Attention Monitoring and Hazard Assessment with Bio-Sensing and Vision: Empirical Analysis Utilizing CNNs on the KITTI Dataset

Siddharth and Mohan M. Trivedi

Abstract— Assessing the driver's attention and detecting various hazardous and non-hazardous events during a drive are critical for driver's safety. Attention monitoring in driving scenarios has mostly been carried out using vision (camera-based) modality by tracking the driver's gaze and facial expressions. It is only recently that bio-sensing modalities such as Electroencephalogram (EEG) are being explored. But, there is another open problem which has not been explored sufficiently yet in this paradigm. This is the detection of specific events, hazardous and non-hazardous, during driving that affects the driver's mental and physiological states. The other challenge in evaluating multi-modal sensory applications is the absence of very large scale EEG data because of the various limitations of using EEG in the real world. In this paper, we use both of the above sensor modalities and compare them against the two tasks of assessing the driver's attention and detecting hazardous vs. non-hazardous driving events. We collect user data on twelve subjects and show how in the absence of very large-scale datasets, we can still use pre-trained deep learning convolution networks to extract meaningful features from both of the above modalities. We used the publicly available KITTI dataset for evaluating our platform and to compare it with previous studies. Finally, we show that the results presented in this paper surpass the previous benchmark set up in the above driver awareness-related applications.

I. INTRODUCTION

With the development of increasingly intelligent and autonomous vehicles it has been possible to assess the criticality of a situation much before the event actually happens. It has also become possible to monitor the driver's responses to various events during the drive. While computer vision continues to be the preferred sensing modality for achieving the goal of assessing driver awareness, the use of bio-sensing systems in this context has received wide attention in recent times [1], [2]. Most of these studies have used electroencephalogram (EEG) as the preferred bio-sensing modality. While these studies have shown that EEG can prove to be very useful for assessing fatigue and attention in the driving context, it generally suffers from low spatial resolution. Furthermore, the use of high-density EEG systems is impractical in a real-world driving context.

Driver awareness depends highly on the driver's physiology since different people react differently to fatigue and to their surroundings. This means that a single fit-for-all type of approach using computer vision based on eye blinks/closure etc. might not scale very well across drivers. It is here that the use of EEG may play a useful role in

Siddharth and Mohan M. Trivedi are with The Laboratory for Intelligent and Safe Automobiles (LISA) at the University of California San Diego, La Jolla, CA, USA, 92093. ssiddhar@eng.ucsd.edu, mtrivedi@eng.ucsd.edu

assessing driver awareness by continuously monitoring the human physiology. Furthermore, EEG may prove to be very useful for detecting hazardous vs. non-hazardous situations on short time scales (such as 1-2 seconds) if such situations do not register in the driver's facial expressions in such short time periods.

The advent of deep learning has translated very well towards vision-based systems for many applications, among them, driver behavior and attention monitoring [3], [4]. But, these advances have not translated towards the data from bio-sensing modalities such as EEG. This is primarily due to the difficulty in collecting very scale bio-sensing data. Collecting bio-sensing data on a large scale is costly, laborious, and time-consuming, whereas for collecting image/videos even a smartphone's camera may suffice. Hence, we explore the use of pre-trained image-based deep learning networks to extract meaningful features for both sensor modalities.

This study focuses on driver awareness and his/her perception of hazardous/non-hazardous situations from bio-sensing as well as vision-based perspectives. We utilize the KITTI dataset [5] for evaluation and individually use features from EEG and image data to compare the performance of these modalities. We also use the fusion of features and show how in certain cases, the use of multiple modalities may be advantageous. We also show how pre-trained deep neural networks can be utilized to extract features from these modalities for boosting the performance even in the absence of very large scale data. To the best of our knowledge, this study is the most comprehensive view of using EEG and vision modalities towards assessing driver awareness and hazard assessment. Finally, we would like to emphasize that the data collection set up used in this study is very practical to use in "real-world" i.e. it is compact in design, wireless, and comfortable to use for prolonged time intervals.

II. RELATED STUDIES

Driver monitoring for assessing attention, awareness, behavior prediction, etc. has usually been done using vision as the preferred modality [6], [7], [8]. This is carried out by monitoring the subject's facial expressions and eye-gaze [9], [10] which are used to train machine learning models. But, almost all such studies utilizing "real-world" driving scenarios have been conducted during daylight when ample ambient light is present. Even if infra-red cameras are used to conduct such experiments at night, vision modality suffers from occlusion and widely varying changes in illumination [6], both of which are not uncommon in driving scenarios.

