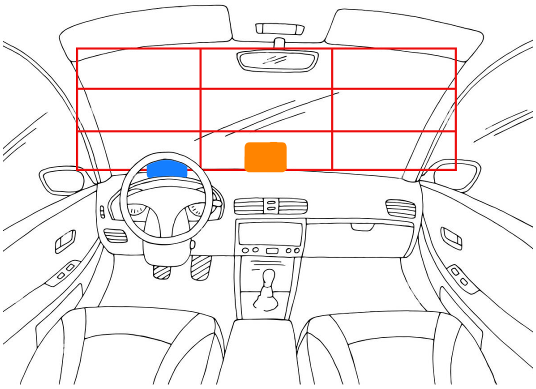

Figure 3: The setup of the city-car environment used during the collection of images and videos. In red, the virtual grid represents how the gaze position area is divided into ROIs, each associated with a region number from 1 to 9 (the area outside the virtual grid is labeled 0). In orange, the GoPro Hero10 used to record the external street. In blue, the iPhone 15 used to record the driver.

GDHPD dataset specifications

Classes	Male	Female	Total
0	50	64	114
1	50	64	114
2	51	71	122
3	51	71	122
4	48	60	108
5	48	60	108
6	50	50	100
7	51	50	101
8	51	61	112
9	50	61	111
Total	500	612	1012

Table 1
The Gaze Directions and Head Posture Dataset (GDHPD) dataset collects the gaze positions of different drivers. The collection contains a total of 1012 images, 500 for males and 612 images for females.

grid facilitated the association of head and eye positions with specific regions of the external images, enabling the identification of Regions Of Interest (ROIs) during experimentation.

The dataset comprises ten classes: the nine cells of the grid, alongside an additional class representing situations where the subject's attention is not directed towards the road (e.g., face turned sideways, gaze directed upwards or downwards, etc.).

The PDD dataset comprises videos captured during driving sessions, utilizing the same recording setup as the previous dataset. Subjects were filmed while driving under various conditions, capturing both the driver and the road view simultaneously. To ensure synchronization, the videos underwent pre-processing using a third-party software, DaVinci Resolve. Synchronization was achieved through voice cues, guaranteeing precise alignment between internal and external footage.

Subsequently, the videos were segmented into subclips of 30 seconds each to streamline subsequent processing steps. Each of the extracted sub-clips was annotated with labels indicating whether the driver exhibited a "CAREFUL" or "NOT CAREFUL" driving style, along with information regarding the driver's use of glasses. After pre-processing, the external images displayed a resolution of 1440x1080, while the internal images were resized to 1080x1920.

For traffic sign detection and recognition, the primary dataset utilized was the Traffic Object dataset from the Mapillary Traffic Sign Dataset, encompassing tens of thousands of images sourced from roads worldwide. Focusing solely on Italian/European traffic signs, around 3,000 images were selected from the dataset after filtering out images with significantly different sign shapes or contents. The chosen images offer a varied range of brightness, positioning within frames, and contextual variations.

The corresponding JSON files were then converted into TXT files and formatted to suit YOLO's training model requirements. This conversion process involved extracting relevant information such as bounding box coordinates and class labels of traffic signs, facilitating model training for sign recognition and localization. To simplify the classification task, the dataset's labels were modified to include three super categories: 'PROHIBITORY', 'DANGER', and 'MANDATORY'.

These categories encapsulate the majority of relevant traffic signs essential for driving safety, thus streamlining the training process. Similarly, the Traffic Signs Dataset in YOLO format (TSDY) was used as a refined version of the larger GTSDB dataset. Comprising 750 images with labels already expressed in YOLO format, each image had a resolution of 1360x800. The number of classes was reduced to four: 'PROHIBITORY', 'DANGER', 'MANDATORY', and 'OTHER', simplifying the classification task and enhancing the focus on critical sign types relevant to driver safety.

3.2. Gaze Classification

Several algorithms, were analyzed to identify the approach with the best trade-off between accuracy in gaze direction prediction and generalization capabilities, allowing to efficiently recognize images with varied contrast and/or brightness or different drivers. The structure takes as input an image (either singular or an extracted frame from the recording) and generates a label prediction through two different but subsequent sub-models. The head-pose estimation algorithm is common for all tested approaches, while the classification algorithm has been severely varied in model type, structure, and input during the search for the optimal solution.

