Video-based driver emotion recognition using hybrid deep spatio-temporal feature learning

Varma, Harshit, Ganapathy, Nagarajan, Deserno, Thomas

Harshit Varma, Nagarajan Ganapathy, Thomas M. Deserno, "Video-based driver emotion recognition using hybrid deep spatio-temporal feature learning," Proc. SPIE 12037, Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, 1203709 (4 April 2022); doi: 10.1117/12.2613118

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Video-based driver emotion recognition using hybrid deep spatio-temporal feature learning

Harshit Varma[a], Nagarajan Ganapathy[b], and Thomas M. Deserno[b]

[a]Indian Institute of Technology Bombay, Mumbai, India
[b]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

ABSTRACT

Road traffic crashes have become the leading cause of death for young people. Approximately 1.3 million people die due to road traffic crashes, and more than 30 million people suffer non-fatal injuries. Various studies have shown that emotions influence driving performance. In this work, we focus on frame-level video-based categorical emotion recognition in drivers. We propose a Convolutional Bidirectional Long Short-Term Memory Neural Network (CBiLSTM) architecture to capture the spatio-temporal features of the video data effectively. For this, the facial videos of drivers are obtained from two publicly available datasets, namely Keimyung University Facial Expression of Drivers (KMU-FED), a subset of the Driver Monitoring Dataset (DMD), and an experimental dataset. Firstly, we extract the face region from the video frames using the Facial Alignment Network (FAN). Secondly, these face regions are encoded using a lightweight SqueezeNet CNN model. The output of the CNN model is fed into a two-layered BiLSTM network for spatio-temporal feature learning. Finally, a fully-connected layer outputs the emotion class softmax probabilities. Furthermore, we enable interpretable visualizations of the results using Axiom-based Grad-CAM (XGrad-CAM). For this study, we manually annotated the DMD and our experimental dataset using an interactive annotation tool. Our framework achieves an F1-score of 0.958 on the KMU-FED dataset. We evaluate our model using Leave-One-Out Cross-Validation (LOOCV) for the DMD and the experimental dataset and achieve average F1-scores of 0.745 and 0.414 respectively.

Keywords: Driver Emotion, Emotion Recognition, Facial Expression, Video Processing, Convolutional Neural Network, Long Short-Term Memory, Classification, Deep Learning

1. INTRODUCTION

According to the World Health Organization, nearly 1.3 million people die from road traffic crashes, and more than 30 million people suffer non-fatal injuries. Moreover, road traffic crashes have become the leading cause of death for young people.[1] Various studies have demonstrated that emotions affect driving performance.[2–5] Emotions influence attention levels, aggression, and risk perception in drivers. Recently, in an exciting study, Dingus et al. found that driving while observably angry, sad, crying, and emotionally agitated increases the risk of a crash by 9.8 times.[6] Thus, automated recognition of a different emotional state is necessary. It can be used to improve human-computer interaction (HCI) and can also be incorporated in advanced driver assistance systems (ADAS).[7]

Several studies have been reported to differentiate and recognize emotional states in drivers.[8] However, most of these approaches are evaluated for laboratory environments.[9] Recently, deep learning-based techniques have achieved state-of-the-art results on popular emotion recognition benchmarks like CK+ for in-the-wild emotions.[10] However, these datasets are collected in a laboratory environment, with emotions being acted out, good illumination conditions, fixed head poses, and little to no occlusions. Models performing well on these datasets are not generalizable to real-world in-the-wild driving scenarios with adverse illumination conditions resulting from sunlight, poorly lit roads, and garages.[9]

Further, in driving, the head pose varies significantly with conditions such as talking with passengers, interacting with the vehicle dashboard, and viewing different mirrors.[9,11] Moreover, in contrast to the popular benchmarks dataset, in a real-world driving scenario, the emotions experienced by the driver are subtle and exist for variable durations. The facial occlusions like sunglasses, hats, mobile phones, water bottles, and the driver's

Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications,
edited by Thomas M. Deserno, Brian J. Park, Proc. of SPIE Vol. 12037,
1203709 · © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2613118

own hands make the task even more challenging.[11] Furthermore, most of these image-sequence datasets only have a single emotion class and focus on sequence-level classification. But in a real-world scenario, multiple emotions may arise in the same time duration, motivating the need for frame-level classification instead of sequence-level classification.[11,12]

Limited studies have been reported on general-purpose facial expression recognition with popular benchmarks datasets, which depend exclusively on spatial information available in the video/image sequence data.[12] However, in real-world driving scenarios, considering temporal context may improve the recognition rate. In conditions such as sequences with facial regions occluded, the image-based approach may provide erroneous results. However, suppose we include temporal context along with spatial information. In that case, we have higher chances of classifying the occluded frames correctly by utilizing the information contained in the previous un-occluded frames.

