2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)
October 8-12, 2022, Macau, China

Challenges for Driver Action Recognition with Face Masks

Elvio G. Amparore, Marco Botta, Idilio Drago, Susanna Donatelli and Giuseppe Mazzone
Università degli Studi di Torino, Italy
e-mail: {elviogilberto.amparore, marco.botta, idilio.drago, susanna.donatelli}@unito.it, giuseppe.mazzone@edu.unito.it

2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) | 978-1-6654-6880-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITSC55140.2022.9921833

***Abstract*— Advanced Driver Assistance Systems (ADAS) are enabling technologies in Intelligent Transportation Systems. Modern ADAS include algorithms to classify drivers' actions and distractions, aiming at identifying situations in which the driver is inattentive. Such systems typically include components for *Driver Action Recognition* (DAR) and *Visual Distraction Classification* (VDC), which prevent risky situations during semi-autonomous driving. DAR and VDC often rely on cameras that track the driver and classify actions based on image recognition algorithms. The COVID-19 pandemic has changed several common social behaviours, including the widespread use of face mask even during driving. In some cases (taxi, bus) face covering policies are compulsory in many legislations. We here show that these behavioural changes challenge state-of-the-art DAR and VDC systems, with the average F1-score in some scenarios dropping by around 30% when exposed to images of drivers wearing masks. Noting a lack of public datasets to update the ML classifiers performing such tasks, we contribute MASKDAR, a dataset for *Action Recognition of Drivers wearing face Masks*. Finally, using MASKDAR we show the importance of including subjects with face masks in datasets for DAR.**

I. INTRODUCTION

A key enabling technology in modern Intelligent Transportation Systems are the *Advanced Driver Assistance Systems* (ADAS) that assist human drivers. They help to reduce risks in critical situations and increase the overall driving safety. ADAS are designed to operate at different automation scenarios – i.e., the L0—L5 autonomy levels [17]. In each of these levels, the commitment required from the human driver varies, going from no automation (L0) to fully automation (L5). In scenarios with partial automation (L2 and L3), the driver attention to the main driving task is still required, as distractions may result in reduced reaction times and incorrect manoeuvres [24].

Modern ADAS include mechanisms to monitor driver inattention using several integrated technologies. One possible way of achieving that is by employing *Human Action Recognition* (HAR) systems, i.e. methods for automatically classifying actions of human subjects from images or video sequences [5]. In the automotive context, the problem is known as *Driver Action Recognition* (DAR), as it aims to detect various forms of secondary tasks the driver may be performing, which may result in degraded driving performance.[1] The task of classifying driver distraction is a more general task than DAR, since it includes both cognitive and visual distraction classification (VDC) along with secondary task detection. All these tasks have been faced in previous works [16], with multiple classification methods, in particular based on Convolution Neural Networks (CNN), shown to deliver good performance [3]. Equally, several public datasets have been introduced for both HAR, DDC and VDC tasks [18].

Since the start of the COVID-19 pandemic, several human behaviours have changed to counter the emergency [20]. Among the various changes, non-pharmaceutical interventions (NPIs) like face masks have become common, both in outdoor, indoor and even car environment. This sudden change in driver's habits calls for a reappraisal of the performance of Human Action Recognition software, since many systems have not been trained to recognize drivers wearing a face mask. Moreover, to the best of our knowledge, no dataset among the publicly available ones for the DAR and VDC tasks considers the duality of drivers with/without face masks as a relevant subjects' feature [18].

Research question: To understand whether it is adequate to use classification models trained on datasets that do not include drivers wearing face masks (as most available datasets) to classify drivers that wear a mask, and to quantify the actual degradation of such classifiers, if any.

Contributions:

1) A dataset for *Action Recognition of Drivers wearing face Masks*, named MASKDAR, that can be used to train machine learning (ML) models to identify 15 distinct actions associated to drivers' distraction. It includes 30 subjects performing driving tasks, both with and without face masks.
2) An experimental analysis to quantify the impact of considering face masks when training/testing ML models for DAR and VDC.

Results: The analysis shows that the presence of face masks significantly degrades the classification performance for ML models that are not explicitly designed to take face masks into consideration. The average F1-scores for models trained without taking into account face masks is reduced from 0.90 to 0.69 in the VDC scenario, and from 0.76 to 0.60 in the DAR scenario. By analysing how image features are used in the classification, we identify the importance of the face regions covered by the masks for the classification. We then show that the inclusion of drivers wearing face masks in the training set restores part of the lost classification performance, with F1-scores reaching 0.84 and 0.71 for the VDC and DAR scenarios, respectively.