3.2.1. Head-Pose Estimation Part

The face detection is performed through a pre-trained Multi-task Cascaded Convolutional Network (MTCNN) model, used for both face detection and alignment in literature [35]. MTCNN consists of a cascade of convolutional networks (P-Net, R-Net, and O-Net), for face landmarks identification. The model first identifies the bounding box of the face region through candidate generation with P-Net and refinement with R-Net and then extracts the main 5 landmarks of the face (left and right eye, nose tip, and mouth corners) with O-Net. Among similar methods, such as Haar Cascade Classifiers [36], MTCNN has shown the best results even in the presence of glasses, partially occluded eyes and beards, and has therefore been selected as our chosen method.

The identified landmarks are then used to analytically calculate the roll, pitch and yaw angles of the driver's head, while the extracted pupil positions will be used as input features for the final classifier to determine the observed ROI. This feature is fundamental for our classification task compared to other facial features and is therefore particularly important to determine accurately.

3.2.2. Classification Part

Our approach involves classifying observed ROI and road signs through a classification method. The field of view is divided into nine sections, with an additional label for identifying any gaze position outside these sections (e.g., distracted driving or maneuvering). We've explored various methods for analysis, ranging from traditional SVM to CNNs, all accurately adapted for our application. We will first introduce our novel method, followed by an overview of other models considered in the analysis.

The standout classification model is HEGClass (Head-Eyes-Gaze Classifier), a hybrid approach outlined in this paper. It takes cropped face images from head-pose estimation, along with head rotation angles and pupil center coordinates, as inputs. This combined approach has yielded high precision in classifying the Region of Interest toward which the gaze is directed. In the HEGClass network, as depicted in Figure 4, initial features are extracted from cropped face RGB images using a pre-trained VGG-16 network. The features are then flattened and concatenated with a normalized array containing head roll-pitch-yaw and pupil center coordinates. This combined feature vector of dimension 4096+7 passes through two fully connected linear layers, followed by ReLU activation functions, and finally through a last fully connected linear layer with Softmax activation to determine class membership among the 10 possibilities (9 ROIs for frontal regions and 1 for others). Model training utilized our GDHPD dataset, with 10 epochs to mitigate overfitting, 32 samples per batch, Cross-Entropy loss function, and Adam optimizer.

The first classical model used for comparison is Support Vector Machine (SVM). In our scenario, where we're classifying 10 distinct classes, we employed an SVM with a polynomial kernel of degree 4, regularization parameter set at 100, and coefficient set to 10. Training exclusively used images from the GDHPD dataset. We extracted roll, pitch, yaw, and pupil centers from the images with MTCNN. Then, using Haar Cascade Classifier [36], we isolated eye patches from each grayscale image and passed them through a pre-trained ResNet to obtain 2048 features for each eye. These two sets of features were then averaged to create a unified array of 2048 elements containing information from both eyes. The resulting samples underwent L2 normalization before being fed into the SVM for both training and testing phases.

Using the same 2048 features extracted ResNet and roll, pitch, yaw and pupil centers, we also trained the ClassNet network. This Classifier Network architecture consists of two convolutional layers with ReLU activation functions, followed by a max pooling layer, and culminates in two fully connected layers with ReLU and Softmax activation functions. Training of ClassNet spanned 300 epochs, employing the MSE loss function (Mean Squared Error) and Adam optimizer.

Finally, the last experimental iteration involved employing the VGG-16 architecture. Training data consisted of 1012 samples from the GDHPD dataset, each composed of the face image tensor, alongside head rotation angles (roll, pitch, yaw), and pupil centers. In this setup, features from each image were directly extracted within the classification network from the RGB image of the entire face, rather than solely from the eyes. Additionally, other features (roll, pitch, yaw, and pupil centers) obtained previously through MTCNN were incorporated alongside the 512 features of the image in the first fully connected layer. Training was executed over 100 epochs, using similarly to the HEGClass model the Cross-Entropy Loss function and Adam optimizer.

3.3. YOLO Training for Traffic Signs

For the detection and recognition of traffic signs, we start from the pre-trained YOLOv8 model, with experimentation also conducted using the YOLOv5 model prior to transitioning to the v8 version. Fine-tuning of YOLOv8 was carried out using two distinct datasets: Traffic Objects and TSDY. The final set of weights chosen for the application was derived from the dual fine-tuning of YOLOv8 with both datasets.

