2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

DeepDriver: Automated System For measuring Valence and Arousal in Car Driver Videos

Rajkumar Theagarajan, Bir Bhanu and Albert Cruz*
Center for Research in Intelligent Systems, University of California, Riverside, CA - 92521
*Computer Perception Lab, California State University, Bakersfield, CA 93311

Email: rthea001@ucr.edu, bhanu@cris.ucr.edu, acruz37@csub.edu

***Abstract*—We develop an automated system for analyzing facial expressions using valence and arousal measurements of a car driver. This information is used by Motor Trends magazine to provide car manufacturers a report on how the drivers felt at each moment on the race track. The reason for this is that, the drivers remember only a brief description of the emotions they felt after test driving a car. Our approach is a data driven approach and does not include any pre-processing done to the faces of the drivers. The motivation of this paper is to show that with large amount of data, deep learning networks can extract better and more robust facial features compared to state-of-the-art hand crafted features. The network was trained on just the raw facial images and achieves better results compared to state-of-the-art methods. Our system incorporates Convolutional Neural Networks (CNN) for detecting the face and extracting the facial features, and a Long Short Term Memory (LSTM) for modelling the changes in CNN features with respect to time. The system was evaluated on videos from the Motor Trend Magazines Best Driver Car of the Year 2014-16 and the AFEW-VA dataset. We compared our approach with state-of-the-art methods and show that our approach achieves the better results compared to seven other methods.**

I. Introduction

Facial expressions are an important indicator of a persons emotional state and intentions. Since the seminal work on automatic recognition of facial expressions and emotions in 1974 [1], the field has garnered increasing attention due to its numerous practical applications beyond the prediction of basic emotional states. Automatic facial expression recognition software has a myriad of applications such as human behavior analysis [2], medical applications [3] [4], and human-computer interfaces. Current algorithms tend to perform well in controlled environments. However, in the real world, there are no such videos where the environment is controlled and people act out expressions in an overt way. Real world videos tend to be unconstrained in settings and the emotions displayed by the people in these videos are usually complex and difficult to recognize. Currently, one of the biggest challenges to the field of facial expression and emotion recognition is the dynamic analysis of expressions in unconstrained, continuous videos for real life applications.

Feature extraction is the most important aspect of computer vision and the classification of facial expressions is no exception to this. After the ImageNet competition [5] entry of Krizhevsky et al. [6], state-of-the-art for feature extraction for many computer vision applications shifted towards Convolutional Neural Networks (CNN).

In this paper, we present a data driven approach to measure valence and arousal of car drivers. Our approach uses CNNs to localize the face of the car driver, extract the features and uses LSTM to model the changes in these features with respect to time. We evaluate our approach on two data sets collected from the Motor Trend Magazines Best Driver Car of the Year 2014- 16 and the AFEW-VA dataset. We compared our results with other state-of-the-art approaches on the same datasets and show that our approach achieves better results.

Our data consists of unconstrained videos taken of a individuals driving multiple cars around the Laguna Seca race track. Our goal is to understand and capture the emotions that the driver is feeling at the very moment in time as he is driving different sports cars. In general, humans tend to forget the emotion they feel during a given moment in time. At the end of an event, humans form biased opinions on the overall experience instead of remembering the emotions felt moment-to-moment [19].

In our approach we measure the Arousal and Valence of the race car driver by analyzing their dynamic changes in facial expressions to predict their emotional state. Arousal is the measure of how energetic one feels and valence is the measure of how pleasant a feeling is. Many previous studies have shown that high-arousal events are remembered better than low-arousal events [20]. In our case, the race car drivers experience a lot of high arousal events as they drive through the race track, which causes them to forget a lot of events and remember only a few important events. This leads to car manufacturers not able to get a proper feedback on how the driver felt driving the car. In the current scenario, experts need to go through hours of video data to analyze how the car drivers reacted to the performance of the car at each part of the race track.