[1] In the literature the A of the acronyms DAR and HAR stays equally likely for "activity" or "action". We prefer to use only the term "action".

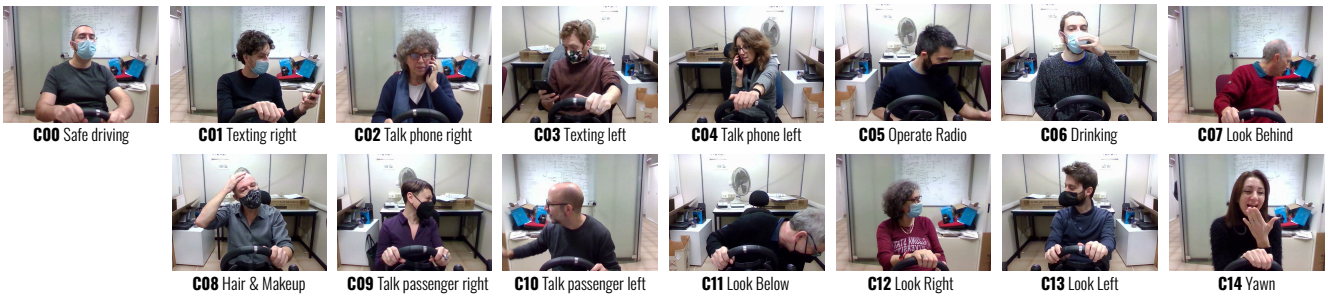

Fig. 1: Samples of pictures of the 15 action classes recorded in lab conditions, with a mix of subjects with/without masks.

Our results thus demonstrate the relevance of considering NPIs in datasets for action recognition. We believe the dataset, scripts and models can contribute to efforts aiming to build better DAR and VDC systems, able to cope with drivers in novel (common) setups.

Availability: MASKDAR is available with no charges under a research agreement with our research group (under GDPR regulations). Contact the authors to set up the agreement.

II. RELATED WORK

Action recognition of drivers has been studied extensively in the last few decades [16]. A crucial component needed to perform these studies is the availability of public datasets for DAR/VDC tasks. The survey in [10] compares and classifies many of these datasets (in the broader context of HAR), using a structured taxonomy. One of the first publicly-available dataset for DAR is found in [27], consisting in four action classes taken on 20 subjects (resulting in 80 images). Such dataset is not large enough for training *Convolutional Neural Network* (CNN) models, and thus subsequent datasets have significantly scaled up in size.

A popular dataset that has sparked a lot of interest and several studies is the AUC-V2 datasets [6], successor of AUC [1]. AUC-V2 is an image-based dataset of about 15 k images from 44 subjects, comprising 10 action classes. In this paper we consider these classes and the general methodology of AUC-V2 as the starting point of our research. However, AUC-V2 does not consider face masks and their impact as obstruction devices for the classification tasks. A multi-modal, multi-sensor day/night dataset is the 3MDAD [9]. While more advanced in the setup than AUC-V2, this dataset does not consider face masks either. Other datasets, as in [21], consider low lighting conditions and infrared images. We here omit these cases for simplicity and because they are orthogonal to the face mask obstruction problem.

The problem of face masks (and other NPIs) identification using CNN is studied in [3] in the context of policy implementation tracking. However, the target is not action recognition, but detection of subjects not wearing masks. A benchmark of CNN architectures for the AUC-V2 dataset is found in [15], while a qualitative comparison of the techniques used to solve the general HAR problem is described in [10]. A survey focused instead on the methods to address the VDC problem is given in [11]. Other approaches not based on CNN but on cue tracking are reviewed in [13].

A work with a setup similar to ours is reported in [23], where the authors use an assisted-driver testbed to simulate an automotive environment using the same 10 action classes of AUC-V2. However, their goal is to benchmark different CNN architectures for the DAR task in a controlled environment. Face masks have not been considered in that research.

The classification of videos (like in [19]) may also constitute a relevant improvement for a DAR classifier. We however do not report related work in this field since we restrict the scope of this paper to the task of classifying still images. The analysis of the impact of face masks on classifiers based on videos is left for future work.

III. DATASET FOR DAR WITH FACE MASKS

Although there are several public datasets for DAR (see [18, Table 1]), to the best of our knowledge no dataset considers the problem of drivers wearing face masks for coronavirus protection. To answer our research question we have developed a dataset, named MASKDAR, described in the following.

