System for detecting driver's drowsiness, fatigue and inattention

Conference Paper · November 2021

DOI: 10.1109/TELFOR52709.2021.9653243

4 authors, including:

Aleksa Arsic
Synchrotek
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Velibor Ilic
The Institute for Artificial Intelligence Research and Development of Serbia
58 PUBLICATIONS **223** CITATIONS
SEE PROFILE

Bogdan Pavkovic
RT-RK Computer Based Systems
48 PUBLICATIONS **348** CITATIONS
SEE PROFILE

System for detecting driver's drowsiness, fatigue and inattention

Aleksa Arsić, Velibor Ilić, Bogdan Pavković, Dragan Samardžija

Abstract – ***High percentage of road accidents with even fatal outcomes are caused by the driver's drowsiness and other factors that can be controlled by the driver himself. Thus, for modern Advanced Driving Assistance Systems (ADAS) it would be of great use to implement a reliable system for detecting driver's drowsiness, fatigue, and inattention. One way this could be achieved is by using convolutional neural networks (CNN) and machine learning (ML) principles. In this paper, we present academic research on the topic which is based on three CNN's used for monitoring the driver with a possibility to dispatch notification when concluded that his state of attention is not suitable for operating a motorized vehicle. Each of the three CNN's processes different parts of the image from the inside of the vehicle and are connected in a series in a way that the outputs of the previous are used as the inputs for the next CNN model in the series.***

Keywords **— machine learning, attention, deep neural networks, convolutional neural networks (CNN), advanced driver assistance systems (ADAS)**

I. Introduction

In the Republic of Serbia in 2019 and similarly, in many other European countries, 25% of road accidents with fatal outcomes were caused by driver's drowsiness, fatigue, inattention, and other similar psychophysical conditions [1]. Such accidents make the third largest cause of road accidents with fatal outcomes in the country. Hence, the Driver Monitoring System would be of great use in modern vehicles since it can lower these numbers and increase passengers', other traffic participants, and environmental safety.

In this paper, we propose a driver monitoring system based on three Deep Convolutional Neural Networks (CNN). The first CNN, face detection network (FDN), as input uses images captured from the camera, this network is trained to determine the frame that contains the driver's face. The second CNN network for selecting the region of eyes (REN), as input uses extracted image with face, and this network is trained to determine the frame that contains the eyes of the driver. The third CNN network, a network for monitoring single eye (SEN), as input uses extracted image with eyes, and this network monitors individual eyes. After processing the image through these three neural networks, it is assessed whether the driver lacks attention. By determining a lack of attention, an option of giving a sound warning is integrated to avoid potentially dangerous or even fatal situations. For experimental purposes, we have trained another experimental CNN to justify the three-network approach and consider the possible exclusion of the REN model. The paper is organized as follows. Section II shows related work on the topic and the main difference between those approaches and our proposed approach. Section III contains a description of used CNN models and the structure of software for monitoring the driver's attention. Section IV contains a description of generating needed datasets for the training and testing of CNNs. After that, section V represents the achieved result and comparison between the three-network approach and experimental CNN. Finally, the last section contains conclusions, and we summarize the presented work and give directions for future work. We included a web link to the short video demonstration at the end of this paper.

Aleksa Arsić, RT-RK, Institute for Computer Based Systems, Novi Sad, Serbia, (e-mail: aleksa.arsic@rt-rk.com).

Velibor Ilić, RT-RK, Institute for Computer Based Systems, Novi Sad, Serbia, (e-mail: velibor.ilic@rt-rk.com).

Bogdan Pavković, RT-RK, Institute for Computer Based Systems, Novi Sad, Serbia, (e-mail: bogdan.pavkovic@rt-rk.com).

Dragan Samardžija, RT-RK, Institute for Computer Based Systems, Novi Sad, Serbia, (e-mail: dragan.samardzija@rt-rk.com).

II. Related work

Systems that monitor driver's attention levels are one of the main components of Advanced Driver Assistance Systems (ADAS) and there have been numerous research papers on this subject. Some of them, older ones, are based on some of the classic, more traditional approaches in computer vision (e.g. Kalman filter, RANSAC, POSIT) [2][3], while some of them took a different approach using modern approaches with Machine Learning (ML) principals and Deep Neural Networks (DNN). [4][5]

S. Park et al [5] proposed a method for detecting driver's drowsiness state using three CNN's where they have adopted pre-trained AlexNet for extracting image features, VGG-FaceNet for extracting facial feature representation, and FlowImageNet for behavior feature representation. After which outputs of the three networks were used and combined to create a single prediction on the drowsiness state. The drowsy state is detected from the whole face, not facial features, such as eyes that are used in our paper. On the other hand, R. Jabbar et al [4] presented a method for detecting driver's drowsiness in real-time based on Multilayer Perceptron Classifier (MPC) with three hidden layers. In their proposition 68 landmark coordinates are extracted from the images using the external library to map the facial structures of the face and are used for training the used MPC model.

