DIF : Dataset of Intoxicated Faces for Drunk Person Identification

Devendra Pratap Yadav
Indian Institute of Technology Ropar
2014csb1010@iitrpr.ac.in

Abhinav Dhall
Indian Institute of Technology Ropar
abhinav@iitrpr.ac.in

ABSTRACT

Traffic accidents cause over a million deaths every year, of which a large fraction is attributed to drunk driving. Automated drunk detection systems in vehicles are necessary to reduce traffic accidents and the related financial costs. Existing solutions require special equipment such as electrocardiogram [27], infrared cameras [17] or breathalyzers. In this work, we propose a new dataset called DIF (Dataset of Intoxicated Faces) containing RGB face videos of drunk and sober people obtained from online sources. We analyze the face videos to extract features related to eye gaze, face pose and facial expressions. A recurrent neural network is used to model the evolution of these multimodal facial features. Our experiments show the eye gaze and facial expression features to be particularly discriminative for our dataset. We achieve good classification accuracy on the DIF dataset and show that face videos can be effectively used to detect drunk people. Such face videos can be readily acquired through a camera and used to prevent drunk driving incidents.

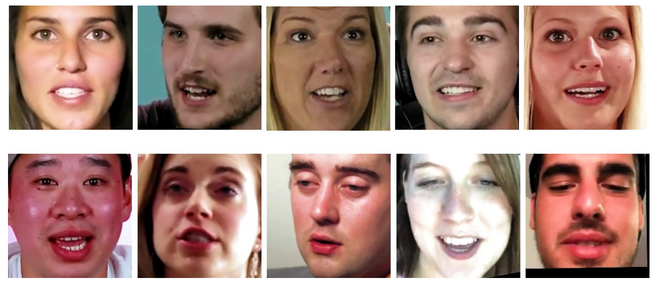

Figure 1: Video frames from the proposed DIF dataset. The top row contains sober subjects and the bottom row contains intoxicated subjects, respectively.

1 INTRODUCTION

Alcohol impaired driving poses a serious threat to the driver as well as pedestrians. Over 1.25 million road traffic deaths occur every year. Traffic accidents are the leading cause of death among those aged 15-29 years [24]. Drunk driving is responsible for around 40% of all traffic crashes [8]. Strictly enforcing drunk driving laws can reduce the number of road deaths by 20% [24]. Modern vehicles are being developed with a focus on smart and automated features. Driver monitoring systems such as drowsiness detection [1] and driver attention monitoring [18] have been developed to increase safety. Incorporating automated drunk detection systems in vehicles is necessary to reduce traffic accidents and the related financial costs.

Drunk detection systems can be divided into three categories :
1. **Direct detection** - Measuring Blood Alcohol Content (BAC) directly through breath analysis.
2. **Behavior based detection** - Detecting characteristic changes in behavior due to alcohol consumption. This may include changes in speech, gait, or facial expressions.
3. **Biosignal based detection** - Using Electrocardiogram signals [27] or face thermal images [17] to detect intoxication.

Direct detection is often done manually by law enforcement officers using Breathalyzers. Biosignal based detection also requires specialized equipment to measure signals. Behavior based detection can be performed passively by recording speech or video of the subject and analyzing it to detect intoxication. We focus on behavior based detection, specifically using face videos of a person. While speech is the most significant behavioral change in drunk people, we can also detect changes in eye movement and facial expressions. To the best of our knowledge, no existing work in literature addresses the problem of detecting intoxication using RGB face videos only, analyzing facial features to discriminate between drunk and sober people. Neither do we have a large dataset exhibiting changes in behavior of drunk and sober people.

The key contribution of our work is a new dataset containing face videos of drunk and sober people: DIF (Dataset of Intoxicated Faces). The face videos contain people with some amount of rigid and non-rigid face movement. We extract features related to eye gaze, face pose and facial expressions for each face video. A sequence of these features is given as input to a RNN with Long short-term memory (LSTM) units, which classifies it as drunk or sober. We obtain 75.54% accuracy on our DIF dataset, determining the eye gaze and facial landmark features to be most discriminative for drunk person identification.