Hazard assessment is a problem that has been tackled with vision as the primary modality [11] and evaluated on the KITTI dataset [5]. Additionally, it has also been shown that the use of EEG can classify hazardous vs. non-hazardous situations over short time periods which is not possible with images/videos [12]. We take this recent study on the KITTI dataset as the benchmark for our evaluation. The KITTI dataset contains vision data (along with other sensors) collected in rural areas and on highways in Karlsruhe city of Germany. The dataset contains many driving scenarios with up to 15 cars and 30 pedestrians visible per image. This makes it ideal for our study since the scene complexity in the dataset varies a lot more than many other such datasets.

The study [12] suffers from three major limitations. First, the study superimposes non-existent stimulus over the real-world driving images which introduce an uncanny aspect since such stimuli are not what one expects during a real-world drive. Second, the study does not use the most common sensor modality to assess the driver's awareness i.e. vision. Hence, no baseline comparison can be done between the two kinds of sensor modalities. Third, the EEG features used in the research are the most commonly used ones and are in no way tuned to the specific application at hand. No attempt has been made to extract higher-level EEG features that are more relevant to human cognition. Finally, the study contains a pool of only five subjects which is quite small since the usual norm is to use at least ten subjects for bio-sensing based applications.

III. RESEARCH METHODS

In this section, we discuss the various research methods that we employed to pre-process the data and extract features from each sensor modality used in this study.

A. EEG-based Feature Extraction

The cognitive processes pertaining to attention and mental load such as while driving are not associated with only one part of the brain. Hence, our goal was to map the interaction between various regions of the brain to extract relevant features related to attention. The EEG was initially recorded from 14-channel Emotiv EEG headset at 128 Hz sampling rate. The locations of the EEG channels according to the International 10-20 system were: AF3, AF4, F3, F4, F7, F8, FC5, FC6, P7, P8, T7, T8, O1, and O2. This EEG headset was chosen because it is compact, wireless, and easy to use in real-world settings. We first pre-processed the data using EEGLAB [13] toolbox. We used the artifact subspace reconstruction (ASR) pipeline in the toolbox to remove artifacts related to eye blinks, movements, line noise, etc. [14]. Then, the cleaned EEG data was band-pass filtered between 4-45 Hz to preserve the data from the most-commonly used EEG frequency bands. We then employed two distinct and novel methods to extract EEG features that captured the interplay between various brain regions to map human cognition.

1) Features based on Mutual Information: To construct the feature space that can map the interaction of EEG information between various regions of the brain, we calculated the mutual information between signals from different parts of the brain [15], [16]. Such features were opted for since they measure the changes in EEG across the various regions of the brain which might be more expressive of human cognition rather than spatially local features. The mutual information $I(X;Y)$ of discrete random variables X and Y is defined as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

The desired feature of conditional entropy $H(Y|X)$ is related to the mutual information $I(X;Y)$ by

$$I(X;Y) = H(Y) - H(Y|X) \quad (2)$$

We calculated the conditional entropy using mutual information between all possible pairs of EEG electrodes for a given trial. Hence, for 14 EEG electrodes, we calculated 91 EEG features based on this measure.

2) Features based on Deep Learning: The most commonly used EEG features are the power-spectrum density (PSD) of different EEG bands. But, these features in themselves do not take into account the EEG-topography i.e. the location of EEG electrodes. Hence, we try to exploit EEG-topography to extract information regarding the interplay between different brain regions. We also wanted to utilize pre-trained convolution networks since training a deep neural network from scratch needs a very large amount of data which is difficult to acquire. Thus, we sought to convert time-domain EEG data to image domain for utilizing such pre-trained convolution networks.

We calculated the PSD of three EEG bands namely theta (4-7 Hz), alpha (7-13 Hz) and beta (13-30 Hz) for all the EEG channels. The choice of these three specific EEG bands was made since they are the most-commonly used EEG bands and have been shown in multiple studies to contribute significantly towards human cognition. We averaged the PSD for each band thus calculated over the complete trial. These features from different EEG channels were then used to construct 2-D EEG-PSD heatmap for each of the three EEG bands using bicubic interpolation. These heat-maps now contain the information related to EEG topography in addition to spectrum density at every location [17].

