2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)
September 12-14, 2018, Madrid, Spain

Deep Learning Based Real-Time Driver Emotion Monitoring

Bindu Verma
School of Computer and System Sciences
Jawaharlal Nehru University New Delhi, India
Email:binduverma67@gmail.com

Ayesha Choudhary
School of Computer and System Sciences
Jawaharlal Nehru University New Delhi, India
Email:ayeshac@mail.jnu.ac.in

***Abstract*—In this paper, we propose a novel, real-time driver emotion monitoring system "in the wild" based on face detection and facial expression analysis. A camera is placed inside the vehicle that continuously looks at the driver's face and monitors the driver's emotional state at regular time intervals. Camera based monitoring of the driver's attentiveness based on the driver's emotional state in naturalistic driving environments is a non-intrusive approach and an important part of an automated driver assistance system (ADAS). Our work employs a face detection model based on mixture of trees with shared pool of parts to robustly detect the drivers face in varied environmental conditions. We also extract facial landmark points, and use them to enhance our emotion recognition system. In our proposed work, we use convolution neural networks. In the first, we use VGG16 to extract appearance features from the detected face image and in the second VGG16 network, to extract geometrical features from the facial landmark points. We then combine these two features using an integration method to accurately recognize the emotions. Based on the recognized emotional state of the driver, the driver can be made aware of his emotional state in case necessary . Experimental results on publicly available driver and face expression datasets show that our system is robust and accurate for driver emotion detection.**

I. Introduction

In this paper, we propose a novel driver monitoring algorithm that automatically detects the driver's face and recognizes the facial expression to make an informed decision about the driver's emotional state. A camera is placed inside the vehicle that continuously looks at the driver. According to psychological studies, a driver's emotions are important in safe driving [1]–[4]. Anxiety, anger, contempt lead to dangerous driving causing road accidents. Similarly, inattentive, distracted driving also causes road accidents. When drivers are in a high emotional state, their minds are distracted and the alertness and ability to judge that is required for driving may be inadequate, leading to dangerous driving. Globally, 2.2% of deaths are because of road crashes and 90% of road mishaps are because of human error [5]. Studies have shown [6] that there are fewer incidents of accidents when drivers are alerted by co-passengers about the unseen hazards which were likely to cause collisions ($30\% - 43\%$). Therefore, it is necessary that the next generation vehicles should have an additional safety feature to alert the driver based on his or her emotional state. Therefore, monitoring drivers emotional state to ensure the driver's alertness becomes a key aspect of an Advanced Driver Assistance System (IDAS) and any non-alertness of driver can be reported for preventing mishaps.

Keeping the driver in an emotional state that is best suited for driving is very important. For safe driving, the driver's emotion should support capabilities like attention, accurate judgment of traffic situations, fast and correct decision making and appropriate communication with other drivers. Impaired emotions of the driver play a significant role in affecting the driving performance. It has been observed that emotions such as dejection, anger, disgust, sadness, frustration, etc., negatively impact the alertness of the driver [7]. Facial expressions are an outward display of a driver's emotions. There are seven basic emotions namely, sadness, happiness, surprise, anger, fear, disgust and contempt [8]. The behavior of dangerous driving can be observed by the facial expressions of the driver. Therefore, to automatically discover the driver's emotional state, it is important to automatically recognize the driver's facial expression. In this work, our aim is to discover the alertness of the driver in real-time, in his natural setting while driving, based on the emotions being experienced by the driver. The main challenges of a camera based driver monitoring system is that the drivers face should be correctly detected in spite of varied illumination conditions in naturalistic driving, occlusions and other passengers in the vehicle. Facial movements including head movements and changes in emotions across time makes it difficult to recognize emotions in real-time. Moreover, the driver monitoring system should be generic enough to accommodate different drivers.