2 PRIOR WORK

While RGB face videos have not been used for detecting intoxication, several other techniques have been proposed based on behavior analysis. Fazeen et al [13] utilized three-axis accelerometer on an Android smartphone to record and analyze driving patterns as well as external road conditions, detecting potentially hazardous drunk drivers. Speech analysis has been widely used to identify drunk people. The INTERSPEECH 2011 Speaker State Challenge [22] provides an Alcohol Language Corpus dataset for intoxication detection. Biadsy et al [4] treat intoxicated speech as a different accent of a language and use phonetic, phonotactic and prosodic cues to detect intoxication. Koukiou et al. [17] utilized thermal images and analyzed the difference in thermal illumination in several areas of the face to detect intoxication.

Studies have shown that alcohol consumption leads to changes in facial expression of emotions [5] and significant differences in eye movement patterns [23]. Drowsiness and fatigue are also observed after alcohol consumption [14]. These changes in eye movement and facial expressions can be analyzed using face videos of drunk and sober people. Affective analysis has been successfully used to detect complex mental states such as depression, psychological distress and truthfulness [6] [11].

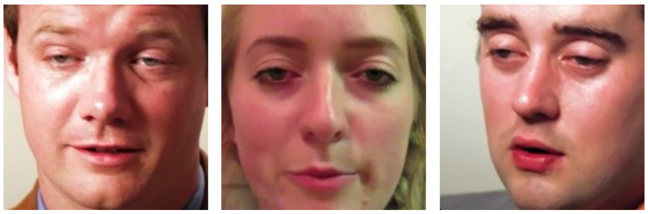

Figure 2: Symptoms of fatigue observed in faces from the drunk category in DIF dataset.

A review of face expression and emotion recognition techniques is necessary in order to implement a system for detecting intoxication using face videos. The early attempts to parameterize facial behavior led to the development of Facial Action Coding System (FACS) [10] and Facial Animation parameters (FAPs) [25]. The Cohn-Kanade Database [15] and MMI Facial Expression Database [19] provided videos of facial expressions annotated with action units. Recent submissions to Emotion Recognition In The Wild Challenge by Dhall et al. [7] have focused on deep learning based techniques for affective analysis. In EmotiW 2015, Kahou et al. [9] used Recurrent Neural Network (RNN) combined with a CNN for modelling temporal features. In EmotiW 2016, Fan et al. [12] used RNN in combination with 3D convolutional networks (C3D) to encode appearance and motion information, achieving state-of-the art performance. Using a CNN to extract features for a sequence and classifying it with RNN perform well on emotion recognition tasks. A similar approach can be used for drunk person identification. The OpenFace toolkit [3] by Baltrusaitis et al. can be used to extract features related to eye gaze [26], facial landmarks [2] and face pose.

3 DATASET COLLECTION

Our work consists of collection, processing and analysis of a new dataset of drunk and sober face videos. We present DIF (Dataset of Intoxicated Faces) for drunk person identification. The dataset urls and features will be made publicly available.

By collecting this dataset, we aim to analyze the difference in facial features of a drunk and sober person. Hence, we try to search for online videos of people exhibiting facial movements and expressions in drunk and sober conditions. We use search queries such as 'drunk reactions', 'drunk review', 'drunk challenge' etc. on YouTube (www.youtube.com) and Periscope (www.pscp.tv) to obtain videos of drunk people. Similarly, for the sober category, we collect several reaction videos from YouTube. Since the videos were recorded in unconstrained real world conditions, our dataset represents drunk and sober people 'in the wild'. We use the video title and caption given on the website to assign class labels. In some cases, the subject labeled as 'drunk' might only be slightly intoxicated and not exhibit significant changes in behavior. In these cases, the drunk class labels are considered as weak labels. There is no such ambiguity in the sober class labels. We collect 30 videos in the sober category with an average length of 10 minutes. The drunk category contains 50 videos with an average length of 6 minutes. The sober category consists of 81 unique subjects, 36 male and 45 female. The drunk category consists of 47 unique subjects, 27 male and 20 female.

