

Occlusion-Aware Driver Monitoring using VLM-Enhanced Situational Understanding

Paola Natalia Cañas^{1,2}, Alexander Diez², Marcos Nieto¹, Igor Rodríguez², Oumayma Sghairi¹, Martí Sánchez^{1,2}
Connected & Cooperative Automated Systems of organization

¹ *Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastian, Spain*

² *University of the Basque Country (UPV/EHU), Donostia-San Sebastian, Spain*

{pncanas, [mnieto](mailto:mnieto@vicomtech.org), [osghairi](mailto:osghairi@vicomtech.org), [msanchez](mailto:msanchez@vicomtech.org)}@vicomtech.org, adiez134@ikasle.ehu.eus, igor.rodriguez@ehu.eus

Abstract—This paper presents a robust, occlusion-aware driver monitoring system (DMS). The system performs driver identification, gaze estimation by regions, distraction detection and face occlusion detection and understanding using Vision-Language Models (VLMs) to categorize the cause of obstruction (e.g., hand, sunglasses, looking away) under varying lighting conditions. Aligned with EuroNCAP recommendations, the inclusion of occlusion detection enhances situational awareness and system trustworthiness by indicating when the system's performance may be degraded. The system employs separate algorithms trained on RGB and infrared (IR) images to ensure reliable functioning. These algorithms used the multimodal Driver Monitoring Dataset (DMD). We detail the development and integration of these algorithms into a cohesive pipeline, addressing the challenges of working with different sensors and real-car implementation. Evaluation on the DMD and in real-world scenarios demonstrates the effectiveness of the proposed system, highlighting the superior performance of RGB-based models and the pioneering contribution of robust occlusion detection in DMS.

Keywords—*Driver Monitoring Systems, Occlusion Detection, Visual Language Models, Driver Monitoring*

I. INTRODUCTION

Significant advancements in autonomous driving, classified as SAE [1] Level 3 and above, allow drivers to delegate responsibilities and temporarily become passengers. However, the system may require a safe and timely intervention from the driver during uncertainties or exceptional circumstances, necessitating careful management of the transition of control. Additionally, in lower levels of autonomy (e.g. < SAE L3), monitoring the driver remains essential to prevent human error during manual operation. Driver Monitoring Systems (DMS) are therefore vital for assessing the driver's attention, alertness, and readiness to drive, thereby enhancing overall road safety.

Modern DMS perform several critical functions, including assessing driver's state, like fatigue, distraction, driver's comprehension of vehicle alerts, among others. EuroNCAP establishes core guidelines for DMS, primarily focusing on driver distraction detection based on the direction of their gaze by regions [2]. They consider different zones of the vehicle's interior, with distraction being defined as the driver not looking at the road or using the phone.

Beyond safety, modern DMS also integrate features like driver identification. This functionality enhances vehicle security by preventing unauthorized access and significantly improves comfort by automatically adjusting personalized settings, such as seat positions, mirror angles, and climate

control, based on the identified user. These personalized settings improve the driving experience by tailoring the vehicle environment to individual preferences. As a result, such applications are gaining relevance in the automotive industry, with additional innovative uses continually being developed.

A major challenge for visual-sensor-based DMS is occlusion, where visual obstructions of the driver's face or eyes—caused by accessories (sunglasses), environmental factors (glare), or hands—can lead to performance degradation or prediction failures. EuroNCAP explicitly recommends that DMS must either be robust against occlusions or alert the user when not functioning correctly.

To address these challenges, we present a novel, robust, and occlusion-aware DMS that functions effectively under varying lighting conditions using both RGB and Infrared (IR) images. Furthermore, this work introduces a pioneering occlusion analysis. Our system moves beyond simple binary detection to implement occlusion cause classification using Vision-Language Models (VLMs), diagnosing the specific cause of obstruction (e.g., hand, accessory). By identifying the cause, the system enhances trustworthiness and situational awareness, enabling a more appropriate alert and ensuring continuous monitoring. Our system aligns with EuroNCAP by providing gaze estimation by regions and by identifying when the system's performance may be degraded by occlusions.

This document details the development and integration of: a **driver gaze estimation by regions** and a **driver identification system**, prepared for both RGB and IR images. A novel, **VLM-based system to categorize the cause of occlusion**. This is all integrated in a cohesive system logic that switches between RGB and IR modalities and leverages occlusion type classification to ensure continuous and reliable functioning in real-world driving conditions. The system also integrates **distraction detection** from previous research [3].