Thus, in this study, we proposed an approach to improvising the recognition of the driver's emotional states using hybrid deep Spatio-temporal feature learning. The proposed approach focuses on video-based categorical emotion recognition in drivers at frame level using the Face Alignment Network (FAN) and convolutional bidirectional Long Short-Term Memory neural network (CBiLSTM).[12,13] The video sequences of drivers' emotional states are obtained from two public databases and an experimental dataset. To the best of the author's knowledge, there are no publicly available datasets collected in a real-world driving environment with natural emotion annotations. Our research questions (RQ) are:

- RQ1: Can drivers' emotional states be detected accurately using the proposed methodology?
- RQ2: Can the proposed methodology be used to detect emotional states in multiple cameras (color and Infrared)?

2. MATERIALS AND METHODS

The pipeline of the proposed methodology to recognize driver's emotional states is shown in figure 1. In the first stage, the videos are preprocessed. The preprocessing involves downsampling of the video to 10 frames per second (FPS) to reduce computational complexity. The clips are applied to the Face Alignment Network (FAN) to detect the face in the second stage. The facial regions are further segmented and divided into clips. The segmented clips are employed in the CBiLSTM network for hybrid deep spatio-temporal feature learning in the third stage. The Leave-One-Out Cross-Validation (LOOCV) technique is used to evaluate the robustness of the proposed approach.

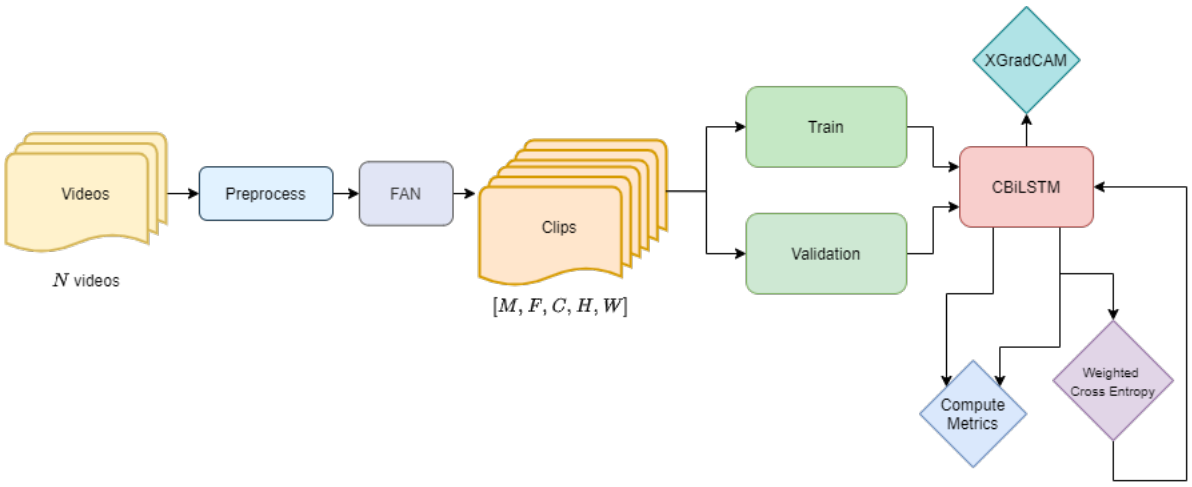

Figure 1. Overall proposed algorithm pipeline: N is the total number of videos, M is the total number of clips, F is the number of frames in each clip, C is the number of channels, and $H \times W$ the frame dimensions

2.1 Datasets

The Keimyung University Facial Expression of Drivers (KMU-FED) dataset consists of 12 participants and 1100 NIR frames, divided into 110 clips of 10 frames each. The clips are annotated for six basic emotions: *anger, disgust, fear, happiness, sadness, surprise*.[14] The DMD has 41 hours of video data and 31 participants; however, only five participants have been released and are considered in this study.[11] The videos have occlusions from objects such as glasses, water bottles, and hands. They also have varied illumination conditions and head pose variations. It has not been annotated for emotions. In addition to the above datasets, we have recorded 13 videos with 7 participants: 2 videos for 6 participants and one video from the last participant. The videos are sampled at 30 frames per second using Intel RealSense cameras.