In our proposed driver monitoring system, we assume that the camera mounted in the vehicle will continuously look at the drivers face. Moreover, the drivers face will be contained in a pre-specified region of the image within a certain range of resolution. To detect the drivers face in the natural setting ("in the wild") we use the face detection algorithm proposed in [9] and also extract facial landmark points (FLP). Our proposed work is an integration of features extracted from two VGG16 convolution neural networks. In the first network, the input to the fine-tuned VGG16 network are the detected region of interest (ROI) face images, to extract the appearance features at the fully connected layer. In the second VGG16 network, input to the fine-tuned VGG16 are the corresponding facial landmark points and high level features are extracted at the fully connected layer. Then, features from both the networks

978-1-5386-3543-8/18/$31.00 ©2018 IEEE

are integrated using weighted summation given in Equation 5 to classify the driver's emotion. Based on the recognized emotional state of the driver, warns the driver about his/her emotional state or performs an action such as play music to calm the driver.

We discuss the related work in Section II. Our proposed approach is explained in Section III and discuss experimental results and their comparisons with state-of-the-art methods in Section IV. Finally, we conclude in Section V.

II. Related Work

Camera based driver alertness monitoring has been an active area of research. However, most methods for driver alertness detection are based on the eye features of the driver, such as to check whether the eyes are open or not and where the gaze of the driver is. Mbouna et al. [10] proposed a method for detecting driver non-alertness by using the features of eye and head positions extracted using Adaboost algorithm. Then features are fed into a trained support vector machine to get the state of driver's alertness. Wang et al. [11] track a face in each frame and find the features of eye, mouth and head. Then, these facial features are analyzed to detect the driver emotions. Gao et al. [12] proposed a real-time framework to detect the driver's emotional state by analyzing facial expressions. A face tracker is used to track facial landmarks, SIFT descriptors are extracted for each landmark and linear SVM is used to classify the expression. Similarly Zhang et al. [13] proposed driver fatigue recognition using facial expression. They extract local facial features using local binary pattern (LBP) and SVM is used to classify the expressions.

Expression classification is a very challenging problem because slight variation in expressions may result in a totally different emotion because of the similarity of facial expressions when different emotions are being felt. For example, facial expressions in the case of sadness or fear are very similar. Convolution Neural Networks (CNN) extract very high level facial features and classify the expression. Mollahosseini et al. [14] propose a deep neural network architecture to classify facial expressions. They use two convolutional layers followed by max pooling and then four inception layers. Hong-Wei et al. [15] use transfer learning in CNN with supervised fine tuning to classify the facial expression. They follow two stages of fine tuning. At one stage they fine-tune the network on one dataset and at the second stage they train the network on training data of other datasets. Lopes et al. [16] proposed a five layered CNN with two convolutional layers, two sub-sampling layers and one fully connected layer and apply pre-processing for features specific for expressions to overcome the problem of lack of availability of large datasets for face expression. Heechul et al. [17] use two deep networks to recognize facial expression. In the first deep network, they extract temporal appearance features from image sequences while in the other deep network, temporal geometry features are extracted from face fiducial points. Finally, these two models are combined for classifying the facial expressions.

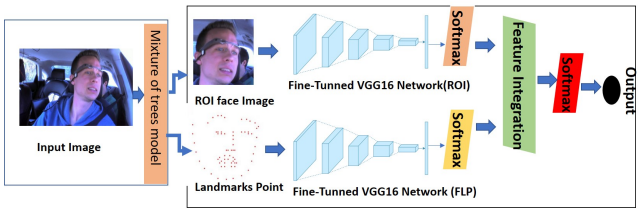

Fig. 1. An overview of our proposed approach: Detected face image and facial landmark points are inputs to the two VGG16 networks respectively. Then features of fully connected layer from both the networks are integrated for emotion classification.