A. Gaze Estimation Based on Regions

Gaze estimation systems operate by identifying the driver's gaze direction to estimate where they are looking at any given moment. This information provides valuable insights into the driver's cognitive state and their engagement with the driving task. Knowing this can provide information about the driver's distraction level [4], awareness of other road actors [5], and predicting driver's maneuvers [6].

Various methods have been employed for gaze estimation. Highly accurate techniques involve estimating eye vectors to pinpoint the driver's focus as a precise 3D point (x,y,z) [7], often

requiring rigorous calibration and eventually translating the vector into a zone (e.g., front, rear-view mirrors). Alternative methods focus on directly classifying the region of interest where the driver's attention is directed [8] without an intermediate gaze vector estimation. These systems typically infer gaze direction by analyzing the driver's head pose and eye appearance using image processing and computer vision algorithms to extract relevant features [9].

In this research, we developed a driver's gaze estimation algorithm based on gaze regions. This approach was selected because it does not require calibration and directly outputs a zone, which is what the EuroNCAP considers when assessing gaze direction.

Similar to our work, Lollett et al. [10] also classified gaze based on regions by defining nine combined head and eye direction states to enhance robustness against non-aligned situations. However, their approach relies on complex image pre-processing involving 3D facial reconstruction from a single 2D image and is limited to the RGB modality.

B. Driver Identification

Driver identification is implemented to personalize car features [11] such as setting preferred seat positions, adjusting rear-view mirrors, climate settings, and even suggesting routes based on past destinations [12] once the driver is re-identified. Furthermore, the driver ID can unlock new third-party applications, including payments [13] or assurance services [14]. In the context of fleet management, driver identification proves useful for tracking work schedules, among other applications [15].

The identification of drivers has been achieved through the analysis of various signals, including image [16], voice [17], or driving behaviours [18]. However, our approach focuses on analyzing the driver's face for identification. The system developed in this research serves as a foundation for other developments, primarily related to comfort. While it is not crucial to implement additional measures to prevent spoofing [19] or other identity theft violations in this context, we prioritize the precision of re-identifying registered drivers. Additionally, we ensure the image of the driver is not degraded by occlusion or low light conditions to be correctly compared against registered IDs, avoiding false rejections.

C. Occlusion Detection and Understanding

Occlusions, whether caused by environmental factors like glare, personal accessories (e.g., sunglasses, hats), or the driver's hand, impair the algorithms' ability to accurately detect unsafe situations. As a result, EuroNCAP recommends that DMS should: a) be robust against occlusions or b) if the system is degraded, alert the user with visual or audible signals in less than 10 seconds after the occlusion. They list some categories of possible occlusion in drivers: Daytime (100,000 lux), nighttime (1 lux), one hand on wheel at 12 o'clock position, facial occlusion (Face mask, hats, long head hair fringe obscuring eyes), sunglasses with a <15% transmittance, thick eyelash makeup and long facial hair.

Driver monitoring systems traditionally employ infrared cameras to manage the challenges of difficult and varying lighting, especially in dark conditions where visibility and image

quality issues affect conventional methods. While IR-only systems provide excellent visibility in darkness where RGB fails, they lose critical color data. Given that most available datasets are RGB-centric, we have designated RGB as the primary operational mode, with IR serving as a crucial fallback for low-light situations. Our research presents a resilient system specifically designed to handle difficult lighting conditions. To ensure resilience across different operational environments, this system combines both RGB and infrared IR cameras.

Prior work, such as that by Lollett et al. [10], addressed self-occlusions by using 3D face reconstruction to detect and standardize the pixel data of an occluded eye. This technique successfully aids the downstream gaze classifier but only performs a binary detection of occluded vs. non-occluded features. Other systems lack this feature entirely. To the best of our knowledge, this research represents the first attempt to introduce robust occlusion analysis into a DMS.

We implement occlusion cause understanding using vision-language models to provide superior system trustworthiness and analysis of the in-cabin scene. Instead of a simple alert, our system diagnoses the specific cause of the obstruction (e.g., hand, sunglasses, looking away). This VLM-based approach allows aligning with EuroNCAP recommendations while enhancing the overall system's diagnostic capability.