A. Subjects and Face Masks

The dataset is composed of 30 subjects (23 males, 7 females), each recorded in two rounds of simulated driving: R1=without face mask, R2=with face mask.[2] In each round, the subject performs a sequence of 15 actions, each lasting 10 seconds. Actions are summarized in Table I and include safe driving as well as multiple actions related to distraction, such as texting, looking away from the road, drinking etc. Therefore, each subject performed the requested actions two times, in the R1 and R2 conditions (without and with masks). Different types of masks (surgical, cloth, N95/FFP2) have been used by the subjects during the experiments.

B. Image capturing

For each action/round/subject, the dataset contains 10 images randomly selected among the video frames captured during the interval in which the subject was performing the requested action. Pictures are taken in a lab scenario

[2]We use CARLA Simulator (`https://carla.org/`), but our results do not depend on the simulation framework and setting, since the ML models rely only on the drivers' pictures.

TABLE I: Action class definitions for the MASKDAR dataset.

Class	Name	Description
C00	Safe driving	The subject simulates a normal, safe driving, looking straight at the road or at the mirrors.
C01	Texting right	The subject holds the cellphone in her/his right hand, and is writing a text message.
C02	Talk phone right	The subject holds the cell phone in her/his right hand, and is performing a phone call.
C03	Texting left	Same as C01 with the left hand.
C04	Talk phone left	Same as C02 with the left hand.
C05	Operate radio	The subject gaze is directed at the car radio, and she/he plays with it using the right hand.
C06	Drinking	The subject is drinking (or about to drink) from a bottle/glass using one hand.
C07	Look behind	The driver is stretching to grasp some object on the back seat.
C08	Hair & makeup	The subject is adjusting her/his hair or doing some makeup.
C09	Talk passenger right	The driver is looking at her/his right and talking.
C10	Talk passenger left	Same as C09 on the left.
C11	Look below	The driver is stretching to grasp something on the car floor.
C12	Look right	The driver is looking at her/his right, but she/he is not talking.
C13	Look left	Same as C12 on the right.
C14	Yawn	The driver is yawning, perhaps inattentive to the driving task.

(similar to [23]) using a frontal camera with a 80° *field-of-view* (FoV). Frontal cameras are a more realistic choice than lateral cameras (as in the AUC-V2 dataset), since the latter may easily be obstructed by a passenger or by luggage (see [7] for a similar discussion). Originally, pictures have been taken at 640×480 resolution, but are then rescaled to 320×240 for faster processing.[3]

C. Action classes

Fig. 1 shows a sample image for each of the 15 action classes described in Tab. I, taken from either R1 or R2 rounds. Action classes C00–C09 directly correspond to the ones in [6]. Action classes C10–C14 have been added following the experimental setup of [18]. Action classes that have left/right variations (C01/C03, C02/C04, C09/C10, C12/C13) have been acquired on the *right* variation only, while the left variation is generated by mirroring. Each action class has 600 images (30 subjects × 2 rounds × 10 samples), resulting in 9000 images for the whole dataset. Since the dataset is acquired in lab conditions, we have considered both the talk left/right actions, even if they are usually exclusive in real setups, following the position of the driver/passenger in the car.

IV. METHODOLOGY

We now describe our experimental methodology, detailing our experiment configurations (Sect. IV), the used classification models (Sect. IV-B) and methods used to evaluate features and explain results (Sect. IV-C).

A. Experiment Configurations

To experimentally test our research question, we consider three configurations of the action classes. Fig. 2 summarizes the configurations. In the first configuration (CONF15), classifiers are trained to identify all driver actions listed in Table I. This is the most challenging scenario, since multiple driver actions are characterized by similar body movements – e.g., C10 (Talk to passenger left) and C13 (Look left). In the second configuration (CONF08), we group together classes that are expected to be confused in pictures, thus creating a coarser classifier able to identify macro-categories of driver actions. In the last scenario (CONF02) we collapse all non-safe actions into a single generic class (E01 Distracted), thus evaluating a binary classifier that can be considered as an instance of the VDC task.

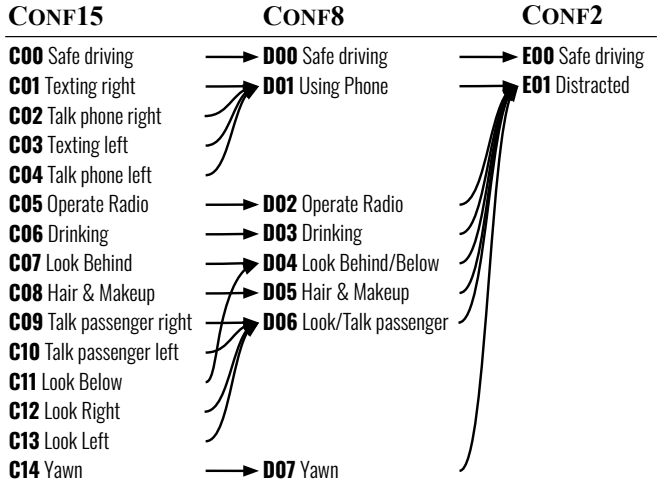

Fig. 2: Action class grouping for the three configurations.

