Innovative Saliency based Deep Driving Scene Understanding System for Automatic Safety Assessment in Next-Generation Cars

Francesco Rundo
ADG, Central R&D
STMicroelectronics
Catania, Italy
francesco.rundo@st.com

Sabrina Conoci
University of Messina
Messina, Italy
sabrina.conoci@unime.it

Sebastiano Battiato
IPLAB
University of Catania
Catania, Italy
battiato@dm.unict.it

Francesca Trenta
IPLAB
University of Catania
Catania, Italy
francesca.trenta@unict.it

Concetto Spampinato
Perceve Lab
University of Catania
Catania, Italy
cspampinato@dieii.unict.it

Abstract—Visual saliency is the human attention mechanism that encodes such visio-sensing information to extract features from the observation scene. In the last few years, visual saliency estimation has received significant research interests in the automotive field. While driving the vehicle, the car driver focuses on specific objects rather than others by deterministic brain-driven saliency mechanisms inherent perceptual activity. In this study, we propose an intelligent system that combines a driver’s drowsiness detector with a saliency-based scene understanding pipeline. Specifically, we implemented ad-hoc 3D pre-trained Semantic Segmentation Deep Network to process the frames captured by automotive-grade camera device placed outside the car. We used an embedded platform based on the STA1295 core (ARM A7 Dual-Cores) with a hardware accelerator for hosting the proposed pipeline. Besides, we monitor the car driver’s drowsiness by using an innovative bio-sensor installed on the steering wheel, to collect the PhotoPlethysmoGraphy (PPG) signal. Ad-hoc 1D Temporal Deep Convolutional Network has been designed to classify the collected PPG time-series in order to assess the driver’s attention level. Finally, we compare the detected car driver’s attention level with corresponding saliency-based scene classification in order to assess the overall safety level. Experimental results confirm the effectiveness of the proposed pipeline.

Index Terms—Drowsiness, Deep learning, D-CNN, Deep-LSTM, PPG (PhotoPlethysmography)

I. INTRODUCTION

Drowsiness is a physiological state characterized by a low level of consciousness and difficulty in maintaining the wakeful state, which could lead to serious road traffic accidents. For these reasons, several studies have investigated the ability to monitor the driver’s attention level. These researches pointed out the relationship between the level of human attention (or Drowsiness Monitoring) and Heart Rate Variability (HRV) index. HRV measures heart activity over a beat-to-beat interval to provide useful information concerning the drowsiness level [1]. More specifically, HRV is a non-invasive method to retrieve the Autonomous Nervous System (ANS) activity [1]. The link between the subject’s attention level and the ANS activity is well known in scientific literature [1]–[3]. Therefore, from the analysis of the HRV we can indirectly trace the level of attention of the subject. However, the driver’s attention level must be related to the current driving scenario. For instance, driving with low traffic and low speed is a low-demanding task than a driving scenario that includes risky maneuvers (over-taking, lane changes, etc.). In this study, the authors propose an algorithm which combines saliency analysis (for driving scene understating) and physiological car driver monitoring (for retrieving the attention level). Saliency analysis has been widely used in many applications, including automotive. It allows the identification of significant parts in a video or image [2], [3]. The driving scene analysis and drowsiness monitoring algorithms require a high computational load. By means of the proposed saliency based approach we can characterize the driving scene from the analysis of the salient elements (cars, pedestrians, etc.), reducing the computational load and improving the performance of the algorithm. The remainder of this paper is structured as follows. In Section II, previous related works exposing physiological signal analysis employed for automatic drowsiness detection are described and reviewed. The description of blocks composing our pipelines is presented in Section III. Section IV explains our experimental setup providing details about the dataset and the hardware devices used to carry out our pipeline. The conclusions and future works are presented in Section V.