II. DATA: DRIVER MONITORING DATASET (DMD)

The data used for this research is entirely from the Driver Monitoring Dataset (DMD) [20]. With previous explorations of the DMD, using data related to distraction [3], this document presents the first analysis of gaze-related data of this dataset. The DMD is a public video dataset containing recordings of 37 drivers engaged in various activities while driving. Annotations cover several aspects, including distraction, drowsiness, hand and wheel interaction, and gaze estimation. The recordings were captured by three cameras focusing on the driver's face, body, and hands, with each camera recording in RGB, IR, and depth data. This dataset is one of the largest datasets focused on driver monitoring systems, with more than 1,000 requests.



Fig. 1. Gaze regions of the DMD

Focusing solely on the nine gaze region annotations defined by the DMD (Fig. 1), we first extracted every frame from the video dataset, resulting in 221,172 RGB images and their corresponding IR counterparts. To mitigate redundancy from contiguous, identical frames and to remove instances with annotation discrepancies (often during gaze zone transitions), we applied deduplication and manual cleaning. This resulted in a final set of 6,424 images per channel. The final dataset was strictly partitioned to ensure images of the same person did not

appear in more than one group, with splits for training (63.39%), testing (20.73%), and validation (15.88%).

III. GAZE REGION CLASSIFIER

We employed the Efficientnet_b0 [21] architecture (74.31%), chosen for its strong performance and computational efficiency, modifying the classification layers to output nine neurons. Initial experiments with mobilenet_v3_small [22] underfitted the data (59.12% accuracy).

To maximize feature extraction and significantly improve the signal-to-noise ratio, the optimal method involved using the MTCNN [23] face detector to crop the input images and pass only the detected face to the classifier, which resulted in a notable accuracy increase (84.31%). Final optimization, including tuning the learning rate and weight decay, along with the application of AugMix [24] and TrivialAugment [25] for data augmentation, yielded a final test accuracy of 86.29%. From the confusion matrix (Fig. 2), the primary challenge observed is the model's difficulty in differentiating between the side mirrors, windows, and the 'front right' region, an expected result given the minimal head movement required to switch between these visually similar zones.

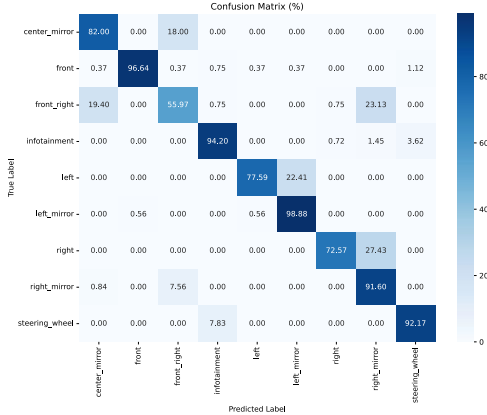


Fig. 2. Confusion matrix of RGB region gaze estimation algorithm

The IR model was trained using the same face-cropped methodology as the optimized RGB model, initially achieving an accuracy of 84.70%. Real-world testing revealed poor performance, primarily due to extreme lighting conditions in the IR dataset, ranging from faces too dark to distinguish to images too bright to discern any features.

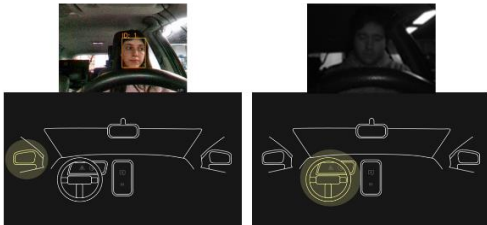


Fig. 3. Gaze classifier predictions of a subject looking at the left mirror (on the left) and a subject looking at the steering wheel (on the right).

To create a more representative training set, we implemented a data filtering step, removing images with a mean pixel brightness above 235 or below 20. Retraining on this filtered

dataset resulted in a final test accuracy of 76.76%. Despite the lower quantitative accuracy, this model was selected for the final pipeline due to its significantly better performance in real-world conditions. The observed difference in accuracy between the RGB and IR models is primarily attributed to lighting issues affecting IR model learning and the quality dependency on the IR sensor used. A robust sensor capturing clear images in low light is vital for improving IR model performance. A graphical representation of the predictions, shown in Fig. 3, was developed.

IV. DRIVER IDENTIFICATION SYSTEM

The driver identification system uses a dual RGB and IR camera setup to enhance vehicle security and personalization. The RGB camera is the primary sensor. The IR camera serves as a backup for conditions like low-light illumination or RGB camera malfunction.