We try to solve this problem by training a convolutional neural network to predict the Arousal and Valence scores of a car driver on a frame by frame level. The features learned by the CNN is then passed to a LSTM that tries to model the dynamic changes in the facial expression to further improve the accuracy of the prediction and hence classify the emotional state of the driver. Fig. 1 shows some examples of facial expressions from the Motor Trend's Dataset.

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

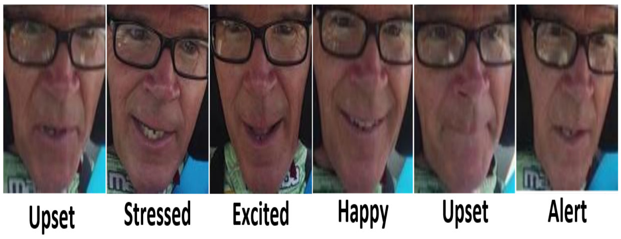

Fig. 1. Example facial expressions from the Motor Trend's Dataset.

II. Related Works and Contributions

One of the most commonly used appearance-based feature methods is Local Binary Pattern (LBP) [7]. However, LBP is a static texture descriptor that only captures the features of an image only at a single moment in time.

A variation of LBP called Volume Local Binary Patterns (VLBP) [7], was developed to capture dynamic textures. VLBP is an extension of LBP to the spatiotemporal domain and is capable to capturing dynamic textures, which are textures in the spatiotemporal domain.

The dimensionality of VLBP is 2^{3n+2} , where n is the number of neighboring pixels, which makes it impractical to use as the size of the neighborhood increases. An alternate solution to VLBP is the Local Binary Patterns in Three Orthogonal Planes (LBP-TOP). The dimensionality of LBP-TOP ($3x2^{n}$) is significantly lower than VLBP, accomplished by extracting two dimensional LBP patterns in the XY, YZ and XZ dimensions respectively.

The other major type of appearance feature is based on the Gabor filter, a set of band-pass filters that can approximate the low- level behavior of the human visual cortex [8]. Almaev et al. [9] combined the original Gabor filter with LBP-TOP and noticed increased accuracy in classification of facial expressions compared to using only LBP-TOP.

Traditional Gabor filters are too sensitive in unconstrained settings because they capture all of the edges within an image, noise included. Cruz et al. [10] proposed Anisotropic Inhibited Gabor Filter (AIGF) that is robust to background noise and computationally efficient. Theagarajan et al. [11] developed a novel feature representation that utilizes the background suppressing ability of AIGF [10] and the compact representation of LBP-TOP [9]. The authors evaluated their approach on the Motor Trends Magazine dataset and achieved a correlation of 0.598 for Valence and 0.494 for Arousal. They further used a decision level filter to fuse their approach with a CNN and reported improved performance.

Recently, there have been many works for emotion recognition using deep learning. Liu et al [21] used a boosted deep belief network for performing training in three stages in a unified loop. The authors mention that by training in iterative stages the network was able to characterize expression-related facial appearance/shape and strengthen the discriminative features of the related emotion.

Liu et al [22] used the Action Units (AU) to recognize facial expressions. The authors used a traditional CNN followed by an AU aware receptive field to enhance the discriminative features of individual AU's. Finally the learned features are passed through a restricted Boltzman machine to learn the hierarchical relationships between individual AU's to predict the emotion.

All these approaches assume the input images to be distinctive from each other. In our case, the facial expressions of the race car drivers are changing at a very fast pace and are not very distinctive. Hence, we measure physiological factors such as the Arousal and Valence to predict the drivers emotional state.

In light of the related works, the key contributions of this paper are:

- A data driven approach measure arousal and valence of car drivers.
- Evaluation on the Motor Trends Magazine dataset for the year 2014-16 [12].
- Comparison with existing state-of-the-art feature descriptors.
- Cross-evaluation on the AFEW-VA [13], dataset to show the generalizability of our approach.

