INCORPORATING GAZE BEHAVIOR USING JOINT EMBEDDING WITH SCENE CONTEXT FOR DRIVER TAKEOVER DETECTION

Yuning Qiu, Carlos Busso

The University of Texas at Dallas
Richardson, TX 75080, USA

Teruhisa Misu, Kumar Akash

Honda Research Institute USA, Inc.
San Jose, CA 95134, USA

ABSTRACT

Despite the recent advancement in driver assistance systems, most existing solutions and partial automation systems such as SAE Level 2 driving automation systems assume that the driver is in the loop; the human driver must continuously monitor the driving environment. Frequent transition of maneuver control is expected between the driver and the car while using such automation in difficult traffic conditions. In this work, we aim to predict driver takeover timing in order for the system to prepare transition from automation to driver control. While previous studies indicated that eye gaze is an important cue to predict driver takeover, we hypothesize that traffic condition as well as the reliability of the driving automation also have a strong impact. Therefore, we propose an algorithm that jointly consider the driver's gaze information and contextual driving environment, which is complemented with the vehicle operational and driver physiological signals. Specifically, we consider joint embedding of traffic scene information and gaze behavior using *3D-Convolutional Neural Network* (3D-CNN). We demonstrate that our algorithm is successfully able to predict driver takeover intent, using user study data from 28 participants collected in simulated driving environments.

Index Terms— Driver's trust, Autonomous driving, Multimodal system, Joint embedding

1. INTRODUCTION

With the recent advancement of intelligent autonomous driving technologies, more vehicles have been equipped with SAE *Level 2* (L2) autonomous driving functions [1]. With L2 automation, the driver shares control of the vehicle with the autonomous driving system, and the driver is required to keep monitoring the driving environment and be ready to take over control of the vehicle at any time for safety and responsibility. Although such transitions are not very frequent during freeway driving, with the increased complexity of target situations envisioned for driving automation, we expect more frequent transitions of control between the driver and the automation. While transition of operation can be smoothly achieved even in the existing systems, it does not mean that the driver is mentally prepared to drive (i.e., fully aware of the surrounding situation). In order to achieve a smooth transition, it is important to predict drivers' takeover intention in *automated vehicle* (AV) and help the driver regain the awareness on the driving situation. In addition, by detecting driver's takeover intention, the system can also adjust the aggressiveness of the driving style if the main cause of the takeover is AV's driving style, which may result in reduced unnecessary takeovers in non-risky situations.

Previous studies demonstrated that driver's gaze and physiological information are effective behavioral signals for driver's status estimation, including trust [2,3], attention [4,5] and anomaly detection [6,7]. Yet, those factors are highly dependent on scene conditions, due to the fact that tasks to increase the driver's awareness in complicated situations naturally increase driver's workload [8,9]. Therefore, we propose to combine contextual driving environment information and the driver's gaze information to detect signs of driver's takeover behavior robustly against different traffic conditions.

This study considers driver's behavior under urban traffic conditions, where the complexity of the driving conditions change frequently depending on the surrounding traffic participants. In our driving simulator, we implement driving scenarios with different driving reliability levels based on the complexity of the traffic conditions. For example, when the reliability option is set to a low level, the automatic driving simulator system occasionally makes abrupt brakes. We collect driver behavioral and physiological signals while driving the traffic environment. Participants are instructed that the system is L2 automation, thus the driver are required to monitor the driving environment and be ready to take over the control to avoid accidents or breaking traffic rules. We predict driver's takeover using this data. The key contributions of this paper are summarized below.

- We collect a dataset for driver takeover detection tasks, consisting of driver behavioral (gaze and takeover behavior) and physiological signals for different driving conditions and automation reliability[1].
- Using the dataset, we demonstrate that our proposed method, which uses a 3D-CNN module to extract joint embedding of the driver's gaze heatmaps and road scene images, outperforms several baseline methods.

