

Driver Attention Estimation Based on Temporal Sequence Classification of Distracting Contexts

Raluca Didona Brehar^a, George Cobişan, Attila Fűzes^b and Radu Dănescu^c

Technical University of Cluj-Napoca, Romania

Keywords: Object Detection, Distracted Driving, Driver Monitoring.

Abstract: A framework for distracted driving level or the degree of attention which a driver pays to the act of driving, is presented in this paper. It uses visual based action recognition models applied on color images that capture the driver's face and hands. The proposed approach contains a temporal sequence model that aggregates information from two object detectors which recognize distracting contexts generated by (1) distracting objects that appear in the images such as mobile devices and (2) the face orientation of the driver, the hands and their position with respect to the wheel. The driver's attention score is predicted using the temporal sequence classification model, a long short term memory, that considers time series features computed based on object detection information.

1 INTRODUCTION


According to the United States Department of Transportation¹, "distracted driving is any activity that diverts attention from driving". It includes "talking or texting on the phone, eating and drinking, talking to people in the vehicle, fiddling with the stereo, entertainment or navigation system, anything that takes the attention of the driver away from the task of safe driving. The emergence of mobile devices and other distracting objects that can be used inside the car has led to an increased level of careless driving. According to the World Health Organization² drivers using mobile phones are approximately four times more likely to be involved in a crash than drivers not using a mobile phone because this type of distraction slows down the reaction time and diminishes the drivers' capability to keep the correct lane or to estimate distances with respect to other cars or pedestrians. A direction of contributions for decreasing the number of accidents due to low driver attention resides in the development of systems that can detect and monitor drivers' behaviour while driving and alert them if the attention is


significantly reduced and they are distracted by various actions or objects.


This paper focuses on predicting the driver's level of attention. Even if in general "attention" is a complex cognitive process, an essential process in humans' every day life, for the context of this paper, **the drivers' attention degree (score) is considered to be inversely proportional to the distraction degree of the driver**. If the driver is not distracted then he pays full attention to the driving process, having an attention level of 1. If the driver talks to the phone, then he is distracted and his attention level is smaller.

The main contribution of the paper resides in the development of a temporal multi-view deep learning based framework for driver attention estimation based on the aggregation of multiple single view context classification models that rely on object detection data from yolo-based (Wang et al., 2022) object detectors. They provide face orientation information, hands on wheel information and distracting objects presence. The steps performed in achieving the original contributions consist in

1. The augmentation of a benchmark dataset used for driver attention monitoring with information about distracting contexts that are related on one hand to object's presence (mobile phones, bottles, cans), and on the other hand are related to the position of the hands with respect to the wheel and the face orientation (looking or not looking at the road).

^a  <https://orcid.org/0000-0003-0978-7826>

^b  <https://orcid.org/0000-0002-9330-1819>

^c  <https://orcid.org/0000-0002-4515-8114>

¹ <https://www.nhtsa.gov/risky-driving/distracted-driving>

² <https://www.who.int/news-room/factsheets/detail/road-traffic-injuries>

2. Training of object detection models based on yolov7 (Wang et al., 2022) for inferring (i) location information of distracting objects, (ii) information about hands location with respect to the wheel and (iii) data about the driver's face location.
3. The development of a multi-view information aggregation model that fuses the data provided by the object detectors and forms a feature vector of object locations and their detection scores.
4. The development of a temporal multi-view deep learning based framework for driving attention score prediction that comprises a long short term memory trained on the feature vector provided by the information aggregation model.
5. The augmentation of the benchmark dataset with driver attention score information.

The proposed model is able to identify the contexts that distract the driver such as holding one hand on the steering wheel or both hands on the wheel, the presence of objects used by the driver such as a phone, a bottle or a comb, and whether the driver is looking at the road or at other elements inside the car. This data can be further processed and analysed to obtain information about the driving behaviour, such as detailed statistics related to the attention level reported over a time interval that aim to illustrate all identified actions in accordance with their duration. The developed model provides good results for the benchmark dataset reaching mean squared error of 0.01.

2 RELATED WORK

Most of the work in the field of driver monitoring comprises multi view multimodal approaches that analyse the drivers' face, body posture and actions captured with several types of sensors that provide color or greyscale data, infrared or depth data. From this data, features are extracted over time and the attention level is predicted by means of standard or deep based classification methods.