III. Technical Approach

This section describes the framework and architecture of the individual model used in our approach. Fig. 2 shows the overall architecture of our approach.

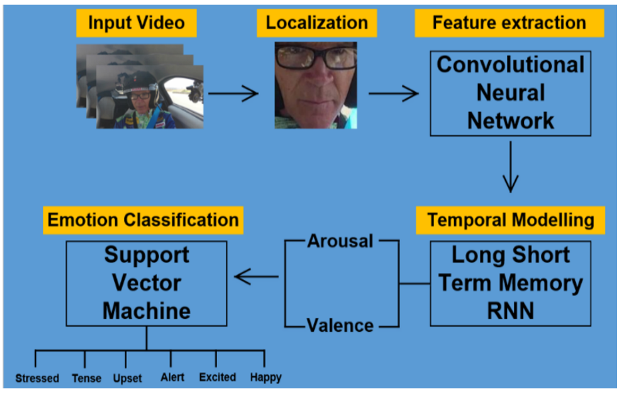

Fig. 2. Overall architecture of our approach.

A. Localization

The localization is done by training a YOLO V2 [14] real time object detector. The network was trained on the Pascal VOC dataset [15] and the FDDB dataset [16].

During training, each image is divided into a 11x11 grid. Each grid predicts a bounding box with two probabilities, which describe how confident it is that an object is present within the grid P(Object). Moreover, each grid also predicts what class is present in the grid given an object is present P(class|Object).

During testing, the image is divided into 11x11 grids and the bounding box inside the grid that has confidence level above a given threshold is selected as location of the target.

B. Feature Extraction

In our approach we perform feature extraction using state-of-the-art Convolutional Neural Networks. We employ the residual network architectures by [17]. We experimented with 3 different architectures namely: ResNet18, ResNet34 and ResNet50 pre-trained on the Imagenet dataset. The network takes a 224x224 patch of image and batch normalization is performed for faster training convergence. Rectified Linear units (ReLu) are used as non-linearities and we used the Mean Squared Error loss function. We used a Tanh after the final fully connected layer.

After training the network, the features are extracted after removing the last fully connected layer, which results in a 512x1 dimensional vector for ResNet18 and ResNet34 and 2048x1 dimensional vector for Resnet50.

C. Temporal Modelling

In order to detect the emotional state of the car driver in the video, we need to take into account the dynamic variations in facial expression. To do this we used a Long Short Term Memory Network (LSTM) [18].

LSTM networks contain units which remember information for either long or short time periods. Each LSTM units contain an Input gate, Forget gate and Output gate. The Input gate controls the extent to which new information flows into the memory. The Forget gate controls the amount of information that flows out of the memory and the Output gate controls the extent to which the information in memory is used to compute the output.

In our approach, the input to the LSTM is a M x N feature vector, where M is the length of the feature vector extracted from the CNN and N is the length of the image sequence we want to model. In our approach we chose N = 7 consecutive frames.

IV. Experimental Setup and Parameters

We evaluated our approach on 2 different datasets: Motor Trends Magazine Best Driver Car of the Year 2014-16 and the AFEW-VA. The overall framework of our approach is implemented on PyTorch framework with 2 TITAN X GPUs with 7 TFlops single precision, 336.5 GB/s of memory and 12 GB of RAM memory per board.

A. Datasets

We performed experiments on the Motor Trends Magazine Best Driver car of the Year 2014-16 and the AFEW-VA dataset. The videos in the first dataset consist of a driver test driving different sports cars. The camera used to record the driver is a GoPro that records videos at 24 frames a second. Each frame in all the videos are annotated with the Valence and Arousal measure. Additionally the 2014 dataset is also annotated with the emotional state of the driver. Currently the Motor trend's dataset is not yet available to the public. The IRB exemption for human testing was obtained and the dataset will be released soon.

The 2014 dataset consists of 10 videos, the 2015 dataset consists of 10 videos and the 2016 dataset consists of 12 videos. All the videos in the 2014 and 2016 dataset are approximately 100 seconds long, while the videos in the 2015 dataset are approximately 20 minutes long.