The initial fine-tuning with the Traffic Objects dataset involved 1802 images for training and 919 for validation. Despite starting with 3000 images, adjustments were made to the training and validation sets due to imbalance issues within the original dataset, which persisted even after categorizing labels based on sign categories as described in the Dataset section. Subsequently, proceeding from the fine-tuned weights, the model underwent retraining with images from the TSDY dataset, utilizing 600 images for training and 141 for validation.

The dual fine-tuning approach resulted in enhanced performance, as evidenced by improved final accuracy and heightened generalization capabilities in detecting road signs, even in images sourced from the PDD dataset and exhibiting varied lighting conditions. To further enrich the diversity and generalization capabilities of the fine-tuned YOLO model, diverse image augmentation techniques from the Albumentations library were employed during training to simulate real-world conditions, including Blur, MedianBlur, and CLAHE (Contrast Limited Adaptive Histogram Equalization). To streamline the training process, the Stochastic Gradient Descent (SGD) optimizer was utilized, with an initial learning rate of

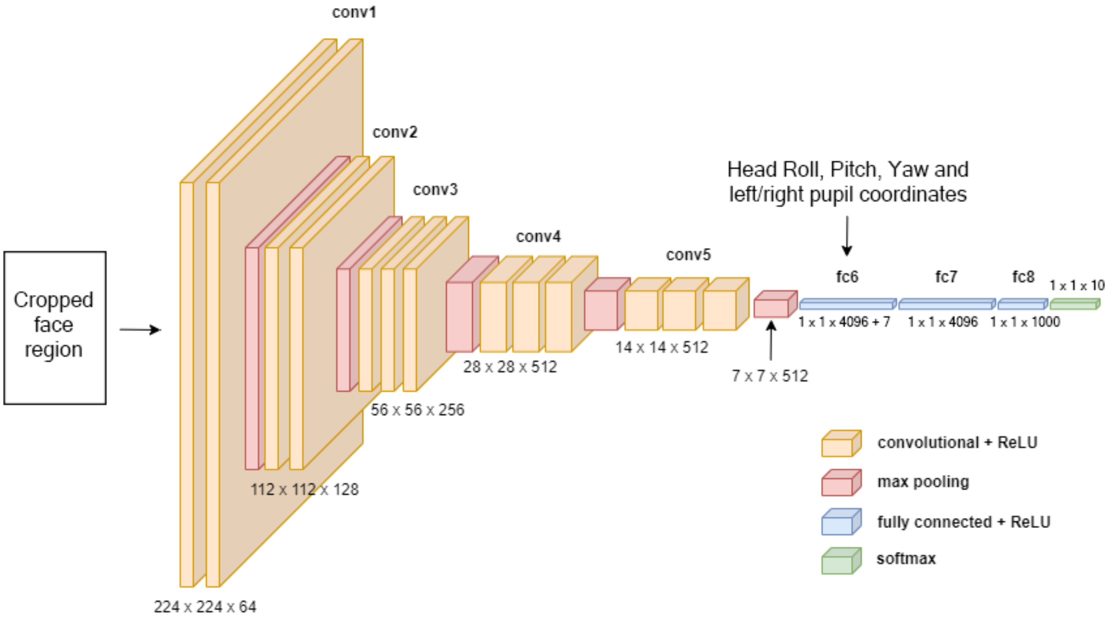

Figure 4: Model of our novel approach HEGClass. The base of the model is a standard pre-trained VGG-16 network, which receives as input the cropped image of the subject's face. In the first fully connected layers 7 additional features [Roll, Pitch, Yaw, rx, ry, lx, ly] are added to arrive at the final ROI prediction (with 10 classes with values [0,9]).

0.01.

Post-training, the model returns a text file for each image processed, containing one line per detected sign alongside its position. By processing the pixel coordinates from this file, sign position information was extracted to reconstruct the sign's center and edges within grid cells. In instances where signs spanned multiple cells, multiple coordinates were necessary for accurate identification.

3.4. Application: Merging the Methods

The final application, depicted in Figure 5, comprises two primary components and generates a CSV report detailing the overall behavior of a driver. It requires as input an internal video capturing the driver and an external video recording the street view. For ease of analysis, synchronization of a 30-second video between the two components is necessary.

