A MULTI-CAMERA DEEP NEURAL NETWORK FOR DETECTING ELEVATED ALERTNESS IN DRIVERS

John Gideon[⋆], *Simon Stent*[†], *and Luke Fletcher*[†]

University of Michigan, Ann Arbor, MI, USA[⋆]
Toyota Research Institute, Cambridge, MA, USA[†]

ABSTRACT

We present a system for the detection of elevated levels of driver alertness in driver-facing video captured from multiple viewpoints. This problem is important in automotive safety as a helpful feedback signal to determine driver engagement and as a means of automatically flagging anomalous driving events. We generated a dataset of videos from 25 participants overseeing an hour each of driving sequences in a simulator consisting of a mixture of normal and near-miss driving events. Our proposed system consists of a deep neural network which fuses information from three driver-facing cameras to estimate moments of elevated driver alertness. A novel aspect of the system is that it learns to actively re-weight the importance of camera inputs depending on their content. We demonstrate that this approach is not only resilient to dropped or occluded frames, but also has significantly improved performance compared to a system trained on any single stream.

Index Terms— affective computing, autonomous driving, deep learning, multi-camera systems

1. INTRODUCTION

The path to fully autonomous driving will, counter-intuitively, involve a significant increase in the research and development of driver-facing technology [1, 2, 3]. One important aspect of this technology is the estimation of a driver's state, which includes factors such as drowsiness, attention and cognitive load [4, 5]. Estimating this state accurately and efficiently is critical to a safe human-machine interface in the intermediate Level 2 and Level 3 [6] stages of autonomous driving.

In this work, we consider the specific problem of detecting instances of elevated alertness in driver-facing video. This is of interest for two reasons. Firstly, it can provide an important cue for a partially automated vehicle to understand if a driver is in an able state to take over control. Secondly, it might allow for improved mining of driving datasets for events with high uncertainty that provoke surprise, such as near-misses, which are otherwise difficult to identify. Locating scenarios with high uncertainty among the large volume of more predictable scenarios is important, since focusing training and testing of perception and planning algorithms on them can help to generate robust models more efficiently [7].

While a number of driver-facing datasets have been published in previous work [8], there is a notable lack of datasets inducing affective information such as surprise. While some affective datasets [9, 10] do capture this information, they are often conversational in nature and do not necessarily reflect the manifestation of affect in driving scenarios. In this work, we introduce the Toyota Research Institute Affective Driving (TRIAD) dataset, which consists of reactions from 25 individuals observing a set of 90 driving scenarios. Compared to previous work, we incorporate multiple cameras, on the basis that sensory redundancy from different viewpoints can lead to improved system robustness and therefore safety.

The problem of surprise detection in video has been well-studied [11, 12]. The classical approach involves the extraction and use of hand-coded facial action units as the basis for building models [13]. However, recent progress in deep convolutional neural networks has allowed the learning of richer image and video representations for facial surprise detection [14]. We build on recent work that demonstrates the efficacy of end-to-end learning for emotion recognition [15]. However, we take this work a step further with the introduction of a novel technique to dynamically combine together different camera viewpoints for the task. Our approach demonstrates the value of using seemingly redundant information from multiple camera streams, not only for this task but for any such task where robust performance is critical.

The dataset and experimental models used in this work are available for download at [URL to be provided].

2. TRI AFFECTIVE DRIVING (TRIAD) DATASET

2.1. Dash-cam Video Selection and Preprocessing

To create our dataset, we first acquired a collection of dash-cam videos from YouTube, featuring a diverse variety of routine and eventful driving scenarios. We filtered the videos to ensure that they were front-facing and of high resolution, were consistently right-hand drive, and did not exhibit excessive camera-shake (to prevent inducing motion sickness). Ambient traffic noise was added to all event videos missing

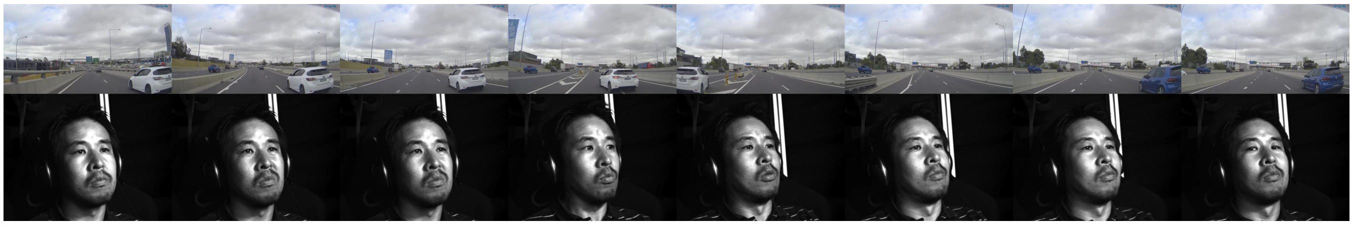

Fig. 1. Top: sub-sampled frames from a randomly selected event video containing a sudden cut-in from a lead vehicle (visible around the fourth frame from the left). Bottom: associated frames from one of the driver-facing reaction videos to this event.