In order to make our system more generalizable, we also used the AFEW-VA dataset. The dataset consists of 600 movie clips. Each frame in all the videos is annotated with the 68 facial keypoints as well as the Valence and Arousal measure. For all our experiment we did 3 fold cross validation. Table I shows the data distribution for our training and testing dataset.

k-fold	Training dataset	Testing dataset
Fold 1	Motor Trend's 2014-15 + AFEW-VA (20+500 videos)	Motor Trend's 2016 + AFEW-VA (12+100 videos)
Fold 2	Motor Trend's 2015-16 + AFEW-VA (22+500 videos)	Motor Trend's 2014 + AFEW-VA (10+100 videos)
Fold 3	Motor Trend's 2014, 2016 + AFEW-VA (20+500 videos)	Motor Trend's 2015 + AFEW-VA (10+100 videos)

TABLE I
DATA DISTRIBUTION FOR THE TRAINING AND TESTING DATASET

B. Metrics and Ground truth

The metrics used for classifying the emotions are Arousal and Valence. Arousal is the measure of how energetic one feels and valence is the measure of how pleasant a feeling is. The range of emotion intensity varies from -1 to +1.

The ground-truth labelling for the valence and arousal values for the Motor Trends Magazine dataset was done by 7 experts. The final ground-truth for each video was obtained by taking the average of each individual ground-truths. Both the video and audio were analyzed while labeling the ground-truth.

The ground-truth for the AFEW-VA dataset was done by 2 experts who are FACS AU certified. Both annotators annotated all videos together, therefore, they discussed all disagreements and coming up with a unique solution.

C. Localization Parameters

In our approach we set the learning rate for the YOLO V2 to 0.005 and the learning rate is annealed by a factor of 2 every {2000, 5000, 8000, 10,000} iterations.

The network was pre-trained on the Pascal VOC dataset and further fine-tuned on the FDDB dataset. The FDDB dataset consists of 5,171 faces in 2845 images. To augment more data, all the images were randomly horizontally flipped. We evaluated the network on 100 test videos from the AFEW-VA dataset. The network was able to achieve a detection rate of Intersection over Union (IoU) of 81.3% even in the presence of occlusions.

D. CNN Parameters

For all the experiments we evaluated the performance of the CNNs based on the correlation and RMS error. The network (ResNet34) was trained and evaluated based on the data 350 distribution as shown in Table I. The input to the network was the cropped facial image resized to 224x224 pixels. . We chose a mini-batch size of 128 and during every epoch the training data was randomly shuffled and randomly flipped.

We did random hyper-parameter selection for the network to find the best learning rate, momentum and weight decay. All the networks are optimized using the Stochastic Gradient Descent algorithm. Table II shows the best hyper-parameters for the individual networks.

Network	Learning rate	momentum	weight decay
ResNet18	6x10^{-3}	0.9	1x10^{-4}
ResNet34	6x10^{-3}	0.9	1x10^{-4}
ResNet50	1x10^{-3}	0.9	5x10^{-3}

TABLE II
BEST HYPER-PARAMETERS FOR DIFFERENT CNN'S

The networks were trained with minimum squared error loss function. The learning rate is decreased by a factor of 2 after every 4 epochs. After every epoch we check if the RMS error is decreasing and in that case we save the model. If the RMS error has not decreased after 3 consecutive epochs, we employ early stopping and the training has been completed. During training it was found that the best results were obtained using ResNet34 and hence throughout this paper, we used ResNet34 for extracting the features.

E. LSTM Parameters

The input to the LSTM is the feature extracted for the sequence of images from the CNN. We can think of this feature as a low dimensional representation of the sequence in the feature space.