2. RELATED WORK

It is important to predict the driver's intention to provide a safe and comfortable driving experience. Various approaches have been proposed to predict driver's maneuvers by analyzing and understanding the driver's behaviors and mental state [10–13]. Studies suggested that trust in and usability of autonomous driving systems play an important role in the interaction between the driver and the system [14–17].

Molnar et al. [17] used the time for a participant to place her/his hands back on the steering wheel after a vehicle's request for take-back alert (also known as *takeover request* (TOR)) to understand the driver's trust level on the automated technology. Du et al. [18] extracted features from the driver's physiological signals (i.e., heart rate and *galvanic skin response* (GSR)) and driver's gaze information to predict the driver's takeover performance after the vehicle raises a TOR in L3 autonomous driving. They also considered the traffic density (i.e., light or heavy oncoming traffic) as environmental factor in their system. In this work, we jointly consider driver's gaze location, the contextual driving environment, the driver's physiological signals and vehicle's CAN-Bus data to detect driver's takeover behavior under different traffic conditions. Our system also uses the

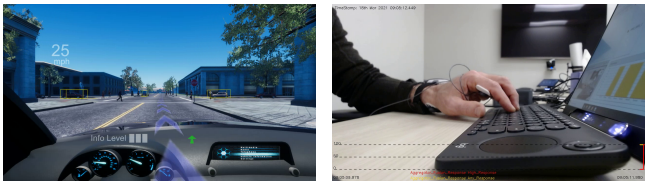
[1]The dataset will be made available for research purposes at `https://usa.honda-ri.com/datasets`.

information displayed in the *human-machine interface* (HMI) as an input for the system, since previous studies have demonstrated that this information is crucial for driver's trust in automated vehicle and the collaboration between the driver and system [19–21].

In multi-modal learning, a joint embedding is considered to be useful for downstream use cases, because it can combine information from multiple modalities. Ngiam et al. [22] used an autoencoder to reconstruct the audio spectrogram and video frames. They extracted the bottleneck representation as the joint embedding for visual speech classification tasks. Mithun et al. [23] applied joint embedding for a video-text retrieval task by projecting the features from the video, audio and text into a common embedding space. They minimized the differences between matching video-text pairs while maximizing the differences between non-matching pairs. Dahnert et al. [24] introduced a 3D-CNN based approach to extract a joint embedding that represents complementary information between 3D scan geometry and CAD models. They used the embedding space for CAD model retrieval. In this work, we robustly detect signs of driver's takeover behavior against different traffic conditions, by leveraging the environmental information from the contextual driving scenes together with driver's visual perception information from gaze locations. A 3D-CNN module is used to learn the joint embedding of the driver's gaze heatmaps and semantic segmentation images from road scenes.

3. DATASET

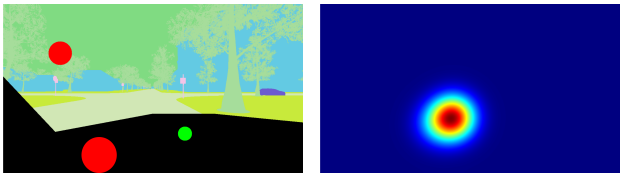
We collected the human-AV interaction data from 32 participants (15 females and 17 males) using our in-house driving simulator implemented using Unreal Engine. Their age ranged from 19 to 66 years old (mean = 36.18, SD= 14.05). The interaction comprised of monitoring the L2 automated driving through multiple intersections in an urban environment. Data from four participants were removed due to errors in the collected data. Participants were requested to report their intent to take over by pressing the Spacebar on the keyboard when they felt uncomfortable or they felt the automation was unsafe The automated driving system was simulated via the "Wizard of Oz" technique by replaying a past researcher's drive data. Each participant recorded 4 sessions, each consisting of 10 intersections that last approximately eight minutes. At each intersection, the vehicle is expected to stop following the signal light and stop signs. Then the vehicle turns left/right or go straight following the preset route. Most of the time the vehicle drives normally between intersections. 17.0% of the intersections involve a driver's takeover, and 1.0% of data points are labeled as "takeover". Figure 1(a) shows an example of a driving scene and the in-vehicle HMI. The interface displays the vehicle's dashboard, including speedometer and tachometer. The heads-up display presents the vehicle's speed on the top left and the next maneuver information (navigation arrow) was presented on the center stack. The HMI system additionally overlaid bounding boxes and predicted paths for the detected objects in the scene (e.g., pedestrians, vehicles and stop signs) to assist the driver's awareness about the situation and automated systems intentions and capabilities. The presented information simulates the AV's sensing results; the AV behavior is actually linked with the information presented. For example, bounding boxes on the pedestrians are not turned on before the vehicle start deceleration (e.g., Fig.1(a)). During the experiment, the participants experienced three different transparency levels (low - speed only; middle - speed + bounding boxes; high - all information), three different maneuvers at intersections (forward; right turn; left turn) and two different weather conditions (sunny; snowy). We collected the participants gaze coordinate on the screen using a Tobii Pro Nano eyetracker, and physiology signals (heart rate and galvanic skin response; GSR) using NeuroLynQ (see Figure 1(b)). We also recorded the video of the driving scenes and vehicle's CAN-Bus signals (e.g., speed and steering).