II. RELATED WORKS

Several studies have attempted to provide reliable solutions in order to monitor the car driver’s attention level alongside...
analyzing driving scenarios. In [4], the authors implemented a pipeline to detect driver drowsiness by collecting the ElectroCardioGraphy (ECG) signal from which to measure alterations in HRV. By performing ECG signal stabilization and using a classical linear discriminant analysis, the overall pipeline effectively discriminates the drowsy status from the wakeful one. One of the main issues concerning the drowsiness detection is the ECG sampling. The minimal setup for sampling a subject’s ECG signal requires the use of at least three electrodes in contact with the human skin. This configuration is known as Einthoven’s Triangle [5]. Consequently, the driver has to put the hands over electrodes, placed in the steering wheel, to collect ECG signal. The third electrode is usually located in the driver’s seat, where the acquisition is more sensitive to noise, affecting the robustness of the HRV signal, which comes out from the so collected ECG signal. [5]. In this respect, many authors have developed methods based on the PhotoPlethysmGraphic (PPG) signal analysis. The PPG signal requires a single sampling contact point making it more suitable than the ECG [6]. In [6], the authors analyze parasympathetic nervous activity by determining HRV from the PPG signal with the aim of classifying the car driver’s fatigue level. The resulting outcomes confirmed the robustness of the proposed approach. In [7], the authors proposed a reliable indicator coming from PPG signal. In [8], [9], the authors reported promising results by using algorithms based on Pulse Rate Variability (PRV) data processing. The authors demonstrated that PRV is a useful measurement for the ANS analysis, which helps in assessing the drowsiness of the subject. Ryu et al. [10] used a red organic light-emitting diodes (OLEDs) and organic photodiodes (OPDs) for collecting the photoplethysmographic (PPG) signal. The benchmark evaluation confirmed the effectiveness of the proposed flexible PPG sensor in detecting the drowsiness level, reporting a higher performance in terms of accuracy than the standard PPG probe.

In the last few years, several studies have pointed out the effectiveness of Deep Learning approaches for estimating the drowsiness of a subject from bio-signals and image analysis. In [11], [12], the authors implemented methods based on image processing to track the driver’s face, gaze, and emotions to assess his inattentiveness and drowsiness. Despite being promising, the image processing performance is affected by the condition of the in-vehicle environment scenario (e.g., light condition, occlusions, etc.). As introduced, the continuing evolution of technologies in the automotive industry has led to the development of promising Drowsiness Detection systems such as the pipeline proposed in [13]. Moreover, researchers have developed various Deep Learning algorithms to evaluate the fatigue level using “data fusion,” processing both the visual and physiological data of the car driver. Altun and Celenk [14] presented a vision-based driver assistance system for scene awareness through visual saliency analysis. The results confirmed that the overall system improves the driver’s situational awareness (SA) by enabling adaptive road surface classification. In [15], the authors assess the driver’s drowsiness by performing eye-tracking during driving. To carry out experiments, the authors collected 40 subjects, including non-drivers and experienced ones. Their findings report that the driver presents full attention to the road’s vanishing points. The authors suggested that the vanishing point of the road can be considered a valuable resource to estimate the road area in a traffic saliency detection mode. The authors proposed a solution that takes the drowsiness monitoring system into account to extract information about the driving scenario [15].

III. METHODS AND MATERIALS

In this section, we provide more details about the proposed pipeline. In Fig. 1, the overall scheme of the proposed solution.
A. The Drowsiness Monitoring System

In this work, we implemented a system to acquire the car driver’s PPG signal to assess and monitor the related attention level. As anticipated, the PPG signal is a non-invasive method for the analysis of the heart pulse rate. Indeed, PPG signal allows to monitor both heart pulse and respiratory rate along with vascular and cardiac disorders [6]. Specifically, the PPG waveform consists of two components: a pulsatile ‘AC’ physiological signal that depends on the cardiac-synchronous changes in the blood volume, and a ‘DC’ component that shows minor changes due to the respiration and thermoregulation process [6]. When pumping blood to the periphery, the heart produces a specific pressure that distends the arteries and arterioles in the subcutaneous tissue. For this reason, we used a device composed of a light-emitter and a detector in contact with the subject skin to sample PPG signal. As heart pressure pulse can be seen as a peak in the PPG waveform, the changes in volume are detected by illuminating the skin and then measuring the amount of back-scattered light [6]. We used the PPG sampling device composed of the Silicon Photomultiplier sensor [16]–[19]. The proposed PPG probes show an array detector device, called Silicon Photomultipliers (SiPMs) [17], with a total area of $4.0 \times 4.5 \, \text{mm}^2$ and 4871 square microcells with 60 $\mu$m pitch. The devices have a geometrical fill factor of 67.4% and are packaged in a surface mount housing (SMD) with about $5.1 \times 5.1 \, \text{mm}^2$ total area [20]. We propose a Pixelteq dichroic bandpass filter with a pass-band centered at about 540 nm with a Full Width at Half Maximum (FWHM) of 70 nm and an optical transmission higher than 90 – 95% in the pass-band range was glued on the SMD package by using a Locite 352TM adhesive.

