An advanced multimodal driver-assistance prototype for emergency-vehicle detection

Original
An advanced multimodal driver-assistance prototype for emergency-vehicle detection / Gabrielli, L.; Migliorelli, L.; Cantarini, M.; Mancini, A.; Squartini, S.. - In: INTEGRATED COMPUTER-AIDED ENGINEERING. - ISSN 1069-2509. - 31:4(2024), pp. 381-399. [10.3233/ICA-240733]

Availability:
This version is available at: 11566/333852 since: 2024-09-02T08:35:05Z

Publisher:

Published
DOI:10.3233/ICA-240733

Journal Title 0 (2023) 1–0
IOS Press

An advanced multimodal driver-assistance prototype for emergency-vehicle detection

Leonardo Gabrielli, Lucia Migliorelli, Michela Cantarini, Adriano Mancini and Stefano Squartini [*],[**]
Department of Information Engineering, Università Politecnica delle Marche, via Brecce Bianche 12, 60131 Ancona, Italy
E-mails: l.gabrielli@univpm.it, l.migliorelli@univpm.it, michela.cantarini@pm.univpm.it, a.mancini@univpm.it, s.squartini@univpm.it

Abstract.

In the automotive industry, intelligent monitoring systems for advanced human-vehicle interaction aimed at enhancing the safety of drivers and passengers represent a rapidly growing area of research. Safe driving behavior relies on the driver's awareness of the road context, enabling them to make appropriate decisions and act consistently in anomalous circumstances. A potentially dangerous situation can arise when an emergency vehicle rapidly approaches with sirens blaring. In such cases, it is crucial for the driver to perform the correct maneuvers to prioritize the emergency vehicle. For this purpose, an Advanced Driver Assistance System (ADAS) can provide timely alerts to the driver about an approaching emergency vehicle. In this work, we present a driver-assistance prototype that leverages multimodal information from an integrated audio and video monitoring system. In the initial stage, sound analysis technologies based on computational audio processing are employed to recognize the proximity of an emergency vehicle based on the sound of its siren. When such an event occurs, an in-vehicle monitoring system is activated, analyzing the driver's facial patterns using deep-learning-based algorithms to assess their awareness. This work illustrates the design of such a prototype, presenting the hardware technologies, the software architecture, and the deep-learning algorithms for audio and video data analysis that make the driver-assistance prototype operational in a commercial car. At this initial experimental stage, the algorithms for analyzing the audio and video data have yielded promising results. The area under the precision-recall curve for siren identification stands at 0.92, while the accuracy in evaluating driver gaze orientation reaches 0.97. In conclusion, engaging in research within this field has the potential to significantly improve road safety by increasing driver awareness and facilitating timely and well-informed reactions to crucial situations. This could substantially reduce risks and ultimately protect lives on the road.

Keywords: Advanced driver-assistance system, emergency siren detection, in-vehicle driver monitoring, audio-visual signal processing, deep learning

1. Introduction

In the past few years, there has been a notable increase in automotive research focusing on technologies aimed at enhancing the safety of both drivers and passengers. This includes the development of intelligent vehicles that are equipped with advanced driver-assistance systems. There has been a notable increase in automotive research focusing on technologies aimed at enhancing the safety of both drivers and passengers [1]. This includes the development of vehicles equipped with advanced driver-assistance systems. (ADASs) [2, 3]. ADASs consist of sensor-equipped electronic devices intended to streamline operations and aid the driver during potentially hazardous situations [4, 5]. These systems are classified into six levels of automation according to the Society of Automotive Engineers (SAE) standard J3016 [6, 7]. In the lowest levels (0 to 2), the environment inside and surrounding the vehicle is controlled by the drivers,

---

[*] Corresponding author. E-mail: s.squartini@univpm.it.

[**] L. Gabrielli and L. Migliorelli equally contributed to the manuscript draft.

and the device merely supports them without acting directly. On the other hand, the system partially or fully replaces human intervention in the higher levels (3 to 5), where ADASs monitor the environment and handle multiple safety devices up to the ultimate goal of autonomous driving [8]. Almost all current vehicles are equipped with ADASs with automation levels between 0 and 2, relying primarily on vision [9] and other sensing technologies such as laser imaging detection and ranging (LiDAR) [10, 11], RADAR [12, 13], and ultrasonic [14, 15]. The most common camera-based solutions include automatic lighting in tunnels or at dusk, adjustment of windshield wipers depending on rain intensity, traffic-sign recognition, lane-change warning, and surround view. Additionally, obstacle detection and distance estimation are facilitated through LiDAR, RADAR, and ultrasonic technologies, enabling functionalities such as adaptive cruise control, emergency braking, parking sensors, and enhancing street navigation accessibility for individuals with impairments [16].

Recent technological advances leverage information from multimodal data to detect and identify events and scenarios [17], alert drivers to distractions and support them to make well-considered decisions on the road, preventing traffic accidents. Hearing and vision are primary factors in human driving, and deep learning applied to audio-video streams has enabled technologies to “listen” and “see” via sensors, understand the cues and respond accordingly [18]. In some driving scenarios, audio and video data assume a complementary role and mutually exhibit a more effective context representation capability, as illustrated in several case studies. In narrow spaces, confined layouts, or dense obstacles (like alleys in historic villages) or densely built-up areas with abundant vegetation, audio data can often provide more insightful information than video one. This is particularly true for identifying and localizing both static and moving sound sources such as cars, bicyclists, or pedestrians [19]. Audio-based systems also detect road moisture and contribute to road safety by assessing pavement roughness, deterioration, speed, and traffic density [20–24]. In addition, audio is particularly effective in identifying weather conditions, such as varying precipitation intensities and low daylighting. This is essential for functions such as automatic activation of windshield wipers and speed control for which rain sounds on different surfaces are automatically analyzed [25, 26].

On the other hand, vision sensors, being non-invasive, are ideal for in-vehicle monitoring [27]. Indeed, video data, analyzed via computer vision or deep-learning algorithms, successfully address various driver’s behaviour such as: fatigue, distraction and attention level [28]. Particularly the latter, can be assessed by analysing facial expressions or head/eye movements [29]. This information assumes considerable importance in enhancing road safety, as it enables the development of ADASs capable of alerting in real time or automatically activating safety devices with inattentive drivers [30].

1.1. Motivation and scope of the work

Currently, ADASs that do not rely on measuring a physical quantity but rather on understanding the surrounding environment are not off-the-shelf equipment in commercially available cars. This scenario includes emergency-vehicle detection systems, thus devices designed to detect vehicles in an emergency state such as ambulances, police cars or fire trucks. Recognizing these vehicles is critical when they approach at high speed. In these situations, drivers must be aware of the approaching emergency vehicle, understand what actions might be appropriate based on traffic conditions.

Vehicular ad hoc networks (VANETs) [31] have been proposed as a step forward for accident prevention, but they are not currently employed in commercial vehicles. To the best of our knowledge, the first proposed implementation of a VANET for emergency vehicles dates back to 2009 [32]. To the same extent, a system employing the Radio Data System protocol (RDS) to broadcast the presence of an emergency vehicle to other vehicles has been proposed in [33]. Unfortunately, to the best of our knowledge, radio communication technologies for accident prevention are not yet found in vehicles, therefore the only standardized system to alert drivers and pedestrians of the presence of emergency vehicles is the use of lights and sirens.

Unfortunately, high quality soundproofing in vehicles and drivers’ hearing impairments can reduce the ability of the driver to detect incoming emergency vehicles [34, 35]. To provide an example, the vehicle used in [36], attenuates external sounds by 45 dBA and was shown to delay the hearing of a siren by more than 5 s. Other factors such as driver response and environmental conditions, can contribute to potential collisions or accidents involving emergency vehicles, emphasizing the need for comprehensive research and technological advancements in this domain [33]. Although, to the best of our knowledge, no statistical study showed how misheard sirens can lead to crashes with emergency vehicles, some evidence arises from the literature. Studies have been previously conducted for siren detection using a smartphone in [37], while [38] proposed ways to improve siren sound and horn positioning to make it easier to spot. Several patents have been filed for emergency vehicle avoidance using acoustic sensors (see, e.g., US patents [39] and [40]) and a study has been commissioned in 2017 by the U.S. Department of Transportation Office of Emergency Medical Services (EMS) [41] showing that lights and sirens are not always perceived by drivers of other vehicles.

This article, which extends previous work described in [42], presents a driver-assistance prototype for emergency-vehicles detection that combines audio and video data to detect the presence of a rescue vehicle nearby and monitor the driver's awareness. With respect to our previous work, which introduced the architecture of the prototype, here we provide an in-depth overview of it and we describe computational algorithms as well as datasets, providing experimental details and results. The core of the prototype lies in recognizing the siren sounds emitted by electronic devices in emergency vehicles. Although there may be minor discrepancies in sound emission parameters across different countries, the high-intensity acoustic alarm proves to be pivotal in notifying citizens and drivers. This sound has the ability to effectively reach intended receivers, grabbing their attention even when they are at significant distances or there are intervening obstacles. Following siren recognition, the system assesses whether to alert the driver based on behavioral analysis, particularly monitoring eye status and gaze orientation, providing crucial cues about their level of alertness.

In comparison to existing literature contributions such as [43–45], which were predominantly focused on emergency-vehicle detection, our work adopts a comprehensive approach by engineering a prototype to seamlessly integrate it into cars. This involved the selection of hardware components readily available in the market, tailored for in-vehicle installation. Additionally, software logic based on deep-learning algorithms was developed for both siren-sound detection and driver-attention monitoring.

The rest of the paper is organized as follows. Section 2 presents the state of the art of emergency siren detection and driver's attention monitoring systems. The hardware and software architectures of the driver-assistance prototype are described in Section 3, and Section 4 explains the deep-learning methodologies employed in this work by analyzing the workflows of audio and video systems. The experimental protocol is detailed in Section 5, and the results of the experiments are summarized and discussed in Section 6. Finally, Section 7 concludes the article and outlines our study limitations, future challenges and perspectives.