Our proposition does not include any pre-trained CNN model architectures, thus, we have designed our own CNN architectures and trained them with our own generated training datasets. These models are used in a similar way, where each model is capable of extracting coordinates of interest starting from larger image elements (drivers face) to smaller ones (points of interest on drivers eyes) and where outputs of the previous are used as the inputs for the next CNN model in the series. Thus, partially mocking the humans brain cognitive behavior of focused and selective attention. [6]

978-1-6654-2585-8/21/$31.00 ©2021 IEEE

III. DRIVER ATTENTION SOFTWARE

The software uses three CNNs whose architecture is similar to one another (Fig. 1). The inputs of each CNN are grey-scaled images that have 100×100 pixels dimension whose values are normalized to the range of [0, 1], i.e. images must be preprocessed before being propagated through CNN models. The dimensions of the final fully connected layer are 12, 10, and 15 fully connected nodes for the FDN, REN, and the SEN model respectfully.

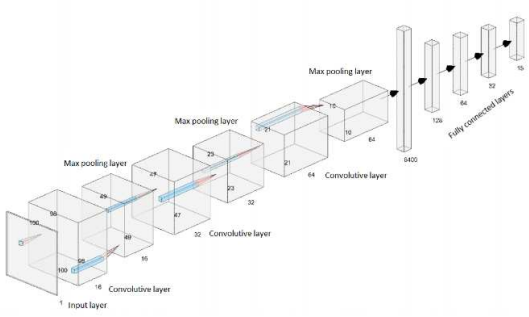

Fig. 1 Architecture of used CNN's

The source of images was a simple webcam that can capture RGB frames with a resolution of 640×480 pixels. The preprocessed frame is propagated through the FDN model and if the predictions confirm that the driver's face is detected, denormalized predictions of the center of the face are used to cut the face from the whole frame with dimensions of 300×200. After preprocessing the face frame is propagated through the REN model. If at least one eye is found in the predictions of the REN network, the eye frame is extracted from the face frame in a dimension of 100×100 pixels. Preprocessed eye frame is propagated through the SEN model.

As several papers suggest, the potential drowsy state of the person can be determined with great confidence solely from analyzing his eyes. [7][8] Whereas one of the most commonly used methods is the monitoring of percentage eye openness tracking (PERCLOS). [9]

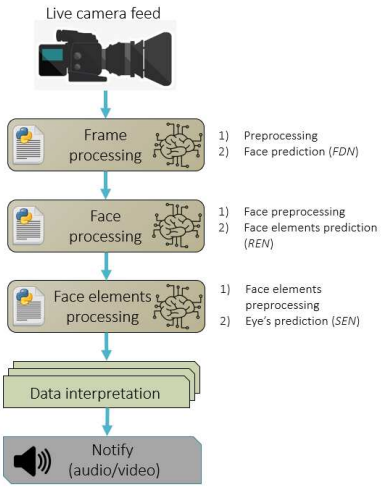

Fig. 2 Driver attention software dataflow

The input size of all three models is the same, where FDN and REN models do not preserve the aspect ratio of height and width of the input frame. It was possible to preserve the aspect ratio of the input images which would result in three completely different architectures of our CNNs. The described approach is mainly taken because of common practice when designing CNNs to have a square-shaped input layer. [10] The data flow of Driver Attention Software is presented in Fig. 2.

Alongside points of interest, every CNN model can predict several other states of the driver's face and face elements, as shown in Fig. 3.