Our work focuses on recognizing the driver's emotional state through recognition of the facial expressions of the driver in a natural environment. In our system, a camera is placed in the car in such a manner that the driver's face is continuously monitored. At fixed intervals of time, the image is analyzed for monitoring the emotions of the driver. To reduce the space time complexity, we use the fact that the driver's face will be present in a fixed (pre-specified) region of the image. Then, the region and orientation of the driver's face is detected in the pre-specified image region while he/she is driving using the face detection "in the wild" method from [9]. The detected face region image and the landmark points are separately used in two CNN to get the appearance and geometric features respectively. These features are then combined to develop a real-time facial expression based emotion recognition system.

III. Proposed Work

We propose a driver monitoring system where the drivers emotional state is monitored in real-time by processing the input images on-the-fly. Each image goes through the pre-processing steps and the output is the recognized emotional state of the driver. Flowchart of our proposed approach is shown in Figure 1.

A. Driver's Face Detection and Landmark Points

We assume that the camera is placed such that it continuously looks at the driver at a certain resolution range and propose a pre-specified window for face detection such that the drivers face is continuously available within this window. The driver's face can be in any orientation and the size of the face in the image can vary depending on the location of the camera center. We need to detect the drivers face despite these variations. We assume that the resolution of faces is within a certain range and use the method of [9], which detects faces "in the wild" with an accuracy of more than 95% for detecting the drivers face in the pre-specified window on the image. It also gives the landmark points and the orientation of the head. In the next stage, we use the landmark points to extract the geometrical features of facial expressions. Orientation of the face is used as the first criterion for detecting the drivers state of alertness because, if the head orientation is more than a certain range, such as extreme side profile, it implies that the driver is not attentive. Such images are clustered as *non-alert*

and do not go through further processing. If the orientation of the face is within a certain range, then the detected face regions and landmark points in these images are passed into the network for expression recognition.

The face detection model used in our system is based on mixture of trees with shared pool of parts [9]. Each facial landmark is modeled as a part and with the use of global mixtures, the topological viewpoint changes are captured. A part will only be visible in certain mixtures/views. Different mixtures share part templates in order to decrease the complexity for modeling many viewpoints. Finally, the model is discriminatively trained in a max-margin framework. This model uses histogram of oriented gradient (HOG) descriptor as the representation for each part. Head pose (yaw angle) is discretized for every 15 from -90 to $+90$ leading to 13 different viewpoints. Each such viewpoint is represented by a tree. For simplifying the underlying concept, let us select a frontal face representing tree $T = (V, E)$ where, V represents parts/vertices's and E represents edges. For an input image I, let $l_i = (x_i, y_i)$ be the location of the pixel for the part V_i and L denotes all such part locations. The scoring function scores the parts located at L on image I as given by Equation 1.

$$Score(I, L) = Appearance(I, L) + Shape(L) + \alpha \quad (1)$$

$$Appearance(I, L) = \Sigma w_i.\Phi(I, l_i) \quad (2)$$

$$Shape(L) = \Sigma a_{ij}dx + b_{ij}dx^2 + c_{ij}dy + d_{ij}dy^2 \quad (3)$$

Equation 2 gives the score of the local match of the template w_i with the HOG feature $\Phi(I, L_i)$ calculated at a particular location given by l_i. Equation 3 gives the shape score which places spatial constraints for each pair of parts V_i and V_j located by L where $dx = x_i - x_j$ and $dy = y_i y_j$ and can be interpreted as springs which help deform the facial landmarks to account for elastic deformation of the face. The parameters a_{ij}, b_{ij}, c_{ij} and d_{ij} of Equation 3 represent factors of springs such as the rest location and the rigidity, which were learnt by training on a set of images [9]. The bias term is associated with the corresponding viewpoint of the face. The inference corresponding to the model is to maximize $Score(I, L)$ over L as given in Equation 4.

$$Score^*(I, L) = max_L[Score(I, L)] \quad (4)$$

With the help of dynamic programming, the inner maximization of the tree $T = (V, E)$ can be efficiently done [9]. Assume that L^* corresponds to the L for which $Score(I, L)$ is maximum, that is, $Score^*(I, L) = Score(I, L^*)$. Then, a face is considered detected if, for a spatial arrangement of parts L^*, $Score(I, L^*) \geq \theta_1$, where θ_1 is a pre-defined threshold. After face detection, ROI image and facial landmark points are stored for further processing.