2. Related work: emergency-vehicles detection and driver-attention monitoring systems

The ADA prototype combines emergency-siren detection and drivers' monitoring into a unique solution, creating an in-vehicle device designed to heighten the driver's awareness in situations where they might be inattentive to an approaching emergency vehicle.

This section discusses the state-of-the-art technologies behind emergency-vehicle detection and driver attention monitoring systems focusing on solutions based on audio and video data, respectively.

2.1. Emergency-vehicles detection systems

Emergency-vehicles detection has been an ongoing research topic, resulting in the development of several models of emergency-vehicle detection systems that have evolved along with sensing technologies. Emergency-vehicle detection systems have upgraded from basic electronic devices to advanced digital systems, with the main goal of helping rescue vehicles reach their destination more quickly and safely. The literature reports a wide range of patents of emergency-vehicle detection systems that base emergency vehicle identification on different approaches, such as radio frequency and electromagnetic data detection, image recognition, and GPS tracking [46–49].

Audio data processing and analysis have always played an important role in the emergency-vehicle detection field due to characteristic alarms emitted by embedded electronic devices. In the 1960s and 1970s, early audio-based emergency-vehicle detection systems used electrical circuits equipped with analog filters to select and amplify sounds recorded with external microphones in the range of siren frequencies, also combined with frequency-voltage converters to detect the slow and continuous variations of the siren signal [50, 51]. Since the 1980s and 1990s, more advanced emergency-vehicle detection systems based on digital signal processing applications have been developed. In several patents, emergency siren detection is performed with digital devices that convert audio signals into discrete time-frequency representations. After spectrogram computation, the system finds the match with the siren frequencies or analyzes the peaks of the signal, also applying a band-pass filter to select the siren tone frequency range [52, 53]. Similar technologies are described in [54], in which a pitch detection algorithm based on the module difference function and peak searching has been implemented on a low-power microprocessor, and [55], in which a two-times Fast Fourier Transform algorithm for siren detection has been programmed on a microcontroller. The limitations of these approaches lie in the performance decay at low signal-to-noise ratios and in the presence of the Doppler shift, as they prevent the recognition of the match between the acquired signal and the reference signal [56]. In addition, some of these algorithms require the completion of the entire siren sound pattern to provide a classification result, with consequent slow detection response [57, 58].

Emergency-vehicles detectors are becoming increasingly sophisticated today, employing deep learning to detect and classify sirens. In particular, the capability of convolutional neural networks to identify the features of the emergency siren at low signal-to-noise ratios, in the presence of the Doppler effect, and on short audio frames (e.g., between 0.5 and 1.5 seconds) has been thoroughly investigated in several contributions [43, 59, 60]. Recent fully audio-based emergency-vehicle detection systems deploy deep learning techniques to detect the emergency siren sound. In [61], the equipment comprises microphones to acquire external sounds in real time and a computing device to perform audio signal segmentation, spectrograms computation and analysis using a convolutional neural network pre-trained for emergency siren recognition. More complex studies integrate computational audio processing and computer vision techniques to generate an audio-visual emergency-vehicle detection system. In [44], multimodal data consisting of siren sounds and ambulance images are analyzed on two separate branches, an audio-based stream and a vision-based stream, which produce independent predictions and merge the results to output a single decision at the final stage. This strategy is employed in patents [62, 63], in which a vehicle-mounted system consisting of audio and video sensors and a computational unit designed to process, concatenate audio-visual feature vectors and generate a response on the presence of an emergency vehicle in the surrounding environment is presented.

2.2. Driver-attention monitoring systems

Modern ADASs, developed to actively or passively support the driver, include in-vehicle devices designed to monitor their level of attention or, in general, situation awareness. This status, defined as “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” [64], enables the driver to make appropriate decisions on the road avoiding hazardous situations both in the context of non-automated driving and in the transition phases of conditionally automated vehicles [65].

Research on driver-attention monitoring systems has developed solutions relying on biological and physiological parameters, vehicle parameters, and visual features of the driver’s facial expressions and movements [66]. Systems that employ sensors to monitor biological and physiological parameters (e.g., electroencephalogram, electrocardiogram, skin temperature, electro-dermal activity, electromyography, and electrooculography) have the advantage of being accurately informative about the driver’s psychophysical state [67]. However, physiological sensors involve skin-contact electrodes that can be perceived as invasive and annoying during driving operations. For this reason, most of the solutions found in commercial cars operate on parameters linked with the vehicle or visual patterns of the driver [68]. Vehicle-oriented technologies can model and recognize the driving style behavior to create a personalized profile [69, 70]. The vehicle speed, longitudinal and lateral acceleration, steering wheel angle, indicator and pedal usage, and some driver control actions in situations of crosswind or uneven road surfaces provide information on anomalous behaviors. However, the complexity of the variables involved in the driving style recognition task and the need to create customized profiles for each driver represent the disadvantages of these systems [71].

The non-intrusiveness in data acquisition, high resolution, low cost, and ease of installation and maintenance have favored the development of camera-based driver-monitoring systems to assess the vigilance state of the driver. Current setups employ one or more RGB or RGB-depth (D) cameras focused on the driver’s face and eyes to acquire eyelid, gaze, and head information, then elaborated by a processing device with computer vision techniques [72]. The eyelids provide feedback on the driver’s drowsiness by detecting a slowdown in blink frequency or eye closure for an excessive duration, assessed through indicators such as PERCLOS and AVECLOS [73]. In addition, eye movements, gaze direction, and head orientation are indicative of perception and awareness of signals outside and inside the vehicle. Systems developed by car manufacturers combine driver attention monitoring and ADASs in response to an inattentive state of the driver. Commercially available solutions employ warning tones or seat vibration signals if the blinking frequency is below the danger threshold [74, 75]. Other technologies enable gaze and face tracking whenever an obstacle is detected on the road, activating pre-crash warnings or, if necessary, automatic braking when the gaze trajectory and head orientation are not directed toward the road [76]. Several challenges involve artificial vision in automotive applications, such as real-time processing, robustness to light changes, and privacy. The use of edge processors as edge Tensor Processing Units (eTPU) [77] supports the deployment of neural networks on embedded devices. Modern approaches to the above-mentioned tasks rely on deep learning that sets complex computing requirements [78–80].

While the state-of-the-art contributions are predominantly centered on algorithms pertaining to emergency-vehicle detection or support systems for driversâĂŹ monitoring, this study takes a different approach. It comprehensively delineates the entire architecture of the assistive prototype which includes both siren identification and driver assessment while posing emphasis on the explanation of the seminal multimedia-data analysis algorithms and datasets used. Full details and experimental results are provided, shedding light on critical contributions that significantly increase our understanding of the subject. In particular, we conduct an in-depth exploration of the challenges related to multimodal driver-assistance systems and their potential impact on real-world scenarios. To the best of our knowledge, this work represents one of the first attempts in literature to propose such a comprehensive integration, marking a substantial step toward the advancement of this technology. We acknowledge, however, that further research is essential for the development of a fully engineered prototype.

3. Proposed prototype

Our driver-assistance prototype is intended to both detect a wailing siren from an emergency vehicle and alert the driver if they are not watchful of the approaching emergency vehicle. The prototype was installed in a Mercedes A-Class car that served in both the design and testing phases. In particular, during the design phase, part of the data relevant to the training, validation and testing of the deep-learning algorithms for multimedia-data analysis were acquired by traveling on the roads of our region (Marche, Italy) with the car equipped with the prototype for data collection [36, 81].

In the following, we outline the prototype algorithmic flow, which is further graphically rendered in Figure 1:

1. The prototype needs to monitor the presence of emergency vehicles constantly. This is done by automatically detecting sirens via microphones;
2. The detection of a siren triggers the driver's awareness-monitoring phase, which relies upon the RGB camera. This latter deals with gaze-fixation estimation and eye-status monitoring;
3. Warning signs – in the form of visual and aural cues – need to be sent to the driver when they are found not being watchful.

It is worth mentioning that both eye and gaze detection are meant to avoid alarming the driver if they are aware of the incoming emergency vehicle, therefore they are a means to improve the user experience.

Timing constraints must be carefully considered, given the application at hand. Consider a regular vehicle and an emergency vehicle, both running in the same direction at different speeds. If the regular vehicle must give way, it needs several seconds to take action. Let, e.g., the difference in speeds be 20 km/h and the distance to notice the siren be 50 m. It takes 8.9 s for the emergency vehicle to overtake the regular vehicle. In this time span, the following actions should take place: (a) the system completes the audio-data analysis and the video-data analysis; (b) the driver becomes aware of the situation; (c) due action is taken to let the emergency vehicle pass. Since giving way requires a few seconds, we constraint our application to take an order of magnitude less than 8.9 s to operate, i.e., less than a second. This time must include the latencies of the audio and video systems, and a period of time to wait for the user to show signs of awareness (corresponding to the last phase in Figure 1, i.e., "Is the driver aware?" box).

Given the objectives, the following sections highlight the hardware and software components of our driver-assistance prototype.

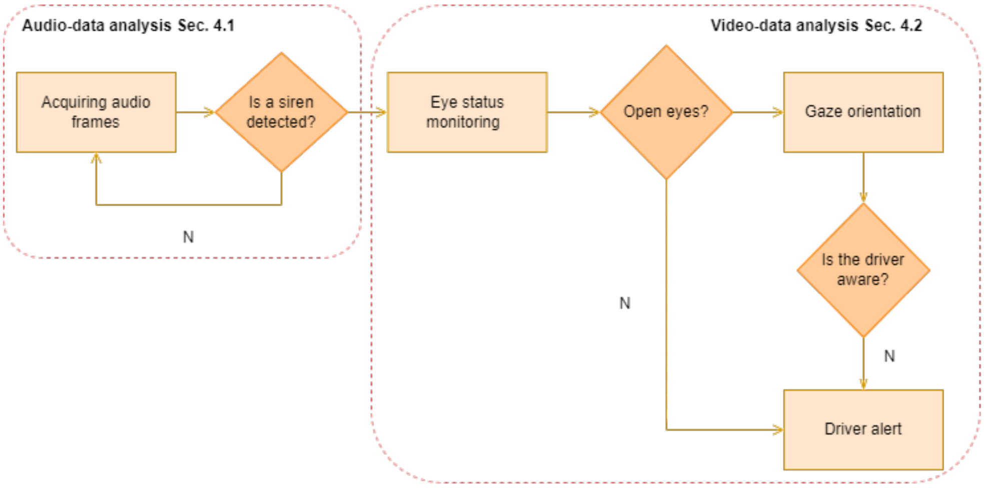


Fig. 1. Flow diagram of the driver-assistance prototype. Firstly the audio-acquisition system automatically detects the presence of a siren, which activates the camera to evaluate the driver's awareness.

