Detection of Distraction-related Actions on DMD: An Image and a Video-based Approach Comparison

Paola Natalia Cañas[a], Juan Diego Ortega[b], Marcos Nieto[c] and Oihana Otaegui[d]
Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Spain

Keywords: Driver Monitoring Systems, Deep Learning, Action Recognition, Distraction Detection, Distracted Driver, ADAS, Autonomous Driving.

Abstract: The recently presented Driver Monitoring Dataset (DMD) extends research lines for Driver Monitoring Systems. We intend to explore this dataset and apply commonly used methods for action recognition to this specific context, from image-based to video-based analysis. Specially, we aim to detect driver distraction by applying action recognition techniques to classify a list of distraction-related activities. This is now possible thanks to the DMD, that offers recordings of distracted drivers in video format. A comparison between different state-of-the-art models for image and video classification is reviewed. Also, we discuss the feasibility of implementing image-based or video-based models in a real-context driver monitoring system. Preliminary results are presented in this article as a point of reference to future work on the DMD.

1 INTRODUCTION

Currently, cars that are known commercially as autonomous still require human participation. These yet belong to level 2 of the standard Driving Automation Classification from the Society of Automotive Engineers (SAE International, 2018); it means that the car has partial automation, but the human is still responsible for driving. Is in Level 3 where the driver can engage in other activities inside the car while the vehicle is driving by its own. To make this transition, go up from level 2 to level 3, driver's state or condition must be an input to the autonomous driving systems. This is because, in level 3, mode transitions between manual and automated driving still will occur; identifying if the person is capable of regaining vehicle's control is important to make decisions.

All the efforts that we can direct to characterize and detect the driver's condition are necessary, because it could help a smart system to anticipate potential risk scenarios. Hence the relevance of Driver Monitoring Systems (DMS). This task of driver monitoring may include various scenarios and multiple aspects of the driver, from hands position to distraction level, on this research we focus on distraction detection. The NHTSA establishes 3 types of distraction: visual, manual and cognitive. To detect driver distraction, without implementing intrusive detection methods, it is required to identify a list of activities known to have a certain cognitive load (like having a conversation on the phone), or any other visual and/or manual distraction; allowing us to infer if the driver is distracted through the identification of these activities. Therefore, distraction detection becomes an action recognition task.

For monitoring the driver inside the car, computer vision techniques are quite convenient as they are non-intrusive methods. Convolutional Neural Networks (CNN's) have demonstrated their advantages and outstanding results on image analysis (Krizhevsky et al., 2012). Therefore, these algorithms of Deep Learning have become the first option for computer vision problems. Since images of the driver are the principal source of information for these systems, CNN's are considered in this research.

As we intend to detect distractions, we want to explore and get some initial metrics of the recently published Driver Monitoring Dataset (DMD) (Ortega et al., 2020) to find if the temporal dimension of actions is worth to consider or a sole image-based analysis should be enough. Also, how some models used for general action recognition or video classification can perform with the available data of this dataset.

[a] https://orcid.org/0000-0002-9752-6724
[b] https://orcid.org/0000-0001-5539-106X
[c] https://orcid.org/0000-0001-9879-0992
[d] https://orcid.org/0000-0001-6069-8787

Cañas, P., Ortega, J., Nieto, M. and Otaegui, O.
Detection of Distraction-related Actions on DMD: An Image and a Video-based Approach Comparison.
DOI: 10.5220/0010244504580465
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 458-465
ISBN: 978-989-758-488-6

2 RELATED WORK

The lack of public and appropriate datasets for the detection of distracted drivers has limited the development of methods for this specific action detection application. The DMD proposes to be robust enough to extend research around driver monitoring. Also, it comes in video format, offering the temporal dimension for action detection.

2.1 Datasets

Researchers often construct their own datasets only for their studies purposes, many of which are never published. Among the most important and comparable datasets with the DMD distraction-dedicated material are:

-State Farm's Kaggle Competition (StateFarm, 2016). In a competition on Kaggle website, the State Farm insurance company published a dataset on the platform. This one contains side-view images from drivers performing a list of 10 distraction-related activities, including normal driving, talking on the phone, and operating the radio, among others. However, the dataset is restricted for the competition and is not allowed to be used for other purposes.