(a) User Interface (b) Wearable devices

Fig. 1. The user interface of driving scenarios, and the devices for driver's physiological and gaze data collection.

We normalize the participants' heart rate and GSR signals using Z-normalization ($z = \frac{x-\mu}{\sigma}$) to compensate for the difference among participants. Here, μ and σ are the mean and standard deviation estimated from the entire data during the four sessions. We synchronize and down-sample all the signals to 10Hz. We interpolate the road scene images and driver's gaze points using nearest-neighbor interpolation. We also interpolate vehicle's CAN-Bus and driver's physiological signals using linear interpolation. To encode only the contextual information about the scene, we use the ground truth semantic segmentation images for the corresponding road scene images that are extracted using the driving simulator software, as shown in Figure 2(a). Since the participants can check the speed of the car from either the speedometer or the heads-up display, we allot the same semantic label in both regions in the image (i.e., regions in red). To mimic driver's visual perception that not only includes a single gaze point as captured by the eyetracker but also the peripheral area around the gaze point, we plot the driver's gaze as a sum of Gaussian distributions as shown in Figure 2(b). With the eye tracker collecting driver's gaze points at 50 Hz, we plot the gaze heatmap for each timestamp using five consecutive gaze points to preserve historical information. Each Gaussian distribution has the same standard deviation for the X axis and Y axis (the effect of the Gaussian parameters are discussed in Section 5.2). We use the equation $f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\frac{1}{2}\left(\frac{(x-x_0)^2}{\sigma_x^2} + \frac{(y-y_0)^2}{\sigma_y^2}\right)\right]$, where x_0 and y_0 are the position of the gaze coordinate, and σ_x and σ_y are the standard deviations, which are assumed to be equal (e.g., $\sigma_x = \sigma_y = \sigma$). Based on previous studies [25], the probability values are based on a kernel density estimation (KDE), which takes the gaze points and produces a probability distribution with a Gaussian kernel across the image. Then, we use the summation over the five distributions of each timestamp, each centered around the gaze point locations. We represent them as heat maps. We calibrate the segmentation images and gaze heatmaps with respect to (1) timestamp: we record the video player and eye tracker timestamps and use the nearest timestamp gaze coordinate; and (2) coordinate: the eye-tracker is a screen-based system, creating a gaze coordinate (x,y) in the screen playing the video. Therefore, the gaze locations are in the same coordinate system as the semantic segmentation image. For computational efficiency, we down-sample the resolution of the road scene semantic segmentation images and the gaze heatmaps to 90 x 160, converting the segmentation images into gray scale images (i.e., each "intensity" corresponds to a different semantic class). We use the throttle, RPM, steering and speed as features from the CAN-Bus signals, which are normalized in the range 0 to 1. We also take the complementary information (i.e., HMI transparency level, navigation and weather) into consideration, transferring them into a one-hot embedding.