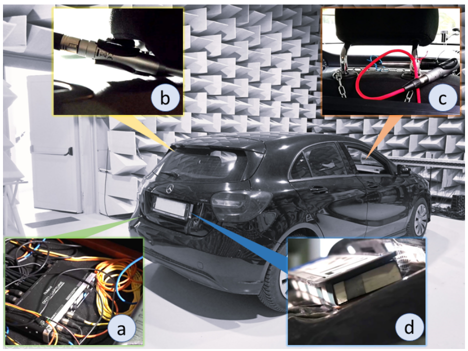


Fig. 2. Details of the audio setup: a) sound card inside the trunk, b) rear internal microphone, c) front internal microphone, d) external microphone behind the license plate.

3.1. Hardware

Omnidirectional Behringer ECM8000 condenser microphones, which can detect sounds in all directions, were chosen for acoustic scene analysis. The placement and number of microphones follow our previous studies that provided the advantages and disadvantages of each installation inside and outside the car [36]. The general audio configuration included a total of eight microphones, four inside the passenger compartment, two in the trunk, and two behind the license plate. Microphones inside the passenger compartment and trunk are not the optimal solution for recording sounds from outside. In-cabin sensors can pick up interference from conversations or radios, while those in the trunk are affected by mechanical noise. In both internal installations, the soundproofing power of the cabin atten-

Fig. 3. Detail of the dashboard of the car equipped with the camera and the global navigation satellite system (GNSS) receiver.

uates some frequency components of the signal. The external placement of the sensors, such as behind the license plate, is suitable for detecting sounds from outside, particularly from the rear, where the driver may have difficulty sensing an oncoming emergency vehicle. An eight-channel Roland Octa-Capture sound card was used to capture sounds, as a maximum of eight microphones is adequate to acquire audio signals functional for multiple tasks [26, 82]. Figure 2 shows some details of the audio setup of the research car.

Concerning the camera, we used the IDS UI-3160CP-C-HQ RGB camera. This USB 3.0 camera is equipped with a 2/3âĂİ global shutter CMOS sensor PYTHON2000 from Onsemi, providing a full resolution of 2.3 MP (1920×1200 pixels) with up to 165 FPS. This has been synchronized with the Pulse per second signal generated from an external GPS receiver based on a u-blox NEO-M8 GNSS device with an external antenna. Figure 3 shows the position of the camera and GNSS inside the car.

A glue-logic software has been included in the loop for handling processes outlined in Figure 1. Considering the need to visually alert the driver unaware of the arrival of an emergency vehicle, the hardware architecture also integrates a heads-up display (HUD) to send warning signs. Given the application scenario, the device to manage data streams and run the deep-learning algorithms for multimedia-data analysis in real time must have sufficient computing power while both fitting into the small space of the car and having limited power availability.

All the requirements guided us in opting for off-the-shelf components and, specifically, for an x86 computer with the ability to host a graphics processing unit (GPU) to obtain a power-efficient execution of deep-learning algorithms. Among the x86 suites, the one with the smallest footprint is the Intel NUC. This processing unit has reduced size and power requirements and, in our elected version, can host an external GTX 1650 GPU.

The maximum power of the system is 60 W (as its worst case), which, considering the prototyping stage, is provided by a power inverter that converts the 12 V DC of the car to a 230 V AC source for supplying the equipment. The NUC also has USB and HDMI connectors for the sensors and drivers to connect the HUD and can run any GNU/Linux distribution, enabling effortless software development.

Figure 4 schematizes the audio-video setup of the ADA prototype.

3.2. Software

The software architecture of the prototype requires several tasks to be executed in parallel. Therefore, a system based on parallel threads distinct in their functionality has been designed. We decided to use Python – apart from the ability to program a multi-threaded architecture – because (i) it allows us to implement flexible graphical user interfaces (GUIs), (ii) it can be ported to all common operating systems, and (iii) it has bindings for the most common libraries for deep learning, audio processing, and image processing.

As visible in Figure 5, the main process generates the GUI, the audio processing task (which soon involves the deep-learning-based analysis), and the video task. The main application is built with the *Kivy* library, an open-source application-development framework for Python. The GUI simulates a car dashboard, and besides some service buttons, it shows the emergency-vehicles detection status (Figure 6). When the automatic detection assistive system intercepts a siren, a warning message is displayed on the GUI to provide a visual cue to the driver. The sound thread uses the Python *sounddevice* library, based on the widely adopted *PortAudio C* cross-platform library. *Sounddevice* enables registering a callback function to process an audio frame by frame. The callback, in turn, invokes the forward method of a *Tensorflow*-based deep-learning model trained to detect sirens in traffic and noisy conditions. Following the siren detection, a signal is passed to the video thread via the business-logic thread that handles the entire system. The video-data analysis pipeline evaluates whether the driver's gaze is directed toward a mirror when the siren is active and, thus, if the driver is watchful.

4. Deep-learning methodologies

4.1. Audio-data analysis

The ADA prototype bases the detection of an incoming emergency vehicle on the sound recognition of its active siren. For this purpose, the audio acquisition system captures audio signals through one or more microphones mounted on the car. Pre-processing operations are applied to the audio data stream before being analyzed by the deep learning algorithm to match the requirements of the pre-computed neural model. Thus, the complete workflow of the audio data analysis system includes a standardization phase, an acoustic feature calculation phase, and a classification phase.

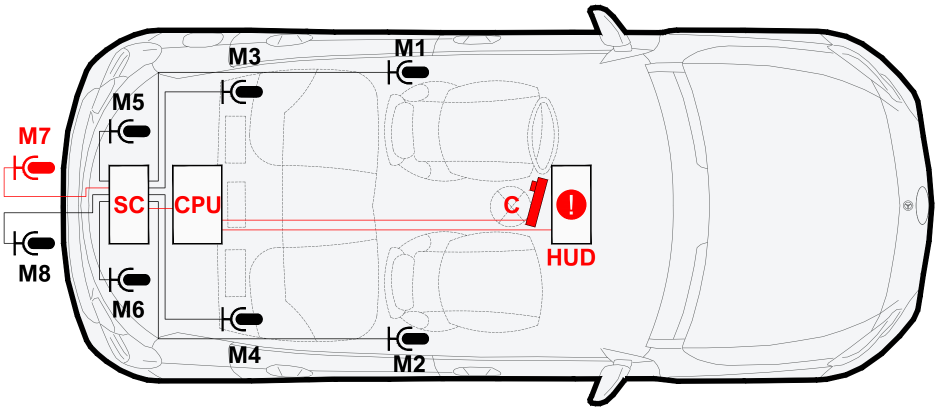


Fig. 4. The hardware setup of the ADA prototype: Behringer ECM8000 condenser microphones (M1–M8), Roland Octa-Capture sound card (SC), IDS UI-3160CP-C-HQ RGB camera (C), Intel NUC (CPU), and heads-up display (HUD). The components in red represent the basic equipment of the prototype.

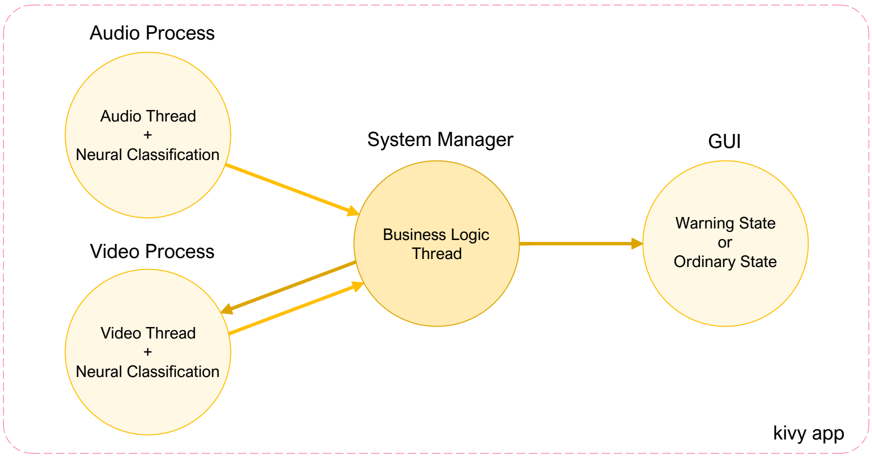


Fig. 5. Overview of the software nodes of the prototype system.

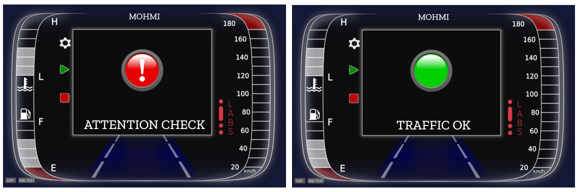


Fig. 6. GUI of the ADA prototype: warning state (on the left) and ordinary traffic state (on the right).