After inizialization, external frames undergo analysis using the YOLOv8 model trained on road signs, producing a text file detailing the detected signals along with their specifications, including the Regions of Interest (ROIs) where these signals are located. For frames with at least one detected sign, the corresponding internal image is used to extract information via the GDHPD module. The image and information are then passed through the network to classify the driver's gaze position. Upon obtaining the prediction of the observed cell in a given frame, it is compared with the position of the corresponding sign. In frames with multiple detected road signs, each corresponding ROI is considered active, thus rendering the driver alert when focusing on any of them without favoring any type of signal.

Given that a single road sign can span multiple cells, 5 points characterize the object's position: the four corners and the center. If the driver looks at a cell containing a partial view of the sign, accounting for the peripheral vision of human eyes we consider them attentive. Moreover, in the absence of signals or when the driver's gaze is directed to cells 4/5/6 (representing the entire road surface), they are still deemed attentive to the street. If gaze is directed to cells 7 or 8, indicating focus on the car's dashboard or infotainment system, the driver's engagement is noted accordingly.

Finally, a CSV file is generated to store the analysis results from the 30-second videos. Each row contains data including the frame count (consistent across internal and external videos), the number of road signs detected in that frame, the cell number(s) housing the detected signals, the predicted cell value from the driver's gaze network, the number of observed signals following ROI

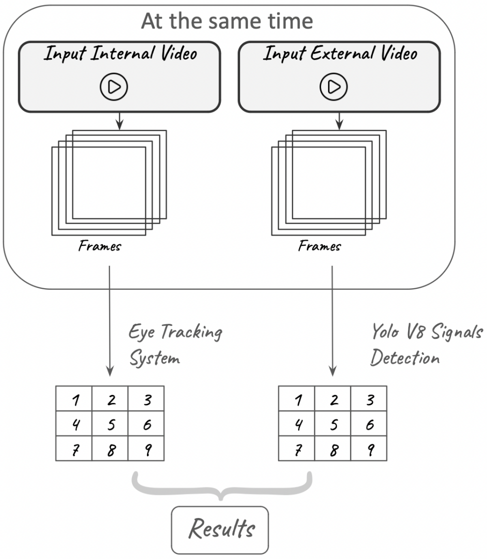

Figure 5: Pipeline of the application presented in the paper. The internal and external videos are simultaneously processed to extract relevant features (facial landmarks and road signs bounding boxes respectively). The features are then used to determine the active ROIs, which are then compared to generate the attention level of the driver.

matching, and an indication of the driver's attentiveness.

4. Results

The objective of this research is to develop a comprehensive system capable of analyzing an individual's attention while driving using only two synchronized videos as input. Given the scarcity of references on the simultaneous analysis of internal and external perspectives, all subsequent evaluations and comparisons will focus on the individual components constituting the final system. Nonetheless, through extensive testing conducted with the PDD dataset, comprising approximately 194 videos each lasting 30 seconds, the final application demonstrates commendable performance.

4.1. Gaze Classification

For what concerns face detection and landmark extraction for facial rotation angle calculation, the MTCNN model outperformed the Haar Cascade Classifier. This superiority stems from MTCNN's ability to handle various facial orientations, which is crucial for our GDHPD dataset as it contains images with rotated or profiled faces. Additionally, MTCNN's prediction of landmarks, including the center of the pupil, proved vital for training the final classification network. However, occasional failures in face detection or slight misplacements of landmarks introduce minor errors in this initial phase.

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. Our novel approach shows the overall best results between all the analyzed methods.

For predicting gaze direction, a zone-based classification approach was chosen over regression due to the difficulty in precisely determining the exact point on the road the individual is looking at, coupled with the human eye's ability to perceive a broad area. Despite testing various methods, the SVM-based approach struggled to exceed a 70% accuracy level, likely due to the similarity in feature values across the 1012 samples, particularly those derived from ResNet for eye images.

Transitioning to neural network-based methods, the ClassNet network yielded lower accuracy than SVM, even after experimenting with different feature combinations. Training a network based on VGG-16 architecture from scratch yielded better results with an accuracy level of 81%. However, the limited size of our dataset and computational constraints hindered achieving satisfactory performance through this approach. Hence, we adopted the hybrid HEGClass approach, achieving an impressive 96% accuracy and 94.3% F1-score without additional data. Comprehensive accuracy and F1-score results are shown in Table 2.