For each configuration, we train the CNN models to identify the drivers' actions in three experimental setups:

- T1 *No face masks*: The model is both trained and tested with drivers wearing no masks. This experiment is our baseline.
- T2 *Training without masks, Test with masks*: The model is trained with drivers wearing no face masks and then tested with drivers wearing masks. This case assesses the performance degradation of missing face masks in the training data.
- T3 *Mixed*: Both the training and test sets include subjects with and without face masks. This case tests the performance of a classifier that is properly trained to classify drivers wearing face masks.

[3]Subjects have provided explicit consent for the release of their pictures. MASKDAR will be made publicly available for research purposes.

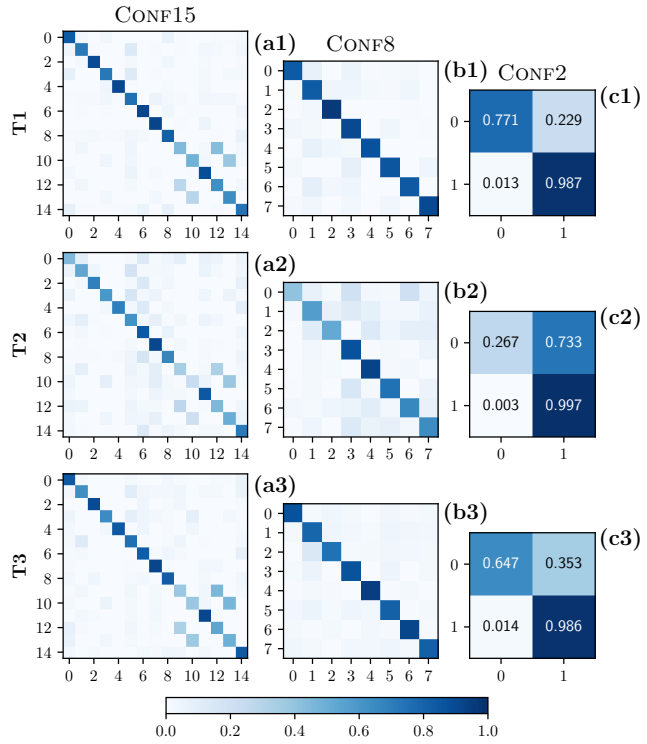

Fig. 3: Confusion matrices for the nine experiments.

TABLE II: F1-scores by action classes.

Class	F1-score T1	T2	T3
C00	0.82	0.49	0.76
C01	0.74	0.55	0.68
C02	0.91	0.78	0.91
C03	0.72	0.68	0.73
C04	0.93	0.81	0.92
C05	0.73	0.54	0.71
C06	0.92	0.62	0.81
C07	0.91	0.74	0.88
C08	0.80	0.68	0.82
C09	0.50	0.38	0.40
C10	0.54	0.32	0.40
C11	0.87	0.77	0.88
C12	0.58	0.48	0.52
C13	0.61	0.52	0.49
C14	0.76	0.64	0.77
$\mathbb{E}$[F1]	0.76	0.60	0.71

(a) CONF15 scores.

Class	F1-score T1	T2	T3
D00	0.85	0.50	0.82
D01	0.83	0.69	0.84
D02	0.94	0.51	0.72
D03	0.88	0.57	0.83
D04	0.91	0.76	0.92
D05	0.85	0.70	0.75
D06	0.86	0.74	0.92
D07	0.91	0.52	0.80
$\mathbb{E}$[F1]	0.88	0.62	0.83

(b) CONF8 scores.

Class	F1-score T1	T2	T3
E00	0.81	0.41	0.70
E01	0.98	0.97	0.98
$\mathbb{E}$[F1]	0.90	0.69	0.84

(c) CONF2 scores.

B. CNN Model

We use CNN models to perform our classification task. In particular we employ the state-of-the-art EfficientNet-B1 architecture (see [22] for the network structure), as used also in [12]. Training is performed using transfer learning [26] from a checkpoint [8] of a model trained using the *ImageNet* dataset [4]. The transferred model is then fine-tuned against the MASKDAR dataset. This approach has been shown to lead to fast model convergence and high classification accuracy in multiple image recognition tasks [26]. Depending on the configuration, our model has 15, 8 or 2 outputs, each associated with the corresponding class in Fig. 2.