Previous studies concerning emergency siren detection have revealed the issues and challenges related to the task. The most significant are retrieving real-world siren audio data, implementing neural architectures with low computational cost, exploring strategies for reducing background noise and developing cross-domain adaptation techniques. In this study, we combine transfer learning and fine-tuning approaches using YAMNet, a pre-trained neural network that employs the MobileNetV1 [83] architecture to predict 521 audio event classes according to the definition of the AudioSet corpus [84]. Although this large-scale dataset includes several typologies of sirens, such as civil de-

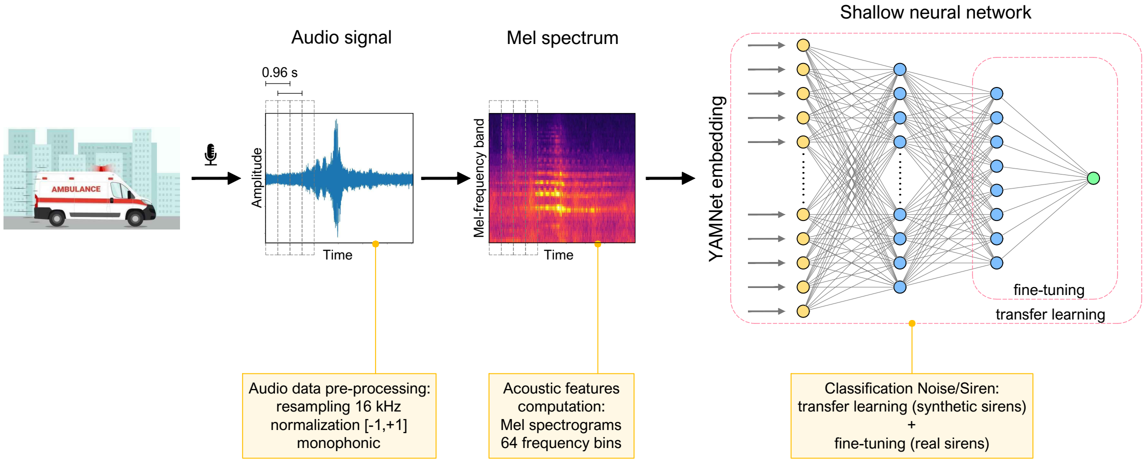


Fig. 7. The algorithmic flow involves processing audio signals captured by onboard microphones for emergency siren detection. The audio data is pre-processed and transformed into a Mel spectrogram. A pre-trained YAMNet model is used to extract embeddings, creating a 1024-dimensional feature vector. This was the input for a shallow neural network designed for noise/siren classification.

fense, police car, fire engine, ambulance, and generic, our study focuses on ambulance sirens according to Italian law [85], which are not covered in AudioSet. For this reason, the pattern recognition of an Italian ambulance siren, consisting of two alternating tones at 392 Hz and 660 Hz, required the development of transfer learning strategies rather than the application of the YAMNet model directly at inference time.

To implement transfer learning, data from the source and target domains should have the same format. The input accepted by the pre-trained model is a 1D-tensor containing a waveform of arbitrary length, represented as a mono audio file with a sampling rate of 16 kHz and normalized in the range [-1.0, +1.0]. The model extracts frames from the audio signal of 0.96-second duration with a 50% overlap for the sliding frame and computes Mel spectrograms as acoustic features. The Roland Octa-Capture soundcard part of the prototype acquires audio signals with a sampling rate of 44.1 kHz and a maximum of eight channels. Therefore, to meet the source format, the audio data are resampled to 16 kHz, made monophonic by averaging the amplitude of each audio channel (if more than one recording sensor is employed), and normalized in amplitude. From pre-processed audio data, Mel-scale spectrograms are extracted on the basis of a triangular filter bank consisting of 64 Mel bins. The Mel spectrogram calculation begins with applying the short-time Fourier transform to the signal divided into frames of 0.025 seconds with a hop of 0.01 seconds. The triangular filter bank is then applied to the power spectra to generate a Mel spectrogram (or log-Mel spectrogram in the logarithmic scale). These acoustic features are widely used for detecting and classifying sound events, as the Mel filter bank simulates the selectivity of the human auditory system using frequency warping.

Once Mel spectrograms are computed, they are transformed into 1024-dimensional feature vectors by the YAMNet model. The advantage of a model pre-trained on a multi-class dataset that includes 2 million clips lies in its use as embedding extractor. By initializing the pre-calculated weights of the 1024-unit dense layer of the MobileNetV1, high-level features are computed and fed into a fully-connected neural network customized to the task. We structured the feed-forward model for emergency siren detection with two hidden layers and one output unit indicating the probability that the frame includes an ambulance siren sound. Specifically, the new model consists of a 16-unit hidden layer, a dropout layer with a drop rate of 0.5, an 8-unit hidden layer, and a single-unit output layer. The exponential linear unit activation function was set in each hidden layer, and in the output layer, the sigmoid function returns values in the range [0,1] since the task is a binary noise/siren classification. Table 1 shows the configuration of the neural architecture for emergency siren detection.

The training strategy was implemented in two distinct phases. A first model (Siren-TL model) exploiting transfer learning from YAMNet was computed with partially synthetic data consisting of vehicular traffic noise recorded with the sensor-equipped vehicle and

Table 1

Configuration of the model for emergency siren detection.

Layer	Units	Activation	Output shape	# Params
Input	–	–	(N, 1024)	0
Dense	16	elu	(N, 16)	16 400
Dropout	–	–	(N, 16)	0
Dense	8	elu	(N, 8)	139
Output	1	sigmoid	(N, 1)	9
Total params				16 545

audio files of simulated ambulance sirens added to the real noise. Then, the weights of the last hidden layer were fine-tuned with samples of real siren recordings to bridge the mismatch between the source and target domains. The final fine-tuned model (Siren-FT model) was then employed for the inference on real audio data. The overview of the algorithmic workflow for emergency siren detection is illustrated in Figure 7.

4.2. *Video-data analysis*

The video data acquired with the RGB camera are relevant to monitoring the driver's alertness level when an emergency vehicle with an active siren is approaching. To define which actions should be monitored by deep-learning algorithms, we first conducted an experiment under static conditions with the car parked inside a semi-anechoic chamber (Figure 3). The trial involved 14 volunteers between the ages of 25 and 39, and the objective was to qualitatively assess the reaction of drivers sitting in the parked car when hearing a siren simulating the arrival of an emergency vehicle. The siren stimuli – synthetically produced – occurred randomly and simulated the arrival of a siren (with fading amplitude due to distance and Doppler effect). The signal amplitude, at its peak, was strong enough to be clearly heard inside the car cabin. This was necessary to make sure that the subject's response was clearly correlated to the stimuli. The subjects were introduced in the laboratory without knowledge of the experiment objectives, to make their response as spontaneous as possible. They were only instructed to sit in the car and put their hands on the driving wheel as if they were driving. The rationale behind the test was explained to them afterward. The goal of this preliminary trial is to gather some basic understanding of possible cues of drivers being aware of the emergency vehicle presence from its siren. These may be useful for the purpose of avoiding unnecessary alerts, thus improving the user experience. In the future more extensive trials should be conducted to further verify our findings on a larger subjects base and, possibly, in real environments.

From the analysis, we understood that more than half of the subjects (11 out of 14 volunteers involved) do not move their heads but rotate their gaze by pointing their attention to the left rear-view mirror; only a minority sample also rotates the head (3 out of 14 volunteers involved). Therefore, the automatic pipeline for monitoring drivers' behavior at the approach of an emergency vehicle, to be integrated into the prototype, dealt with eye status estimation (i.e., open or closed) and gaze orientation assessment when the eyes were open. An overview of the algorithmic pipeline for driver monitoring using video data is shown in Figure 8.

As shown in Figure 8, our pipeline first involves the identification of the driver's face and then automatically implements a crop in the area around the eyes. This was performed with the MediaPipe face mesh [86], which is aimed at estimating the 3D position of 468 facial landmarks from monocular images. We chose MediaPipe for its ability to work in real time, even on mobile devices. It employs two deep-learning architectures to infer the geometry of the face surface without the need for a dedicated depth sensor. The first architecture acts as a detector and operates on the entire image to compute face location (this, specifically, allowed us to exclude faces other than the driver). Then, a 3D face reference model operates on the previously identified 2D-landmarks locations to regress the 3D surface geometry. Landmarks related to eyes are then used to generate the corresponding bounding boxes (this is visible in Figure 8, too).

From the RGB images of the cropped eyes, we implemented a first pre-trained MobileNetV2 [87], which classifies if the driver's eyes are open or closed. We initialized MobileNetV2 convolutional kernels' weights with those of the pre-training on the Imagenet dataset. We defined three dense layers with 1024, 512 and 2 neurons, respectively, and we initialized them with Glorot weights initialization [88]. The dense layers were activated via the ReLU (in the first two layers) and the softmax (in the last layer) activation functions. We chose MobileNetV2 as it seeks to achieve good performance on mobile devices. The network relies upon an inverted residual structure, in which residual connections lie between bottleneck layers to allow gradients to flow through the network without passing through non-linear activation functions. The intermediate layers implement depth-wise separable convolutions with stride 2 to lower the computational burden

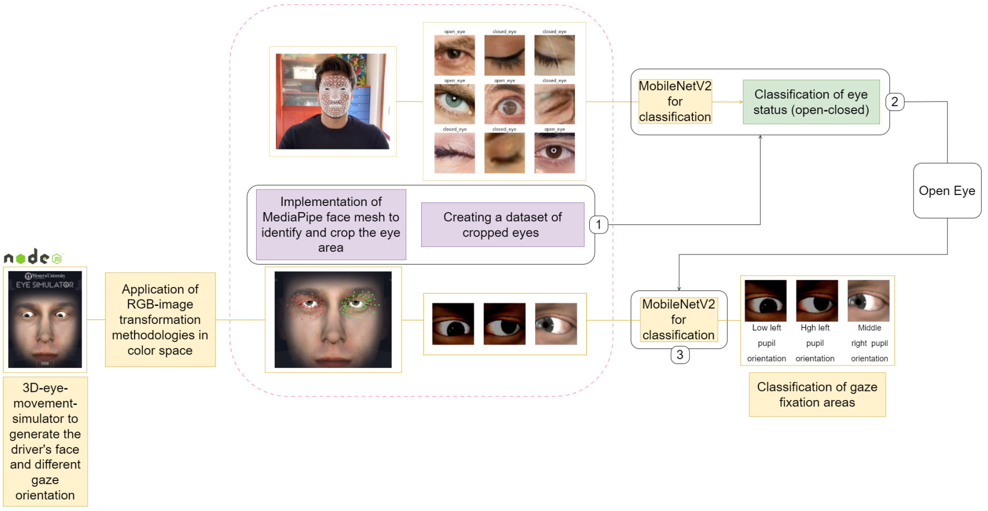


Fig. 8. Algorithmic flow to assess driver's attention level. In black, we reported the main flowchart following Figure 1. Two streams are also shown to distinguish experiments conducted with a simulated (bottom) and a real dataset (top) that share the same two main steps (dotted box).