-AUC Distracted Driver Dataset (Abouelnaga et al., 2018). Given the use limitations of the previous dataset, Abouelnaga et al. team of researchers carried out their own distracted driver dataset. It complies with the same characteristics as the one from State Farm and is available to the scientific community under a usage agreement. It has the same list of distraction activities as State Farm's Dataset but this one is bigger in size.

-Drive&Act Dataset (Martin et al., 2019). This dataset has a slightly different approach than the DMD, since it aims to support the identification of driver actions in autonomous driving scenarios. Within this context, it is understood that the driver does not actively participate in the driving task, so his/hers activities become those of a common passenger and not of a driver. Therefore, there are multiple not-related-to-driving activities, few of them like "Normal driving" are shared with the DMD, but others belong outside a driving context like "working on a laptop". Drive&Act offers data in video format from 6 inside-car perspectives and 3 channels of information (RGB, infrared and Depth). The material of this dataset is available publicly also under the acceptance of a usage agreement.

2.2 Algorithms

For action recognition, two approaches have been determined: an image or a video analysis. For image-based algorithms, the intention is to identify an action (understood as a time-dependent sequence) from a still image. In video-based algorithms, spatial and temporal dimensions are both taken into account.

Image-based. For driver's distraction detection, the authors of the AUC Dataset implement a CNN's ensemble and extract features from images, performing image classification. These methods of prior feature extraction have been adopted on many investigations for driver state identification algorithms: skin-segmentation (Xing et al., 2019), hands detection (Rangesh and Trivedi, 2018), face detection (Yuen et al., 2016), head pose estimation (Borghi et al., 2017) and body landmarks detection for driver posture estimation (Deo and Trivedi, 2018). Distraction detection by image classification with CNN's architectures like VGG-16 (Baheti et al., 2020), AlexNet, GoogleNet, and residual network also have been studied (Tran et al., 2018).

General human action recognition problems have been approached with still-image-based algorithms, giving good results (Zhang et al., 2016). Some of the strategies involve human pose estimation (Yang et al., 2010), human and/or object detection to find human-object interactions (Girish et al., 2020) and combinations with general scene understanding (Chan et al., 2019).

Video-based. This research is supported by advances in video detection of general human actions, since there is not much evidence for driver distraction detection in video.

2D CNN's are widely used for learning spatial features from each video frame. However, with an arrangement of 2D CNN's and calculations of optical flow, spatial-temporal information can be considered and accomplish action recognition tasks (Chen et al., 2020).

An alternative to analyse videos is to have a 2D CNN (for spatial features), followed by an LSTM (for temporal dependencies). The combination of these two have proven to be a good option for action recognition (Donahue et al., 2014).

On the other hand, 3D CNN's or 3DConvNets have proven to out-stand 2D CNN's in action recognition tasks (Tran et al., 2015). This extra dimension of the kernels or filters allows the network to capture the motion information encoded in multiple contiguous frames. This means that the computed features are both spatial and temporal.

3 DMD: DRIVER MONITORING DATASET

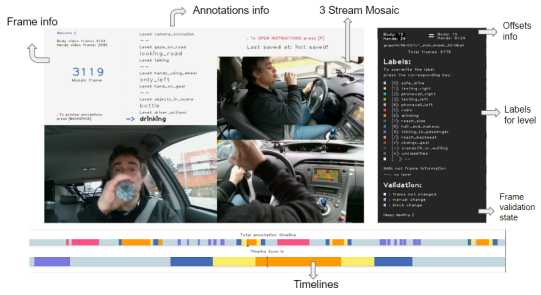


Figure 1: Annotation tool (Tato) interface.

The DMD[1] was created to fill the gap of a multi-purpose dataset for driver monitoring. It promises to support driver state estimation and the analysis of various aspects of driver monitoring; since it presents material of unfavourable driving situations. This dataset covers different fronts like driver fatigue detection, driver gaze estimation, hands-on-wheel tracking and what we made use of: distracted-driver-related material in video format.

All recordings were made from 3 in-vehicle perspectives with 3 cameras strategically positioned to capture the face, hands and body of the driver. Each camera offers 3 channels, they include RGB, infrared and depth information. A total of 37 volunteers participated in the creation of the DMD, where 27% were women and 73% were men and it has 40:45 hours of video material. For all these advantages, it has been demonstrated that this is a very all-rounded dataset and deserves to be explored; giving many insights for future work in driver monitoring systems development.