(a) Semantic segmentation image (b) Driver's gaze heatmap

Fig. 2. The semantic segmentation images of the road scenario and in-vehicle scene, and the heatmap of driver's gaze points. We regard those two "gray scale" images as two channel image and use to extract joint embedding.

4. PROPOSED DRIVER TAKEOVER DETECTION

Our goal is to robustly detect signs of driver's takeover behavior against different traffic conditions. While previous studies separately take driver's gaze data and environmental information [18,26], we aim to fuse these two modalities by learning a joint embedding space that carries complementary information. Specifically, we prepare a series of two-layer images, one layer for the scene (semantic segmentation) information and the other layer for the gaze information. We apply a convolution over those two channels as well as time series inputs to better capture the interaction between those modalities. Moreover, this work considers the vehicle's CAN-Bus signals and driver's physiological signals to fully represent signs of driver takeover under different driving conditions. Therefore, the network considers the driving behavior of the autonomous system and driver's physiological reactions. We also use the HMI information level, navigational information and weather information.

Figure 3 shows the proposed 3D-CNN based multimodal driver's takeover detection system. This system extracts features from multi-modal data of a driving segment, and fuse the features to predict whether the driver takeover at the final timestamp of the segment. We define the length of the driving segment as *look-back*, and set the default value as 3 seconds. The first stage of our system is the feature extraction module. We implement a 3D-CNN network [27] to extract joint embedding for the driver's gaze heatmaps and road scene semantic segmentation. The detailed architecture of the 3D-CNN module is shown in Table 1, where all the 3D-CNN layers and linear layers are activated by Rectified Linear Unit (ReLU). To extract feature embedding from the vehicle's CAN-Bus and driver's physiological data, we first flatten the data to a vector with 180 (6 signals x 3 seconds x 10 Hz) data points, and input them into a dense layer module that consists of three fully connected layers activated by ReLU, where the numbers of neurons of each layer are 128, 64 and 32, respectively. These embeddings are concatenated with the one-hot embeddings of the HMI complementary modalities introduced in Section 3, as fused representation of the selected driving segment. The second stage is to use a DNN module that takes the fused representation as input, to predict driver's takeover actions. The DNN module consists of four fully connected layers, where the numbers of neurons for each layer are 512, 512, 256 and 1, respectively. The first three layers are activated with the ReLU function, while the last layer is activated with the Sigmoid function. We train our model for 20 epochs, using Adam optimizer with an initial learning rate equal to $\eta = 0.001$, and a *batch size* equal to 32.

5. EXPERIMENTAL RESULTS

5.1. Comparison with baselines

Since predicting drivers' takeover is a binary classification task, we use the *receiver operating characteristic* (ROC) curves and *area under ROC curve* (AUC) as the metric to compare the model performance. We used the data from 28 participants, with the data from the multiple modalities synchronized perfectly, and we split the data in participant independent partitions, with data from 20 participants as the training set, 4 participants as the validation set, and 4 as the testing set. We prepare 5 groups of [training, validation, testing] set by shuffling the partitions. We train five models using these five partitions. We report the ROC curve and AUC values using the results across the five models. We calculate the TP and FP rates by considering all the prediction results from the 5 models.

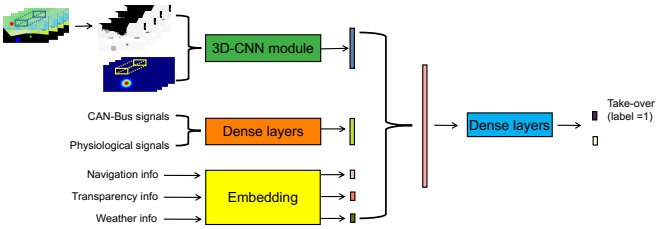


Fig. 3. The proposed 3D-CNN based multi-modal driver's takeover detection system, with a 3D-CNN module extracting joint embedding of driver's gaze heatmaps and road scene semantic segmentation images.