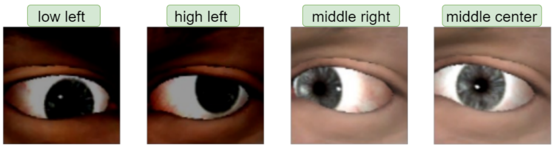


Fig. 9. Samples of classes for assessing the gaze-fixation areas.

(with respect to standard convolutions) and the feature map sizes. The last layer of the network has two neurons to classify whether the eye in the image is open or closed.

The open-eye classification was used to conduct the assessment of the driver's gaze and, specifically, whether or not the driver was looking at the left rear-view mirror. To this goal, we implemented a MobileNetV2 to classify the area in which the driver's gaze fell among nine possible: high center, high left, high right, low center, low left, low right, middle center, middle left, and middle right. A sample of the classes is shown in Figure 9. For our purposes, we follow the same training paradigm, except that the last dense layer has nine neurons as the classes of interest.

5. Experimental protocol

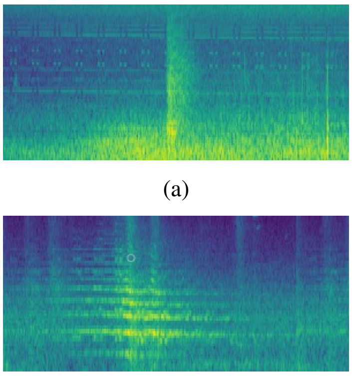
5.1. Audio-data analysis

5.1.1. Dataset

In the experiments, we used three datasets, the first to train the transfer learning model (A3S-Synth-TL), the second to fine-tune it (A3S-Aug-FT) and the last to test its performance (A3S-Rec). The audio collections consist of recordings made during two acquisition campaigns with the research vehicle in May 2021 and October–November 2022 [36, 81]. Also, synthetic data were created to address the amount of siren audio files required to train the transfer learning model.

The first dataset (A3S-Synth-TL) is partially synthetic and includes 200 audio files equally balanced between traffic noises and ambulance sirens of 60 seconds duration each. The noise audio files were randomly selected from diverse acquisition contexts (e.g., urban, suburban, rural, highways) and weather conditions (dry or wet), carefully checking that they did not contain ambulance sirens. To obtain a collection of siren audio files of adequate size and controlled quality, siren events that include the phases of an ambulance approaching, overtaking, and departure from the reference vehicle were generated via algorithm. The Doppler effect was simulated over a 60-second duration with the procedure described in [89]. Several source speeds and coordinates of the starting point relative to the observer were set, also considering attenuation by distance. The sirens thus generated were combined with real noise recordings at signal-to-noise ratios (SNRs) between [0,-30] dB. The A3S-Synth-TL dataset was split into training, validation and test sets with a 70:15:15 ratio and used for transfer learning from the pre-trained YAMNet model to the shallow model for Italian ambulance siren recognition.

Real-world data were deployed to fine-tune the model and test its performance. At the inference stage,


(a)

(b)

Fig. 10. Samples of a) synthetic and b) real siren spectrograms.

we employed the A3S-Rec dataset, a small collection of audio files recorded in several environmental contexts, comprising six siren events of variable duration and six noise audio files of 60 seconds each immediately preceding the siren occurrences. This dataset was chosen to compare the performance of transfer learning with YAMNet with techniques investigated in previous studies [36]. To fine-tune the neural model trained with synthetic siren data, we used a single 20-second additional siren event not present in the A3S-Rec dataset, increased to 5 minutes using data augmentation techniques (amplification, inversion, noise reduction, and noise addition). We randomly selected 5 noise recordings, each lasting 60 seconds, to create a 10-minute balanced dataset (A3S-Aug-FT). All the audio files of real-world sirens and noises belong to the recordings of the left external sensor behind the license plate (microphone M7). Audio datasets can be available upon request. Figure 10 shows examples of simulated and real siren spectrograms.

5.1.2. *Training settings and evaluations metrics*

To compute a model with high generalization capability during the testing phase, we investigated the performance of different architectures through a grid search approach [90]. We varied the number of hidden layers and neurons in each of them, evaluating several activation functions and the effect of dropout to reduce the overfitting on training data. The experiments for defining the transfer learning architecture were performed with a learning rate equal to 0.001 and Adam [91] optimizer for 500 epochs with early-stopping regulated by the validation accuracy.

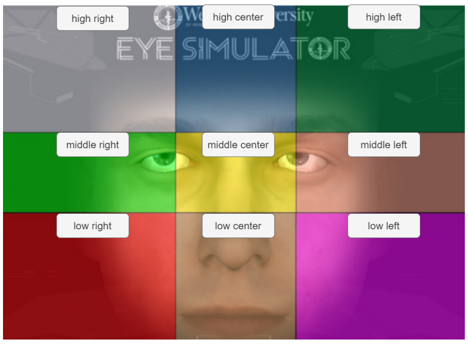


Fig. 11. Identification of the nine classes of gaze orientation.

The model that provided the best results in testing both with the A3S-Synth-TL and A3S-Rec datasets was fine-tuned with the real – plus augmented – instances of siren sounds and traffic noises of the A3S-Aug-FT dataset. This model is composed of two hidden layers, so we froze the first dense layer and retrained only the last linear layer with a low learning rate and few epochs to avoid rapid overfitting. The fine-tuning process was carried out with a learning rate of 0.0001, Adam optimizer, and 100 epochs with early stopping controlled by the training loss. In all experiments, the batch size was set equal to 4, and the binary cross-entropy loss was defined as the loss function.

Testing performance was evaluated with the area under the precision-recall curve (AUPRC). This metric, ranging between 0 and 1, is used for binary classification with unbalanced data and focuses on positive examples.

5.2. *Video-data analysis*

5.2.1. *Dataset*

The first dataset was a collection of real RGB images automatically acquired online via the Google search API SerpApi. Specifically, we focused our attention on the faces of people with open or closed eyes. Indeed these were the two classes of interest for our purposes which were further the search queries to download the images from the SerpApi. Before training our MobileNetV2, we conducted a data-cleaning step by removing biased samples (e.g., images without people's faces). Thus we derived a cleaned version of the final dataset, and we applied to this the MediaPipe face mesh to identify the bounding box related to the eyes of the person. Next, we automatically cropped the identified bounding-box area and applied a resize to the image making it 160×160 pixels. We derived a dataset of 6100 images per class (i.e., open and closed eyes). Then the totality of images (12200) was divided into training (70%), validation (20%) and testing (10%). Color-space transformations were implemented on-the-fly on the training dataset to simulate varying illumination conditions over the course of a day.

The second dataset we collected was generated through the simulator proposed by Western University[1]. At this prototype stage, we decided to take advantage of the tool as it enables us to decide how to orient the gaze in a face. This allows us to simulate how a driver's gaze moves when they hear the siren of an emergency vehicle. To automatically collect the images through the simulator, we programmed a bot in Node.JS. By moving the mouse cursor on the main window, we had the possibility to vary the gaze orientation and thus acquire images of subjects looking in different directions. Given the premises, we identified nine areas within the window and noted the gaze direction accordingly. The nine classes were: high center, high left, high right, low center, low left, low right, middle center, middle left, and middle right, as reported in Figure 11. Once the images were obtained, we first implemented transformations in the color space to increase the variability of the data (e.g., changing iris color) while mitigating possible biases [92]. Then, as for the dataset described above, we extracted via the MediaPipe face mesh pipeline the RGB frames of the eyes to collect a balanced dataset of 5376 images, of which 80% was for the training set, 15% was for the validation set and 5% for the test set. Video datasets can be available upon request.

Considering the pre-training strategy, each color channel of the images in both datasets was zero-centered with respect to the ImageNet dataset and pixel values were normalized between [-1.0, +1.0].

5.2.2. Training settings and evaluations metrics

For training both the MobileNetV2s, we used square images of size 160×160. Cross entropy was used as the loss function, and the optimizer was Adam [91]. The batch size was 64, while the learning rate was 0.001. The network was trained for 100 epochs, and the best combination of weights was selected among the 100 epochs with early stopping controlled by the accuracy on the validation set. All these training settings result from a grid-search analysis which allows us to find the best combination between loss, optimizer and learning-rate scheduling in terms of networks' efficacy.

Concerning the evaluations on the test set, we exploited the confusion matrix and its related metrics (e.g., accuracy).

6. Results and discussion

The proposed work presents a driver-assistance prototype for emergency-vehicles detection that uses audio data to detect the presence of an emergency vehicle approaching and video data to monitor the driver's awareness. Specifically, the prototype picks up, via an audio acquisition system, and automatically recognizes, through a deep learning algorithm, the sound emitted by the emergency-vehicle siren. After detecting the siren, the system decides whether or not to alert the driver through the analysis of behaviors that provide clues to the driver's alertness, particularly by monitoring eye status and gaze orientation.

As for audio analysis algorithms, the aim is the creation of a model capable of recognizing Italian ambulance siren sounds and generalizing them to different environmental contexts. In the first experimental phase, the Siren-TL model resulting from the configuration that proved to be the best performing of the several analyzed, described in Table 1, achieved an accuracy of 0.97 in testing on the A3S-Synth-TL dataset and an AUPRC of 0.81 on the A3S-Rec dataset without fine-tuning. In the second experimental phase, the Siren-FT model was computed by fine-tuning the weights of the last hidden layer, involving in this process only 145 trainable parameters. The test of the final model on the A3S-Rec dataset produced an AUPRC of 0.92. Table 2 compares the performance in noise/siren classification of different neural architectures on the A3S-Rec dataset and the number of parameters involved in the training process.