2.2 Annotation and Emotion Classes

Due to limited emotion annotations in DMD and the experimental database, we created a command line based interactive annotation tool to annotate the DMD manually and our experimental dataset for three emotion classes: *happy*, *neutral*, and *bothered*.

In-the-wild categorical emotion recognition studies have considered *anger, disgust, fear, happiness, sadness, surprise*, and *neutral* as emotion classes. However, most of these emotional states do not frequently occur while driving. The unreleased DEFE dataset considers the emotions: *anger*, *happiness*, *neutral*.[15] The dataset collected by Ma et al. considers the emotions: *happy*, *bothered*, *concentrated*, *confused*.[16] They assume *bothered* to contain *anger* and *disgust* as well. In our case, we consider the emotion classes: *happy*, *bothered*, and *neutral*. We assume the *bothered* emotional state to contain *anger, disgust, fear, sadness*, and *confused* states.

2.3 Preprocessing

To reduce computational complexity, we reduce the frame rates of the video datasets from 30 to 10 FPS. We then detect and crop out the faces present in each frame. We then divide these cropped frames into fixed-sized clips of 32 frames; each frame resized to 224×224. Each clip consists of a consecutive sequence of frames, creating a segment of the original video.

2.4 Face and Landmarks Detection

We use FAN for extracting facial landmarks.[13] The face bounding box is then created using the landmark coordinates, using the range of the coordinate values along each direction. We dilate this bounding box by 10% along each side to include some background context.

2.5 The CBiLSTM Network

Our main framework is a CBiLSTM network that accepts a 32-sized video clip as the input. Each frame in this clip is first encoded using a CNN architecture. We use the lightweight SqueezeNet model, as for a similar task of driver gaze detection, it demonstrated better performance and localization when compared to other standard CNN models.[17,18] The CNN-encoded frames are then passed on to a two-layered BiLSTM network to incorporate the temporal information contained in the clip (see figure 2). For each time step, the output of the BiLSTM network is passed through a final fully connected layer that yields the class-wise softmax probabilities.

2.6 Visualization

To enable the interpretability of the results and to analyze the model's performance, we utilize the XGrad-CAM framework, using the last layer of the SqueezeNet model.[19] On a related task, Grad-CAM was shown to provide interpretable insights into the model's performance.[18,20] We use XGrad-CAM, which is an enhanced version of Grad-CAM in terms of conservation and sensitivity.