Continuing with the distraction approach of the DMD, it has been planned to have a list of 13 distraction-related activities performed by the driver, two of them belong to normal driving behaviour (Safe Driving and Standstill/Waiting). These are: Safe Driving, Texting(Right hand), PhoneCall(Right hand), Texting(Left hand), PhoneCall(Left hand), Operating the Radio, Drinking, Reach side, Hair&Makeup, Talking to Passenger, Change Gear, Reach Backseat, Standstill/Waiting. The activities are then annotated by frame intervals. The boundaries for each activity were defined, meaning that the beginning and end of each activity are established under an annotation criteria.

The annotations come in Video Content Description (VCD) format [2]. This is the first annotation toolset compliant with the ASAM OpenLABEL standard. VCD is defined with JSON schemas and supports spatio-temporal annotations for the description of objects, events, actions, contexts and relations from a scene or a data sequence.

[1]https://dmd.vicomtech.org/

4 EXPERIMENTS

Only a lite version of the DMD is available under request as of the date of this paper. We explore the possibilities of this dataset implementing models used for action recognition and see how they perform, as we wait for the full dataset to be public.

We examine two approaches for driver distraction detection in a real context scenario: an image-based and a video-based analysis. Both methods have been in discussion for action recognition tasks as presented in Section 2.

4.1 Dataset Preparation

The material used in this research contains the distraction-related material from 1 group; this includes the recordings from 5 subjects with a size of about 43,8 GB. The data used is in RGB format and belongs only from the side-view camera that captured the driver's body.

4.1.1 Labelling

For the annotation of the temporal distraction-related activities of the DMD, we have created a Temporal Annotation Tool (TaTo). It was developed in Python using the OpenCV library and creates annotations in VCD format. With this software, frame intervals of a video can be labelled from a list of classes the user defines. It shows a timeline for more efficient navigation through the video and better visualization of annotations (see Figure 1), supporting a frame-per-frame annotation and frame-block annotation using key-frames.

It is expected that the community uses this tool to better adequate the DMD to their research requirements if needed. TaTo can be found on the web[3], is open-source and can be adapted to other temporal annotation use cases. On the same repository, a Dataset Exploration Tool (DEx) can also be found.

[2]https://vcd.vicomtech.org/

[3]https://github.com/Vicomtech/DMD-Driver-Monitoring-Dataset/tree/master/annotation-tool

This tool offers to prepare the DMD material for training, that includes exporting data in videos or images and cutting material by frame intervals.

4.1.2 Sub-datasets and Data Splits

We have prepared the data to serve as input to our models, this involves resizing images from 1280×720 pixels to 224×224 and 112×112 pixels (only for Conv2DLSTM model), cutting material into specific frame length videoclips and splitting data by the following subset proportions: 80% for training and 20% for testing. We defined 3 sequence (action) lengths: 70 (2,35s), 50 (1,68s) and 30 (1,00s) frames. The resulting number of observations for each sub-dataset by data split is shown in Table 1.

Table 1: Number of videoclips and images per data split for each of the sub-datasets created for analysis.

Data split	30f	50f	70f	Images
Train	3772	2164	1623	89956
Valid	894	508	379	22488
Test	940	535	402	28111
Total	5606	3207	2404	140555

4.1.3 Classes

Because of the nature of some actions, most of their corresponding frame intervals in the videos are not sufficiently long to reach 70 frames, meaning that the driver did not take a minimum of 70 frames long to perform that activity. This causes that, when cutting frame intervals by the sequence length, some classes end up having a very small quantity of videoclips if not none. Besides, few classes already have little representation within this first group of the DMD. For these reasons, we only work with the first 9 activities. They are identified with "0" to "8" labels, taking into account the order as they were presented in Section 3. Figure 2 shows the resulting class distribution of each of the sub-datasets material.

4.2 Image Approach

We propose to use transfer learning to potentiate our models, using MobileNet and InceptionV3 as feature extractors. Models were trained with a batch size of 32 and a learning rate of $1e{-}3$ with Adam optimizer.

MobileNetV1-based & InceptionV3-based Models. We defined the architecture shown in Figure 3 which uses a MobileNetV1 model (Howard et al., 2017) and an InceptionV3 (Szegedy et al., 2015) model, pre-trained with ImageNet dataset, as feature extractor.