For example the upper-body pose is used by (Borghini et al., 2017) to monitor the driver's attention level. They propose a regression neural network composed of three independent convolutional nets which are fused by a single fusion layer whose purpose is to determine the upper body pose by depth information. This pose estimation network is completed by a model that reconstructs gray-level face images directly from depth maps. Their work is extended by (Borghini, 2018) that combine deep learning methods and depth maps for head pose estimation and facial

landmark detection for driver attention monitoring. A multi-modal dataset is provided by (Jha et al., 2020) that use a Fi-Cap device that continuously tracks the head movement of the driver for providing annotations for head pose algorithms, RGB cameras and a time-of-flight depth cameras for recording the scenes where the driver performs common secondary activities such as navigation using a smart phone and operating the in-car infotainment system. The solution provides deep learning based approaches for gaze estimation and head pose estimation.

Other approaches are directed towards gaze analysis for driver state estimation. The challenges of such methods reside in the difficulty of a robust gaze estimation due to large head movement. A gaze zone random forest classifier is proposed by (Wang et al., 2017) that use head vectors computed with pose from orthography and eye image features extracted from facial landmarks and 3D face models. A multi-state driver's face monitoring system is proposed by (Hu et al., 2022) that recognize blinking and yawning behaviours, and also use deep learning based architectures for head pose estimation and gaze estimation. Gaze is also used to predict the attention map of the driver. (Rong et al., 2022) integrate an attention prediction module into a pre-trained object detection framework and predict the attention in a grid-based style using ResNet3D and having as input the front and top depth and infrared driver images.

(Muhrrer and Vollrath, 2011) performed a study in order to investigate how different distraction conditions influence the anticipation of events in a car-following scenario, considering also different manoeuvres of a preceding car, in order to generate various anticipations and therefore a different adaptation of the driving behaviour. Additionally, a cognitive and a visual secondary task were introduced.

Recently the AI City Challenge (Naphade et al., 2023), (Naphade et al., 2022) introduced as one of the main tracks the naturalistic driving action recognition having as objective the classification of distracted behaviour activities executed by the driver in a given time frame. To achieve the goals of the challenge several solutions were proposed. (Zhou et al., 2023) describe a solution based on large model fine-tuning based on Vision Transformers combined with a multi-view multi-fold ensemble to produce fine-grained clip-level classification. (Alyahya et al., 2022) propose a temporal driver action localization framework that consists of three stages: (i) preprocessing, which performs driver tracking and video segmentation, (ii) action classification based on SlowFast as an action classifier with Resnet50 as the backbone; and (iii) the temporal action localization. A key point based ap-

proach is described by (Vats and Anastasiu, 2022) that extract complex static and movement-based features for predicting a sequence of key-frame activities. An improved version of multi-scale vision transformer network, which learns a hierarchy of robust representations is employed by (Liang et al., 2022). They also use a sliding-window classification strategy to facilitate temporal localization of actions-of-interest. Another approach proposed by (Li et al., 2023) contains three modules: snippet-level action recognition based on a lightweight X3D model, a training-free probability calibration method that generates frame-level action probability scores from snippet-level results and temporal action localization.

3 PROPOSED METHOD

The proposed overall pipeline of the vision based method for performing driver attention estimation is shown in Figure 1.

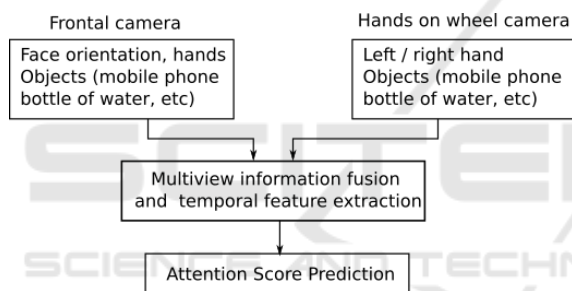


Figure 1: Processing pipeline of the proposed driver attention estimation algorithm.

It has as main modules the frontal and hands on wheel object detectors that provide the location of the drivers' face, hands and also the orientation of the face: looking or not looking at the road, and the location of distracting objects.

The detector is a YOLO based model (Wang et al., 2022). It was trained for detecting the following classes: 0: "Hand not on wheel"; 1: "Hand on bottle"; 2: "Hand on hair comb"; 3: "Hand on wheel"; 4: "Hand on phone" when it processes the images from the hands' camera. On the other hand the detector architecture was also trained on images provided by the face camera and recognizes the following classes: 0: "Looking at the road"; 1: "Not looking at the road"; 2: "Object"; 3: "Phone".