We process these videos using the pyannote-video library [21]. First, we perform shot detection on the video and process each shot separately in subsequent stages. Then, we perform face detection and tracking to extract bounding boxes for faces present in each frame of the video. We also get a unique id for each tracked face. Using these ids and bounding boxes, we crop the tracked face from the video. Hence, we obtain a set of face videos where each video contains a drunk or sober person exhibiting facial motion. These face videos are of 224x224 size, maintaining the aspect ratio of the original videos. We also perform face alignment on each frame of the video using the OpenFace toolkit [3].

Each aligned face videos is split into non-overlapping sequences of 75 frames, which corresponds to 5 seconds of video at 15 frames per second. While extracting these sequences, we reject those in which a face was not detected accurately. We also apply a threshold based on the variance of facial landmark points to remove sequences having low facial movement. Our final dataset consists of 1294 sequences for the drunk category and 1443 sequences for the sober category. We have a total of 205,275 frames corresponding to sober and drunk people in the dataset.

4 BASELINE AND EXPERIMENTS

Convolutional Neural Network (CNN) - Recurrent Neural Network (RNN) models have been successfully used for modeling and classifying facial affect by Kahou et al. [9] and Fan et al. [12] in recent EmotiW challenges. Since we need to analyze facial expressions for the task of drunk person identification, we use a CNN-RNN model as our baseline.

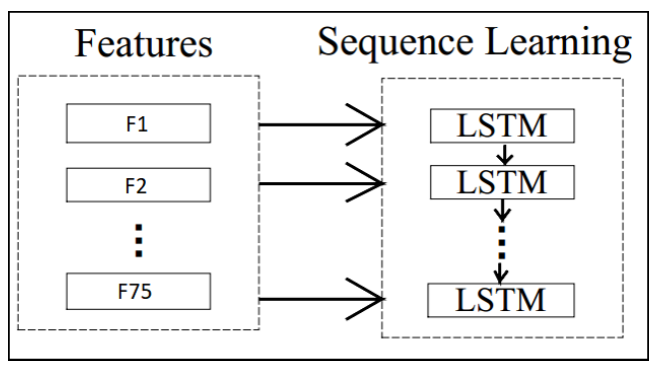

Figure 3: Representation of the LSTM RNN used for classification.

We use a CNN for feature extraction, which gives us a feature vector for each frame in a sequence. In order to model the spatio-temporal changes which can discriminate between sober and drunk faces, we learn a RNN with Long short-term memory (LSTM) units. The LSTM network takes the features extracted for a sequence of video frames as input to classify the person as drunk or sober.

We use a single layer LSTM network with 256 dimensional output. This is connected to a single dense layer with 256 nodes followed by a dense layer with 2 nodes (number of classes). We use softmax activation with cross entropy loss for classification. Dropout is used for regularization. We tune hyperparameters such as dropout rate, number of hidden nodes and batch size to obtain best performance.

4.1 Experimental Protocol

The DIF dataset contains 2737 sequences across two classes. The drunk class contains 1294 sequences and the sober class contains 1443 sequences. Classification accuracy is used as the performance measure. However, since we have only two classes, poor classification accuracy for one class might still result in high overall accuracy. Hence, the recall values are also reported for each class. Five-fold cross validation is used to evaluate a model's classification performance. The test set contains 20% of the dataset, and train set contains the remaining 80%. 10% of the training set is used for validation. It is ensured that both classes have equal number of instances in train, test and validation sets.

While performing five-fold cross validation, the training, testing and validation sets are not split randomly. If sequences extracted from the same face video, containing the same person, are present in both train and test set, models performing face recognition rather than drunk person identification will give higher test accuracy. We perform face embedding and clustering to ensure that the identities present in the train, validation and test set are different. However, the face clustering procedure is not 100% accurate and there may be some identity overlap between the sets. While training, we save the model with the highest validation accuracy and use it for testing. We perform five-fold cross validation 5 times on the DIF dataset and report the average results.

4.2 VGG-Face Features

We use a CNN based on VGG-Face [20] is used to extract powerful feature representation for face images. VGG-Face has previously been successfully used for face recognition tasks. However, this may cause the classification model to be biased towards face recognition and perform poorly for drunk person identification.