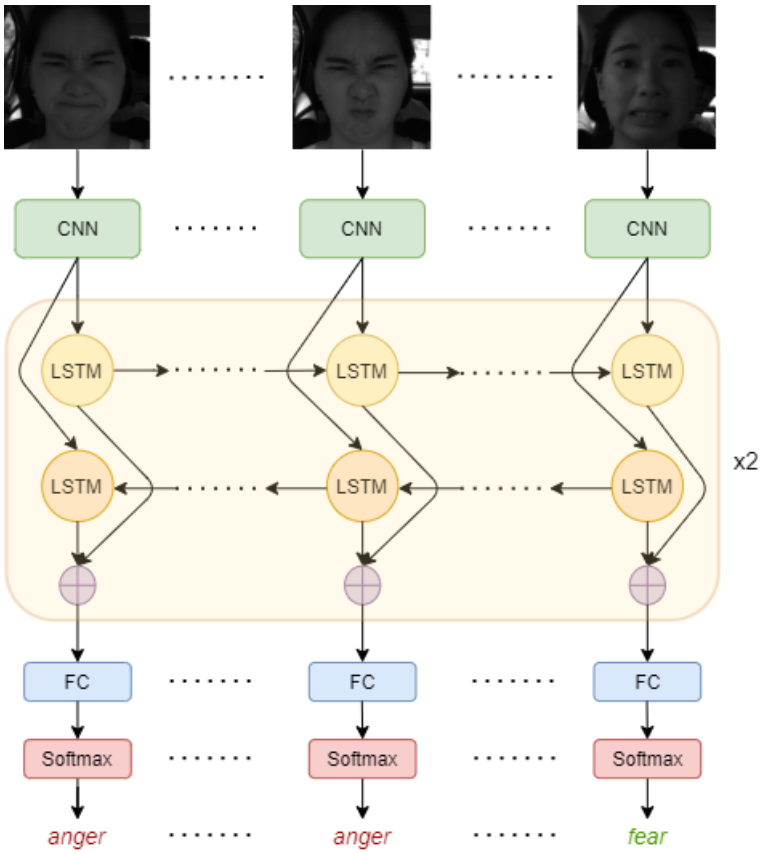

Figure 2. The CBiLSTM network

3. RESULTS

3.1 Performance on the KMU-FED Dataset

We achieve a validation accuracy of 95.8 and an F1-score of 0.958, considering six basic emotion classes: *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*. We observe that the highest confusion is between fear and sadness. Qualitatively, through XGradCAM visualizations shown in figure 4, we observe that our model learns discriminative features for the six emotions, effectively localizing different regions of the face. We train our network for 27 epochs, with a batch size of 8, using a learning rate of 10^{-4}. We fix the clip size as ten and the LSTM hidden dimension as 512.

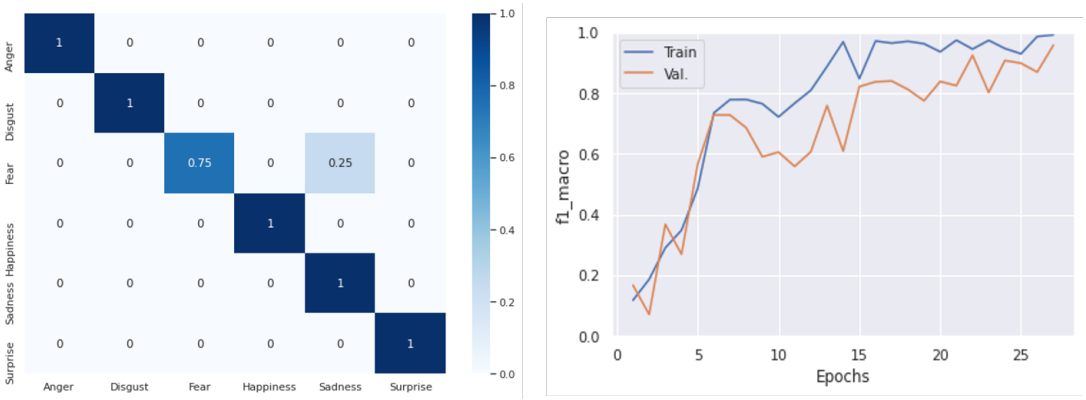

Figure 3. The obtained validation confusion matrix and the average F1-score variation

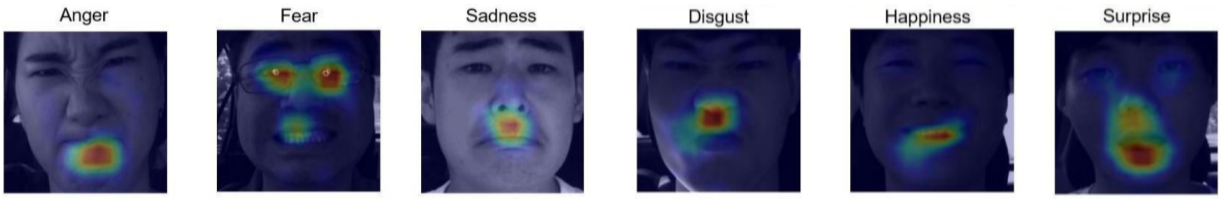

Figure 4. The obtained XGrad-CAM visualizations for each emotion category on the KMU-FED dataset

3.2 Performance on the DMD Dataset

For this dataset, we evaluate our models using Leave-One-Out-Cross-Validation (LOOCV), as this dataset has only five videos. We achieve individual F1-scores of 0.988, 0.618, 0.650, 0.731, 0.740 and an average F1 score of 0.745. This demonstrates that even when the data size is small, our model remains effective. The confusion matrices in figure 5 show that the largest confusion exists between the *neutral* and the *bothered* states. We train our network for 20 epochs, with a batch size of 8, using a learning rate of 10^{-4}. We fix the clip size as 32 and the LSTM hidden dimension as 256.