The multi-view information fusion module realizes the temporal synchronization of the two data cues: the frontal camera and the hands-on wheel camera. The temporal synchronization of the multi-view images is achieved by grouping all detected ac-

tions/objects of the two cameras based on the image acquisition timestamp.

The temporal feature extraction module builds the feature vector that is further used by the sequence classification module. This vector contains the image identifier, or more precisely, the associated frame identifier, the camera from which the detected actions originate, which can be 'hands' or 'face'. It also contains the category index, which is the unique identifier of the detected class represented as an integer, the coordinates for the detected bounding boxes, represented by a vector of 4 elements. Last but not least, it includes the confidence score of the detected class resulting from the YOLO model detection, represented as a real number in range 0 and 1, which represents the probability that the model's output is correct.

The attention score prediction module uses a long short term memory that estimates the drivers' degree of attention (which is inversely proportional to the driver's degree of distraction).

Three temporal models are trained for predicting the attention level. The first model uses only the feature vector from the hands view camera. The second model uses only the feature vector from the face view camera. While the third model uses the combined information from both cameras (hands and face) as shown in Figure 2.

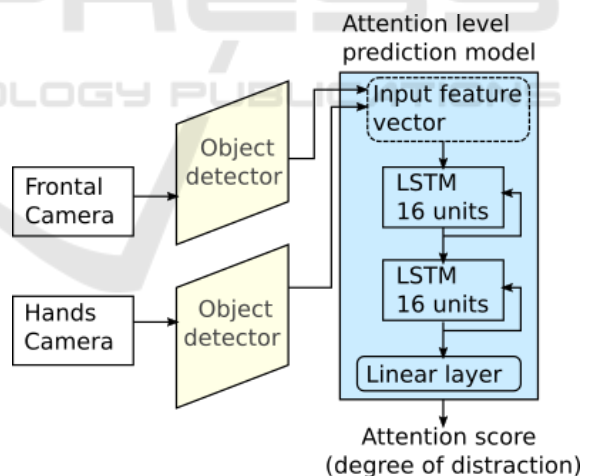


Figure 2: Diagram of the model used for predicting the attention level based on the information from the two cameras.

The architecture for each individual long short term memory model consists of two recurrent layers with 16 hidden units, followed by a linear layer with one output used for predicting the attention score, as depicted in Figure 2. The models have been trained for 30 epochs, using a batch size equal to 256, a mean squared error loss function and the Adam optimizer.

4 EXPERIMENTS AND RESULTS

4.1 Dataset Description

For training and evaluating the proposed model we have inferred annotations to a benchmark dataset, DMD Driver Monitoring Dataset for Attention and Alertness Analysis(Ortega et al., 2020). This dataset contains recordings captured from various positions inside a vehicle, among which two strategically positioned cameras capture the face and the hands of the driver. Some activities were recorded with the car in motion, with the car stopped, and in a simulator.

The original dataset contains approximately 150 videos for each of these cameras, and for the experiments of this paper 50 of them were selected for annotation, each being different in terms of filming conditions, driver, and objects present. Table 1 describes how many images were annotated for each camera position, and the data was split to 70% for training, 15% for validation, and 15% for testing.

Table 1: Number of annotated images for each class in the hands and face dataset.

Hands Dataset	
Class	# Annotated images
Hand on wheel	2.427
Hand not on wheel	366
Hand on hair comb	136
Hand on bottle	128
Hand on phone	271
Face dataset	
Looking at the road	3.838
Not looking at the road	822
Object	365
Phone	590

The driver’s attention level is annotated automatically in the first phase. It is as a real number in the range $0 \dots 1$, where 1 means the driver is extremely attentive while 0 means the driver is very distracted. The attention score is established automatically based on the reasoning mechanisms in algorithms 1 and 2. As the object detection score decreases, the value of the attention score will increase or decrease by 0.1 depending on the type of action (distracting or not). For modelling the variation of the score we use the p variable in the algorithm.

Additionally, there may be situations where certain frames in a video do not contain annotations due to occlusions that may occur in front of the camera. These situations are addressed by setting the attention level to 0.5 (as even the human user is uncertain about the driver’s distraction level).