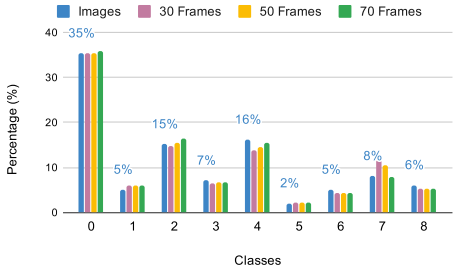


Figure 2: Data distribution by classes. Activity labels are: 0:Safe Driving, 1:Texting(Right), 2:PhoneCall(Right), 3:Texting(Left), 4:PhoneCall(Left), 5:Operating the radio, 6:Drinking, 7:Reach Side, 8:Hair&Makeup.

The first 20 layers (28 in total) of the base model are frozen and the rest are set as trainable to perform fine-tuning. The 2 fully-connected layers with 1024 filters have a dropout of 0,5 and the one with 512 filters has a dropout of 0,2.

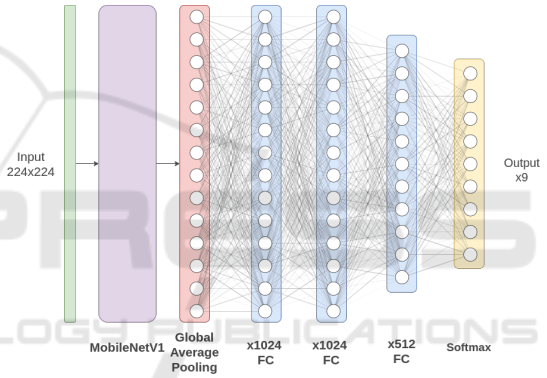


Figure 3: Model architecture based on MobileNetV1 for image classification. Same architecture for models based on InceptionV3.

4.3 Video Approach

In this research, we want to avoid additional preprocessing or previous manual feature extraction processes (optical flow calculation or skin-segmentation). Therefore, we only will consider end-to-end learning models. All these models were trained with a learning rate of $1e{-}3$ with Adam optimizer for all, 70-frames, 50-frames and 30-frames sequence lengths.

MobileNetV1 + LSTM Arrangement. To capture both spatial and temporal information from videos, we propose to use a time-distributed arrangement of MobileNets that will extract the features of each of the frames from the input video in parallel, followed by an LSTM for the temporal dependencies. Due to the good results and light-weight advantages of Mo-

bileNets, this model was included in the proposed action recognition architecture that is shown in Figure 4.

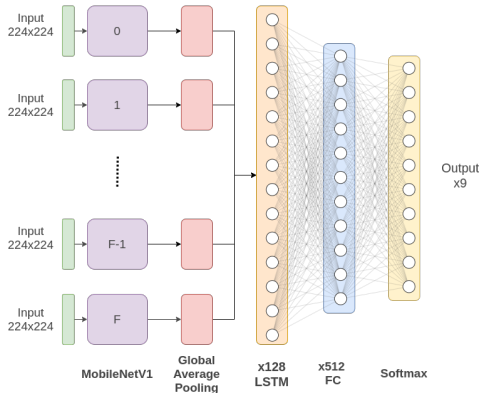


Figure 4: Model architecture with LSTM arrangement for video classification. F represents the sequence length (30, 50 or 70).

Conv3D-based Architecture. For a Conv3D-based architecture, we took as reference the one presented in (Tran et al., 2015). We reduced it to half the number of layers and filters in each layer. As a result, we ended up with the number of layers and filters specified in Figure 5.

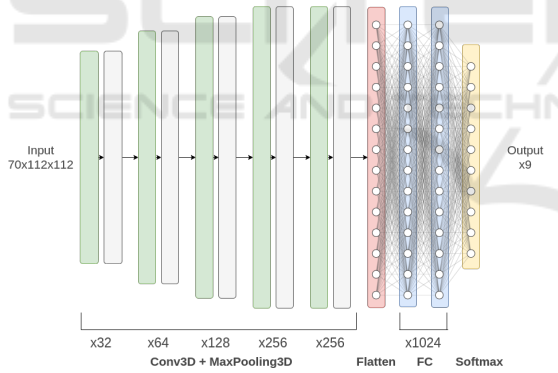


Figure 5: Model architecture based on reduced Conv3D for video classification.