The SiPM has a maximum detection efficiency of about 30% at 565 nm and a PDE of about 27.5% at 540 nm (central wavelength in the filter pass-band). As described, the PPG detector is composed of a light emitter in combination with a detector based on SiPM technology. The OSLAM LT M673 LEDs were used by SMD package and InGan technology [20]. The used LEDs devices are characterized by an area of $2.3 \times 1.5 \, \text{mm}^2$ with a 120° angle view, a spectral bandwidth of 33 nm and a lower power emission (mW) in the standard range. To optimize the use of the PPG probe, we designed a printed circuit board (PCB) consisting of a user-interface based on NI (National Instruments) instrumentation. The PCB comprises a 4V portable battery, a power management circuits, a conditioning circuit for output SiPMs signals, along with several USB connectors for PPG probes and related SMA output connectors.

In [16]–[19], we reported a further details about the hardware used to acquire PPG signal. The PPG Sensor Probe includes the SiPM sensor and the mentioned LEDs. The power consumption of the SiPM device is managed by Power management circuit [16]–[18]. As introduced, we placed several PPG probes on the vehicle’s steering wheel, in order to acquire the PPG signal. To collect the physiological signal, it is enough that only one hand of the driver is placed over the PPG sensor. Once acquiring the PPG raw signal, the NI device processes the collected data through internal 24 bits ADC. In addition, the NI device includes a Windows-based operating system with a LabView software framework [6]. We implemented a LabView algorithm to perform a preliminary process to filter the PPG raw data. In this respect, we defined ad-hoc FIR (Finite Impulse Response) filtering both low-pass and high-pass band. Furthermore, we perform such pre-processing operation by computing the first and second derivatives of the collected PPG signal to evaluate the min and max value for each waveform. Finally, we rendered the PPG signal on the monitor attached to the NI device. In [6] and [19], we provide more details about the used NI device and LabView software framework. The end solution of the proposed pipeline will be ported over STA1295 embedded platform replacing the NI device used for development purpose only [21], [22]. For the purpose of this study, we developed an effective pipeline to process PPG raw data by using MATLAB framework, as reported in the validation session. In Table I and II, we reported details about the hyper-filtering setup used to process the sampled raw PPG signal [17]–[19]. Hyper-filtering approach is an innovative method (firstly introduced by the authors) similar to hyper-spectral for imaging but applied to a signal [19]. With hyper-filtering the source PPG signal will be filtered and processed with different frequency setup (see Table I and II) in order to get discriminative patterns related to drowsiness level of the analyzed subject. More details in [19]. Once we have collected the hyper-filtered PPG signal patterns [17], ad-hoc 1D Temporal Dilated Convolutional Neural Network (1D-CNN) will be used to classify the so generated patterns. The proposed Temporal Convolutional Network consists of a Dilated Causal Convolution layer that processes the time steps of a given sequence. Indeed, the term “causal” means that the activations computed at time $t$ derive only from inputs from time $t − 1$. The architecture of our proposed network is composed of multiple residual blocks, each containing two dilated causal convolution layers with the same dilation factor, followed by normalization, ReLU activation, and spatial dropout layers. More specifically, we implemented a 1D-CNN

### Table I

**HYPER LOW-PASS FILTERING SETUP (IN HZ)**

<table>
<thead>
<tr>
<th>F</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>f7</th>
<th>f8</th>
<th>f9</th>
<th>f10</th>
<th>f11</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP</td>
<td>0.5</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>LP</td>
<td>0</td>
<td>1.4</td>
<td>2.9</td>
<td>2.5</td>
<td>3.8</td>
<td>3.9</td>
<td>4</td>
<td>4.5</td>
<td>5</td>
<td>5.3</td>
<td>6.9</td>
</tr>
</tbody>
</table>