Data: Object detection score O_s ,
hands category id: hc

$hc = 0 \rightarrow$ One Hand On Wheel;

$hc = 1 \rightarrow$ Hand Not On Wheel;

$hc = 2 \rightarrow$ Hand On Bottle;

$hc = 3 \rightarrow$ Hands On Wheels;

$hc = 4 \rightarrow$ Hand On Phone;

$hc = 5 \rightarrow$ Hand On Hair Comb.

Result: Attention score: A_s

$A_s \leftarrow -1$;

if $O_s > 0.7$ **then** $p \leftarrow 0$;

if $O_s \in [0.3, 0.7]$ **then** $p \leftarrow 0.1$;

if $O_s < 0.3$ **then** $p \leftarrow 0.2$;

if $hc == 0$ **then** $A_s = 0.8 - p$;

if $hc \in [1, 2, 5]$ **then** $A_s = 0.6 - p$;

if $hc == 3$ **then** $A_s = 1 - p$;

if $hc == 4$ **then** $A_s = 0.1 + p$;

if $hc == -1$ **then** $A_s = 0.5$;

Algorithm 1: Automatic annotation of attention score for images captured by the camera facing the hands of the driver.

Data: Object detection score O_s ,

face category id: fc

$fc = 0 \rightarrow$ Looking at the road;

$fc = 1 \rightarrow$ Not looking at the road;

$fc = 2 \rightarrow$ object is detected;

$fc = 3 \rightarrow$ phone is detected

Result: Attention score: A_s

$A_s \leftarrow -1$;

if $O_s > 0.7$ **then** $p \leftarrow 0$;

if $O_s \in [0.3, 0.7]$ **then** $p \leftarrow 0.2$;

if $O_s < 0.3$ **then** $p \leftarrow 0.2$;

if $fc == 0$ **then** $A_s = 1.0 - p$;

if $fc == 1$ **then** $A_s = 0.1 + p$;

if $fc == 2$ **then** $A_s = 0.4 + p$;

if $fc == 3$ **then** $A_s = 0.2 + p$;

if $fc == -1$ **then** $A_s = 0.5$;

Algorithm 2: Automatic annotation of attention score for images generated by the camera capturing the face of the driver.

The ground truth score for the combined model (that uses information from both hands and face camera) equals the average of the attention scores from the hands and from the face annotations.

In the second phase of the annotation process the attention scores are adjusted by the human annotators that either increase or decrease the ground truth scores depending on their own interpretation of situation.

4.2 Multimodal Object Detectors

A YOLO v7 network (Wang et al., 2022) was trained for detecting the hands and face related actions and objects. The evaluation metrics of the object detector are presented in Table 2. It can be noticed that a high mean average precision is recorded for most of the classes that capture the position of the hand with respect to the wheel, while the class with the smallest mean average precision is 'hand on hair comb' because it appears in very few sequences and is very similar in appearance with hand on bottle. Yet this situation does not affect the attention score as both actions (hand on comb or hand on bottle) denote a decrease in the driver's attention level.

Table 2: The analysis of the YOLO model for the hands and face dataset (Precision, Recall and Mean Average Precision).

Class	Prec.	Rec.	mAP
All (hands)	0.91	0.70	0.86
Hand on wheel	0.98	0.99	0.98
Hand not on wheel	0.94	0.73	0.86
Hand on hair comb	0.76	0.7	0.64
Hand on bottle	0.84	0.95	0.88
Hand on phone	0.95	0.96	0.96
All (face)	0.95	0.92	0.96
Looking at the road	0.96	0.96	0.99
Not looking at the road	0.90	0.87	0.91
Object	0.95	0.89	0.94
Phone	1	0.98	0.96

4.3 Evaluation of the Driver Attention Prediction Model

Predicting the degree of attention through the LSTM model is done on sequences of lengths in [5, 10, 15, 20, 25, 30] frames. The attention level is predicted for various offset frames. An offset=1 means the attention level is predicted for the next frame, offset=2 means the model estimates what will be the attention level in the second, and offset=5 means the model will predict what will be the attention score after 5 frames. The inference time for one frame takes an average of 35ms for the individual models for hands and face, while the combined model has an average processing time of 40ms per frame.