In our experiment we chose the length of the sequence to be 30 consecutive images. We chose 30 frames as a sequence as this was the best window length that would capture the complete dynamic changes in the facial expression as they transition to different emotions. The sequences are chosen such that each frame is the last frame of a sequence exactly once, for example, if the first sequence is $s_0 = \{v_1, v_2, ..., v_{29}, v_{30}\}$, the next sequence is $s_1 = \{v_2, v_3, ..., v_{30}, v_{31}\}$. Each sequence is labeled with two labels t1 and t2, which is the Valence and Arousal label corresponding to the last frame in the sequence. Since each sequence has only 2 labels (Valence and Arousal) which corresponds to the last frame of the sequence, only the hidden state of the last time step h tn is used to compute the output of the network.

The network is optimized using the ADAM algorithm. The learning rate for the network was set to be 6x10 -5 with weight decay of 1x10 -1 , dropout of 60% and minimum squared error loss function.

V. EXPERIMENTAL RESULTS

A. CNN Results

Table III shows the results using just the CNN (ResNet34).

Dataset	Arousal Correlation	Arousal RMS	Valence Correlation	Valence RMS
MTD	0.601 ± 0.071	0.064 ± 0.018	0.570 ± 0.083	0.082 ± 0.021
AFEW-VA	0.587 ± 0.65	0.090 ± 0.034	0.623 ± 0.82	0.101 ± 0.047

TABLE III
RESULTS USING THE CNN

In Table III, MTD stands for the Motor Trend's Dataset and was tested using 3-fold cross validation and AFEW-VA dataset was tested using 100 unseen videos.

B. LSTM Results

We evaluated the LSTM such that each sequence is an individual video. The reason for this is that the network is trained with 1 sequence at a time, thus the network predicts the output label for the sequence by looking only at N frames and not the whole video. Table V shows the LSTM results.

Fig. 2 shows the Valence and Arousal plots with respect to their ground-truth for a video from the Motor Trends Magazine 2016 test dataset. Table IV shows the comparison of our results with the current state-of-the-art approaches.

C. Emotion Classification Results

Table VI shows the predicted emotion results. Currently, only the Motor Trends Magazine 2014 dataset is annotated with the emotional state of the driver.

Fig. 4 shows the Valence vs. Arousal plane with the data points corresponding to Table V overlaid on it. The predicted valence and arousal values are converted from Cartesian to polar coordinates to fit in the pie chart.

The emotions that the driver displayed in the dataset are classified as stressed **(ST)**, tense **(TE)**, upset **(UP)**, alert **(AL)**, excited **(EX)** and happy **(HA)**. To predict the emotions, we trained a SVM classifier using the ground-truth Arousal and Valence values from the Motor Trends Magazine 2014 dataset. The trained SVM was then tested using the LSTMs predicted output for the Motor Trends Magazine 2014 dataset.

The output of the LSTM is a single prediction of Arousal and Valence for every 30 frames. We were able to achieve an accuracy of 83.78% with false alarm of 8.56%. The authors of [11] used the C4.5 binary tree using the Weka machine learning tool and achieved an accuracy of 70.28% and false alarm of 18.02%. Thus, it is clearly seen that our results are better than the authors of [11].

Fig. 5 and Fig. 6 are examples when the driver was doing a corkscrew turn and a normal turn respectively. In Fig.5 and Fig. 6, the image to the left is the raw facial image of the driver and the image to the right is his current location on the Laguna Seca Racetrack. In Fig. 5, the driver was "Tense" as