Analyzing the comparative results in Table 2, the CNN without fine-tuning entailed training a convolutional model on synthetic data and testing it on the A3S-Rec dataset without domain adaptation. In this case, although the convolutional neural network showed the advantage of a small number of trainable parameters, the AUPRC of only 0.65 underlined the discrepancy between synthetic and real data. The experiments using the CNN + fine-tuning of the two last hidden layers with only 50 siren and 50 noise frames of 0.5 seconds each, chosen randomly within the A3S-Rec dataset, confirmed the requirement for domain adaptation between the synthetic siren data and those recorded with the equipped vehicle. In fact, fine-tuning

---

[1] https://edtech.westernu.edu/3D-eye-movement-simulator/

Table 2

Comparison of the classification performance on the A3S-Rec dataset obtained with other systems.

Model/Method	Features	# Params	AUPRC [0,1]
CNN without fine-tuning [36]	Mel spectrograms 128 bands	19 948	0.65
CNN + fine-tuning [36]	Mel spectrograms 128 bands	35 340	0.84±0.02
Prototypical Networks [36]	Mel spectrograms 128 bands	111 936	0.86±0.02
Siren-TL	Mel spectrograms 64 bands	16 545	0.81
Siren-FT	Mel spectrograms 64 bands	16 690	0.92

operations with data belonging to the target domain significantly improved the classification performance in testing (average AUPRC equal to 0.84±0.02), despite an increased number of trainable network parameters. At last, the methodology offered by prototypical networks [93] had proven to be the most effective in conditions of limited availability of training data. Training the algorithm with synthetic data and using few samples of the real dataset for the embedding computation at the inference stage resulted in an average AUPRC equal to 0.86±0.02. While the prototypical architecture and episodic training strategy ensured excellent similarity learning among examples of the same class, the complexity of the network implied a larger number of trainable parameters than in previous studies.

The better results of transfer learning with YAMNet compared to the other techniques highlights the advantages of this strategy. This pre-trained model on a large-scale dataset can extract features useful for classifying a wide range of sound events without any adaptation. We observed that direct application of the YAMNet model to our siren data assigned them to the generic siren class, demonstrating its ability to identify the event as an alarm sound. For this reason, the adaptation of the model to a specific task, in our case, the recognition of the Italian ambulance siren, can be performed with a shallow neural network trained on a small-sized dataset. The improved performance also comes from the window size of the signal. The YAMNet model analyzes a 0.96-second time window with a 0.48-second shift, facilitating recognition of the siren tone sequence with respect to 0.5-second frames without overlapping employed in the previous studies. Finally, the reduced number of trainable parameters makes the model the most suitable for real-time applications in embedded devices.

Video analysis has two main steps and is triggered to monitor the driver's awareness when an emergency vehicle is recognized via the audio-based pipeline described above. The first step implemented an algorithmic pipeline relevant to assessing whether the driver's eyes are open or closed. The related results are shown in Figure 12 via the confusion matrix. The experiment was conducted using the MobileNetV2 trained, validated and tested on a dataset of real RGB images of faces from which we cut out the area around the eyes using the MediaPipe face mesh algorithm. As visible from the plot of the confusion matrix, the pre-trained convolutional neural network on Imagenet is able to generalize over the task of interest by achieving an accuracy of 0.99. In case the eyes are open, the second step of the video analysis involves assessing the orientation of the driver's gaze. This serves to assess whether the driver is paying attention to the mirrors (with particular attention to the left rear-view one). With this goal, a pipeline similar to the first step was implemented to classify the driver's gaze orientation. Specifically, the area around the eyes of a simulated face was derived via the MediaPipe face mesh algorithm then a MobileNetV2 was trained, validated and tested on RGB images of eyes with nine different orientations. The confusion matrix in Figure 13 shows the results of this experiment. The accuracy was equal to 0.97, with two samples belonging to the middle center class classified as middle right, probably due to the fact that the two classes are very similar to each other.

7. Conclusion

The presented research illustrates a multimodal ADAS prototype that aims to detect approaching emergency vehicles and promptly warn the driver if their attention appears to be lacking. Our prototype was designed and deployed inside a real vehicle located in a semi-anechoic chamber which, above all, has the main limitation of hampering the possibility of assessing the driver's behaviour who hears an emergency vehicle's siren. Although the results of our audio and video-data analysis algorithms are satisfactory, we acknowledge that further extensive research is needed to rigorously transition and evaluate the proposed system from controlled laboratory scenarios to real roads.

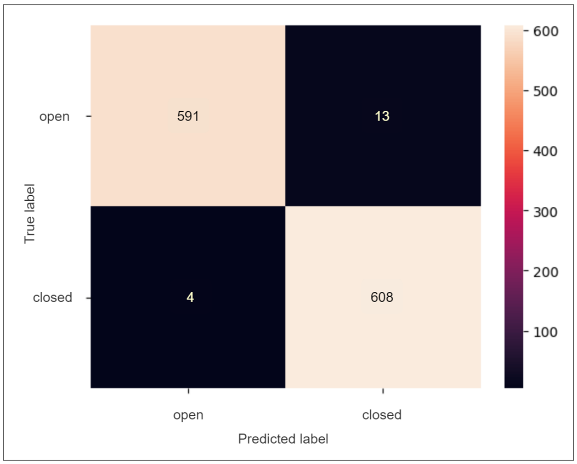


Fig. 12. Confusion matrix for ocular opening/closing classification task from RGB images.

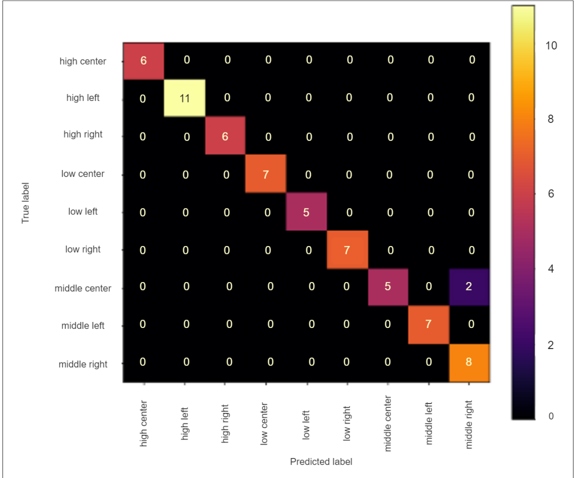


Fig. 13. Confusion matrix for the classification of gaze orientation from RGB images.

Specifically, we foresee notable challenges related to the seamless integration of hardware and software architectures, even though the miniaturization of acoustic and visual sensorsâĂŤalready standard in the latest car modelsâĂŤdoes not present substantial impediments for future developments. Instead, inspired by contributions in the closer literature [94], the focus of our efforts will be devoted to the innovative design and development of deep learning algorithms that promise effectiveness and efficiency when applied to real-world data.

With reference to efficiency, the latencies of all algorithms must be carefully assessed. In this regard, as a future development, we are working on re-engineering both the video and audio-analysis pipelines. Concerning audio, we are evaluating smaller overlaps between analysis windows to obtain a faster response from the algorithm and the application of filtering techniques to reduce traffic noise. For the video data, we are devising a single multi-task and lightweight (i.e., optimized for real-time computation on single-board computer-type devices) approach for directly estimating the driver's gaze orientation. While for both tasks of our interest, we are keen to explore new machine-learning paradigms inspired by closer fields of research [95–99].

To move beyond the prototype phase, we are also working on consistent data collection to train and validate all the algorithms on real use cases. Such a collection involves both the acquisition of new data on the road and the use of generated data to increase the number and variability of the samples. Indeed, regarding the audio, notwithstanding that we worked on real data collected by driving the car on Italian roads, the algorithms must recognize any sound produced by a siren per a specific country regulation, so other tones must surely be acquired to devise the large-scale distribution of the system.

On the other hand, for the video, we trained the deep-learning approach only on fictitious datasets. To this data we applied color-space transformation techniques to simulate, for example, varying lighting conditions throughout the day. Although pre-training on Imagenet is relevant for increasing the generalization power of the network, the dataset, as it stands, does not take into account any racial and gender bias that might emerge [100] from such a limited data collection. Moreover, for the prototype, we do not consider different head orientations that may affect the performance of the network for assessing the driver's gaze orientation, nor other elements beyond the driver's eye movements were monitored. In addition to the ethical risk, these aspects can seriously undermine any willingness to scale up the prototype.

A crucial solution to remedy the problems could be collecting new data and adopting pre-training techniques on a new, more structured pipeline. Ultimately, a fully approved street-ready prototype should undergo extensive road testing for hundreds of hours to validate its effectiveness in detecting emergency vehicles with diverse scenarios and subjects.

Acknowledgements

This work is financially supported by Marche Region with the programme POR MARCHE FESR 2014-2020. The project name is “Miracle” which is the acronym of Marche Innovation and Research fAcilities for Connected and sustainable Living Environments (CUP B28I19000330007).

We thank ASK Industries S.p.A. for providing the Mercedes A-Class vehicle used for the audio recordings and driver-assistance prototype installation.

References

[1] C. Ding, R. Dai, Y. Fan, Z. Zhang and X. Wu, Collaborative control of traffic signal and variable guiding lane for isolated intersection under connected and automated vehicle environment, *Computer-Aided Civil and Infrastructure Engineering* **37**(15) (2022), 2052–2069.

[2] F. Arena, G. Pau and A. Severino, An overview on the current status and future perspectives of smart cars, *Infrastructures* **5**(7) (2020), 53, doi: 10.3390/infrastructures5070053.

[3] T.K. Chan and C.S. Chin, Review of autonomous intelligent vehicles for urban driving and parking, *Electronics* **10**(9) (2021), 1021, doi: 10.3390/electronics10091021.

[4] V.K. Kukkala, J. Tunnell, S. Pasricha and T. Bradley, Advanced driver-assistance systems: A path toward autonomous vehicles, *IEEE Consumer Electronics Magazine* **7**(5) (2018), 18–25, doi: 10.1109/MCE.2018.2828440.

[5] E. Marti, M.A. De Miguel, F. Garcia and J. Perez, A review of sensor technologies for perception in automated driving, *IEEE Intelligent Transportation Systems Magazine* **11**(4) (2019), 94–108, doi: 10.1109/MITS.2019.2907630.

[6] S. International, Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, *SAE international* **4970**(724) (2018), 1–5.

[7] S.S. Shadrin and A.A. Ivanova, Analytical review of standard Sae J3016 «taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles» with latest updates, *Avtomobil'. Doroga. Infrastruktura.* (2019), 10.