Table 3 presents the evaluation results for the three proposed models. It can be noted that all three temporal prediction models have a mean squared error very close to zero in the cases when they predict the attention score for the next frame. In Table 3 results for sequence lengths up to 20 consecutive frames have been

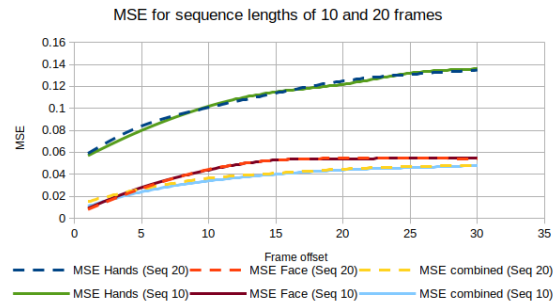


Figure 3: Mean square error evolution for sequence lengths of 10 to 20 frames and various frame offsets.

included, but experiments were also conducted with larger sequence lengths (up to 30 frames), for which the results followed a similar pattern: good performance for small frame offset. It can be noted that the larger the sequence length the better the prediction is, having smaller values for the mean squared error and values close to 1 for R^2 score.

The results indicate that the LSTM models performed well in predicting the attention score individually, when using information only from hands or from face camera object detectors and also when using the combined information from both hands and face cameras. Three models were developed because we wanted to analyse the behaviour of the proposed system in cases when one source of information is missing or may be damaged and in cases when all sources of information are present. Other information cues like outside road information can be integrated in the model for increasing the precision of the results.

The evolution of the means squared error is depicted in Figure 3 for sequences of length 10 and 20 frames considering frame offsets in the range [1, 30]. As it can be noted from Table 3 and Figure 3 as the frame offset increases the prediction errors increase and the R^2 score decreases. So, best results are obtained using a small frame offset of 1, 5 or even 10 frames and higher size of sequence lengths as the history of the temporal evolution of the information from both hands and face camera is more relevant.

The proposed model was also compared with various regression models such as: Random Forest, Adaboost, Support Vector Regressor. Due to the nature of the input required by these regressors, the experiments are done on feature vectors computed for just on frame, no temporal aggregation was considered. Even if the results presented in Table 4 are very comparable with the ones obtained by the proposed model, the temporal model has the advantage of providing good results for larger sequence lengths and frame offsets, being a good base for anticipating the driver's attention level in time.

Table 3: Results of the LSTM model trained on hands, face and combined data: R^2 score, MAE: Mean Average Error, MSE: Mean Squared Error.

(Frame offset Sequence Length)	R^2 hands	MAE hands	MSE hands	R^2 face	MAE face	MSE face	R^2 combined	MAE combined	MSE combined
1, 1	0.592	0.098	0.066	0.866	0.037	0.008	0.761	0.055	0.015
5, 1	0.386	0.130	0.100	0.545	0.086	0.028	0.571	0.086	0.027
10, 1	0.345	0.149	0.107	0.263	0.125	0.046	0.417	0.110	0.037
15, 1	0.266	0.164	0.119	0.151	0.143	0.053	0.330	0.122	0.042
20, 1	0.216	0.174	0.128	0.128	0.149	0.054	0.291	0.129	0.045
25, 1	0.179	0.182	0.134	0.122	0.151	0.054	0.260	0.133	0.047
30, 1	0.145	0.188	0.139	0.116	0.152	0.055	0.232	0.136	0.049
1, 5	0.628	0.097	0.061	0.857	0.039	0.009	0.818	0.050	0.012
5, 5	0.523	0.123	0.078	0.539	0.087	0.029	0.616	0.084	0.024
10, 5	0.379	0.146	0.101	0.260	0.125	0.046	0.439	0.109	0.036
15, 5	0.280	0.161	0.117	0.153	0.144	0.053	0.361	0.121	0.040
20, 5	0.223	0.172	0.126	0.126	0.150	0.054	0.309	0.129	0.044
25, 5	0.191	0.180	0.132	0.116	0.152	0.055	0.266	0.133	0.047
30, 5	0.166	0.187	0.136	0.115	0.153	0.055	0.237	0.137	0.048
1, 10	0.649	0.095	0.057	0.861	0.038	0.009	0.825	0.049	0.011
5, 10	0.509	0.124	0.080	0.552	0.083	0.028	0.628	0.083	0.024
10, 10	0.370	0.146	0.102	0.287	0.122	0.044	0.461	0.108	0.034
15, 10	0.293	0.161	0.115	0.153	0.144	0.053	0.366	0.122	0.040
20, 10	0.248	0.172	0.122	0.128	0.149	0.054	0.311	0.129	0.044
25, 10	0.188	0.179	0.132	0.112	0.152	0.055	0.270	0.133	0.046
30, 10	0.167	0.187	0.136	0.115	0.153	0.055	0.239	0.137	0.048
1, 20	0.64	0.095	0.059	0.871	0.036	0.008	0.760	0.057	0.015
5, 20	0.487	0.126	0.084	0.564	0.083	0.027	0.579	0.087	0.027
10, 20	0.381	0.146	0.101	0.284	0.124	0.044	0.425	0.111	0.036
15, 20	0.299	0.161	0.114	0.149	0.144	0.053	0.324	0.122	0.041
20, 20	0.23	0.171	0.125	0.120	0.150	0.055	0.293	0.130	0.045
25, 20	0.198	0.18	0.131	0.118	0.151	0.055	0.271	0.134	0.047
30, 20	0.173	0.187	0.135	0.131	0.150	0.054	0.238	0.136	0.048