The captured image is sent to the face detector module, which uses MTCNN to accurately detect faces in various conditions. If MTCNN successfully detects the largest face in the RGB image, it is sent to the feature extractor. If the MTCNN model fails to detect a face in the RGB image, a CLAHE filter [26] is applied to the IR image to enhance contrast, and MTCNN attempts face detection. The detected and cropped face is sent to the feature extractor, which uses FaceNet [27] to transform the face into a 128-dimensional embedding, serving as a unique ID. For re-identification, this new embedding is compared against all registered IDs using cosine similarity. Registered IDs store separate mean embeddings for RGB and IR to allow identification across both lighting conditions. The thresholds for a match are set at 0.65 for RGB and 0.575 for IR images. These values were optimized for better re-identification metrics through testing. The architecture of the system is illustrated in Fig. 4.

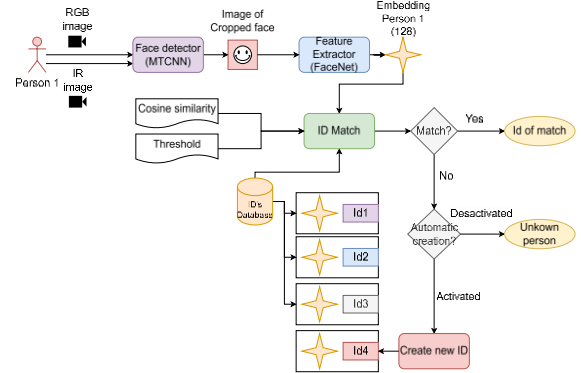


Fig. 4. Driver identification system data flow diagram

To enhance robustness against variations in angle, illumination, and facial expressions, the final embedding for each ID is calculated as the mean of all embeddings generated from multiple captured images. During manual registration, the system captures three images of the driver, ideally with different head poses, ensuring they are sufficiently different by checking that the cosine similarity between each pair exceeds a given threshold. This method ensures that the embedding captures the driver's face from different angles, facilitating easier identification.

The system was evaluated on RGB images from the DMD dataset. The metrics used in these tests were accuracy, False Accept Rate (FAR), and False Reject Rate (FRR). FAR measures the rate at which unauthorized users are incorrectly accepted by the system, while FRR measures the rate at which authorized users are incorrectly rejected. The tests utilized RGB images from the dataset, with ten individuals not included in the database and fifteen individuals included. The system's performance, using the mean embedding approach, was measured using standard security metrics, resulting in: accuracy of 99.38%, a FAR of 0.81%, and an FRR of 0.49%.

V. OCCLUSION DETECTION AND UNDERSTANDING

We utilize the Moondream2 [28] vision language for occlusion classification, chosen for its small size (2B parameters) to avoid high specification hardware requirements compared to larger models like Qwen (7B or 72B of parameters) [29]. This pioneering approach leverages the VLM's zero-shot classification capabilities and extensive knowledge. Unlike challenging traditional multi-class classifiers limited by specific training data, the VLM enables the system to diagnose the specific cause of obstruction (e.g., hand, sunglasses) beyond EuroNCAP's specific examples, providing superior explainability compared to binary detection methods.

The VLM module is strategically activated only after an extended occlusion is suspected. The primary trigger is the failure of the face detector to return a bounding box. Following a 3-second continuous signal loss (no face detected), the system confirms an extended occlusion and triggers the VLM.

This strategy represents a practical and efficient method for incorporating heavy, foundational models like VLMs into real-time safety applications. The VLM is not intended to replace the primary, lightweight perception systems; rather, we rely on them for high frame-rate operation. The VLM is reserved for situations where ordinary models need assistance, such as failing to find a cause, detecting anomalies, or confirming a complex situation. Employing VLMs for frame-by-frame prediction would be computationally infeasible and often monotonous, contributing little to the system's overall diagnostic capability. By utilizing the VLM only when a persistent failure is detected, we achieve high-value, contextual understanding.

The failed frame is resized and passed as input to the Moondream2 model. This VLM is tasked with analyzing the image based on the following prompt: *"You are an AI assistant analyzing a driver's face for occlusion detection. Based on the input image, briefly and concisely explain what is causing the occlusion of the driver's eyes or mouth. Focus on specific behaviors (e.g., looking at the right or left side, drinking,...) or objects (e.g., hands, accessories, or obstructions) that are blocking the driver's eyes or mouth. Provide concise and actionable insights in simple language"*.