[8] J.D. Fernández-Rodríguez, J. García-González, R. Benítez-Rochel, M.A. Molina-Cabello, G. Ramos-Jiménez and E. López-Rubio, Automated detection of vehicles with anomalous trajectories in traffic surveillance videos, *Integrated Computer-Aided Engineering* (2023), 1–17.

[9] G. Velez and O. Otaegui, Embedding vision-based advanced driver assistance systems: a survey, *IET Intelligent Transport Systems* **11**(3) (2017), 103–112, doi: 10.1049/iet-its.2016.0026.

[10] Y. Li and J. Ibanez-Guzman, Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems, *IEEE Signal Processing Magazine* **37**(4) (2020), 50–61, doi: 10.1109/MSP.2020.2973615.

[11] R. Roriz, J. Cabral and T. Gomes, Automotive LiDAR technology: A survey, *IEEE Transactions on Intelligent Transportation Systems* **23**(7) (2021), 6282–6297, doi: 10.1109/TITS.2021.3086804.

[12] G. Hakobyan and B. Yang, High-performance automotive radar: A review of signal processing algorithms and modulation schemes, *IEEE Signal Processing Magazine* **36**(5) (2019), 32–44, doi: 10.1109/MSP.2019.2911722.

[13] C. Waldschmidt, J. Hasch and W. Menzel, Automotive radarâĂŤFrom first efforts to future systems, *IEEE Journal of Microwaves* **1**(1) (2021), 135–148, doi: 10.1109/JMW.2020.3033616.

[14] A. Carullo, M. Parvis et al., An ultrasonic sensor for distance measurement in automotive applications, *IEEE Sensors journal* **1**(2) (2001), 143.

[15] Z. Qiu, Y. Lu and Z. Qiu, Review of ultrasonic ranging methods and their current challenges, *Micromachines* **13**(4) (2022), 520, doi: 10.3390/mi13040520.

[16] A. Mancini, E. Frontoni and P. Zingaretti, Embedded multisensor system for safe point-to-point navigation of impaired users, *IEEE Transactions on Intelligent Transportation Systems* **16**(6) (2015), 3543–3555, doi: 10.1109/TITS.2015.2489261.

[17] Y. Wei, D. Hu, Y. Tian and X. Li, Learning in audio-visual context: A review, analysis, and new perspective, *arXiv preprint arXiv:2208.09579* (2022), doi: 10.48550/arXiv.2208.09579.

[18] A.J. Muñoz-Montoro, P. Revuelta-Sanz, A. Villalón-Fernández, R. Muñiz and J. Ranilla, A system for biomedical audio signal processing based on high performance computing techniques, *Integrated Computer-Aided Engineering* **30**(1) (2023), 1–18.

[19] Y. Schulz, A.K. Mattar, T.M. Hehn and J.F. Kooij, Hearing what you cannot see: Acoustic vehicle detection around corners, *IEEE Robotics and Automation Letters* **6**(2) (2021), 2587–2594, doi: 10.1109/LRA.2021.3062254.

[20] R. Avanzato, F. Beritelli, F. Di Franco and V.F. Puglisi, A convolutional neural networks approach to audio classification for rainfall estimation, in: *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Vol. 1, IEEE, 2019, pp. 285–289, doi: 10.1109/IDAACS.2019.8924399.

[21] C. Ramos-Romero, P. León-Ríos, B.M. Al-Hadithi, L. Sigcha, G. De Arcas and C. Asensio, Identification and mapping of asphalt surface deterioration by tyre-pavement interaction noise measurement, *Measurement* **146** (2019), 718–727, doi: 10.1016/j.measurement.2019.06.034.

[22] F.G. Praticò, R. Fedele, V. Naumov and T. Sauer, Detection and monitoring of bottom-up cracks in road pavement using a machine-learning approach, *Algorithms* **13**(4) (2020), 81, doi: 10.3390/a13040081.

[23] S. Djukanović, J. Matas and T. Virtanen, Acoustic vehicle speed estimation from single sensor measurements, *IEEE Sensors Journal* **21**(20) (2021), 23317–23324, doi: 10.1109/JSEN.2021.3110009.

[24] G. Szwoch and J. Kotus, Acoustic detector of road vehicles based on sound intensity, *Sensors* **21**(23) (2021), 7781, doi: 10.3390/s21237781.

[25] J. Alonso, J. López, I. Pavón, M. Recuero, C. Asensio, G. Arcas and A. Bravo, On-board wet road surface identification using tyre/road noise and support vector machines, *Applied acoustics* **76** (2014), 407–415, doi: 10.1016/j.apacoust.2013.09.011.

[26] G. Pepe, L. Gabrielli, L. Ambrosini, S. Squartini and L. Cattani, Detecting road surface wetness using microphones and convolutional neural networks, in: *Audio Engineering Society Convention 146*, Audio Engineering Society, 2019.

[27] P.K. Murali, M. Kaboli and R. Dahiya, Intelligent In-Vehicle Interaction Technologies, *Advanced Intelligent Systems* **4**(2) (2022), 2100122, doi: 10.1002/aisy.202100122.

[28] X. Zeng, F. Wang, B. Wang, C. Wu, K.R. Liu and O.C. Au, In-vehicle sensing for smart cars, *IEEE Open Journal of Vehicular Technology* (2022), doi: 10.1109/OJVT.2022.3174546.

[29] G. Sikander and S. Anwar, Driver fatigue detection systems: A review, *IEEE Transactions on Intelligent Transportation Systems* **20**(6) (2018), 2339–2352, doi: 10.1109/TITS.2018.2868499.

[30] M.A. Regan, C. Hallett and C.P. Gordon, Driver distraction and driver inattention: Definition, relationship and taxonomy, *Accident Analysis & Prevention* **43**(5) (2011), 1771–1781, doi: 10.1016/j.aap.2011.04.008.

[31] S. Zeadally, J. Guerrero and J. Contreras, A tutorial survey on vehicle-to-vehicle communications, *Telecommunication Systems* **73** (2020), 469–489.

[32] A. Buchenscheit, F. Schaub, F. Kargl and M. Weber, A VANET-based emergency vehicle warning system, in: *2009 IEEE Vehicular Networking Conference (VNC)*, 2009, pp. 1–8. doi:10.1109/VNC.2009.5416384.

[33] B. Lidestam, B. Thorslund, H. Selander, D. NÃd'sman and J. Dahlman, In-Car Warnings of Emergency Vehicles Approaching: Effects on Car Drivers' Propensity to Give Way, *Frontiers in Sustainable Cities* **2** (2020). doi:10.3389/frsc.2020.00019.

[34] K. Choudhury and D. Nandi, Review of Emergency Vehicle Detection Techniques by Acoustic Signals, *Transactions of the Indian National Academy of Engineering* (2023), doi: 10.1007/s41403-023-00424-9.

[35] J. Palecek and M. Cerny, Emergency horn detection using embedded systems, in: *2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, IEEE, 2016, pp. 257–261.

[36] M. Cantarini, L. Gabrielli and S. Squartini, Few-Shot Emergency Siren Detection, *Sensors* **22**(12) (2022), 4338, doi: 10.3390/s22124338.

[37] D. Rane, P. Shirodkar, T. Panigrahi and S. Mini, Detection of Ambulance Siren in Traffic, in: *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2019, pp. 401–405. doi:10.1109/WiSPNET45539.2019.9032797.

[38] C.Q. Howard, A.J. Maddern and E.P. Privopoulos, Acoustic characteristics for effective ambulance sirens, *Acoustics Australia* **39**(2) (2011).

[39] B. Soltanian, M.C. Tsang, S. Gowda and D. Abendroth, Use of sound with assisted or autonomous driving, 2022, US Patent 11,351,988.

[40] S. Tariq, R. Gogna, M. Wimmershoff and S.S. Subasingha, Emergency vehicle detection and response, 2022, US Patent 11,450,205.

[41] D.F. Kupas, Lights and siren use by Emergency Medical Services (EMS): above all do no harm, National Highway Traffic Safety Administration, Office of Emergency Medical Services, Washington DC, 2017.

[42] M. Cantarini, L. Gabrielli, L. Migliorelli, A. Mancini and S. Squartini, Beware the Sirens: Prototyping an Emergency Vehicle Detection System for Smart Cars, in: *Applied Intelligence and Informatics: Second International Conference, AII 2022, Reggio Calabria, Italy, September 1–3, 2022, Proceedings*, Springer, 2023, pp. 437–451, doi: 10.1007/978-3-031-24801-6_31.

[43] V.-T. Tran and W.-H. Tsai, Acoustic-based emergency vehicle detection using convolutional neural networks, *IEEE Access* **8** (2020), 75702–75713, doi: 10.1109/ACCESS.2020.2988986.

[44] V.-T. Tran and W.-H. Tsai, Audio-vision emergency vehicle detection, *IEEE Sensors Journal* **21**(24) (2021), 27905–27917, doi: 10.1109/JSEN.2021.3127893.

[45] S. Kaushik, A. Raman and K.R. Rao, Leveraging Computer Vision for Emergency Vehicle Detection-Implementation and Analysis, in: *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2020, pp. 1–6.

[46] R. Lawson, Emergency vehicle warning system, Google Patents, 2006, US Patent 7,061,402.

[47] L. Chan, K.K. Gaudin, R. Lyons, W.J. Leise, J.A. Nepomuceno, R.C. Shah, E.P. Matesevac III, J.C. Kellett, S.C. Cielocha, J. Myers et al., Systems and methods for providing awareness of emergency vehicles, Google Patents, 2020, US Patent 10,584,518.

[48] D. Agnew, S. Lüke, M. Fischer and D. Krökel, Emergency vehicle detection with digital image sensor, Google Patents, 2017, US Patent 9,576,208.

[49] A.G. Lemmons and S.B. Riley, Emergency vehicle alarm system and method, Google Patents, 2012, US Patent 8,258,979.

[50] L.G. Dill and R.B. Molitor, Siren actuated warning device for automobiles, Google Patents, 1961, US Patent 3,014,199.