A 2048 dimensional feature vector is extracted for each of the 75 frames in a sequence. A learning rate of 0.0001 and batch size of 32 is used for training. After 15 epochs, the model reaches training accuracy of 99.9%, and validation accuracy of 79.3%. However, the test accuracy is much lower at 69.3%. This is due to the VGG-Face features being trained for face recognition rather than drunk person identification. The network learns the identities of the people in drunk and sober classes, and fails when new people are presented in the test set. While the test accuracy seems good, the average recall is 0.451 for the drunk class and 0.936 for the sober class. This suggests that the network is simply predicting random output for the drunk class. Similar observation is true for the high validation accuracy. Moreover, since the face embedding and clustering procedure is not completely accurate, some identity overlap may be present between training and validation set, increasing the validation accuracy. Hence, we conclude that VGG-Face features do not perform well for drunk person identification.

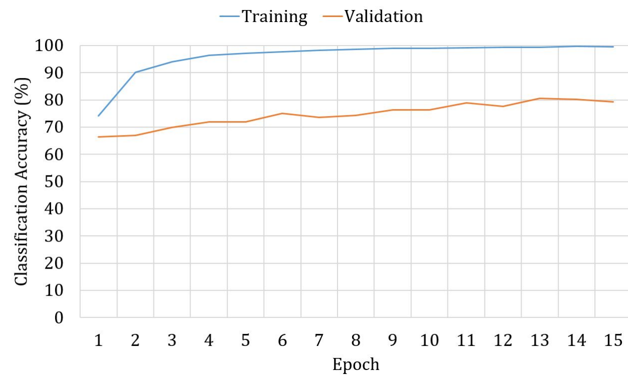

Figure 4: Train and validation accuracy on using VGG-Face features on DIF dataset [3]

4.3 Eye gaze, pose and facial expression features

Using the observations from VGG-Face features, we try to extract facial features which contain minimal amount of identity information. The CNN models in the OpenFace toolkit [3], are used to extract several facial features for each frame in a sequence. These features are:
1. Eye gaze (8 values) - Eye gaze direction for both eyes and an average gaze direction in radians
2. Eye landmarks (56 2D points) - Points denoting location of 2D eye region landmarks
3. Face pose (6 values) - Angle and 3D location of face with respect to camera
4. Face landmarks (68 2D points) - Location of 68 2D face landmarks
5. Point distribution model (40 values) - Parameters of a point distribution model (PDM) describing rigid face shape related to location, scale and rotation of face, as well as non-rigid face shape related to face expression and identity
Combining these, a 302 dimensional feature vector is extracted for each frame. For a sequence of 75 frames, we get 302x75 features.

These mutimodal face features are analyzed to determine their effectiveness in discriminating between drunk and sober faces. The class wise recall values and average classification accuracy for five-fold cross validation is given for various features in Table 1.

We use a learning rate of 0.0003 and batch size of 32 for all further experiments. Using all the features, we get a 302x75 dimensional input and obtain an average test accuracy of 72.2%. However, learning is not same for both classes. The drunk class has a low recall of 0.64 while the sober class has high recall of 0.81. This suggests that the model has failed to learn discriminative features.

On using eye gaze features only, we get a 8x75 dimensional input. This model obtains an average training accuracy of 66.9%. After 30 epochs, the training and validation accuracy remain noticeably low at 71.3% and 65.5% respectively as shown in Figure 5. This suggests that eye gaze provides discriminative features without overfitting on identity information. Using only PDM features, the model reaches training accuracy of 92.7% and validation accuracy of 70.8%. Similarly, using only face landmark features, we get training and validation accuracy of 93.3% and 68.9% respectively. The

Table 1: Classification accuracy and recall values for multi-modal face features on DIF dataset using CNN-RNN model

Sequence features	Recall (Sober)	Recall (Drunk)	Classification accuracy (%)
Eye gaze	0.73	0.61	66.91
Eye landmarks	0.66	0.65	65.88
Face pose	0.70	0.49	59.26
Face landmarks (FL)	0.74	0.70	71.61
PDM	0.86	0.51	68.47
All features	0.81	0.64	72.23
FL+PDM	0.81	0.64	72.55
Eye gaze+FL+PDM	0.75	0.76	**75.54**

high training accuracies suggest that both PDM and face landmark features provide some identity information. However, face landmark features provide a more balanced recall value for sober and drunk classes. This suggests that face landmarks provide important discriminative information related to facial expressions.