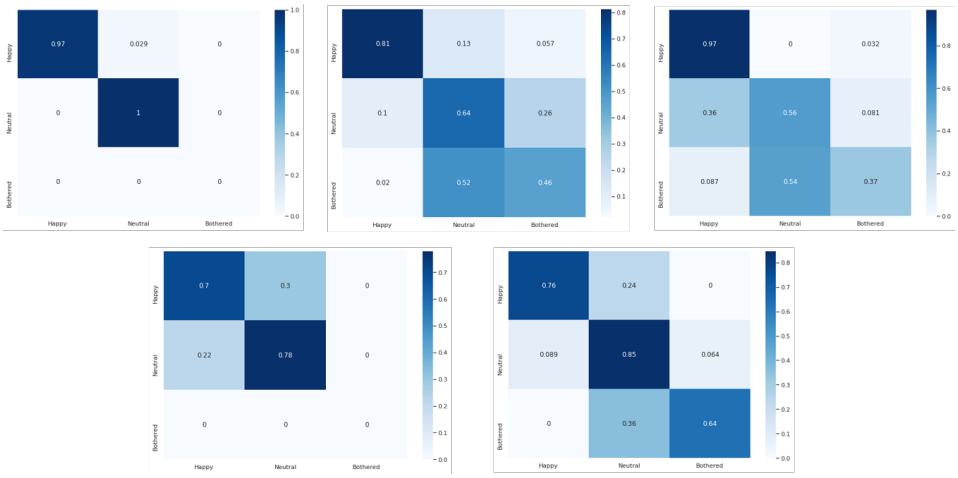

Figure 5. The obtained validation confusion matrices for each of the five videos

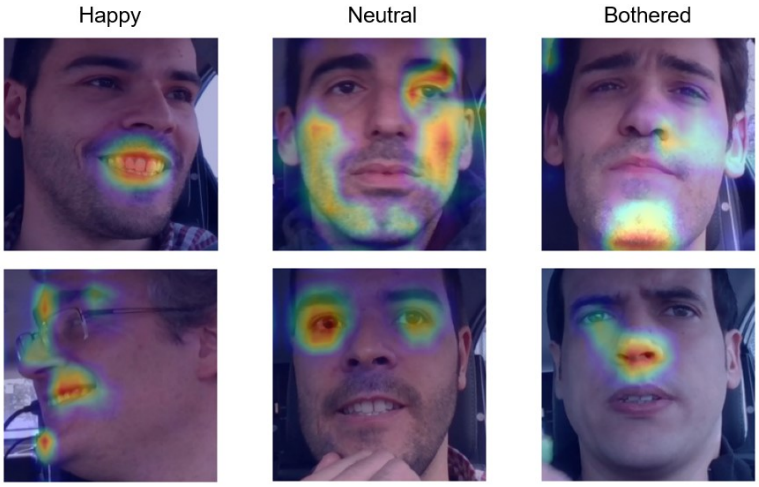

Figure 6. The obtained XGrad-CAM visualizations for each emotion category on the DMD dataset

3.3 Performance on the Experimental Dataset

The results on this data are inferior to those achieved in the other two datasets since this data is highly imbalanced with the following distribution of frames: *happy* - 750 frames, *neutral* - 10,537 frames, *bothered* - 999 frames. We try to mitigate this by subsampling the *neutral* class and using a weighted loss. Although this improves the performance, the final results are still not satisfactory. We evaluate our models using a variant of LOOCV. At a time, we leave out all videos belonging to an individual subject and train on the remaining videos, and then validate on the left-out videos. We achieve individual F1-scores of 0.260, 0.386, 0.362, 0.344, 0.844, 0.320, 0.382, and an average F1 score of 0.414. We train our network for 20 epochs, with a batch size of 8, using a learning rate of 10^{-4}. We fix the clip size as 32 and the LSTM hidden dimension as 256.

4. NEW/BREAKTHROUGH WORK

In this work, we proposed a novel FAN-based CBiLSTM network architecture for dynamic facial expression recognition in drivers in the smart-vehicle environment. The proposed methodology is robust, reliable, and applicable for multiple camera sensors (color and infra-red) (see figures 4 and 6). Our work effectively locates the emotional cues of the face associated with various emotional states for accurate detection (see figures 4 and 6). As seen in section 3.2, our method performs considerably well even in a low-resource setting. Furthermore, our method could be a useful tool to assess emotional states and can be extended to real-life environments.

5. CONCLUSIONS

In this work, we proposed a FAN-based CBiLSTM based architecture that effectively captures the spatio-temporal features of the video data (RQ1). Furthermore, we showed that the proposed methodology could be used for multispectral cameras (RQ2) with the experiment. We demonstrated the effectiveness and robustness of our model on two different datasets: KMU-FED, DMD, and on our experimental videos. We offer a fast, accurate, and applicable solution for various camera sensors with the proposed methodology.