[51] B. Stefanov, Wailing siren detecting circuit, Google Patents, 1979, US Patent 4,158,190.

[52] B.E. Warren, Vehicle warning system, Google Patents, 1986, US Patent 4,587,522.

[53] B. Bernstein and G.L. Sohie, Emergency signal warning system, Google Patents, 1990, US Patent 4,956,866.

[54] F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci and P. Desii, A real-time siren detector to improve safety of guide in traffic environment, in: *2008 16th European Signal Processing Conference*, IEEE, 2008, pp. 1–5.

[55] T. Miyazakia, Y. Kitazonoa and M. Shimakawab, Ambulance siren detector using FFT on dsPIC, in: *Proceedings of the 1st IEEE/IIAE International Conference on Intelligent Systems and Image Processing*, 2013, pp. 266–269.

[56] Y. Ebizuka, S. Kato and M. Itami, Detecting approach of emergency vehicles using siren sound processing, in: *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, IEEE, 2019, pp. 4431–4436, doi: 10.1109/ITSC.2019.8917028.

[57] J.-J. Liaw, W.-S. Wang, H.-C. Chu, M.-S. Huang and C.-P. Lu, Recognition of the ambulance siren sound in taiwan by the longest common subsequence, in: *2013 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, 2013, pp. 3825–3828, doi: 10.1109/SMC.2013.653.

[58] S. Kiran and M. Supriya, Siren detection and driver assistance using modified minimum mean square error method, in: *2017 International Conference On Smart Technologies For Smart*

Nation (SmartTechCon), IEEE, 2017, pp. 127–131, doi: 10.1109/SmartTechCon.2017.8358355.

[59] M. Cantarini, A. Brocanelli, L. Gabrielli and S. Squartini, Acoustic features for deep learning-based models for emergency siren detection: An evaluation study, in: *2021 12th International Symposium on Image and Signal Processing and Analysis (ISPA)*, IEEE, 2021, pp. 47–53, doi: 0.1109/ISPA52656.2021.9552140.

[60] L. Marchegiani and P. Newman, Listening for sirens: Locating and classifying acoustic alarms in city scenes, *IEEE Transactions on Intelligent Transportation Systems* **23**(10) (2022), 17087–17096, doi: 10.1109/TITS.2022.3158076.

[61] O. Watkins, N. Pendleton, G. Hotson, C. Fang, R.L. Kwant, W. Gao, D. Ramanan, N. Cebron and B. Browning, Emergency siren detection for autonomous vehicles, Google Patents, 2022, US Patent App. 17/073,680.

[62] X. Kecheng, H. Sun, Q. Luo, W. Wang, Z. Lin, W. Reynolds, F. Liu, J. Hu and J. Miao, Machine learning model to fuse emergency vehicle audio and visual detection, Google Patents, 2022, US Patent App. 17/149,659.

[63] X. Kecheng, H. Sun, Q. Luo, W. Wang, Z. Lin, W. Reynolds, F. Liu, J. Hu and J. Miao, Emergency vehicle audio and visual detection post fusion, Google Patents, 2022, US Patent App. 17/149,638.

[64] M.R. Endsley, Toward a theory of situation awareness in dynamic systems, *Human factors* **37**(1) (1995), 32–64, doi: 10.1518/001872095779049543.

[65] F. Zhou, X.J. Yang and J.C. de Winter, Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving, *IEEE Transactions on Intelligent Transportation Systems* **23**(3) (2021), 2284–2295, doi: 10.1109/TITS.2021.3069776.

[66] M. Doudou, A. Bouabdallah and V. Berge-Cherfaoui, Driver drowsiness measurement technologies: Current research, market solutions, and challenges, *International Journal of Intelligent Transportation Systems Research* **18** (2020), 297–319.

[67] M.Q. Khan and S. Lee, A comprehensive survey of driving monitoring and assistance systems, *Sensors* **19**(11) (2019), 2574, doi: 10.3390/s19112574.

[68] Y. Dong, Z. Hu, K. Uchimura and N. Murayama, Driver inattention monitoring system for intelligent vehicles: A review, *IEEE transactions on intelligent transportation systems* **12**(2) (2010), 596–614, doi: 10.1109/TITS.2010.2092770.

[69] Attention Assist, Available online: www.emercedesbenz.com.

[70] Driver Alert Control, Available online: www.media.volvocars.com.

[71] C.M. Martinez, M. Heucke, F.-Y. Wang, B. Gao and D. Cao, Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey, *IEEE Transactions on Intelligent Transportation Systems* **19**(3) (2017), 666–676, doi: 10.1109/TITS.2017.2706978.

[72] H. Toda, K. Ogawa, K. Ohue and T. Kakegawa, Driver monitoring system for vehicle, Google Patents, 2006, US Patent App. 11/237,978.

[73] L. Barr, S. Popkin, H. Howarth et al., An evaluation of emerging driver fatigue detection measures and technologies, Technical Report, United States. Department of Transportation. Federal Motor Carrier Safety âĂę, 2009.

[74] Driver Attention Warning System, Available online: www.saabnet.com/tsn/press/071102.html.

[75] A. Nabo, Driver attention-dealing with drowsiness and distraction, *Göteborg: IVSS* (2009).

[76] Driver Monitoring System, Available online: www.testdriven.co.uk/lexus-ls-600h/.

[77] K.-C. Hsu and H.-W. Tseng, Accelerating applications using edge tensor processing units, in: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021, pp. 1–14, doi: 10.1145/3458817.3476177.

[78] M. Ngxande, J.-R. Tapamo and M. Burke, Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques, *2017 pattern recognition Association of South Africa and Robotics and mechatronics (PRASA-RobMech)* (2017), 156–161, doi: 10.1109/RoboMech.2017.8261140.

[79] R.A. Naqvi, M. Arsalan, G. Batchuluun, H.S. Yoon and K.R. Park, Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor, *Sensors* **18**(2) (2018), 456, doi: 10.3390/s18020456.

[80] E. Magán, M.P. Sesmero, J.M. Alonso-Weber and A. Sanchis, Driver drowsiness detection by applying deep learning techniques to sequences of images, *Applied Sciences* **12**(3) (2022), 1145, doi: 10.3390/app12031145.

[81] M. Cantarini, L. Gabrielli, A. Mancini, S. Squartini and R. Longo, A3CarScene: an audio-visual dataset for driving scene understanding, *Data in Brief* (2023), 109146, doi: 10.1016/j.dib.2023.109146.

[82] G. Pepe, L. Gabrielli, S. Squartini, L. Cattani and C. Tripodi, Deep learning for individual listening zone, in: *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2020, pp. 1–6, doi: 10.1109/MMSP48831.2020.9287161.

[83] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861* (2017), doi: 10.48550/arXiv.1704.04861.

[84] J.F. Gemmeke, D.P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal and M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 776–780, doi: 10.1109/ICASSP.2017.7952261.

[85] Ministero dei Trasporti, Decreto Ministeriale 17 ottobre 1980 (G.U. n. 310 del 12.11.1980): Modifiche sperimentali delle caratteristiche acustiche dei dispositivi supplementari di allarme da applicare ad autoveicoli e motoveicoli adibiti a servizi antincendi e ad autoambulanze, 1980, Available online: https://croceitalia.it/pdf/dispositivi_supplementari_allarme.pdf (accessed on 28 February 2023).

[86] Y. Kartynnik, A. Ablavatski, I. Grishchenko and M. Grundmann, Real-time facial surface geometry from monocular video on mobile GPUs, *arXiv preprint arXiv:1907.06724* (2019), doi: 10.48550/arXiv.1907.06724.

[87] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[88] X. Glorot and Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[89] J. Smith, S. Serafin, J. Abel and D. Berners, Doppler simulation and the Leslie, in: *Proc. Int. Conf. on Digital Audio Effects, Hamburg*, 2002.

[90] P. Liashchynskyi and P. Liashchynskyi, Grid search, random search, genetic algorithm: a big comparison for NAS, *arXiv preprint arXiv:1912.06059* (2019), doi: 10.48550/arXiv.1912.06059.

[91] D.P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014), doi: 10.48550/arXiv.1412.6980.

[92] S. Hooker, Moving beyond âĂIJalgorithmic bias is a data problemâĂİ, *Patterns* **2**(4) (2021).

[93] J. Snell, K. Swersky and R. Zemel, Prototypical networks for few-shot learning, *Advances in neural information processing systems* **30** (2017), doi: 10.48550/arXiv.1703.05175.

[94] J. Urdiales, D. Martín and J.M. Armingol, An improved deep learning architecture for multi-object tracking systems, *Integrated Computer-Aided Engineering* (2023), 1–14.

[95] K.M.R. Alam, N. Siddique and H. Adeli, A dynamic ensemble learning algorithm for neural networks, *Neural Computing and Applications* **32** (2020), 8675–8690.

[96] D. Li, J. Wu, F. Zhu, T. Chen and Y.D. Wong, Modeling adaptive platoon and reservation-based intersection control for connected and autonomous vehicles employing deep reinforcement learning, *Computer-Aided Civil and Infrastructure Engineering* **38**(10) (2023), 1346–1364.

[97] D.R. Pereira, M.A. Piteri, A.N. Souza, J.P. Papa and H. Adeli, FEMa: A finite element machine for fast learning, *Neural Computing and Applications* **32** (2020), 6393–6404.

[98] H. Shi, Y. Zhou, X. Wang, S. Fu, S. Gong and B. Ran, A deep reinforcement learning-based distributed connected automated vehicle control under communication failure, *Computer-Aided Civil and Infrastructure Engineering* **37**(15) (2022), 2033–2051.

[99] L. Ruiz, S. Díaz, J.M. González and F. Cavas, Improving the competitiveness of aircraft manufacturing automated processes by a deep neural network, *Integrated Computer-Aided Engineering* (2023), 1–12.

[100] A.H. Sham, K. Aktas, D. Rizhinashvili, D. Kuklianov, F. Alisinanoglu, I. Ofodile, C. Ozcinar and G. Anbarjafari, Ethical AI in facial expression analysis: Racial bias, *Signal, Image and Video Processing* **17**(2) (2023), 399–406, doi: 10.1007/s11760-022-02246-8.