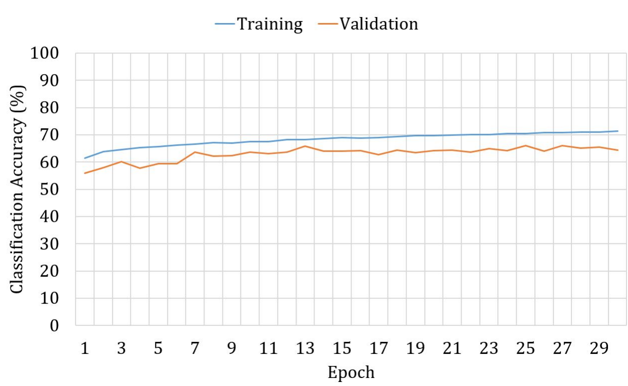

Figure 5: Train and validation accuracy on using Eye gaze features on DIF dataset

Based on classification accuracy alone, eye landmark and face pose features seem to provide poor results. On using all features except the eye landmark and face pose features, we observe a high average test accuracy of 75.54%. The classification accuracies and recall values are comparatively lower for these features. This suggests that face pose and eye landmark features are not particularly discriminative for our dataset. While the average head movement of subjects may change on intoxication as shown in Figure 6, our dataset often focuses on subjects looking into the camera or talking to a person, leading to little change in head position and rotation. Similar to our dataset, a person driving a vehicle does not exhibit much change in face pose. Additionally, removing the 122 eye landmark features improves the accuracy since the eye gaze features provide similar information related to eye movement with just 8 values, making it easier for the LSTM network to learn discriminative eye movement patterns.

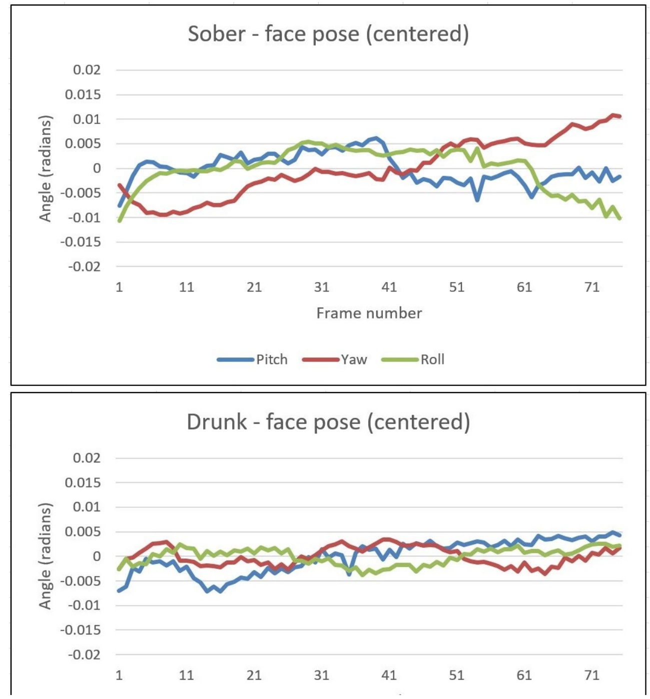

Figure 6: Difference between average centered (mean = 0) face pose in a sequence for drunk and sober categories.

We obtain the best classification accuracy of 75.54% for a combination of eye gaze, face landmark and PDM features. The recall values are balanced at 0.75 and 0.76 for sober and drunk classes respectively. Compared to face landmark and PDM features, eye gaze provides discriminative features without any identity information leading to a more general model. However, face landmark and PDM features still provide some relevant facial expression information and improve the accuracy of the model. Based on our experiments, we find eye movement and facial expression features to be effective for drunk person identification using RGB face videos. Adding more identities to the drunk and sober classes in our dataset may help us to reduce the learning of identity information from face landmark and PDM features, leading to a more general model.