To obtain unbiased results, we perform cross-validation using a leave-one-out process, in which we train the models using data from 29 subjects and test on data from the remaining subject. Thus each of the 9 experiments (three tests T1, T2 and T3, each one with the three defined configurations) requires to train 30 models on 29 subjects, that are then tested on the single excluded subject. This approach helps to verify the model robustness more thoroughly than by just using random sampling of the dataset into train/test sets.

Each model in test T1 has been trained for 50 epochs, in batches of 32 input images, that pass through data augmentation layers (rotation, translation, contrast and zoom). Training might stop earlier if there is no accuracy improvement on the validation set for 10 epochs. Then, for test T3 the trained models from test T1 have been fine tuned on the images of subjects wearing masks for 10 epochs.

For what concerns the performance metrics of each experiment, for each action we generate, from the experiment confusion matrix, a 2 by 2 matrix $\left[\begin{smallmatrix} TP & FN \\ FP & TN \end{smallmatrix}\right]$, from which we compute the per-action scores using the classic performance metrics:

$$Precision = \frac{TP}{TP+FP}, \qquad Recall = \frac{TP}{TP+FN}$$

$$F1\text{-}score = 2 \cdot \frac{Precision * Recall}{Precision + Recall}$$

The score of the model on the single experiment is given by $\mathbb{E}$[F1], i.e. the mean F1-score value over all the classes.

C. Checking Deep Layer Activation (eXplainable AI)

To further inspect the predictions obtained by our CNN models, we use the DeepSHAP method [14]. Given an input image for which we use a CNN model to predict its class, DeepSHAP extracts the relative importance (Shapley values) of the 2D features in input at every convolutional layer of the CNN model, generating an *activation map*. Inspecting the maps obtained for the deepest convolutional layer allows us to get insights on what parts of the input image concurred positively (or negatively) in the final classification score. Fig. 4 shows some examples of the obtained activation maps for MASKDAR. The first column shows input images, while the other two columns depict activation maps. Red areas activated the CNN to increase the classification likelihood of the particular example in the given class, while blue areas decrease such likelihood.

V. RESULTS

We now experimentally assess our research question: how much do face masks degrade the performance of a DAR (or a VDC) system, specifically if it was not trained for that purpose? We describe the results of the three experimental setups in the following subsections.

A. T1 — *Baseline, no face masks*

Results for T1 are summarized in Fig. 3(T1 row), which reports the average confusion matrices (as heat maps) for CONF15, CONF8 and CONF2, respectively. In each matrix rows correspond to the actual class and columns correspond to the predicted class. Each cell shows a color corresponding to the average conditional prediction probability after 30 cross-validation folds. Tab. II complements the confusion matrices by reporting, in each column labelled T1, the values of the F1-scores for every class, as well as the average F1-scores for each of the three configurations.

Starting with CONF15 (Fig. 3($a1$)), we see that the model classifies most images correctly. F1-scores in this configuration – T1 column in Tab. IIa – are high for most classes, reaching 0.93 for C04 (Talk phone left). Classes with lower performance are C09 and C12 (Talk passenger right/Look right) with around 0.5–0.6 F1-scores, and C10/C13 (Talk passenger left/Look left). These classes appear to be indistinguishable by the classifier (as can be observed by the off-diagonal elements of the confusion matrix). Other classes (C05, C14) also appear to be more difficult to classify, according to their F1-scores. Such results are not surprising, since the performance of image-based HAR systems remains somewhat limited with real-world data [10]. This is particularly the case for models built on single images (i.e., as opposed to models built on videos), which is known to be more challenging when classifying complex action dynamics [2] (such as talk/look). Results obtained for this baseline setting are similar to previous (comparable) experiments [25].

As we group some actions together in CONF8 the classifier reduces such mistakes and the overall performance improves, as can be observed in Fig. 3($b1$) and in T1 column of Tab. IIb. Grouping actions into just two classes (CONF2) further increases the performance. Fig. 3($c1$) shows that the classifier has high recall for the Distracted class, while it makes some mistakes for samples of the Safe Driving class. The F1-Scores reaches 0.81 and 0.98 for the Safe Driving and Distracted classes, respectively. The average F1-scores, which capture differences on performance across the multiple classes, are at 0.76, 0.88 and 0.90, respectively.