audio tracks. A loud honking sound was superimposed on the audio during certain events when it made sense in the context of the video and made the experience more surprising. In total, 90 videos were curated of which 20 were deemed eventful or surprise-inducing and 70 were routine or uneventful driving scenarios. Among the events, moments of high alertness such as a near-miss or a road accident occurred on average 14 seconds into the video, giving viewers sufficient temporal context to understand what was happening.

2.2. Driving Simulator Setup

We set up a driving simulator for participants to view the collected dash-cam videos in a vehicle cockpit and to record their reactions. The cockpit consisted of an adjustable seat mounted to a frame with a privacy curtain for better immersion, a large monitor on which to display the video sequences and a force-feedback wheel and pedal set for driver interventions. To record facial reactions, we used three cameras spaced evenly across the cockpit dashboard, capturing 968×728 pixel monochrome video at 15Hz.

The controls and sensors were connected to a central computer to synchronize recordings and playback of the dash-cam videos. Global timestamps were recorded for every sensor frame to ensure data synchronization. A supervisor monitored the experiment remotely to ensure smooth operation.

2.3. Data Collection and Preprocessing

We captured data from 25 participants aged between 18-64. Capture sessions for participants lasted one hour each. During the session, participants were instructed to monitor a partially autonomous or Level 2 vehicle [6]. They would be shown videos of the car driving autonomously, but were required to take over using the wheel and pedals if they believed it was necessary. Over the course of the experiment, each participant observed the complete set of dash-cam videos in a random order. Sample frames from a typical event video and reaction video are shown in Figure 1. Several subtasks were interleaved among the videos to improve participant focus and provide further metadata such as reaction times and gaze tracking.

Following the capture, we sorted and aligned all sensor recordings to extract segments corresponding to each of the dash-cam videos. For the remainder of this paper, we focus only on the three driver-facing camera recordings. Event videos were reduced to 8 second clips containing the event. For non-event videos, a random 8 second clip was extracted. This created the final dataset: 15 hours of video footage, consisting of 2,250 (90 sequences × 25 participants) × 8 second video clips from 3 different cameras.

Face-aligned frames were extracted from each camera clip using OpenFace [16]. This resulted in 120 unique frames per clip. However, face extraction was not always successful, as the view was sometimes obstructed. In such cases we interpolated from the closest possible frame. We resized the face-aligned frames to 48×48 pixels to match the size of the FER2013 dataset, used later for pre-training our system.

2.4. Data Annotation

Annotation was performed by a group of five annotators on driver reaction videos to the 20 events. Each annotator watched 20×25 reaction videos and provided their estimation of the driver's level of surprise by turning the steering wheel in response. Their feedback was continuous both in time and in magnitude, with labels ranging from 0.0 (not turning the wheel) to 1.0 (turning the wheel 90 degrees or more). Videos were grouped by subject and were shown in random order with 2 second breaks in between. This resulted in a continuous label of surprise for each frame of the event videos by averaging across annotator ratings. The non-event videos were not labeled and were assumed to have zero surprise. Binary, clip-level labels were determined based on the maximum rating over the course of the video.

2.5. Dataset Splits

The reaction video dataset was divided across event videos in three splits according to the order in which they were viewed by participants. A participant's reaction to the 2 first events and 8 first non-events make up the test set. These reactions were not conditioned by having observed many other videos, and were therefore considered the closest to the distribution of genuine reactions. Reactions to the next 3 events and 9 non-events make up the validation set. The training set was composed of the reactions to the final 15 events and 53 non-events.

3. METHODS AND RESULTS

The goal of our experiments was to construct a system to detect elevated alertness in driver-facing video, regardless of available camera views. Our initial system used the network structure of Breuer *et al.* [15], as detailed in Figure 2. In keeping with [15], we applied dropout [17] after the final convolutional layer (0.25) and after the first dense layer (0.5).

3.1. Pre-training on FER2013 Dataset

To provide better generalisation performance, we opted to pre-train our baseline model on the FER2013 dataset [18], which was introduced as part of a facial expression recognition challenge and includes 35,887 face images cropped to 48×48 pixels found through Google image searches over seven emotion-related keywords. These keywords were combined with words related to gender, age, or ethnicity to generate a more diverse set of subjects than available in TRIAD.