5 CONCLUSION AND FUTURE WORK

We present a new dataset of face videos of drunk and sober people, DIF : Dataset of Intoxicated Faces, and use it to demonstrate that RGB face videos can be effectively used to detect drunk people. Our approach does not require special equipment or sensors and can be easily used in real world settings. We use a LSTM network using CNN features as input, to classify a face video as drunk or sober. We experiment with various features extracted from face images. Through our experiments on the DIF dataset, we find eye gaze and facial expression features to be most effective.

Our dataset can be further improved by collecting more videos of sober and drunk people from online sources or in controlled environments. The classification network may be further improved by using models which have been shown to be effective for emotion recognition in videos, such as 3D convolutional networks.

Additionally, while our work only considered facial features, body movement and expressions have been successfully used for affect recognition [16]. Future work may incorporate such features to analyze the difference in body movements of drunk and sober people. Speech analysis has also been shown to be effective for detecting intoxication. We can incorporate audio features of the subject and perform multimodal feature fusion to accurately detect drunk people.

REFERENCES

[1] Driver assistance systems for urban areas. 2018. www.bosch-mobility-solutions.com/en/highlights/automated-mobility/driver-assistance-systems-for-urban-areas.

[2] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*. IEEE, 354–361.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[4] Fadi Biadsy, William Yang Wang, Andrew Rosenberg, and Julia Hirschberg. 2011. Intoxication detection using phonetic, phonotactic and prosodic cues. In *Twelfth Annual Conference of the International Speech Communication Association*.

[5] Eva Susanne Capito, Stefan Lautenbacher, and Claudia Horn-Hofmann. 2017. Acute alcohol effects on facial expressions of emotions in social drinkers: a systematic review. *Psychology research and behavior management* 10 (2017), 369.

[6] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. 2016. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence* 38, 8 (2016), 1548–1568.

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

[8] Drinking, Driving: A Road Safety Manual for Decision-Makers, and Practitioners. 2007. Global Road Safety Partnership, Geneva, Switzerland.

[9] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 467–474.

[10] Paul Ekman and Wallace V Friesen. 1978. *Manual for the facial action coding system*. Consulting Psychologists Press.

[11] Rana El Kaliouby and Peter Robinson. 2005. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*. Springer, 181–200.

[12] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

[13] Mohamed Fazeen, Brandon Gozick, Ram Dantu, Moiz Bhukhiya, and Marta C González. 2012. Safe driving using mobile phones. *IEEE Transactions on Intelligent Transportation Systems* 13, 3 (2012), 1462–1468.

[14] Carl-Magnus Ideström and Björn Cadenius. 1968. Time relations of the effects of alcohol compared to placebo. *Psychopharmacologia* 13, 3 (1968), 189–200.

[15] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 46–53.

[16] Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing* 4, 1 (2013), 15–33.

[17] Georgia Koukiou and Vassilis Anastassopoulos. 2012. Drunk person identification using thermal infrared images. *International journal of electronic security and digital forensics* 4, 4 (2012), 229–243.

[18] Driver Attention Monitor. 2018. owners.honda.com/vehicles/information/2017/CR-V/features/Driver-Attention-Monitor.

[19] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 5–pp.

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conference*.

[21] Pyannote-video. 2018. https://www.github.com/pyannote/pyannote-video.

[22] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. 2011. The INTERSPEECH 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*.

[23] Jéssica Bruna Santana Silva, Eva Dias Cristino, Natalia Leandro de Almeida, Paloma Cavalcante Bezerra de Medeiros, and Natanael Antonio dos Santos. 2017. Effects of acute alcohol ingestion on eye movements and cognition: A double-blind, placebo-controlled study. *PloS one* 12, 10 (2017), e0186061.

[24] Global status report on road safety 2015. 2015. World Health Organization: Geneva, Switzerland, 2015.

[25] MPEG Video. 1998. SNHC,âĂIJText of ISO/IEC FDIS 14 496-3: Audio,âĂİ. *Atlantic City MPEG Mtg* (1998).

[26] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3756–3764.

[27] Chung Kit Wu, Kim Fung Tsang, Hao Ran Chi, and Faan Hei Hung. 2016. A precise drunk driving detection using weighted kernel based on electrocardiogram. *Sensors* 16, 5 (2016), 659.