The VLM's diagnostic output directly fulfills and enhances EuroNCAP recommendations for alerting the driver to system degradation. The system requests and displays the VLM's textual explanation of the cause (e.g., "hands blocking the face"). There, an appropriate warning, such as the one in Fig. 6, could be triggered, indicating the extended occlusion.

Only non-occluded frames are forwarded to subsequent modules that rely on unobstructed visual inputs, such as the gaze classifier and driver identification, which maintains downstream reliability.

VI. PIPELINE IMPLEMENTATION AND EVALUATION

We created a cohesive pipeline that integrates the four main components—driver identification, gaze estimation, distraction detection, and occlusion classification—to build a robust DMS. The system uses a dual RGB and IR camera setup positioned behind the steering wheel (for face view) and another camera attached to the A-pillar (for body view). The data flow is conceptually described in Fig. 5.

The distraction detection algorithm, running continuously with the RGB body camera feed, classifies the driver's actions into 5 categories. It alerts the driver of long distraction after 3 seconds, following EuroNCAP recommendations. This module does not employ dual IR/RGB functioning.

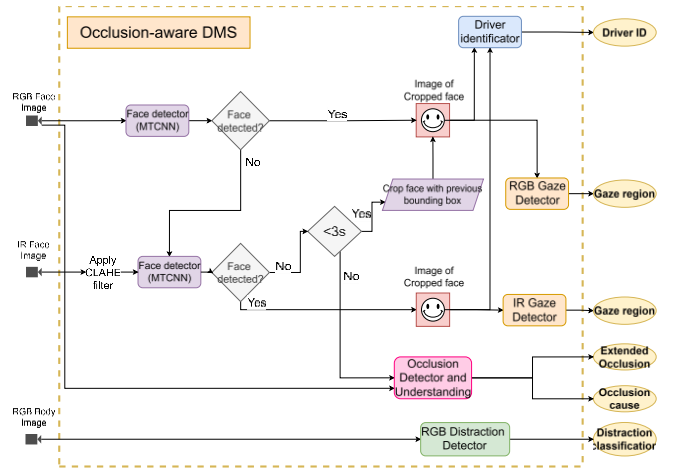


Fig. 5. Occlusion-aware DMS data flow diagram.

The pipeline primarily operates with the RGB image, using the IR image as a fallback. The core logic is as follows: 1) The RGB image is first passed to the MTCNN face detector. If a face is detected, the image is cropped and passed to the RGB Gaze Detector for region classification. The face is also used for driver identification (which runs at the start and periodically thereafter, as it is assumed that the driver will not change). 2) If no RGB face is detected, the system switches to the IR pipeline, applying a CLAHE filter to the IR image before MTCNN attempts face detection. If an IR face is detected, it is passed to the IR Gaze Detector. The system will only revert to the RGB flow after confirming consistent positive detections with RGB images, ensuring a quick and reliable mode switch for transient events like driving through tunnels. 3) If no face is detected in both channels, a timer is initiated. As an optional functionality, during this timer, the image is initially cropped using an "estimated bounding box" (the previous face detection's bounding box extended by 20%) to address potential false negatives from the face detector while accommodating minor head movements. 4) If the face detector failure is continuous for 3 seconds, an extended occlusion is confirmed, and the Occlusion Detector and Understanding module is triggered. The VLM analyzes the RGB image to diagnose the specific cause of the occlusion (e.g.,

"hands blocking the face"), and this information is used to issue an alert. The whole working system is shown in Fig. 6, where an extended occlusion is detected.



Fig. 6. Visualization of the occlusion-aware DMS. A prolonged face occlusion (3+ seconds) caused by the driver drinking water stops gaze prediction (upper right). The VLM diagnoses the cause (lower left) while the distraction detection module simultaneously warns of a "long distraction" (lower right).

A. Real-world Evaluation

To validate the models' effectiveness beyond quantitative dataset results, we conducted tests in real-life scenarios. Subjects were instructed to sequentially look at the gaze regions and occlude their faces in various ways to assess robustness to various types of occlusions. Since distraction is not part of the contribution of this paper, it was not evaluated. The metrics of performance can be found in [3].