Initially, the system was trained on the original FER2013 7-class classification task using balanced batches of size $1,024$ with early stopping after 50 epochs. We used the Adam optimizer [19] with a learning rate of 10^{-3} and a decay of 10^{-5}. Next, we discarded the 7-way output layer and replaced it with a binary output to represent surprise / non-surprise. Dataset annotations were similarly converted into binary surprise / non-surprise. Training was continued until convergence, after approximately another 50 epochs. This resulted in a system with 0.914 unweighted average recall (UAR) when validating on the FER2013 private test set. UAR is an average of each class accuracy and is therefore unbiased by the class imbalance.

To improve model generalisation we made heavy use of random data augmentation throughout training, including horizontal flipping, rotations between -10 and 10 degrees, horizontal and vertical shifts of up to 5%, shear of up to 5%, zoom of up to 5% and gamma corrections between 0.5 and 2.

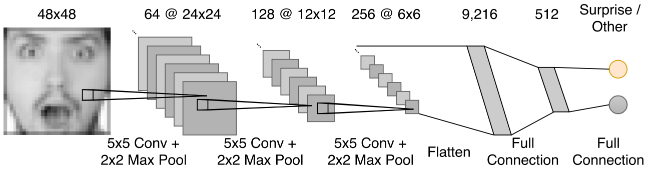

Fig. 2. The initial network pre-trained with the FER2013 dataset consisting of three convolutional layers followed by two dense layers. Each layer was followed by a rectified linear unit activation, besides the final layer which was followed by a two channel softmax function.

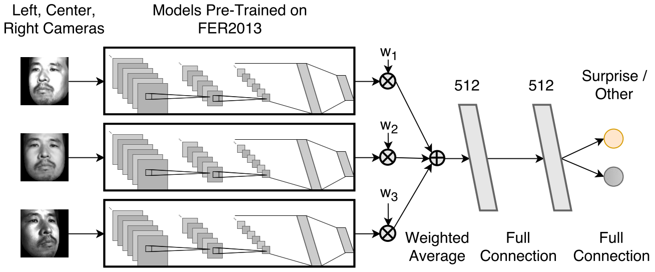

Fig. 3. The network used to predict elevated driver alertness at the frame-level. The three camera view models were merged using a weighted scheme, depending on camera validity.

3.2. Frame-Level Modeling

We next used the pre-trained model to classify frames containing elevated driver alertness within the TRIAD dataset. The goal of this classification was to fit a model to map the three face-aligned frames from the cameras to the continuously annotated levels of driver surprise. A particular frame was considered to contain elevated alertness if the mean of the annotator ratings was above 0.25. This threshold was determined by observing the distribution of surprise over the event annotations. To this end, we used the system shown in Figure 3, in which face-aligned frames were first fed through three identical FER2013 pre-trained sub-networks with shared weights. Feature outputs from the three sub-networks were then combined using one of three methods:

1. *Merge by valid*: Sub-network weights w_1, w_2 and w_3 in Figure 3 were set to 0 if the associated camera view was deemed invalid (determined by the failure of face detection and alignment). The remaining weights were normalized to sum to 1 over the valid cameras.

2. *Merge by mean confidence*: Identical to *merge by valid*, except that valid weights were further weighted by a confidence value. Confidence values were learned by the addition of a 512-unit fully connected hidden layer with sigmoid activation branched from the output of each sub-network, and were normalized to sum to 1 over the valid cameras.

3. *Merge by max confidence*: Identical to *merge by mean confidence*, except sub-network features were merged using an element-wise maximum rather than mean.

Additional experiments were also performed using single camera streams. In each case, this produced a combined 512-dimension vector representation that was followed by another 512 dense layer. Finally, a dense softmax output layer was used to classify between surprise/non-surprise.

The same learning scheme was adopted as in pre-training. We optimized over a ground truth lag between 0 and 15 frames (up to one second) to compensate for the annotator reaction times and found a 6 frame (400 ms) delay to be optimal. The system achieved maximum validation performance of 0.912 UAR when using the center camera and the mean and max confidence merging schemes, as seen in Table 1.

Table 1. Validation results on TRIAD of frame-level experiments using different camera inputs and weighting approaches.

Method	Validation UAR
Left camera	0.882
Center camera	0.912
Right camera	0.851
Merge by valid	0.901
Merge by mean	0.912
Merge by max	0.912

3.3. Temporal Modeling

Finally, to incorporate the temporal nature of the problem, we constructed a system to classify the 8-second video snippets provided by TRIAD. A snippet was considered to contain elevated alertness if it contained a frame with a surprise label above 0.25. We used the frame-level representation of the final dense 512 layer output of Figure 3 as the input to this temporal model. This provided a 512-dimension vector for each of the 120 frames in a given sequence.