Takeaway: *Using a dataset containing only drivers wearing no masks and state-of-the-art ML models leads to classifiers able to identify drivers' actions with good performance (e.g., 0.98 F1-score for the Distraction class).*

B. T2 — *Training without masks, test with masks*

We assess the performances of the same T1 models against pictures of drivers wearing face masks. Fig. 3(T2 row) shows the confusion matrices for the three configurations. Results for T2 are again obtained by cross-validation, where in each cross-validation fold the subject in the test set is the same as in the T1 fold, but wearing a face mask.

Fig. 3(T2 row) shows a performance degradation w.r.t. T1 for all three configurations. In CONF15 we observe a large dispersion on the conditional prediction probability away from the matrix diagonal, pointing to a general reduction of classification power. A similar pattern is observed for

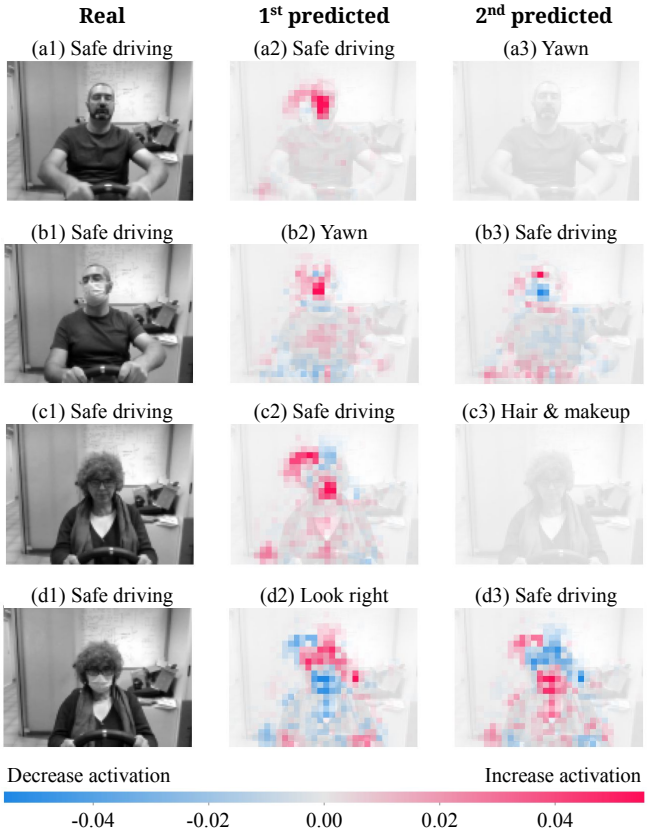

Fig. 4: Examples of activation maps of the CNN models.

CONF8 as well as for CONF2. In the latter configuration, the classifier is clearly unable to distinguish the Safe Driving examples. These results show the importance of updating DAR and VDC systems for the post-COVID19 scenarios.

The third columns in each table of Tab. II details results for T2. Comparing numbers against those obtained in T1, a major performance reduction is confirmed for *all* classes. The F1-score for the best performing class in T1/CONF15 (C04, Talk phone left) is reduced from 0.93 to 0.81 in T2. The F1-score is as low as 0.32 for C10 (Talk passenger left). The average F1-score is reduced from 0.76 to 0.60 in CONF15, from 0.88 to 0.62 in CONF8, and from 0.90 to 0.69 in CONF2.

We investigate the activation maps of the deepest convolutional layers in the classifiers using DeepSHAP, to inspect which visual features are used by the models to make the predictions. Fig. 4 shows a few examples of our inspection. Row (a) and (c) correspond to drivers without a face mask (CONF15/T1 scenario), while rows (b) and (d) are similar pictures of the same subjects wearing face masks (CONF15/T2 scenario). Column 1 shows the input images, column 2 shows the activation maps for the most likely action classes (according to the classifier) and column 3 shows the maps for the second most likely classes. Red (and blue) areas in the activation maps correspond to the areas that influenced positively (and negatively) the classifier output for the given

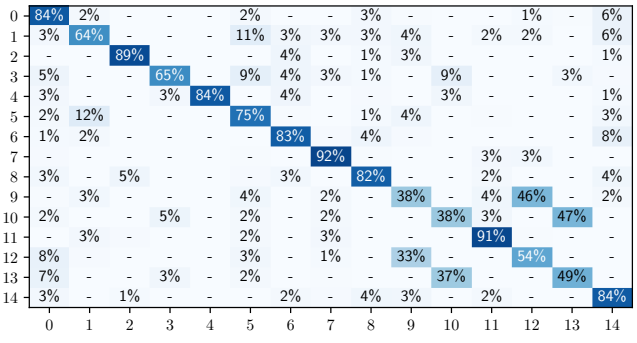

Fig. 5: Detailed confusion matrix for T3/CONF15.

class. We consider models trained without subjects wearing face masks.