Table 1. The 3D-CNN module for joint embedding extraction in this work.

Layer	Channels	Kernel	Stride	[Shape]
Input	2	N/A	N/A	(30, 90, 160, 2)
3D-CNN	32	(3, 3, 3)	(1, 1, 1)	(30, 90, 160, 32)
3D-Maxpooling	N/A	(2, 2, 2)	(2, 2, 2)	(15, 45, 80, 32)
3D-CNN	64	(3, 3, 3)	(1, 1, 1)	(15, 45, 80, 64)
3D-Maxpooling	N/A	(2, 2, 2)	(2, 2, 2)	(7, 22, 40, 64)
3D-CNN	128	(3, 3, 3)	(1, 1, 1)	(7, 22, 40, 128)
3D-Maxpooling	N/A	(2, 2, 2)	(2, 2, 2)	(3, 11, 20, 128)
3D-CNN	256	(3, 3, 3)	(1, 1, 1)	(3, 11, 20, 256)
3D-Maxpooling	N/A	(2, 2, 2)	(2, 2, 2)	(1, 5, 10, 256)
Flatten	N/A	N/A	N/A	10240
Linear	256	N/A	N/A	256
Dropout	p = 0.5	N/A	N/A	(N/A)
Linear	256	N/A	N/A	256

We evaluate our method with three baselines. The first baseline is based on manually selected features from the data across multiple modalities, which are used as input of a *support vector machine* (SVM). For the features of the CAN-Bus and physiological data, we calculate the statistics within a selected driving segment (i.e., maximum, minimum, standard deviation, and mean value). For the gaze and environmental features, considering the objects viewed by the driver, we estimate the gaze area as circle with the gaze point as the center. Since the error range of the eyetracker is 0.5 degree, and the monitor that we used is 17.3 inch display and the distance between user's eye and the monitor is about 30 cm, we set the radius of the circle to 15 pixels (eye tracker output coordinate ± 0.5 degree error range). We calculate the proportion of the objects (i.e., [vehicle, pedestrian, sign, speed, navigation, others]) located in the circle, as a 6-dimension vector for each timestamp. We keep the one-hot embeddings as the features of the complementary modalities. In total, we have a 213D vector as the feature of each driving segment. The SVM classifier identifies driver's takeover using these features.

The second baseline uses the same manually selected features as the input of a *deep neural network* (DNN). The implementation has 5 layers, where the numbers of neurons per layer are 256, 128, 64, 32 and 1. Each layer is activated with ReLU, except the output

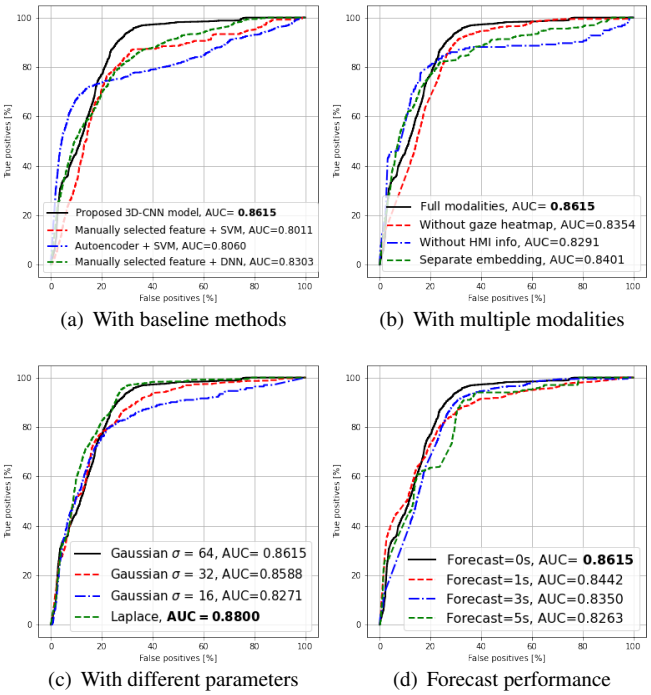
layer that is activated with a Sigmoid. Inspired by Ngiam et al. [22], the third baseline is built with autoencoders and SVM. We build two autoencoders, one for reconstructing the semantic segmentation images and gaze heatmaps, and the other for CAN-Bus and physiological signals. To jointly consider the driving environment and driver's gaze area, we build the image autoencoder based on the 3D-CNN layers. The autoencoder for CAN-Bus and physiological signals is built with dense layers. The bottleneck layer embeddings of the autoencoders are extracted as representative features. We concatenate the features with the one-hot embedding of the complementary modalities to train an SVM classifier to identify driver's takeover. The ROC curves and the corresponding AUC values in Figure 4(a) shows that the proposed model (AUC = 0.8615) outperforms all the three baselines (AUC = 0.8011, 0.8303, 0.8060, respectively). The autoencoder-SVM, which is a combination of the proposed joint embedding and the SVM classifier, only performs better when the *false positive* is less than 20%.