### Table II

**HYPER HIGH-PASS FILTERING SETUP (IN HZ)**

<table>
<thead>
<tr>
<th>F</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>f7</th>
<th>f8</th>
<th>f9</th>
<th>f10</th>
<th>f11</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP</td>
<td>0.5</td>
<td>1.2</td>
<td>2.6</td>
<td>2.7</td>
<td>3.3</td>
<td>3.5</td>
<td>4</td>
<td>4.4</td>
<td>5</td>
<td>5.7</td>
<td>6.4</td>
</tr>
<tr>
<td>LP</td>
<td>7</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>
of 12 blocks, including a downstream softmax layer. Each block structure consists of a dilated convolution layer with a $3 \times 3$ kernel filters, a spatial dropout layer, followed by another dilated convolution layer, a ReLU layer, and a final spatial dropout. In this context, we set the initial dilation size to 2, increasing it at each block. Finally, a softmax layer completes the proposed pipeline. The purpose of this study is to predict the car driver’s drowsiness level by processing the hyper-filtered signal-patterns of the acquired PPG signal through the designed 1D-CNN architecture. The output of the network is a scalar number in the range [0,1] that indicates the attention level of the car driver, discriminating drowsy (0) and wakeful drivers (1). Our findings suggested that it is indeed possible to estimate the drowsiness of the driver with high accuracy. In order to provide efficient support while driving, we ported the proposed 1D Deep CNN backbone over an automotive-grade STA1295 Accordo5 embedded platform [22].

**B. The Video Saliency Scene Understanding Block**

In this section, we provided architectural details of the proposed Deep Network suitable to perform saliency analysis. In the first step, we captured video frames representing the driving scenario with an automotive-grade camera device. We also implemented ad-hoc 3Dto2D Semantic Segmentation Fully Convolutional Network (SS-FCN) to process the collected video frames. The proposed segmentation model follows the encoder/decoder architecture to process the driving scene. The resulting saliency map represents the most salient object in the acquired driving scene. Our proposed SS-FCN architecture is mainly composed of two parts: an encoder and a decoder. We designed the encoder block (3D Enc Net) to extract spatiotemporal features. It is composed of 5 blocks. Convolutional layers are gathered in 2 blocks of 2 layers for the first two blocks. The kernel size of each separable convolutional layer is $3 \times 3 \times 3$. Each block involves a sequence of two $3 \times 3 \times 3$ convolution operations, which is followed by a batch normalization, ReLU layer and a max pooling operation with a pooling size of $1 \times 2 \times 2$. This sequence is repeated two times. Then again, a sequence of $3 \times 3 \times 3$ convolutional layers is performed. Similar to the previous blocks, the succession of two convolution operations is repeated for the remaining three blocks, followed by batch normalization, another convolutional layer with $3 \times 3 \times 3$ kernel, batch normalization and ReLU with a downstream $1 \times 2 \times 2$ max-pooling layer. The decoder backbone (2D Dec Net) composed of 5 blocks and up-sampling layers is designed to decode the visual features. It also includes 2D convolutional layers having $3 \times 3$ kernel. After each convolutional layer, we employ batch normalization layers and Rectified Linear Unit (ReLU). The residual connections are added through convolutional blocks. The decoder part is implemented to adjust the size of feature maps by up-sampling operations. Finally, the output of the proposed SS-FCN architecture is the feature map of the acquired video frame related to the segmented area of the most salient object in the driving scenario. In semantic segmentation, usually the fixation points comprise the most salient objects of the entire image. In Fig. 2 and Fig. 3 we show such instances of the SS-FCN output reporting the driving scenario saliency maps and corresponding training loss dynamic of the network. We performed the training and testing of the SS-FCN model against the DHF1K dataset [21]. The model was evaluated using the Area Under the Curve (AUC), Similarity, Correlation Coefficient and Normalized Scanpath Saliency performance indexes. The proposed architecture achieves acceptable performance (AUC: 0.885; Similarity: 0.355; Correlation Coefficient: 0.455; Normalized Scanpath Saliency: 2.564) compared with similar architectures [21]. Despite being promising solutions, the architectures reporting the better performances require a high computational cost alongside the use of sophisticated hardware. On the contrary, our proposed pipeline shows low workload without involving specific hardware accelerations [22].