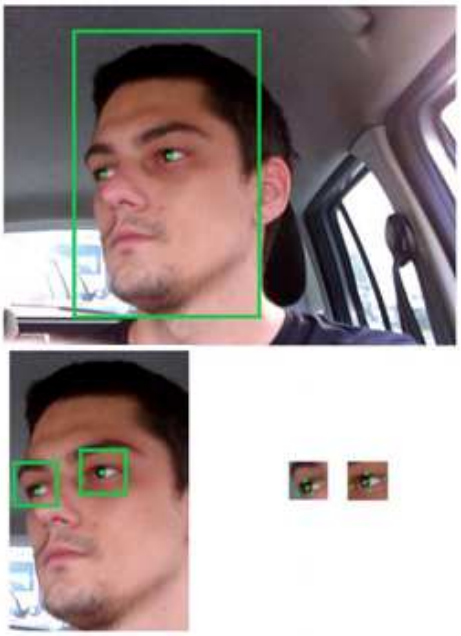

Fig. 3 Points of interest predicted by all three CNN models

Output tensor of the FDN model consists of: $\boldsymbol{p_f}$ - the probability that the driver's face is found on the current frame; $(\boldsymbol{f_x}, \boldsymbol{f_y})$ - x and y coordinates of the center of the driver's face; $(\boldsymbol{e_{lx}}, \boldsymbol{e_{ly}})$ - x and y coordinates of the driver's left eye; $(\boldsymbol{e_{rx}}, \boldsymbol{e_{ry}})$ - x and y coordinates of the driver's right eye, as eye's coordinates are used for later comparison between the results of the FDN and REN model; $(\boldsymbol{p_l}, \boldsymbol{p_r}, \boldsymbol{p_u}, \boldsymbol{p_d})$ - the probability that the driver's face is oriented to the left, right, up, or down, respectively; $\boldsymbol{f_w}$ - width estimate of the driver's face.

Output tensor of the REN model consists of: $(\boldsymbol{p_{le}}, \boldsymbol{p_{re}})$ - the probability that the left and right eye, respectively, are present on the current face frame; $(\boldsymbol{e_{lx}}, \boldsymbol{e_{ly}})$ - x and y coordinates of the driver's left eye; $(\boldsymbol{e_{rx}}, \boldsymbol{e_{ry}})$ - x and y coordinates of the driver's right eye; $(\boldsymbol{p_l}, \boldsymbol{p_r}, \boldsymbol{p_u}, \boldsymbol{p_d})$ - the probability that the driver's gaze is oriented left, right, up, or down, respectively.

Output tensor of the SEN model consists of: $\boldsymbol{p_{ec}}$ - probability that the observed eye is closed; $(\boldsymbol{c_{ux}}, \boldsymbol{c_{uy}})$ - x and y coordinates of the upper center point of the eye; $(\boldsymbol{c_x}, \boldsymbol{c_y})$ - x and y coordinates of the pupil of the eye; $(\boldsymbol{c_{dx}}, \boldsymbol{c_{dy}})$ - x and y coordinates of the lower center point of the eye; $(\boldsymbol{l_x}, \boldsymbol{l_y})$ - the x and y coordinates of the left point of the observed eye; $(\boldsymbol{r_x}, \boldsymbol{r_y})$ - x and y coordinates of the right point of the observed eye; $(\boldsymbol{p_l}, \boldsymbol{p_r}, \boldsymbol{p_u}, \boldsymbol{p_d})$ - the

probability that the driver's gaze is oriented left, right, up, or down, respectively.

Based on CNN's outputs which represent the probability that the driver's head has an angle and the possibility that the eye is closed, the decision of alerting the driver about the absence of attention is made. The decision is made throughout the time interval of 2.5 seconds if any of the following criteria are met:

- FDN model detects 40 frames in which the driver's face has an angle,
- SEN model detects 35 frames in which the observed eye is closed (in the case where both eyes are found and observed, it is sufficient that just one meets these criteria).

Values, 40 for the FDN model and 35 for the SEN model are experimentally taken for the criteria as for those values system gave the best results of the several other values tested.

IV. Generating Datasets

The process of generating our own datasets can be divided into several sections: Individual frames being captured and saved from the video source that are manually labeled and used in the training of the FDN model. This dataset contains numerous images with drivers' face being inside captured frame (partially or whole) and images without the driver in them; Using FDN model predictions from whole frames, face frames are cut, manually labeled, and used to train the REN model; Using predictions of the REN model, face elements (eyes) are cut from face frames, manually labeled, and used to train the SEN model.

Final datasets consist of around ten thousand images for the FDN model and about five thousand images for both REN and SEN models. Datasets contain images of eight volunteers, five males, and three females. The datasets included videos of different distances from the camera, as well as videos from both the inside and outside of a parked vehicle. They were instructed to behave naturally as if they were in a real-life driving situation. Every image was manually labeled using self-implemented labeling software.