Fig. 1. PSD heat-maps of the three EEG bands i.e. theta (red), alpha (green), and beta (blue) EEG bands are added according to respective color-bar range to get combined RGB heat-map image.(Circular outline, nose, ears, and color-bars have been added for visualization only.)

Fig. 1 shows an example of these 2-D heatmaps for each of the three EEG bands. As can be seen from the figure, we plot each of the three EEG bands using a single color channel i.e. red, green and blue. We then add these three color band

images to get a color RGB image containing information from the three EEG bands. The three color band images are added in proportion to the amount of EEG power in the three bands using alpha blending [18] by normalizing each band's power using the highest value in the image. Hence, following this procedure we are able to represent the information in the three EEG bands along with topography through a single color image. The interaction between these three colors (thus forming new colors by adding the three primary colors) in various quantities is representative of the information about the distribution of spectral power across the brain.

This combined colored image representing EEG-PSD with topography information is then fed to a pre-trained deep-learning based VGG-16 convolution neural network [19] to extract features from this image. This network consists of 16 weight layers and has been trained with more than a million images for 1,000 object categories using the Imagenet Database [20]. Previous research studies [21], [22] have shown that using features from such "off-the-shelf" neural network can be used for various classification problems with good accuracy. The EEG-PSD colored image is resized to $224 \times 224 \times 3$ for input to the network. The last layer of the network classifies the image into one of the 1000 classes but since we are only interested in "off-the-shelf" features, we extract 4,096 most significant features from the last but one layer of the network. The EEG features from this method are then combined with those from the mutual information method for further analysis.

B. Facial expression-based feature extraction

The analysis of facial expressions has been the preferred modality for driver attention analysis. Most of the research work in this area has been done by tracking fixed localized points on the face based on face action units (AUs).

First, we extracted the face region from the frontal body image of the person captured by each camera frame. This was done by fixing a threshold on the image size to reduce its extreme ends and placing a threshold of minimum face size to be 50×50 pixels. This was done to remove any false positives and decrease the computational space for face detection. We then used the Viola-Jones object detector with Haar-like features [23] to detect the most likely face candidate. In extremely uncommon cases when the face detector failed due to major occlusion by the subject's hands in front of their face, the frames were discarded.

Fig. 2. Detected face (marked in red) and face localized points (marked in green) for two participants (left and center) in the study, and some of the features (marked in yellow) computed using the coordinates of the face localized points.

1) Facial points localization based features: Face action units (AUs) has been used for a variety of applications ranging from affective computing to face recognition [24]. Our goal was to use face localized points similar to AUs without identifying facial expressions such as anger, happiness, etc. since they are not highly relevant in driving domain and short time intervals. We applied the state-of-the-art Chehra algorithm [25] to the extracted face candidate region from above. This algorithm outputs the coordinates of 49 localized points representing various features of the face as in Fig. 2. The choice of this algorithm was done because of its ability to detect these localized points through its pre-trained models and hence not needing training for any new set of images. These face localized points are then used to calculate 30 different features based on the distances such as between center of the eyebrow to the midpoint of the eye, between the midpoint of nose and corners of the lower lip, between the midpoints of two eyebrows, etc. and angles between such line segments. To remove variations by factors such as distance from the camera and face tilt, we normalized these features using the dimensions of the face region. To map the variation in these features across a trial, we calculated the mean, 95th percentile (more robust than maximum), and standard deviation of these 30 features across the frames in the trial. In this manner, we computed 90 features based on face-localized points from a particular trial.

2) Deep Learning-based features: For the extraction of deep learning-based features, we used the VGG-Faces networks instead of VGG-16 [26]. This was done to extract features more relevant to faces since the VGG-Faces network has been trained on more than 2.6 million face images from more than 2,600 people rather than on various object categories in the VGG-16 network. Similar to the feature extraction methods above, we sent each face region part to the network and extract the most significant 4,096 features. To represent the changes in these features across the trial i.e. across the frames, we calculated the mean, 95th percentile, and standard deviation of the features across the frames in a trial. The features from the above two methods were then combined for further analysis.