**C. The Driver Attention Analyzer**

This block estimates the overall driver’s attention level, coming from the PPG signal analysis combined with the output (i.e., saliency map) of the proposed SS-FCN block. Specifically, this block verifies that the level of attention is coherent with the ”meaning” of the corresponding salience map i.e. with driving scenario dynamic. A static saliency map reflects a low dynamic in the driving scenario, involving a low-
level attention, whereas an increasing variation of the saliency maps require a car driver’s full attention. Formally, let be $S_f(x, y, t)$ the saliency map and $O_S$ the output of the 1D-CNN, the designed block analyses the driver attention by using the following equation:

$$V(S(x, y, t)) = \begin{cases} \frac{\partial S_f(x, y, t)}{\partial t} \leq \vartheta & O_S \leq \varphi \\ \frac{\partial S_f(x, y, t)}{\partial t} > \vartheta & O_S > \varphi \end{cases}$$ (1)

Using Eq. (1), the Driver Attention Block checks the dynamic changes of the saliency map $S_f(x, y, t)$ through a given threshold $\vartheta$. We set the ad-hoc threshold $\varphi$ to verify the level of attention obtained by the Drowsiness Monitoring systems. Both thresholds are determined during the training phase by means of a heuristic calibration i.e. by choosing a setup that maximizes the performance of the overall pipeline during the learning.

IV. EXPERIMENTAL RESULTS

In this section, we first introduce the dataset and then the training details and the main findings of our study. We evaluated the implemented pipeline on the DH1FK dataset [21]. We also collected video sequences of several driving scenarios using a camera with a 2.3 Mpx resolution and a maximum framerate of 60 fps. Under the supervision of a group of physiologists, the recruited subjects show different levels of attention while collecting the corresponding EEG signal [18] to confirm their physiological findings. To carry out our experiments, we involved a total of 43 subjects in the age group of 21-70 years. We acquired the PPG signal of the subject with a sampling frequency of 1 KHz, considering a 5-minute interval. Moreover, we used 70% of the all acquired PPG time-series and video driving scene frames for training, the remaining 30% for testing and validating. To sum up, the overall proposed pipeline assesses the driver’s attention level by combining the PPG signal and the saliency map. Specifically, the Driver Attention Analyzer block compares the level of attention required by the driving scenario with the output of the Drowsiness Monitoring System. The system alerts the driver with an acoustic signal if the driver’s vigilance level is lower with respect to the current driving scenario. In this context, we defined two range of values to classify the attention level. A value (output of the 1D-CNN based Drowsiness Monitoring System) in the range of $0 - 0.6$ indicates a medium-low attention level. On the contrary, a high attention level comprises the values from $0.61$ to $1$. Finally, we set an ad-hoc normalized threshold $\vartheta$ (0.45) to define a static scene-based saliency map (see Eqs. (1)). The driving scene is considered as ‘dynamic’ if the values of the normalized saliency map gradient are greater than 0.45 (requiring a high level of attention). For a normalized saliency map gradient lower than 0.45 (requiring a low level of attention) the driving scene is considered ‘static’. In Fig.4, two instances of the driving scenarios are reported.
The encouraging results have pointed out the possibility to assess the driver’s drowsiness to preserve driving safety. The main benefit of the proposed method is that it does not require frequency domain analysis, compared to other promising approaches based on the HRV analysis [18]. Furthermore, the proposed method requires the use of PPG signal since it can be easily sampled from a wide range of sensors placed on the steering wheel. To evaluate the car driver’s level of attention retrieved from driver’s PPG signal analysis, we designed a fully convolutional deep network that determines a saliency map of the driving scene. The results have highlighted the effectiveness of the proposed pipeline in assessing the driver’s drowsiness level required by the driving scenario. Specifically, the system compares the level of attention reconstructed from driver’s PPG signal with the level of attention properly calibrated to the driving scenario retrieved from saliency analysis alerting the driver if a risk occurs i.e. if there is a mismatch between the so detected attention levels. More investigation are underway to improve the saliency analysis by extending the dataset including further driving scenarios in different domains. The overall pipeline is currently being ported to an automotive designed board based on the SoC STA1295 ACCORDO 5 processor produced by STMicroelectronics [22]. In the future, we will try to include a more robust domain adaptation method, including supervised and unsupervised approaches already tested in other pipelines [23]–[26].

ACKNOWLEDGMENT

The used dataset was collected under the clinical study Ethical Committee CT1 authorization n.113 / 2018 / PO. This research was funded by the National Funded Program 2014-2020 under grant agreement n. 1733, (ADAS + Project).