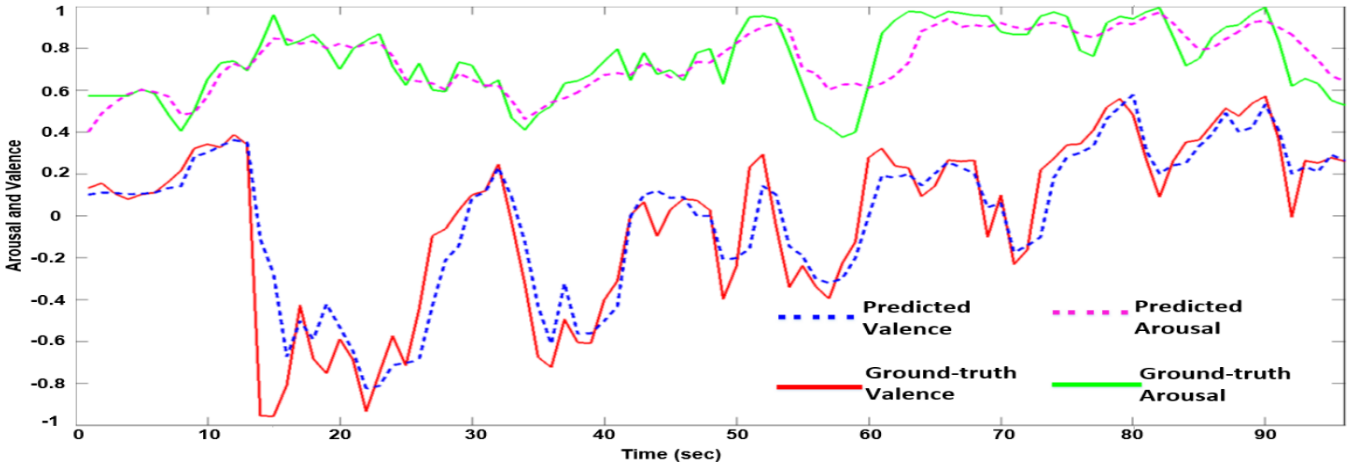

Fig. 3. Predicted Valence and Arousal values for a video from the Motor Trend's Dataset 2014-16.

Correlation	LBP	VLBP	LBP-TOP	Gabor	AI Gabor	LGBP-TOP	LAIBP-TOP	Proposed
Valence	0.255 ± 0.203	0.312 ± 0.156	0.341 ± 0.123	0.393 ± 0.243	0.491 ± 0.081	0.504 ± 0.196	0.633 ± 0.188	**0.642 ± 0.107**
Arousal	0.111 ± 0.095	0.291 ± 0.128	0.321 ± 0.207	0.302 ± 0.214	0.336 ± 0.214	0.370 ± 0.313	0.527 ± 0.201	**0.683 ± 0.087**

TABLE IV
COMPARISON OF OUR CORRELATION WITH STATE-OF-THE-ART ON THE MOTOR TREND'S DATASET 2014-16

Dataset	Arousal Correlation	Arousal RMS	Valence Correlation	Valence RMS
MTD	0.683 ± 0.087	0.079 ± 0.023	0.642 ± 0.107	0.065 ± 0.011
AFEW-VA	0.626 ± 0.079	0.087 ± 0.045	0.639 ± 0.109	0.093 ± 0.038

TABLE V
RESULTS USING THE LSTM

Class	ST	TE	UP	AL	EX	HA
ST	**50**	3	0	6	0	0
TE	2	**47**	0	1	2	0
UP	1	1	**42**	2	0	0
AL	2	4	1	**24**	1	0
EX	0	2	1	0	**12**	2
HA	1	1	1	0	2	**11**

TABLE VI
RESULTS USING THE LSTM

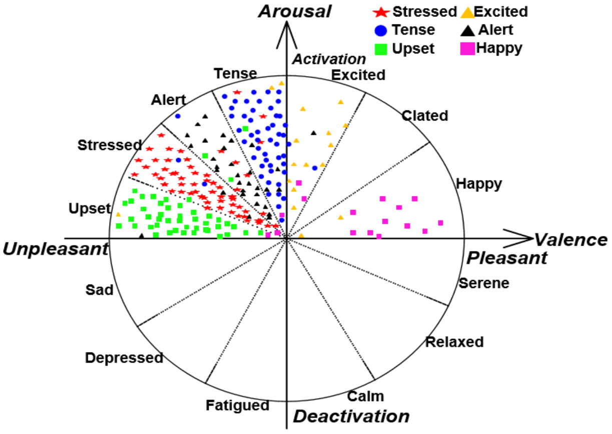

Fig. 4. Valence Vs Arousal Plane with emotion classification results.

he was attempting a dangerous corkscrew turn at a very high speed and our system was able to classify it correctly. In Fig. 6, the driver was "Happy" because the car had a good torque and was able to pick up speed very quickly after a normal turn.