C. Assessing trends of sensor features using Deep Learning

The features discussed in sections III.A and III.B above were computed over the whole trial such as by generating a single EEG-PSD image for a particular trial. This is a special case when the data from the whole trial is being averaged. Here, we propose a novel method to compute the trend of these features i.e. their variation in a trial based on deep learning. To compute features with more resolution we generated multiple EEG-PSD images for successive time durations in a trial. Fig. 3 shows the network architecture for this method. The EEG-PSD images are generated for multiple successive time durations in a trial each of which is then sent to the VGG-16 network to obtain 4,096 most significant features. Similarly, this process was done for conditional entropy features by calculating this over multiple time periods in a trial rather than once on the whole trial.

Fig. 3. Network architecture for EEG-PSD trend based Deep Learning method.

We then use principal components analysis (PCA) [27] to reduce the feature size to 60 to save computational time in the next step. These $60{\times}N$ (N = number of successive time intervals) features were then sent as input to a Long Short Term Memory (LSTM) network [29]. The same process was performed for face-based features.

The LSTM treats each of these features as a time-series and is trained so as to capture the trend in each of them for further analysis. This method could only be applied when the time duration of the trials is fixed since the length of each time series should be the same. Since the duration of KITTI videos varied widely, we applied this method only in the trials used for detecting hazardous/non-hazardous situations.

Fig. 4. Experiment setup for multi-modal data collection. (A) EEG Headset, (B) External camera, and (C) Driving videos displayed on the screen. The subject sits with her/his arms and feet on a driving simulator with which s/he interacts while watching the driving videos.

IV. DATASET DESCRIPTION

In Fig. 4 we show the experimental setup for data collection with driving videos used as the stimulus in our dataset. The participants are comfortably seated equipped with an EEG headset (Emotiv EPOC). Facial expressions of each subject are recorded using a camera in front of him/her. The participants are asked to use a driving simulator which they are instructed to control as per the situation in the driving stimulus. For example, if there is a "red light" or "stop sign" at any point in a driving stimulus video, the participants should press and hold the brake. Twelve participants (most of them in their 20s with two older than 30 years) based in San Diego participated in our study. All the modalities were

synchronized together using Lab Streaming Layer (LSL) software framework [28].

For consistency between our work and other previous studies [12], [11], we used 15 video sequences from the KITTI dataset [5]. These video sequences range from 14 to 105 seconds. These videos in the dataset were recorded at $1242{\times}375$ resolution at 10 frames-per-second (fps). We resized the videos to $1920{\times}580$ to fit the display screen in a more naturalistic manner. Two external annotators marked the above 15 video sequences as requiring low-driver attention or high-driver attention based on the context in the video. For example, the video instances where the car is not moving at all were characterized as low attention instances whereas driving through narrow streets with pedestrians on the road were labeled as instances with high driver attention required. These sequences contained videos from widely varying illumination and driving conditions (in street, highway, city, etc.) as shown in Fig. 5.

Fig. 5. Various image instances with varying illumination conditions and type of road (street, single-lane, highway, etc.) from the KITTI Dataset.

Second, 41 instances, each two second long were characterized as hazardous/non-hazardous. Fig. 6 presents some examples of instances from both categories. As an example, a pedestrian suddenly crossing the road "unlawfully" or a vehicle overtaking suddenly represents hazardous events whereas "red" traffic sign at a distance and a pedestrian at a crossing with ego vehicle not in motion are examples of non-hazardous events. Among these instances, 20 instances were labeled as hazardous whereas rest as non-hazardous. Hence, the goal is to classify such instances in a short time period of two seconds using the above sensor modalities.

V. QUANTITATIVE ANALYSIS OF EEG AND VISION SENSOR MODALITIES

In this section, we present the single modality and multi-modal evaluation results for driver attention analysis and hazardous/non-hazardous instances classification. For each modality, we first used PCA [27] to reduce the number of features from the above algorithms to 30. We then used extreme learning machines (ELM) [30] for classification. These features were normalized between -1 and 1 across the subjects before training. A single hidden layer ELM was used with triangular basis function for activation. For the method with trend based temporal EEG and face features data, we used two layer LSTM with 200 and 100 neurons in respective layers. The network training was done using stochastic gradient descent with a momentum (SGDM) optimizer.