V. Results

The testing accuracy of all three CNN models for FDN, REN, and SEN models are 90.79%, 86.68%, and 72.65%, respectively. The final evaluation of the CNN model is performed over the test data set. Test datasets contain ~ 10% of the total datasets used to train each of the CNN models. The test set for the FDN model consists of around 1000 images while test sets for the REN and SEN models consist of about 500 images each. Output predictions are considered correct if the percentage difference between the expected value and the prediction value is less than 10% when it comes to the parameters that represent the coordinates of the points of interest. In other cases, when the prediction of the CNN model represents the probability, we assume that the prediction is correct if the prediction value exceeds the threshold of 0.5 when the expected value is equal to 1 and less than 0.5 when the expected value is equal to 0.

Simple Moving Average (SMA) with a window of fifteen past frames is used on predictions of every CNN model. The use of SMA achieves greater error tolerance of the whole system. Even if the CNN models make a mistake in a few frames, it will not impact the result of the overall assessment of the driver's attention state.

FDN model and REN model were both trained to predict the central points of the driver's eyes. Based on the obtained results we consider the REN model exclusion possibility, i.e., observation is performed to see if the eyes are too small elements within the whole frame to be found with satisfactory precision by the FDN model.

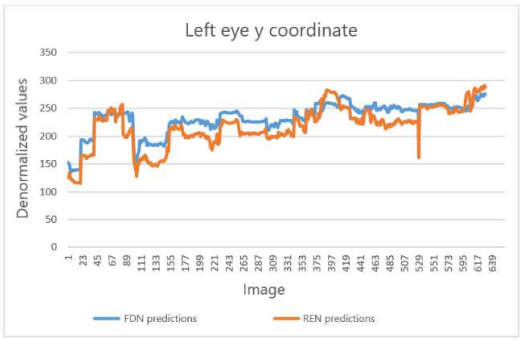

Fig. 4 Left eye y coordinate predictions of the FDN and REN model

In Fig. 4 we present predictions of the FDN model and REN model of the driver's left eye *y* position. It would seem there was no need to introduce another CNN model into the system. The REN model is integrated as one of the system parts, and based on the empirical results, it is known that these results are satisfactory accurate. The difference between predictions of the FDN and the REN model are in the range from 20 to 40 pixels, which is an unacceptable and extremely large error. A conclusion is reached that the FDN model is not capable of reliably finding the central points of the left eye, i.e., it made sense to introduce the second, REN model into the system. Similar results were obtained when analyzing the right eye predictions of the mentioned models.

As mentioned, we have trained another CNN with the intent to confirm discussed problem. Train dataset for this new CNN was around 500 images from whom around 50 images were extracted for the test data set. This new CNN could detect coordinates of the points of interest on the driver's eyes along with the position of the driver's face on the whole frame.

With the fixed position of the camera and in respect of possible minimal and maximal distance of the driver from the camera, the eyes will not represent more than ~9% area of the whole frame. Those dimensions of the driver's eyes should be considered too small for one CNN to accurately predict points of interest and decide if they are closed or not. Given that the SEN model input was an image where the eye is taking a great portion of the frame area, we can tolerate greater error from the output predictions of the SEN model.

In Fig. 5a, the green point represents the accurate, manually labeled, center of the eye, the blue point

represents the prediction of the experimental CNN with an error margin of 3% and the red point represents the prediction of the experimental CNN with the error margin of 6%. Even with the error margin as small as 3%, the predicted center of the eye is greatly displaced from the real center, thus not accurate.

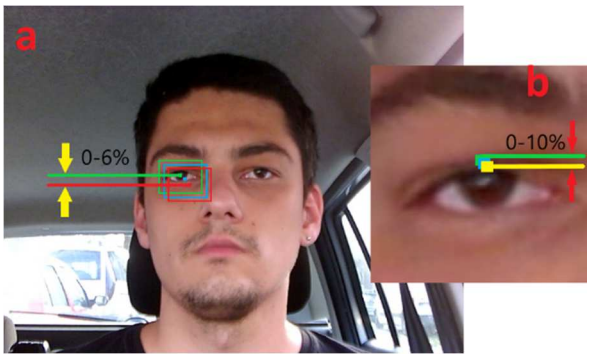

Fig. 5 (a) Different error margins of the new experimental CNN model (b) Different error margins of the SEN model

In Fig. 5b, the green point represents the accurate, labeled, upper center point of the eye, the blue point represents prediction of the SEN model used in our system with an error margin of 5% and the yellow point represents the prediction of the SEN model with the error margin of 10%. Even when the error margin is high as 10% prediction of the upper center point of the eye is borderline accurate.