Table 4: Comparison of various algorithms used for the prediction of the attention score, considering no temporal aggregation. The attention score is predicted is done only on the current frame.

Model	R^2 hands	MAE hands	MSE hands	R^2 face	MAE face	MSE face	R^2 combined	MAE combined	MSE combined
LSTM (proposed)	0.592	0.098	0.066	0.866	0.037	0.008	0.761	0.055	0.015
Random Forest	0.48	0.1	0.08	0.66	0.05	0.01	0.68	0.08	0.02
AdaBoost	0.65	0.09	0.062	0.88	0.047	0.008	0.8	0.1	0.01
SVR	0.92	0.068	0.002	0.96	0.017	0.002	0.97	0.03	0.001

4.4 Demonstrative Results

Figures 4, 5, 6 show driver attention estimation results with the context objects marked on images. Each image depicts the bounding boxes along with the name and confidence score of the detected object. The predicted attention score is displayed in red in the upper left corner of the images. Figure 4 shows a situation of low distraction. One can notice a good attentiveness with the driver being focused on the road and

having both hands on the steering wheel.

Figure 5 depicts a situation of medium level distraction while driving. In this case, the camera perceiving the face captured another object that significantly reduced the driver's attentiveness, increasing his level of distraction. In Figure 5-bottom there is a situation of uncertainty because one hand is obstructed by the driver's body, and the other is on the object detected in the face image.

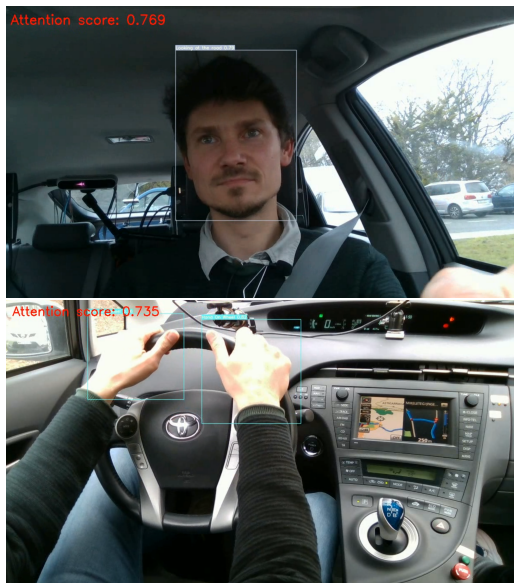


Figure 4: Results: low distraction, high level of attention.

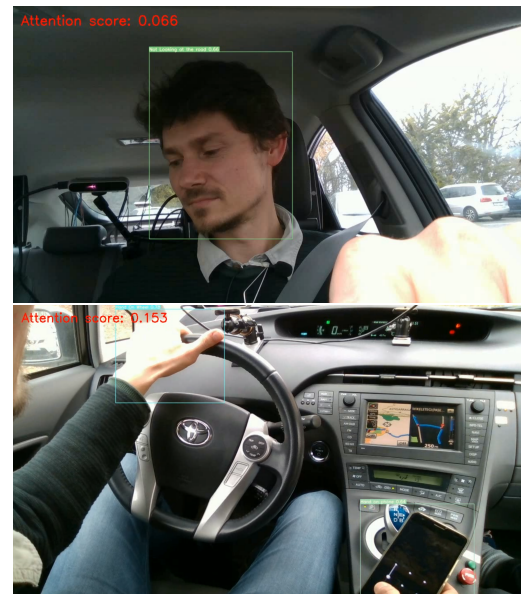


Figure 6: Results: high distraction, low level of attention.