The RGB gaze classifier demonstrated high reliability, correctly classifying all regions except for the front right (Region 5), which was often confused with the center mirror (Region 4). This expected confusion is attributed to the minimal head turn required between these zones. The IR gaze classifier was weaker, correctly predicting only a few regions (left mirror, front, steering wheel, and infotainment) and incorrectly predicting the rest. These real-life results confirmed that the RGB gaze model outperforms the IR model, demonstrating higher reliability and robustness, aligning closely with the expectations set by the test data.

The occlusion classifier showed superior robustness, making accurate predictions in all tested scenarios, including face obstruction with a hand or bottle, as shown in Fig. 7.

VII. LIMITATIONS AND FUTURE WORK

This system is primarily designed as a perception system focused on accurately analyzing potential problematic or unsafe driver situations. It is not a decision-making system for vehicle automation, nor does it provide a proper Human-Machine Interface (HMI) for direct communication with the driver. Its utility lies in providing rich in-cabin situational data—specifically driver identification, gaze region, distraction state and occlusion type—to other advanced driver-assistance systems (ADAS) or vehicle control modules for risk assessment and eventual automated response.

While the DMS components were designed for computational efficiency (e.g., using Efficientnet_b0 for gaze estimation and a small VLM), this paper does not intend to report definitive real-time performance metrics for the complete,

integrated pipeline. The introduction of the VLM for occlusion classification, while enhancing diagnostic capability, poses a significant computational challenge, given today's typical vehicle-embedded systems. Our strategy offers an approach to implement these models, given the constraints.



Fig. 7. Examples of occlusion detection and understanding. From left to right and up to down: Hair covering the face, hand scratching eye, hands on 12 o'clock position, drinking, holding a card and dark sunglasses.

The comprehensive evaluation of the advanced VLM-based occlusion module faces two primary limitations: 1) Driver occlusion dataset: There is currently no publicly available dataset that provides the necessary variety of situations and contexts leading to fine-grained occlusions (e.g., objects, different behaviors, and accessories) for robust testing. The DMD used in this work offers occlusion labels that correspond primarily to camera obstruction due to hand placement. 2) Evaluation Metric: It is inherently difficult to establish a single quantitative metric for evaluating the VLM's understanding and identification of the occlusion cause. Since the occlusion system output is a descriptive text (e.g., "hands blocking the face"), this cannot be easily reduced to a single label for a simple accuracy score, necessitating human analysis and comparison. Therefore, a comprehensive quantitative test, beyond the qualitative real-world testing performed, could not be conducted. Still, our real-world validation gave good results.

These limitations incentivize several key directions for future research: 1) Our next critical objective is to extend the DMD to consider a broader, more extensive range of occlusion types, aligning with and expanding upon situations highlighted by EuroNCAP (e.g., sunglasses, hair, hands on the wheel). 2) Evaluation Methodology: We will define a formal evaluation methodology for assessing the descriptive output of the VLM, moving beyond simple accuracy to measure the semantic quality of the cause identification. 3) Complete in-cabin scene and risk assessment: Future work will aim to incorporate additional DMS functionalities, such as drowsiness detection or occupant monitoring; also, improving distraction detection to include a wider range of behaviours. This integration of signals, will allow for a unified driver's state and general in-cabin situation estimation and risk assessment.

VIII. CONCLUSIONS

We presented a comprehensive DMS pipeline, integrating driver gaze estimation by regions, driver identification, and driver distraction detection, with a pioneering focus on occlusion awareness. The system utilized the DMD to train specialized algorithms capable of operating with both RGB and IR images, showing another successful case of using the DMD.

This research introduces the first application of VLMs for fine-grained occlusion analysis in a DMS. By using VLMs, the system moves beyond simple binary detection to diagnose the specific cause of obstruction (e.g., hand, accessory). This demonstrates an effective strategy for integrating heavy VLMs into real-time applications by triggering the VLM on-demand to enhance the existing lightweight perception system, rather than running it continuously.

The DMS is versatile, requires no calibration for gaze estimation, and functions effectively in challenging environments, including low-light conditions. Crucially, the system can alert the driver and stop gaze predictions when performance is degraded due to occlusion, thus adhering to and enhancing EuroNCAP recommendations. This effort marks a pioneering step in introducing robust occlusion detection functionality into a DMS, significantly enhancing its reliability.