VI. CONCLUSIONS

We developed a system for detecting the stress and inattention of car drivers from the Arousal and Valence values using a single front facing camera. The system is tested on the Motor Trends magazine Best Car of the Year 2014-16 and the AFEW-VA dataset. Our data driven approach shows improved results over the current state-of-the-art methods, on both the Arousal and Valence measurements as well as the emotion classification for the Motor Trends Magazine dataset.

VII. ACKNOLEDGEMENT

Acknowledgment: This work was supported in part by NSF grant 1330110 and ONR grant N00014-12-1-1026 . The contents of the information do not reflect the position or policy of US Government.. We also thank Dr. Albert Cruz and Motor Trend's Magazine for providing us with the dataset.

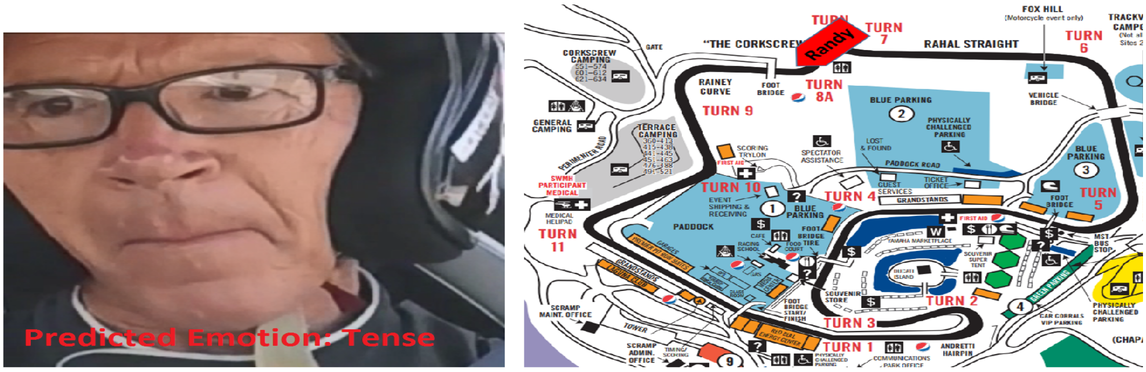

Fig. 5. Correctly predicted “Tense” sequence.

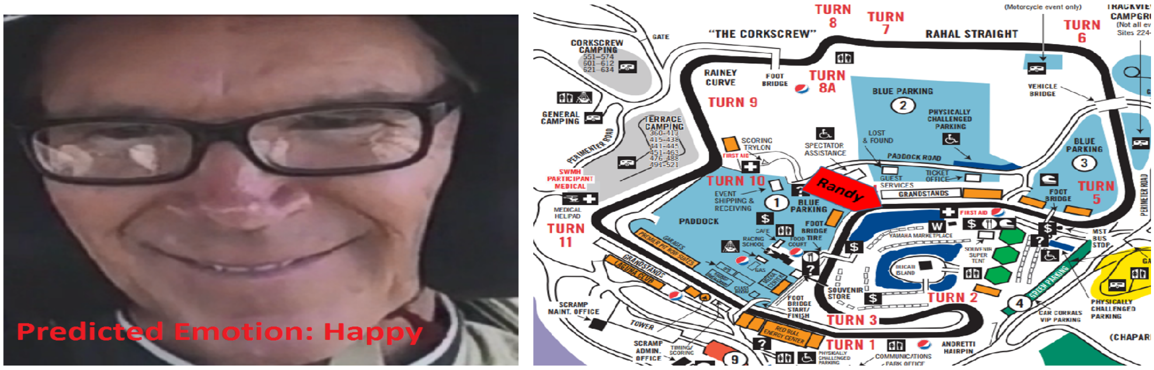

Fig. 6. Correctly predicted “Happy” sequence.