Figure 5: Results: medium distraction, medium level of attention.

A high distraction situation is displayed in Figure 6. The driver is not paying attention to the road and he is looking at the phone from his hand. The predicted attention score is very low.

5 CONCLUSIONS

The paper presents a framework for assessing the driver’s distraction level while driving by determining the distracting actions and the objects present in the

indoor environment of the car. The driver’s distraction level refers to how much attention does the driver pay to the driving process. A temporal regression model is trained using various features extracted on top of object detection performed on images coming from two monocular color cameras: one camera is recording the driver’s face, and the another one captures the driver’s hands and the steering wheel. The degree of attention was annotated on a benchmark dataset and the temporal sequence model was trained for predicting the attention level based on the evolution of information captured from face and hands monitoring cameras.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Ministry of Research and Innovation, CNCS—UEFISCDI, project number PN-III-P4-ID-PCE2020-1700.

REFERENCES

Alyahya, M., Alghannam, S., and Alhussan, T. (2022). Temporal driver action localization using action classification methods. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3318–3325.

Borghi, G. (2018). Combining deep and depth: Deep learning and face depth maps for driver attention monitoring. *ArXiv*, abs/1812.05831.

Borghi, G., Venturelli, M., Vezzani, R., and Cucchiara, R.

- (2017). Poseidon: Face-from-depth for driver pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hu, Z., Zhang, Y., Xing, Y., Li, Q., and Lv, C. (2022). An integrated framework for multi-state driver monitoring using heterogeneous loss and attention-based feature decoupling. *Sensors*, 22(19).
- Jha, S. K., Marzban, M. F., Hu, T., Mahmoud, M. H., and Busso, N. A.-D. C. (2020). The multimodal driver monitoring database: A naturalistic corpus to study driver attention. *IEEE Transactions on Intelligent Transportation Systems*, 23:10736–10752.
- Li, R., Wu, C., Li, L., Shen, Z., Xu, T., Wu, X.-J., Li, X., Lu, J., and Kittler, J. (2023). Action probability calibration for efficient naturalistic driving action localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5270–5277.
- Liang, J., Zhu, H., Zhang, E., and Zhang, J. (2022). Stargazer: A transformer-based driver action detection system for intelligent transportation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3160–3167.
- Muhrer, E. and Vollrath, M. (2011). The effect of visual and cognitive distraction on driver's anticipation in a simulated car following scenario. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(6):555–566. Special Issue: Driving Simulation in Traffic Psychology.
- Naphade, M., Wang, S., Anastasiu, D. C., Tang, Z., Chang, M., Yao, Y., Zheng, L., Rahman, M. S., Venkatchalapathy, A., Sharma, A., Feng, Q., Ablavsky, V., Sclaroff, S., Chakraborty, P., Li, A., Li, S., and Chellappa, R. (2022). The 6th ai city challenge. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3346–3355. IEEE Computer Society.
- Naphade, M., Wang, S., Anastasiu, D. C., Tang, Z., Chang, M.-C., Yao, Y., Zheng, L., Rahman, M. S., Arya, M. S., Sharma, A., Feng, Q., Ablavsky, V., Sclaroff, S., Chakraborty, P., Prajapati, S., Li, A., Li, S., Kunadharaju, K., Jiang, S., and Chellappa, R. (2023). The 7th ai city challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Ortega, J. D., Kose, N., Cañas, P., Chao, M.-A., Unnervik, A., Nieto, M., Otaegui, O., and Salgado, L. (2020). Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In Bartoli, A. and Fusiello, A., editors, *Computer Vision – ECCV 2020 Workshops*, pages 387–405. Springer International Publishing.
- Rong, Y., Kassautzki, N.-R., Fuhl, W., and Kasneci, E. (2022). Where and what: Driver attention-based object detection. *Proc. ACM Hum.-Comput. Interact.*, 6(ETRA).
- Vats, A. and Anastasiu, D. C. (2022). Key point-based driver activity recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3273–3280.
- Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors.
- Wang, Y., Zhao, T., Ding, X., Bian, J., and Fu, X. (2017). Head pose-free eye gaze prediction for driver attention study. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 42–46.
- Zhou, W., Qian, Y., Jie, Z., and Ma, L. (2023). Multi view action recognition for distracted driver behavior localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5375–5380.