Examples (a) and (c) show cases where the classifier correctly predicts the Safe Driving class. The second column in Fig. 4 shows a high concentration of red areas over the face of the subjects – i.e., face traits have activated the classifier for the correct class in these examples. The third column in Fig. 4 shows no strong activation area for the action identified by the classifier as second most likely ones for these examples. These examples suggest that the classifier relies on face traits to make the predictions.

Examples (b) and (d) are incorrectly classified as Yawn and Look Right, while the pictures actually correspond to the Safe Driving class. In both cases, the classifier has Safe Driving as its second most likely option. In (b3) and (d3) we observe patterns that are clearly opposed to the one seen for Safe Driving on the Examples (a2) and (c2). Large areas over the face negatively contribute to the prediction in (b3) and (d3). Instead, (b2) shows that the area covered by the mask has activated the classifier for the (incorrect) Yawn class, whereas (d2) shows other areas of the image activating the (incorrect) Look Right class. These examples suggest that the classifier is confused by the face obstruction.

We have evaluated several examples, and they show similar activation patterns, supporting our conjecture that the CNN models are confused because of the lack of facial traits.

Takeaway: *Drivers wearing masks reduce classification performance significantly if DAR and VDC rely on models trained without subjects wearing face masks. Activation map inspection suggests that CNN models are confused by the missing features of subjects' faces. In the VDC scenario (*CONF2*) the average F1-scores drop from 0.90 to 0.69.*

C. T3 — *Training and testing with and without masks*

Last, we repeat the cross-validation experiments, but this time models are both trained and tested using pictures of drivers with/without face masks.

Fig. 3(T3 row) shows that model performance metrics improve from T2, and are closer to the ones in T1. For example, the classifier built in CONF2 is able to identify Safe Driving again – compare numbers in the matrix in Fig. 3($c3$) to those in Fig. 3($c2$). Similar patterns can be observed for CONF8 and CONF15. The confusion matrix for the T3 experiment in CONF15 is also reported in details in Fig. 5.

Yet, the performance observed in T1 is not fully recovered in T3. Numbers in the fourth columns in the tables in Tab. II quantify these aspects. F1-scores for most classes in T3 are close to those in T1, but with small reductions, even if there are increases for some classes (C05, C08, C11 and C14). On average, the F1-score for CONF15 jumps from 0.60 in T2 to 0.71 in T3, remaining slightly below the 0.76 obtained in T1. Similar patterns are observed for the other configurations.

Again, manual inspection of the activation maps (see discussion for Fig. 4) suggests that the lack of facial features for subjects wearing masks is behind the performance degradation when comparing T3 to T1. Proving such conjecture as well as searching for other CNN models able to compensate for the presence of the face masks are left for future work.

Takeaway: *Including drivers wearing face masks in training sets recovers most of the performance degradation observed for baseline models (without face masks). Yet, the tested CNN models are still penalized. We conjecture that the lack of important facial features prevents results from returning to the levels of the baseline scenario.*

VI. CONCLUSION

We investigated whether drivers wearing face masks have an impact on the classification performance of systems supporting DAR and VDC tasks. For performing a fair and thorough comparison, we introduce MASKDAR, a dataset for action recognition of drivers wearing face masks. Our results showed that the presence of drivers wearing face masks degrades the performance of classifiers built without taking such new scenario into account. The inclusion of subjects wearing face masks mostly recovers the classifiers' performance, even if some performance degradation still remains.

While MASKDAR allows to test our research question, we also acknowledge the fact that it is recorded in lab conditions, with a limited diversity of subjects. Moreover, we restricted our research on the classification of still images. The impact of face masks in video classification (like in [19]) is also a relevant question.

Inspection of the activation maps (see discussion for Fig. 4) suggested that the lack of facial features for subjects wearing masks is behind the performance degradation of T2. Systematically proving such conjecture as well as searching for other CNN models able to compensate for the presence of the face masks are left for future work.

ACKNOWLEDGEMENT

This work has received funding from the European Union's Horizon 2020 research&innovation program ECSEL Joint Undertaking (JU) under Grant Agreement No. 876487 (NextPerception). `https://www.nextperception.eu/`

REFERENCES

[1] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. "Real-time distracted driver posture classification". In: *arXiv:1706.09498* (2017).

[2] Paola Cañas et al. “Detection of Distraction-related Actions on DMD: An Image and a Video-based Approach Comparison.” In: *VISIGRAPP (VISAPP)*. Vol. 5. 2021, pp. 458–465.