ACKNOWLEDGMENT

This work was funded by the Horizon Europe programme of the European Union, under grant agreement 101203230 (project CERTAIN). Funded by the European Union. Views and opinions expressed here are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

- [1] SAE International, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," J3016_202104, Apr. 2021.
- [2] Euro NCAP, "Assessment Protocol – Safety Assist Safe Driving," Version 10.4, European New Car Assessment Programme, 2024.
- [3] HIDDEN FOR DOUBLE-BLIND REVISION
- [4] Z. Hu, C. Lv, P. Hang, C. Huang, and Y. Xing, "Data-Driven Estimation of Driver Attention Using Calibration-Free Eye Gaze and Scene Features," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 2, pp. 1800-1808, 2022.
- [5] J. Girgis, M. Powell, B. Donmez, J. Pratt, and P. Hess, "How do drivers allocate visual attention to vulnerable road users when turning at urban intersections?" *Transportation Research Interdisciplinary Perspectives*, vol. 19, 2023, Art. no. 100822.
- [6] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena, "Car that Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models," in *IEEE ICCV*, 2015, pp. 3182-3190.
- [7] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4318-4322, 2020.
- [8] S. Ghosh, A. Dhall, G. Sharma, S. Gupta, and N. Sebe, "Speak2Label: using domain knowledge for creating a large scale driver gaze zone estimation dataset," in *Proceedings IEEE/CVF Workshops*, 2021.
- [9] S. Jha, N. Al-Dhahir, and C. Busso, "Driver Visual Attention Estimation Using Head Pose and Eye Appearance Information," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 4, pp. 216-231, 2023, doi: 10.1109/OJITS.2023.3258184.
- [10] C. Lollett, H. Hayashi, M. Kamezaki and S. Sugano, "A Robust Driver's Gaze Zone Classification using a Single Camera for Self-occlusions and Non-aligned Head and Eyes Direction Driving Situations," 2020 IEEE

- International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 4302-4308
- [11] V. Vučićić, L. Eidel, M. Tang, and E. Ax, "USID - Unsupervised Identification of the Driver for Vehicle Comfort Functions," in *Human Interaction and Emerging Technologies (IHET-AI 2024)*, 2024.
- [12] P. Chen, J. Wu, and N. Li, "A Personalized Navigation Route Recommendation Strategy Based on Differential Perceptron Tracking User's Driving Preference," *Computational Intelligence and Neuroscience*, vol. 2023, Art. no. 8978398, 2023.
- [13] K. Dahiya, J. Goel, A. Kaushik, K. Rai, K. Jain, and A. Gambhir, "Evolving Payment Security: A Facial Recognition-Based Credit Card Reader with A Multifunctional Cascade Neural Network," in *IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, 2024, pp. 1630-1634.
- [14] L. Liu, "Facial Authentication System Design of Online Interactive Platform for Innovation and Entrepreneurship Courses for Mobile Platform Terminals," in *3rd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, 2022.
- [15] T. R. Gresham, "Safe and Sustainable Fleet Management with Data Analytics and Training," in *Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, 2021, pp. 1-6.
- [16] A. Makrushin, "Visual recognition systems in a car passenger compartment with the focus on facial driver identification," 2014.
- [17] K. Chan and C. Chao, "DriverID: Driver Identity System Based on Voiceprint and Acoustic Sensing," in *IEEE International Conference on Consumer Electronics 2022, Taipei*, 2022, pp. 45-46.
- [18] L. Bonfati, J. Junior, H. Siqueira, and S. Stevan, "Correlation Analysis of In-Vehicle Sensors Data and Driver Signals in Identifying Driving and Driver Behaviors," *Sensors*, vol. 23, 2022, Art. no. 263.
- [19] Z. Yu, "Searching Central Difference Convolutional Networks for Face Anti-Spoofing," in *IEEE/CVF CVPR*, Seattle, 2020, pp. 5294-5304.
- [20] HIDDEN FOR DOUBLE-BLIND REVISION.
- [21] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., 09-15 Jun 2019, pp. 6105-6114.
- [22] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [23] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.
- [24] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.
- [25] S. G. Müller and F. Hutter, "TrivialAugment: Tuning-free yet state-of-the-art data augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 774-782.
- [26] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2014, pp. 2392-23.
- [27] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE CVPR*, 2015, pp. 815-823.
- [28] Korrapati, V. moondream2. 2024. Retrieved from <https://moondream.ai/>
- [29] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., ... Lin, J. (2024). Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv [Cs.CV]*. Retrieved from <http://arxiv.org/abs/2409.12191>