5.2. Ablation studies

We conduct ablation studies to evaluate the effectiveness of individual factors of our joint embedding for gaze and contextual driving environment. We first explore the model performance if we exclude the driver's gaze modality to assess its value in predicting takeover actions. Figure 4(b) shows that when trained without the gaze heatmaps, the model performance (i.e., AUC) decrease from 0.8615 to 0.8354. With the HMI complementary information excluded, the model performance drops to 0.8291. To validate the effectiveness of the joint embedding, we implement the proposed model by separately considering the driving environment and driver's visual perception. We input only the road scene images to the 3D-CNN module, and use a four-layer DNN module to take the manually selected gaze features that are described in Section 5.1 as input. We keep the rest of the proposed model for the CAN-Bus, physiological and HMI complementary information unchanged. The AUC of the model with separate embedding drops to 0.8401, which is 0.0214 lower than the model trained with joint embedding. This result demonstrates the effectiveness of the proposed joint embedding for robustly detecting signs of driver's takeover behavior against different traffic conditions

We also compared parameters to make the driver's gaze heatmap. Figure 4(c) shows the results of the experiments conducted using the driver's gaze heatmaps that are filtered by a Laplace distribution and Gaussian distribution with different σ value, revealing the effect of the size of the driver's gaze area. When σ is set to 16, 32 and 64 pixels (default), the corresponding AUC values are 0.8271, 0.8588 and 0.8615, indicating that a Gaussian distribution with small σ might decrease the model performance. However, when the driver's gaze heatmaps are plotted with a Laplace distribution, the model performance achieves a AUC of 0.8800. This may be caused by the structure of human eyes, where the cone distribution (photoreceptor cells in the retinas) is an approximate Laplace distribution [28].

We also investigate the capability of the proposed model to forecast driver's takeover in advance. The ROC curves in Figure 4(d) show the model performance on forecasting driver's takeover 1 second, 3 seconds and 5 seconds ahead of the driver's pedal pressing. The look-back value is set to 3 seconds. While the 0-second curve, representing the detection task, has a AUC value of 0.8615, AUC of the 1-second curve achieves 0.8442. Even though the performance drops slightly to 0.8365 and 0.8263 for the 3 seconds and 5 seconds, respectively, the proposed method is still capable of forecasting the driver's takeover intention with reasonable performance.



(a) With baseline methods (b) With multiple modalities

(c) With different parameters (d) Forecast performance

Fig. 4. The ROC curves for comparison of model performance.