References

[1] F. I. Parke, A Parametric Model for Human Faces, 1974.

[2] W.J. Yan, Q. Wu, Y.J. Liu, S.J. Wang, and X. Fu, “CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces,” *IEEE Automatic Face and Gesture Recognition*, 2013 pp. 1-7.

[3] V. Bevilacqua, D. D’Ambruoso, G. Mandolino, and M. Suma, “A new tool to support diagnosis of neurological disorders by means of facial expressions,” *IEEE Medical Measurements and Applications Proceedings,* 2011, pp. 544-549.

[4] J.M. Girard, J.F. Cohn, M.H. Mahoor, S. Mavadati, and D.P. Rosenwald, “Social risk and depression: Evidence from manual and automatic facial expression analysis,” *IEEE Automatic Face and Gesture Recognition,* 2013, pp. 1-8.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma and A.C. Berg. “Imagenet large scale visual recognition challenge.”, 2015, *IJCV*, 115(3), pp. 211-252.

[6] A. Krizhevsky, I. Sutskever and G.E. Hinton. “Imagenet classification with deep convolutional neural networks,” *Advances in NIPS*, 2012.

[7] T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2006, 28(12), pp. 2037-2041.

[8] M.J. Lyons, “Automatic classification of single facial images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999, 21(12), pp. 13571362.

[9] T.R. Almaev and M.F. Valstar, “Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition,” *Conference on Affective Computing and Intelligent Interaction, ACII* 2013, pp. 356361.

[10] A. Cruz, B. Bhanu and N.S. Thakoor, “Backgound suppressing Gabor energy filtering,” *Pattern Recognition Letter*, 2015, 52, pp. 40-47.

[11] R. Theagarajan, B. Bhanu, A. Cruz, B. Le and A. Tambo, “Novel Representation for Driver Emotion Recognition in Motor Vehicle Videos,” *ICIP*, 2017.

[12] http://www.motortrend.com/news/the-future-of-testing-measuring-the-driver-as-well-as-the-car/

[13] J. Kossaifi, G. Tzimiropoulos, S. Todorovic and M. Pantic, “AFEW-VA database for valence and arousal estimation in-the-wild”, *Image and Vision Computing*, 2017.

[14] J. Redmon and A. Farhadi, “YOLO9000: better, faster, stronger.” *arXiv preprint* arXiv:1612.08242, 2016.

[15] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *International journal of computer vision*, 2015, 111(1), pp. 98-136.

[16] V. Jain, and E. Learned-Miller, “Fddb: A benchmark for face detection in unconstrained settings,” *Technical Report UM-CS-2010-009, University of Massachusetts,* Amherst, 2010, 88.

[17] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition.” *CVPR*, 2016, pp. 770-778.

[18] P. Rodriguez, G. Cucurull, J. Gonzlez, J.M. Gonfaus, K. Nasrollahi, T.B. Moeslund, and F.X. Roca, “Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification,” *IEEE Transactions on Cybernetics*, 2017.

[19] J.E. LeDoux, “Synaptic self: How our brains become who we are,” *Penguin*, 2003

[20] A.P. Ross, J.N. Darling and M.B. Parent, “High energy diets prevent the enhancing effects of emotional arousal on memory”. *Behavioral Neuroscience*, 2013, pp. 771-779, 127.

[21] P. Liu, S. Han, Z. Meng and Y. Tong, “Facial expression recognition via a boosted deep belief network”, *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805-1812.

[22] M. Liu, S. Li, S. Shan and X. Chen, “Au-aware deep networks for facial expression recognition”, *IEEE Face and Gesture Recognition workshops*, 2013, pp. 1-6.