4.2. YOLOv8 Classification and Detection

Through dual fine-tuning of the YOLOv8 network by using the Traffic Object Dataset and the TSDY dataset, an impressive final F1-score of around 95% was achieved, with an example of prediction on the PDD dataset shown in Figure 6. We have observed an interesting phenomenon when training the network solely with the Traffic Objects dataset, where the F1-score is significantly lower. Specifically, the training of YOLOv8 with the Traffic Objects dataset yielded an overall accuracy of approximately 65%-70%. Performing the same process with YOLOv5, instead, showed unexpectedly a higher accuracy (around 80%), albeit with occasional misclassifications of elements such as empty spaces between tree branches. In any case, for the scope of this project this comparison is not particularly relevant, given the much higher accuracy with dual fine-tuning.

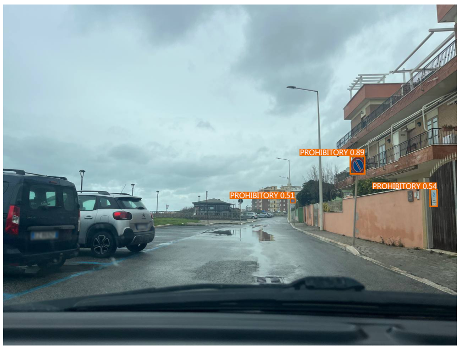

Figure 6: Frame extracted from the PDD dataset predicted by the fine-tuned model YOLOv8. The predicted traffic signs and the correlated label and accuracy are highlighted in orange.

Despite the substantial improvement in generalization capabilities achieved through dual training, errors in sign recognition persist. Certain objects along the road may be mistaken for road signs, such as advertisements containing elements that, with low resolution, could be confused. While this issue is present, its impact on the overall results remains manageable and could potentially be mitigated with a wider variety of images. Another challenge arises from grouping signs of different shapes and colors into the same class, creating a bias in their classification. Additionally, signs containing other signs within them may only become relevant in specific situations, such as parking signs reserved for disabled individuals. For this reason, some signs were excluded from the training phase.

Given the high accuracy values in detection, adjusting the confidence threshold can help alleviate misclassification issues. Signs may be recognized even when rotated, facing the opposite direction of the lane, or located in irrelevant areas. In such cases, they are counted as points of inattention. Due to class imbalance, accurately classifying the type of road sign remains a challenge. Consequently, for our purposes, only information related to the bounding box defining the sign's position is extracted, without specifying the type of sign. Despite attempts to simplify the dataset to recognize only one class, "TRAF_SIGN", challenges persisted between detecting signs and identifying unrelated environmental areas. Therefore, the decision was made to revert to using the original labels.

4.3. Overall Analysis

The final results of the application, pertaining to the prediction of the driver's average attention while viewing a video, exhibit high performance across most cases, with a few notable exceptions. Tests were conducted on 191 videos, each lasting 30 seconds, sourced from the PDD dataset. Following initial evaluations, videos compromised by low light conditions (such as those filmed in almost night-time environments) or excessive blurring of frames, rendering accurate prediction unfeasible, were eliminated.

In instances where images lack clarity, are blurry, or exhibit excessive shaking, the predominant predicted class is 0, indicating the model's failure to accurately identify the correct gaze position. Moreover, during nighttime or low-light scenarios, accurate gaze evaluation is significantly impeded by diminished brightness. Additionally, YOLO struggles with precise detection of relevant signs, often leading to confusion. Classes 5 and 6 are frequently identified as the gaze position during driving, aligning with the fact that these areas correspond to central regions of the windscreen. In some situations, such as when the vehicle is stationary at a traffic light or in traffic congestion, the system may recognize the same signs across multiple frames. However, drivers may not consistently attend to them throughout, as they may have already observed them and they may not be of immediate significance at that moment.

5. Conclusions

This work aimed to develop and assess a comprehensive system for evaluating attention to traffic signs in driving environments. We accomplished this by creating two new datasets (GDHPD, PDD) and modifying two existing ones (Traffic Objects, TSDY) to better suit our task requirements. The final application was divided into two parts, utilizing YOLOv8 for sign prediction and MTCNN + HEGClass for gaze position classification.

Despite encountering challenges during various training and testing phases, as described in the Results section, the overall accuracy of the final system remains very high, notwithstanding the partial errors accumulated by its constituent parts.

These challenges serve as valuable insights for future research endeavors. Opportunities for improvement include implementing mechanisms to track seen and unseen signals, enhancing prediction accuracy in diverse lighting and atmospheric conditions through dataset augmentation or pre-processing techniques, and expanding datasets to ensure greater completeness.