6. CONCLUSION & FUTURE WORK

We presented a 3D-CNN based multi-modal system to detect the driver's takeover intention in L2 autonomous driving. With a 3D-CNN module that extracts joint embedding of the semantic segmentation images of the road scene and driver's gaze heatmaps, this system combines the contextual driving environment, in-vehicle human-machine interface and driver's visual perception, which is effective for detecting driver's takeover intention. Besides driver's visual perception, this multi-modal system jointly considers the vehicle CAN-Bus signals, driver's physiological signals, and in-vehicle HMI complementary information. Through the experiments, we confirmed that our method using the joint embedding for the gaze and scene information outperforms a method that combines independently extracted features (late fusion).

A future work of this study is investigating driver's gaze representations. According to Wandell [28], human's visual perception system consists of rods and cones, which filter the visible lights following an approximate Gaussian distribution and an approximate Laplace distribution, respectively. By building a filter that combines Gaussian and Laplace distribution to model perception capability of the human's rods and cones more accurately, we expect to improve the model performance and reveal more detailed information about the relationship between the driver's gaze, contextual driving environment and the autonomous driving conditions.

One important limitation of this work is the user's learning effect through long-term interactions with the system. While this study focuses on short-term driving behavior just before takeover happens, many studies such as Akash et al. [29] indicate that driver's takeover behavior is largely influenced by driver's trust on the automation established through past experiences. Therefore, we would like to combine our method with those methods that consider long-term user experiences.

7. REFERENCES

[1] M. Galvani, "History and future of driver assistance," *IEEE Instrumentation & Measurement Magazine*, vol. 22, no. 1, pp. 11–16, February 2019.

[2] B. E. Noah, P. Wintersberger, A. G. Mirnig, S. Thakkar, F. Yan, T. M. Gable, J. Kraus, and R. McCall, "First workshop on trust in the age of automated driving," in *Adjunct Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2017)*, Oldenburg, Germany, September 2017, pp. 15–21.

[3] P. Wintersberger, B. E. Noah, J. Kraus, R. McCall, A. G. Mirnig, A. Kunze, S. Thakkar, and B. N. Walker, "Second workshop on trust in the age of automated driving," in *Adjunct Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2018)*, Toronto, ON, Canada, September 2018, pp. 56–64.

[4] M. Miyaji, H. Kawanaka, and K. Oguri, "Study on effect of adding pupil diameter as recognition features for driver's cognitive distraction detection," in *International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP 2010)*, Newcastle Upon Tyne, UK, July 2010, pp. 406–411.

[5] S. Jha and C. Busso, "Estimation of gaze region using two dimensional probabilistic maps constructed using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 3792–3796.

[6] Y. Qiu, T. Misu, and C. Busso, "Driving anomaly detection with conditional generative adversarial network using physiological and can-bus data," in *ACM International Conference on Multimodal Interaction (ICMI 2019)*, Suzhou, Jiangsu, China, October 2019, pp. 164–173.

[7] Y. Qiu, T. Misu, and C. Busso, "Use of triplet loss function to improve driving anomaly detection using conditional generative adversarial network," in *Intelligent Transportation Systems Conference (ITSC 2020)*, Rhodes, Greece, September 2020, pp. 1–7.

[8] K. A. Brookhuis and D. de Waard, "Assessment of drivers' workload: Performance and subjective and physiological indexes," in *Stress, workload, and fatigue*, P.A. Hancock and P.A. Desmond, Eds., Human Factors in Transportation, pp. 321–333. Lawrence Erlbaum Associates Inc., Mahwah, NJ, USA, November 2000.

[9] C.J.D. Patten, A. Kircher, J. Östlund, L. Nilsson, and O. Svenson, "Driver experience and cognitive workload in different traffic environments," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 887–894, September 2006.

[10] M. Itoh, K. Yoshimura, and T. Inagaki, "Inference of large truck driver's intent to change lanes to pass a lead vehicle via analyses of driver's eye glance behavior in the real world," in *Proceedings of SICE Annual Conference*, Takamatsu, Japan, September 2007, pp. 2385–2389.

[11] Y. Murphey, D. S. Kochhar, P. Watta, X. Wang, and T. Wang, "Driver lane change prediction using physiological measures," *SAE International Journal of Transportation Safety*, vol. 3, no. 2, pp. 118–125, July 2015.