Conv2DLstm-based Architecture. We brought Conv2DLSTM into consideration and constructed a simpler architecture based on this type of layer to compare, as is shown in Figure 6. These have convolutional structures in both the input-to-state and state-to-state transitions of an LSTM, making the input size of the cell to be a 3D tensor where the last two dimensions are spatial, representing the rows and columns (Shi et al., 2015) and taking into account spatial and temporal dimensions. This type of layer is denominated as Conv2DLstm in TensorFlow framework.

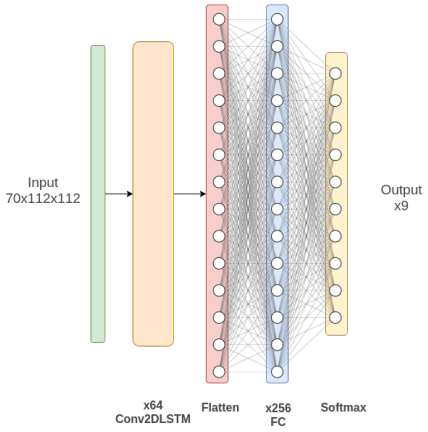


Figure 6: Model architecture based on simple Conv2DLSTM layer for video classification.

4.4 Performance Evaluation

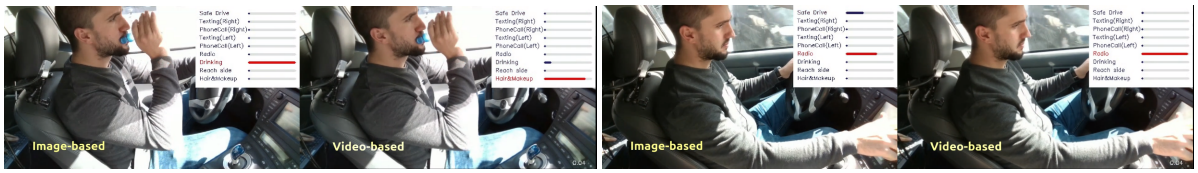
In order to know the computational effective time of prediction of each model, we calculated the elapsed time each model spent on predicting a unit of their correspondent classification target; this means that, for an image classification model, we measure the time it takes to predict one image and, for a video classification model, the time it takes to predict a video. To compare between models, we divided the resulted times from video-based models by their sequence length to have the time of prediction per frame. These tests were run on a server with a GPU Nvidia Tesla T4 with 16GB with no optimization methods.

Besides, we developed a script to perform "real-time" inference. It makes predictions on a specific video or a camera input directly, allowing the installation of this distraction detection models in a real scenario. The interface shows the prediction score of each of the classes through a bar graph (see Figure 7).

We implemented this system in a driving simulator to test the models, also made inference on some videos not included in any of the training data splits; this way, some qualitative appreciations on models performance could be done.

5 RESULTS

Dataset. When generating the sub-datasets, even though the original material is the same, when exporting to images and videoclips, images end up with a greater number of observations. As seen in Table 1, for the image sub-dataset there is a big difference in the number of examples compared to any of the videoclip sub-datasets. However, for video analysis,



(a) "Drinking" misclassification with "Hair&Makeup" activity. (b) "Operating the radio" activity classification.

Figure 7: Comparison of inference on video with image-based and video-based models.

Table 2: Action recognition inference accuracy and computational performance (milliseconds per frame) with image and video-based models. For video-based models, the total time of inference of one video is also presented.

Data	Model	Top-1 Accuracy. (%)	Computational effective time (ms/f)
30-frame videoclips	MobileNetV1 + LSTM	97,3%	96,73ms - (3,22ms/frame)
	Conv3D-Based	97,2%	62,48ms - (2,08ms/frame)
	Conv2DLSTM-Based	95,8%	93,48ms - (3,12ms/frame)
50-frame videoclips	MobileNetV1 + LSTM	96,0%	133,63ms - (2,67ms/frame)
	Conv3D-Based	95,6%	64,45ms - (1,29ms/frame)
	Conv2DLSTM-Based	95,8%	136,87ms - (2,78ms/frame)
70-frame videoclips	MobileNetV1 + LSTM	95,7%	197,69ms - (2,81ms/frame)
	Conv3D-Based	95,5%	**73,55ms - (1,05ms/frame)**
	Conv2DLSTM-Based	93,5%	171,13ms - (2,45ms/frame)
Images	MobileNetV1-Based	**99,5%**	41,79ms/frame
	InceptionV3-Based	99,3%	52,00ms/frame

this version of the DMD still offers a decent amount of videos, the smallest being 2404 videoclips.