Results of the predictions over test dataset for both experimental and three-model approach are presented in Table 1, where the accuracy of predictions vary in the range from ~41% to ~86% for the experimental CNN model. With an error margin of 3%, this CNN model is not capable of reliably predicting points of interest on driver's eyes from the whole frame and thus if the eyes are open or closed. From the other perspective, a 3% error margin is too strict of a requirement for predicting a driver's face as it takes a much greater area of the whole frame as opposed to the area taken by the driver's eyes. This was one of the main reasons why we took the approach with three CNN's where each CNN model has greater error tolerance. The three-model approach we took is more complex, however, it gives more reliable and better results than the one CNN approach we discussed.

Table 1 presents the accuracy of the experimental and the three CNN model approach over the test dataset. For the simplicity of the paper only selected data is shown.

Face el.	**Parameter**	**One net. Acc.**[%]	**Three net. Acc.** [%]
Face	*x* coordinate	73.91	91.19
	y coordinate	80.43	98.72
	Width	52.17	82.98
Left Eye	Center up *y*	67.39	93.45
	Center *y*	73.91	96.42
	Center down *y*	71.74	92.26
	Left point *y*	76.09	85.71
	Right point *y*	86.96	90.47
Right eye	Center up *y*	69.57	89.28
	Center *y*	67.39	91.07
	Center down *y*	71.74	90.47
	Left point *y*	71.74	85.12
	Right point *y*	76.09	86.91

Table 1 Accuracy of the experimental and three CNN models over test dataset

VI. Conclusion

In this paper, we proposed an application that relies on three CNNs that were designed from scratch, and in response to that, we generate our datasets that were used in training. As proposed in various researches [4] [5], deep CNNs have been proven efficient and elegant solution to the initial problem, where a lot can be achieved with bigger and more diverse datasets. We showed that they are exceptionally sensitive and with skillful handling can provide results with great precision.

We conclude and point out that this is only academic research and by no means this can be a commercial solution integrated into modern ADAS systems, however, it could be a good starting reference. The factors used to assess the attention levels are far from enough for a complete assessment and there are many more factors that are not considered. For example, driving at night can be another problem in which there should be an adequate solution for a video source that would assure undisturbed processing done by CNN's.

Proposed System Demonstration

A short demonstration in real-life driving situations, as well as in a controlled environment can be found on the following weblink: youtu (.) be (/) ha-RSszCpQQ

Acknowledgments

This paper has been supported by the Ministry of Education, Science and Technological Development through the project no. 451-03-68/2020-14/200156: "Innovative scientific and artistic research from the FTS domain".

References

[1] Републички завод за статистку Републике Србије, "Статистички избештај о стању безбедности саобраћаја у Републици Србији у 2019. години," 2019.

[2] S. Abtahi, B. Hariri, and S. Shirmohammadi, "Driver drowsiness monitoring based on yawning detection," *Conf. Rec. - IEEE Instrum. Meas. Technol. Conf.*, no. July, 2011

[3] L. M. Bergasa, J. M. Buenaposada, , "Analysing driver's attention level using computer vision," *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, pp. 1149–1154, 2008

[4] R. Jabbar, K. Al-Khalifa, M. Kharbeche, "Real-time Driver Drowsiness Detection for Android Application Using Deep Neural Networks Techniques," *Procedia Comput. Sci.*, 2018,

[5] S. K. and Y. S. Park, "Driver drowsiness detection system based on feature representation learning using various deep networks," no. The ACCV Workshop on Driver Drowsiness Detection from Video 2016, Taipei, Taiwan, ROC, 2016.

[6] E. Commodari, "Novice readers: The role of focused, selective, distributed and alternating attention at the first year of the academic curriculum," *Iperception.*, vol. 8, no. 4, 2017.

[7] J. Xu, J. Min, and J. Hu, "Real-time eye tracking for the assessment of driver fatigue," *Healthc. Technol. Lett.*, 2018

[8] Ó. Cobos, J. Munilla, A. M. Barbancho, I. Barbancho, and L. J. Tardón, "Facial activity detection to monitor attention and fatigue," *Proc. Sound Music Comput. Conf.*, 2019.

[9] O. Of and M. Carriers, "PERCLOS : A Valid Psychophysiological Measure of Alertness As Assessed by Psychomotor Vigilance," *October*, 1998,

[10] Raimi Karim, "A compiled visualisation of the common convolutional neural networks," *Towards Data Science*, towardsdatascience.com/illustrated-10-cnn-architectures-95d78ace614d.