[3] Amit Chavda et al. “Multi-Stage CNN Architecture for Face Mask Detection”. In: *2021 6th Int. Conf. for Convergence in Technology (I2CT)*. 2021, pp. 1–8.

[4] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conf. on computer vision and pattern recognition*. 2009, pp. 248–255.

[5] KB Devika et al. “Driver Distraction Recognition-driven Collision Avoidance Algorithm for Active Vehicle Safety”. In: *2021 IEEE ITSC*. 2021, pp. 237–243.

[6] Hesham M Eraqi et al. “Driver distraction identification with an ensemble of convolutional neural networks”. In: *Journal of Adv. Transportation* (2019).

[7] Boris Faizov, Vlad Shakhuro, and Anton Konushin. “Automatic Detection of Certain Unwanted Driver Behavior”. In: *GraphiCon* (2021).

[8] Yixing Fu. *Image classification via fine-tuning with EfficientNet*. `https://keras.io/examples/vision/image_classification_efficientnet_fine_tuning/`. [Online; accessed Feb-2022]. 2020.

[9] Imen Jegham et al. “A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD”. In: *Signal Processing: Image Communication* 88 (2020), p. 115960.

[10] Imen Jegham et al. “Vision-based human action recognition: An overview and real world challenges”. In: *Forensic Science Int.: Digital Investigation* 32 (2020).

[11] Alexey Kashevnik et al. “Driver distraction detection methods: A literature review and framework”. In: *IEEE Access* 9 (2021), pp. 60063–60076.

[12] Ashlesha Kumar, Kuldip Singh Sangwan, et al. “A Computer Vision Based Approach for Driver Distraction Recognition Using Deep Learning and Genetic Algorithm Based Ensemble”. In: *Int. Conf. on AI and Soft Computing*. Springer. 2021, pp. 44–56.

[13] Wanli Li et al. “A survey on vision-based driver distraction analysis”. In: *Journal of Systems Architecture* 121 (2021), p. 102319.

[14] Scott M Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems* 30 (2017).

[15] Jimiama Mafeni M. et al. “Benchmarking deep learning models for driver distraction detection”. In: *Int. Conf. on ML, Optimization, and Data Science*. Springer. 2020, pp. 103–117.

[16] Negar Moslemi, Mohsen Soryani, and Reza Azmi. “Computer vision-based recognition of driver distraction: A review”. In: *Concurrency and Computation: Practice and Experience* 33.24 (2021), e6475.

[17] Jaswanth Nidamanuri et al. “A Progressive Review-Emerging Technologies for ADAS Driven Solutions”. In: *IEEE Transactions on Intelligent Vehicles* (2021).

[18] Juan Diego Ortega et al. “DMD: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis”. In: *European Conf. on Computer Vision*. Springer. 2020, pp. 387–405.

[19] Chaopeng Pan et al. “Driver activity recognition using spatial-temporal graph convolutional LSTM networks with attention mechanism”. In: *IET Intelligent Transport Systems* 15.2 (2021), pp. 297–307.

[20] Daniela Perrotta et al. “Behaviours and attitudes in response to the COVID-19 pandemic: insights from a cross-national Facebook survey”. In: *EPJ data science* 10.1 (2021), p. 17.

[21] Mohamed H Saad, Mahmoud I Khalil, and Hazem M Abbas. “End-To-End Driver Distraction Recognition Using Novel Low Lighting Support Dataset”. In: *(ICCES)*. IEEE. 2020, pp. 1–6.

[22] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *Int. conf. on Machine Learning*. 2019, pp. 6105–6114.

[23] Duy Tran et al. “Real-time detection of distracted driving based on deep learning”. In: *IET Intelligent Transport Systems* 12.10 (2018), pp. 1210–1219.

[24] Kristie Young, Michael Regan, and Mike Hammer. “Driver distraction: A review of the literature”. In: *Distracted driving* 2007 (2007), pp. 379–405.

[25] Chaoyun Zhang et al. “Driver behavior recognition via interwoven deep convolutional neural nets with multi-stream inputs”. In: *IEEE Access* 8 (2020), pp. 191138–191151.

[26] Deyuan Zhang, Zhenghong Liu, and Xiangbin Shi. “Transfer learning on efficientnet for remote sensing image classification”. In: *ICMCCE*. IEEE. 2020, pp. 2255–2258.

[27] CH Zhao et al. “Recognition of driving postures by contourlet transform and random forests”. In: *IET Intelligent Transport Systems* 6.2 (2012), pp. 161–168.