The temporal network, shown in Figure 4, consisted of three convolutional layers, followed by a Gated Recurrent Unit (GRU) [20] with dimensionality 64 and additional fully connected layers. All layers again used ReLU activation, except for the softmax output. The convolutional layers were found to empirically improve classification performance versus a GRU alone. We used the same learning scheme as before, but trained only over the new layers of the network, keeping the frame-level representation fixed.

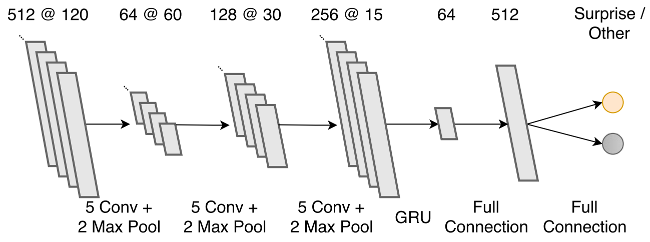

Fig. 4. The network used to predict elevated driver alertness in 8-second video snippets. The input to this network is the 512 dimensional representation learned over each of the 120 frames.

Table 2. The test results of the video-level experiments.

	Cameras at Test (UAR)		
Training Method	One	Two	Three
Left camera	0.766	-	-
Center camera	0.720	-	-
Right camera	0.792	-	-
Merge by valid	0.757	0.815	0.854
Merge by mean	0.773	0.819	0.823
Merge by max	**0.862**	**0.866**	**0.897**

4. RESULTS AND DISCUSSION

We evaluated the performance of six variations of the final video-level system: training with the three single cameras (left, center, or right) and training with all cameras using three different weighting and merging methods. Each system was evaluated by testing on data with just one, two, or all three cameras available. The results on the held-out test set are shown in Table 2.

The results demonstrate that our method of learning to re-weight the camera views by confidence and then taking the element-wise maximum to merge the representations produces the system which is most robust to different test conditions. The highest performance of 0.897 UAR is achieved when all three cameras are available at test time. However, the method is also robust to missing cameras and results in 0.862 and 0.866 UAR with only one or two cameras available at test time, respectively.

5. CONCLUSIONS

This work introduced the TRIAD dataset, a collection of driver facial reactions to dash-cam videos containing instances of elevated alertness. We described a system for the detection of such instances, which was designed to be robust to changes in the available number of camera views at test time be actively re-weighting the importance of different camera feeds depending on their content. This is useful when inputs are sometimes unreliable, as is the case in driver-facing sensing where cameras can be occluded by various driver actions. It is also beneficial when transferring to a system with a different camera setup, including fewer or greater numbers of cameras.

While we focused on detecting discrete events of elevated driver alertness, there are many further avenues to explore in the TRIAD dataset. Due to the continuous nature of the annotations, it should be possible to train a model which outputs driver alertness that is continuous both in time and rating. We also plan to explore how this system generalizes to the detection of anomalous events in naturalistic driver-facing video datasets.

6. REFERENCES

[1] Natasha Merat, A Hamish Jamson, Frank CH Lai, Michael Daly, and Oliver MJ Carsten, "Transition to manual: Driver behaviour when resuming control from a highly automated vehicle," *Transportation Research Part F: Traffic Psychology and Behavior*, vol. 27, pp. 274–282, 2014.

[2] Dehlia Willemsen, Arjan Stuiver, and Jeroen Hogema, "Automated driving functions giving control back to the driver: A simulator study on driver state dependent strategies," in *24th International Technical Conference on the Enhanced Safety of Vehicles*, 2015, number 15-0109.

[3] Edmund Wascher, Stephan Getzmann, and Melanie Karthaus, "Driver state examination – Treading new paths," *Accident Analysis & Prevention*, vol. 91, pp. 157–165, 2016.

[4] Fabian Friedrichs and Bin Yang, "Camera-based drowsiness reference for driver state classification under real driving conditions," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*. IEEE, 2010, pp. 101–106.

[5] Kathrin Zeeb, Axel Buchner, and Michael Schrauf, "Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving," *Accident Analysis & Prevention*, vol. 92, pp. 230–239, 2016.

[6] SAE On-Road Automated Vehicle Standards Committee et al., "Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems," *SAE Standard J3016*, pp. 01–16, 2014.

[7] David A Cohn, Zoubin Ghahramani, and Michael I Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, 1996.

[8] Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen, "DriveAHead – a large-scale driver head pose dataset," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1165–1174.

[9] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[10] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[11] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 427–432.

[12] Dragos Datcu and Leon JM Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," *Emotion Recognition: A Pattern Analysis Approach*, pp. 411–435, 2014.

[13] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010, pp. 94–101.

[14] Pooya Khorrami, Thomas Paine, and Thomas Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 19–27.

[15] Ran Breuer and Ron Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[16] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–10.

[17] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al., "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[19] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," *Syntax, Semantics and Structure in Statistical Translation*, p. 103, 2014.