[12] N. Li and C. Busso, "Detecting drivers' mirror-checking actions and its application to maneuver and secondary task recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 980–992, April 2016.

[13] A. Rangesh, N. Deo, R. Greer, P. Gunaratne, and M. Trivedi, "Predicting take-over time for autonomous driving with real-world data: Robust data augmentation, models, and evaluation," *ArXiv e-prints (arXiv:2104.11489)*, pp. 1–12, July 2021.

[14] K. Zeeb, A. Buchner, and M. Schrauf, "What determines the take-over time? an integrated model approach of driver take-over after automated driving," *Accident Analysis & Prevention*, vol. 78, pp. 212–221, May 2015.

[15] M. Walch, K. Lange, M. Baumann, and M. Weber, "Autonomous driving: investigating the feasibility of car-driver handover assistance," in *International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2015)*, Nottingham, United Kingdom, September 2015, pp. 11–18.

[16] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *International Journal of Human-Computer Interaction*, vol. 31, no. 10, pp. 692–702, October 2015.

[17] L.J. Molnar, L.H. Ryan, A.K. Pradhan, D.W. Eby, R.M. St. Louis, and J.S. Zakrajsek, "Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 54, pp. 319–328, October 2018.

[18] N. Du, F. Zhou, E.M. Pulver, D. M. Tilbury, L.P. Robert, A.K. Pradhan, and X.J. Yang, "Predicting driver takeover performance in conditionally automated driving," *Accident Analysis & Prevention*, vol. 148, pp. 105748:1–11, December 2020.

[19] R. Pokam, C. Chauvin, S. Debernard, and S. Langlois, "Augmented reality interface design for autonomous driving," in *International Symposium on Future Active Safety Technology Toward zero traffic accidents (FAST-zero 2015)*, Gothenburg, Sweden, September 2015, pp. 145–146.

[20] Y. Yang, B. Karakaya, G. C. Dominioni, K. Kawabe, and K. Bengler, "An HMI concept to improve driver's visual behavior and situation awareness in automated vehicle," in *International Conference on Intelligent Transportation Systems (ITSC 2018)*, Maui, HI, USA, November 2018, pp. 650–655.

[21] O. Carsten and M. H. Martens, "How can humans understand their automated cars? hmi principles, problems and solutions," *Cognition, Technology & Work*, vol. 21, no. 1, pp. 3–20, February 2019.

[22] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *International conference on machine learning (ICML2011)*, Bellevue, WA, USA, June-July 2011, pp. 689–696.

[23] N.C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *ACM on International Conference on Multimedia Retrieval (ICMR 2018)*, Yokohama,Japan, June 2018, pp. 19–27.

[24] M. Dahnert, A. Dai, L. Guibas, and M. Niessner, "Joint embedding of 3D scan and CAD objects," in *IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, Seoul, South Korea, October-November 2019, pp. 8748–8757.

[25] S. Jha and C. Busso, "Estimation of driver's gaze region from head position and orientation using probabilistic confidence regions," *IEEE Transactions on Intelligent Vehicles*, vol. to appear, 2021.

[26] N. Du, J. Haspiel, Q. Zhang, D. Tilbury, A. K. Pradhan, X.J. Yang, and L.P. Robert Jr, "Look who's talking now: Implications of AV's explanations on driver's trust, av preference, anxiety and mental workload," *Transportation Research Part C: Emerging Technologies*, vol. 104, pp. 428–442, July 2019.

[27] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, January 2013.

[28] B. Wandell, *Foundation of Vision*, chapter 3: The Photoreceptor Mosaic, Sinauer Associates Inc, 1985.

[29] K. Akash, N. Jain, and T. Misu, "Toward adaptive trust calibration for level 2 driving automation," in *ACM International Conference on Multimodal Interaction (ICMI 2020)*, Utrecht, The Netherlands, October 2020, pp. 538–547.