Fig. 6. (A) Examples of 2-seconds incidents classified as hazardous. Examples include pedestrians crossing the street without a crosswalk while the ego vehicle is being driven and another vehicle overtaking suddenly. (B) Examples of 2-seconds incidents classified as non-hazardous. Examples include stop signs and railway crossing signs.

We performed leave-one-subject-out cross-validation for each case. This meant that the data from 11 subjects was used for training at a time and the classification was done on the remaining 12th subject. This choice of cross-validation was driven by two factors. First, this method of cross-validation is much more robust and less prone to bias than models such as leave-one-sample-out cross-validation that constitutes training data from all the subjects at any given time. Second, since the data contained 180 trials only, as opposed to thousands of trials for any decent image-based deep-learning dataset, it did not make sense to randomly divide such a small number of trials to training, validation, and test sets since it might introduce bias by uneven division across trials from individual subjects.

A. Evaluating attention analysis performance

In this section, we evaluate single and multi-modal performance for assessing the driver's attention across the video trials. For all the four modalities, the features as defined above were calculated for data from each video trial. The ELM-based classifier was then trained based on each video trial divided into one of the two classes representing low-attention and high-attention required by the driver.

Fig. 7 shows the result for each modality for attention classification. The mean performance across the subjects for EEG, faces, and EEG + faces combined are 93.33 %, 81.67%, and 92.78% respectively. Clearly, EEG performs

Fig. 7. Single modality classification performance for driver attention analysis.

better than vision modality, so much so that adding the features from the two together doesn't lead to an increase in the performance. We also see that across most of the subjects, EEG performs better than vision modality. This is perhaps because the slowly-varying facial expressions during a driving task are not so expressive as to map the driver's attention very well. It is also to be noted that for all the cases, the accuracy of our system is significantly higher than chance (50% accuracy). We also note that the accuracy between subjects varies between EEG and faces modality. This may be due to some subjects being facially more expressive than others in the driving scenario.

B. Evaluating hazardous/non-hazardous incidents classification

In this section, we present the results of the evaluation of the modalities over very short time intervals (2 seconds) pertaining to hazardous/non-hazardous driving incidents as shown in Fig. 6. Since it is not possible for the subjects to tag the incidents while they are participating in the driving simulator experiment and hence these incidents were marked by the external annotators as mentioned above in Section IV.

Fig. 8. Multi-modality classification performance for driver attention analysis.

The mean accuracies across the subjects for EEG, faces, EEG + faces were 88.41%, 82.93%, and 90.24%. This shows that again EEG performs better than vision modality but more importantly, the combined performance of these modalities is better than individual ones. This means that on very short time periods (such as 2 seconds here), face features contain information independent and additive to what EEG features have. It is also to be noted that the performance of EEG in our feature extraction and processing pipeline is

significantly better than the previous hazard analysis study (AUC 0.79) [12]. Also, again we see that the accuracies for all cases are significantly higher than chance level.

We also note that on performing analysis with finer temporal resolution using LSTMs, the performance further increases to an average of 96.34%. This shows how the type of traditional and deep learning features we extracted above can be used for such a task with a good performance by utilizing multiple sensor modalities.

VI. CONCLUDING REMARKS

The use of bio-sensing modalities combined with audio-visual ones is rapidly expanding. With the advent of compact bio-sensing systems capable of data collection during real-world tasks such as driving, it is natural that this research area will gather more interest in the coming years. In this work, we showed how a commercially available compact EEG system with the capability of easily collecting data in the driving scenario can be used to monitor driver's attention in both short and long periods of time. We also presented a pipeline to process data from individual modalities for the applications of classifying hazardous vs. non-hazardous incidents while driving. Furthermore, we presented a method to be able to use pre-trained convolution neural networks to extract deep-learning based features from these modalities in addition to traditionally used ones. In future, we would explore the addition of eye-tracking as another sensor modality to study gaze dynamics and automatic stimuli tagging.

ACKNOWLEDGMENT

REFERENCES

[1] Lee, B.G., Lee, B.L. and Chung, W.Y., 2014. Mobile healthcare for automatic driving sleep-onset detection using wavelet-based EEG and respiration signals. Sensors, 14(10), pp.17915-17936.

[2] Guo, Z., Pan, Y., Zhao, G., Cao, S. and Zhang, J., 2018. Detection of Driver Vigilance Level Using EEG Signals and Driving Contexts. IEEE Transactions on Reliability, 67(1), pp.370-380.