Looking at Figure 2, is clear that the material distribution among classes is not balanced. To give an example, the activity "5.Operating the Radio" has a very low representation in the dataset compared to "0.Safe Driving". This is also reflected in test data split, meaning that some classes have more observations for testing than others.

Image Approach. Models for image recognition achieve better results in accuracy compared to video-based models, as can be seen in Table 2. The best accuracy was given by the MobileNet-based model with 99,5% on the test split. Also, the same architecture but with InceptionV3 as feature extractor achieved a very close result. This shows that transfer learning techniques work very well on this problem, meaning that the representation learnt from images from ImageNet help and potentiate the analysis on the DMD.

Due to a slight difference in performance, MobileNet has proven to be a lighter network that can predict faster than InceptionV3.

Video Approach. The best model, in terms of accuracy, is the MobileNet + LSTM model trained with 30-Frame videoclips, having an accuracy of 97,3%. The Conv2DLSTM was always last on the list, with the lowest accuracy results for the 3 sequence lengths.

Transfer learning have also enhanced our MobileNetV1 + LSTM model; this can be evidenced by comparing results with the more basic and convolutional-based neural network: conv2DLSTM. These two models, in the end, share the same architecture, a convolutional network module to process spatial information followed by an LSTM module for temporal information.

Video-based models have a lower accuracy compared with image-based models. Important to also notice that, as the sequence length increases, the accuracy decreases; this could indicate that the activities analysed are not very time dependent and that image-based algorithms could be enough for action recognition.

5.1 Qualitative Results

For video-based models, "Drinking" activity is often confused with "Hair&Makeup" (as illustrated in Figure 7a). Also this models seemed to be more confident when predicting the "Operating the Radio" activity than image-based (as illustrated in Figure 7b).

When performing "Drinking" and "Hair&Makeup" activities, drivers sometimes hold the corresponding objects at wheel level, this causes that the network misclassifies them as "Safe Driving" or "Text Left/Right".Is hard even for humans to determine the exact end of an activity and the beginning of another. The DMD had followed an

annotation criteria that define these limits and with which the model must be consequent with its predictions; but still, there are moments of ambivalence and doubt of when the network should start classifying a certain movement as a specific activity. Besides, some unintentional movements of drivers activate some classes, confusing the network.

5.2 Computational Performance

MobileNet is known to be more efficient since it has fewer parameters and computations. This behaviour is observed when compared with InceptionV3 models in image classification context; however, when added a LSTM layer for video classification, it becomes the least efficient video-based option among the architectures considered.

Image-based models might have better accuracy in this study, but video-based models had shown to be computationally more feasible. It is important to highlight that Conv3D-based models present a decrease of 2-4% in accuracy compared with the MobileNetV1-Based model; but at the same time, they have a major reduction of their computational effective time compared to the image-based models. It is clear that sacrificing a couple of points in accuracy and winning in computational performance is a great trade-off to consider. These are the reasons why video-based, specially Conv3D-based, models are considered the best option to implement in a real-context scenario.

6 DISCUSSION AND FUTURE WORK

This first experiment of distracted driver detection with the DMD opens possibilities of future research in this field, raises issues that deserve discussion and whose definition is crucial to further investigations:

- How much variation in human-pose within an action performance is considered being better analysed by video-based algorithms or image-based? Leaving us with the next point:
- Can an image-based algorithm generalize all variations of an activity? Meaning that the action itself implies changes in time, like body-pose variations. Can an image-based approach recognize that the activity of drinking includes two distinct body positions?: “lifting the bottle” and “holding the bottle up” on the head for drinking.
- How much distance must exist between two actions to be identified differently? Is the activity “Texting Left” divergent enough to “Texting Right”?. Also, the activities that share the same starting movement like “Drinking” and “Hair&Makeup”, which is “lifting”, could be hardly distinguishable at the beginning. This last takes us to the next issue:
- What actions can be decomposed into atomic actions like “Lifting”, which can be a sub-action of “Drinking” or “Hair&Makeup”; or “Holding object”, that could be extracted from “Texting-left/right” activities.
- Then, again, which actions are more appropriate to be analysed from videos and which from still images?. “Lifting” and “Leaving aside” are order or time-dependant. These two activities can not be differentiated from an image, it requires a sequence of frames to determine the action.
- What actions can be supported by object recognition or human-pose detection. Activities like “Drinking” can be separated from “Hair&Makeup” if the bottle presence on the scene is considered as well as the hair comb’s. Also, “Radio” activity might be better recognized if human-pose was an input.
- Wishing on taking this exercise to a larger scale, how many activities must be considered to accomplish a complete driver distraction monitoring?. Or, what changes the course of this exploration:
- Can driver monitoring be only based on 2 classes to discriminate between “Safe driving” and “Not safe driving”? Where the latter would cover the list of activities presented on this paper and any other that the driver performs that is outside limits of safeness.

To further explore the advantages of this dataset, the inclusion of the other 2 camera perspectives (Hands and face cameras) to the analysis and the depth and infrared channels of information, can be contemplated. The extraction of manual features before classification, pre-processing like the calculation of the optical flow or object detection, are some strategies we believe are worth trying and are workable with the DMD.

7 CONCLUSIONS

In this study, we have tested different model architectures of image and video classification for an action recognition task, including transfer learning techniques. The results obtained suggest that distraction

detection gets a better outcome when applying image-based solutions. However, when computational performance must be taken into account, video-based neural networks are more feasible, especially models with 3D convolutions. We have demonstrated the possibilities the DMD offers to the scientific community, extending the discussion for better solutions to action recognition problems applied to a driver monitoring context. Finally, we share some thoughts on some issues this line of research might encounter and propose some future work with the DMD.

ACKNOWLEDGEMENTS

This work has received funding from Basque Government under project AUTOLIB of the program ELKARTEK 2019.

REFERENCES

Abouelnaga, Y., Eraqi, H. M., and Moustafa, M. N. (2018). Real-time Distracted Driver Posture Classification. In *32nd Conference on Neural Information Processing Systems (NIPS 2018)*.

Baheti, B. V., Talbar, S., and Gajre, S. (2020). Towards computationally efficient and realtime distracted driver detection with mobilevgg network. *IEEE Transactions on Intelligent Vehicles*.

Borghi, G., Venturelli, M., Vezzani, R., and Cucchiara, R. (2017). Poseidon: Face-from-depth for driver pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chan, A., Saleem, K., Bhutto, Z., Memon, L., Shaikh, M., Ahmed, S., and Siyal, A. (2019). Feature fusion based human action recognition in still images.

Chen, J.-C., Lee, C.-Y., Huang, P.-Y., and Lin, C.-R. (2020). Driver behavior analysis via two-stream deep convolutional neural network. *Applied Sciences*.

Deo, N. and Trivedi, M. M. (2018). Looking at the driver/rider in autonomous vehicles to predict take-over readiness.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description.

Girish, D., Singh, V., and Ralescu, A. (2020). Understanding action recognition in still images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*.

Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiss, S., Voit, M., and Stiefelhagen, R. (2019). Drive & Act: A Multi-modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In *The IEEE International Conference on Computer Vision (ICCV)*.

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.

Rangesh, A. and Trivedi, M. M. (2018). Handynet: A one-stop solution to detect, segment, localize & analyze driver hands.

SAE International (2018). Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Technical report, SAE International.

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., and Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting.

StateFarm (2016). State Farm Distracted Driver Detection. Online source: https://www.kaggle.com/c/state-farm-distracted-driver-detection.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*.

Tran, D., Manh Do, H., Sheng, W., Bai, H., and Chowdhary, G. (2018). Real-time detection of distracted driving based on deep learning. *IET Intelligent Transport Systems*.

Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E., and Wang, F.-Y. (2019). Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE Transactions on Vehicular Technology*.

Yang, W., Wang, Y., and Mori, G. (2010). Recognizing human actions from still images with latent poses. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Yuen, K., Martin, S., and Trivedi, M. M. (2016). Looking at faces in a vehicle: A deep cnn based approach and evaluation. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE.

Zhang, Y., Cheng, L., Wu, J., Cai, J., Do, M. N., and Lu, J. (2016). Action recognition in still images with minimum annotation efforts. *IEEE Transactions on Image Processing*.