Overall, this work presents significant potential for further refinement and advancement, promising avenues for enhancing the performance and robustness of attention evaluation systems in driving contexts.

References

[1] G. Fitch, S. Soccolich, F. Guo, J. McClafferty, Y. Fang, R. Olson, M. Pérez-Toledano, R. Hanowski, J. Han-

key, T. Dingus, The impact of hand-held and hands-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] A. J. McKnight, A. S. McKnight, The effect of cellular phone use upon driver attention, Accident Analysis & Prevention 25 (1993) 259–265.

[4] J. C. de Winter, D. Dodou, The driver behaviour questionnaire as a predictor of accidents: A meta-analysis, Journal of safety research 41 (2010) 463–470.

[5] K. J. Anstey, J. Wood, S. Lord, J. G. Walker, Cognitive, sensory and physical factors enabling driving safety in older adults, Clinical psychology review 25 (2005) 45–65.

[6] E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies, IEEE access 8 (2020) 58443–58469.

[7] Y. Hou, C. Wang, J. Wang, X. Xue, X. L. Zhang, J. Zhu, D. Wang, S. Chen, Visual evaluation for autonomous driving, IEEE Transactions on Visualization and Computer Graphics 28 (2021) 1030–1039.

[8] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, D. Whitney, Predicting driver attention in critical situations, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14, Springer, 2019, pp. 658–674.

[9] D. Yang, Y. Wang, R. Wei, J. Guan, X. Huang, W. Cai, Z. Jiang, An efficient multi-task learning cnn for driver attention monitoring, Journal of Systems Architecture (2024) 103085.

[10] 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.

[11] G. Capizzi, G. L. Sciuto, C. Napoli, M. Woźniak, G. Susi, A spiking neural network-based long-term prediction system for biogas production, Neural Networks 129 (2020) 271 – 279. doi:10.1016/j.neunet.2020.06.001.

[12] B. A. Nowak, R. K. Nowicki, M. Woźniak, C. Napoli, Multi-class nearest neighbour classifier for incomplete data handling, volume 9119, 2015, pp. 469 – 480. doi:10.1007/978-3-319-19324-3_42.

[13] C. Ciancarelli, G. De Magistris, S. Cognetta, D. Appetito, C. Napoli, D. Nardi, A gan approach for anomaly detection in spacecraft telemetries 531 LNNS (2023) 393 – 402. doi:10.1007/978-3-031-18050-7_38.

[14] A. Alfarano, G. De Magistris, L. Mongelli, S. Russo, J. Starczewski, C. Napoli, A novel convmixer transformer based architecture for violent behavior detection 14126 LNAI (2023) 3 – 16. doi:10.1007/978-3-031-42508-0_1.

[15] V. Ponzi, S. Russo, A. Wajda, R. Brociek, C. Napoli, Analysis pre and post covid-19 pandemic rorschach test data of using em algorithms and gmm models, volume 3360, 2022, pp. 55 – 63.

[16] S. Russo, S. I. Illari, R. Avanzato, C. Napoli, Reducing the psychological burden of isolated oncological patients by means of decision trees, volume 2768, 2020, pp. 46 – 53.

[17] E. Iacobelli, V. Ponzi, S. Russo, C. Napoli, Eye-tracking system with low-end hardware: Development and evaluation, Information (Switzerland) 14 (2023). doi:10.3390/info14120644.

[18] F. Fiani, S. Russo, C. Napoli, An advanced solution based on machine learning for remote emdr therapy, Technologies 11 (2023). doi:10.3390/technologies11060172.

[19] N. Brandizzi, S. Russo, G. Galati, C. Napoli, Addressing vehicle sharing through behavioral analysis: A solution to user clustering using recency-frequency-monetary and vehicle relocation based on neighborhood splits, Information (Switzerland) 13 (2022). doi:10.3390/info13110511.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] J. Terven, D.-M. Córdova-Esparza, J.-A. Romero-

González, A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas, Machine Learning and Knowledge Extraction 5 (2023) 1680–1716.

[26] A. A. Lima, M. M. Kabir, S. C. Das, M. N. Hasan, M. Mridha, Road sign detection using variants of yolo and r-cnn: An analysis from the perspective of bangladesh, in: Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021, Springer, 2022, pp. 555–565.

[27] 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.

[28] I. Dua, T. A. John, R. Gupta, C. Jawahar, Dgaze: Driver gaze mapping on road, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 5946–5953.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.