In future work, we wish to explore better ways of handling the class imbalance. Further, the model architecture can be improved in multiple ways by incorporating attention mechanisms, utilizing better loss functions, like the ArcFace loss, and considering a hierarchical classification setting with the first stage being a *neutral*/non-*neutral* binary classifier.[21] Eventually, the proposed approach can be used to detect the emotional states of the patients and participants in smart-home environments.[22]

REFERENCES

[1] World Health Organization, "Road traffic injuries," (2021). https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, last accessed on 2021-09-05.

[2] Jeon, M., Yim, J.-B., and Walker, B. N., "An angry driver is not the same as a fearful driver: effects of specific negative emotions on risk perception, driving performance, and workload," in [*Proceedings of the 3rd international conference on automotive user interfaces and interactive vehicular applications*], 137–142 (2011).

[3] Steinhauser, K., Leist, F., Maier, K., Michel, V., Pärsch, N., Rigley, P., Wurm, F., and Steinhauser, M., "Effects of emotions on driving behavior," *Transportation research part F: traffic psychology and behaviour* **59**, 150–163 (2018).

[4] Hu, T.-Y., Xie, X., and Li, J., "Negative or positive? the effect of emotion and mood on risky driving," *Transportation research part F: traffic psychology and behaviour* **16**, 29–40 (2013).

[5] Magaña, V. C., Scherz, W. D., Seepold, R., Madrid, N. M., Pañeda, X. G., and Garcia, R., "The effects of the driver's mental state and passenger compartment conditions on driving performance and driving stress," *Sensors* **20**(18), 5274 (2020).

[6] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., and Hankey, J., "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proceedings of the National Academy of Sciences* **113**(10), 2636–2641 (2016).

[7] Brookhuis, K. A., De Waard, D., and Janssen, W. H., "Behavioural impacts of advanced driver assistance systems–an overview," *European Journal of Transport and Infrastructure Research* **1**(3) (2001).

[8] Roidl, E., Frehse, B., and Höger, R., "Emotional states of drivers and the impact on speed, acceleration and traffic violations—a simulator study," *Accident Analysis & Prevention* **70**, 282–292 (2014).

[9] Nowara, E. M., Marks, T. K., Mansour, H., and Veeraraghavan, A., "Near-infrared imaging photoplethysmography during driving," *IEEE Transactions on Intelligent Transportation Systems* (2020).

[10] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I., "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in [*2010 ieee computer society conference on computer vision and pattern recognition-workshops*], 94–101, IEEE (2010).

[11] Ortega, J. D., Kose, N., Cañas, P., Chao, M.-A., Unnervik, A., Nieto, M., Otaegui, O., and Salgado, L., "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in [*European Conference on Computer Vision*], 387–405, Springer (2020).

[12] Du, G., Wang, Z., Gao, B., Mumtaz, S., Abualnaja, K. M., and Du, C., "A convolution bidirectional long short-term memory neural network for driver emotion recognition," *IEEE Transactions on Intelligent Transportation Systems* (2020).

[13] Bulat, A. and Tzimiropoulos, G., "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in [*Proceedings of the IEEE International Conference on Computer Vision*], 1021–1030 (2017).

[14] Jeong, M. and Ko, B. C., "Driver's facial expression recognition in real-time for safe driving," *Sensors* **18**(12), 4270 (2018).

[15] Li, W., Cui, Y., Ma, Y., Chen, X., Li, G., Zeng, G., Guo, G., and Cao, D., "A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," *IEEE Transactions on Affective Computing* (2021).

[16] Ma, Z., Mahmoud, M., Robinson, P., Dias, E., and Skrypchuk, L., "Automatic detection of a driver's complex mental states," in [*International Conference on Computational Science and Its Applications*], 678–691, Springer (2017).

[17] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K., "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size," *arXiv preprint arXiv:1602.07360* (2016).

[18] Vora, S., Rangesh, A., and Trivedi, M. M., "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Transactions on Intelligent Vehicles* **3**(3), 254–265 (2018).

[19] Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., and Li, B., "Axiom-based grad-cam: Towards accurate visualization and explanation of cnns," *arXiv preprint arXiv:2008.02312* (2020).

[20] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*Proceedings of the IEEE international conference on computer vision*], 618–626 (2017).

[21] Deng, J., Guo, J., Xue, N., and Zafeiriou, S., "Arcface: Additive angular margin loss for deep face recognition," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 4690–4699 (2019).

[22] Deserno, T. M., "Transforming smart vehicles and smart homes into private diagnostic spaces," in [*Proceedings of the 2020 2nd Asia Pacific Information Technology Conference*], 165–171 (2020).