[3] Kuefler, A., Morton, J., Wheeler, T. and Kochenderfer, M., 2017, June. Imitating driver behavior with generative adversarial networks. In Intelligent Vehicles Symposium (IV), 2017 IEEE (pp. 204-211). IEEE.

[4] Dwivedi, K., Biswaranjan, K. and Sethi, A., 2014, February. Drowsy driver detection using representation learning. In 2014 IEEE International Advance Computing Conference (IACC) (pp. 995-999).

[5] Geiger, A., Lenz, P., Stiller, C. and Urtasun, R., 2013. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 32(11), pp.1231-1237.

[6] Dong, Y., Hu, Z., Uchimura, K. and Murayama, N., 2011. Driver inattention monitoring system for intelligent vehicles: A review. IEEE transactions on intelligent transportation systems, 12(2), pp.596-614.

[7] Deo, N. and Trivedi, M.M., 2018. Looking at the Driver/Rider in Autonomous Vehicles to Predict Take-Over Readiness. arXiv preprint arXiv:1811.06047.

[8] Ohn-Bar, E. and Trivedi, M.M., 2016. Looking at humans in the age of self-driving and highly automated vehicles. IEEE Transactions on Intelligent Vehicles, 1(1), pp.90-104.

[9] Martin, S., Vora, S., Yuen, K. and Trivedi, M.M., 2018. Dynamics of Driver's Gaze: Explorations in Behavior Modeling and Maneuver Prediction. IEEE Transactions on Intelligent Vehicles, 3(2), pp.141-150.

[10] Doshi, A. and Trivedi, M.M., 2012. Head and eye gaze dynamics during visual attention shifts in complex environments. Journal of vision, 12(2), pp.9-9.

[11] Kolkhorst, H., Tangermann, M. and Burgard, W., 2017, March. Decoding Perceived Hazardousness from User's Brain States to Shape Human-Robot Interaction. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (pp. 349-350). ACM.

[12] Kolkhorst, H., Burgard, W. and Tangermann, M., Decoding Hazardous Events in Driving Videos, Proceedings of the 7th Graz Brain-Computer Interface Conference 2017.

[13] Delorme, A. and Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 134(1), pp.9-21.

[14] Mullen, T., Kothe, C., Chi, Y.M., Ojeda, A., Kerth, T., Makeig, S., Cauwenberghs, G. and Jung, T.P., 2013, July. Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG. In 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 2184-2187). IEEE.

[15] Peng, H., Long, F. and Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), pp.1226-1238.

[16] Siddharth, S., Patel, A., Jung, T.P. and Sejnowski, T., 2018. A Wearable Multi-modal Bio-sensing System Towards Real-world Applications. IEEE Transactions on Biomedical Engineering.

[17] Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. "Multi-modal Approach for Affective Computing." arXiv preprint arXiv:1804.09452 (2018).

[18] Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(3), pp.90-95.

[19] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[20] Russakovsky, O., Deng, J., Su, H., et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV). Vol 115, Issue 3, 2015, pp. 211252

[21] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition, arXiv:1403.6382v3, 2014.

[22] Siddharth, Rangesh, A., Ohn-Bar, E. and Trivedi, M.M., 2016, November. Driver hand localization and grasp analysis: A vision-based real-time approach. In Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on (pp. 2545-2550). IEEE.

[23] Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-I). IEEE.

[24] Tian, Y.I., Kanade, T. and Cohn, J.F., 2001. Recognizing action units for facial expression analysis. IEEE Transactions on pattern analysis and machine intelligence, 23(2), pp.97-115.

[25] Asthana, A., Zafeiriou, S., Cheng, S. and Pantic, M., 2014. Incremental face alignment in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1859-1866).

[26] Parkhi, O.M., Vedaldi, A. and Zisserman, A., 2015, September. Deep Face Recognition. In BMVC (Vol. 1, No. 3, p. 6).

[27] Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), pp.37-52.

[28] Kothe, C., 2014. Lab streaming layer (LSL). https://github.com/sccn/labstreaminglayer. Accessed on October, 26, p. 2015.

[29] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

[30] Huang, G.B., Zhu, Q.Y. and Siew, C.K., 2006. Extreme learning machine: theory and applications. Neurocomputing, 70(1-3), pp.489-501.