B. Expression Recognition

The extracted facial regions from the previous stage are used for facial expression recognition to recognize the driver's emotions. Therefore, we use the extracted face region and landmark points as the input to a pre-trained VGG16 with 16 layers to classify the emotions. Due to the limited availability of training data, transfer learning can be used to train the deep networks.

1) Network Fine-tuning: In our proposed approach we use pre-trained Convolutional Neural Network (CNN) to extract high level features. VGG16 is trained on ImageNet dataset consisting of a large number of images and having 1000 classes. In our framework, we fine-tune two VGG16 networks on ROI images and facial landmark points, separately for the datasets VIVA_FACE [18], DriveFace [19], JAFFE [20], MMI [21] and CK+ [22] to learn facial and landmark features, thereby improving the recognition accuracy. Moreover, fine-tuning a pre-trained network with a few images is generally much faster than training a network from scratch, thus, saving time and increasing recognition accuracy.

In VGG16 network, the last three layers are configured for 1000 classes and need to be configured for the new classification problem. We replace the last three layers in both the networks with a fully connected layer, a soft max layer and a classification output layer. Then, we fine-tune one VGG16 with all the face ROIs in all the datasets and other VGG16 network with facial landmark points for all the above-mentioned datasets. We have fine-tuned the pre-trained models for 1000 epochs with a learning rate of 0.01 and have used stochastic gradient descent with momentum (SGDM) optimizer.

2) Model Integration and Classification: After fine-tunning, we classify the facial expressions. In the pre-processing step, we first detect the face and landmark points in each frame using algorithm [9]. Then, we pass the face detected ROI image and facial landmark points to two separate VGG16 networks. Then, features from the top layer from both the networks are integrated to make the final classification using weighted summation given in Equation 5.

$$F_n^{integrated} = \alpha \times U_n^{vgg16}(\text{Face ROI}) + \beta \times V_n^{vgg16}(\text{FLP}) \quad (5)$$

where, $n = 1, 2, ..., E$ and E are the total emotion classes. Parameter α and β between 0 to 1 and value of α and β depend on the performance of each network. In our work we have experimentally decided the value of α and β as 0.6 and 0.4 respectively. U^{vgg16} and V^{vgg16} are the outputs of VGG16 for face ROI and FLP data respectively and $F^{integrated}$ is the final vector for classification. Though AlexNet and VGG16 network on ROI images give good classification accuracy, the integration of features from networks on face ROI image and facial landmark points gives comparatively better accuracy as explained in Section IV.

The emotions recognized can then be informed to the driver by the system to allow the driver to be aware of his/her emotional state and take an informed decision, of controlling his/her emotions if required, to remain alert while driving. However, if the emotional state is continuously detected for a certain time period, and no action is being taken by the driver or there is no change in his emotional state, specially in case of negative emotions such as anger and fear, the system should warn the neighboring vehicles by blowing a horn/siren and the

same information about the driver can be communicated for help by sending messages.

IV. Results and Discussion

We perform our experiments on Intel Core i7-7500U CPU @ 2.70GHz with 8GB RAM and GeForce 940MX GPU in Matlab. Our framework is for recognizing driver's emotion through facial expression recognition. We have used VIVA_FACE [18] and DriveFace [19] to recognize the driver expression. Moreover, to generalize and validate our proposed framework, we perform experiments on publicly available facial expression datasets such as JAFFE [20], MMI [21] and CK+ [22]. We perform cross-subject validation, such that in each iteration, $2/3$ data is used for training and $1/3$ data for testing.

A. Experiments on VIVA_FACE dataset

We test our model on VIVA faces dataset [18] which contains images of drivers in naturalistic driving environments. By implementing the method as explained in Section III-A, a total of 459 facial regions is detected from the dataset. This dataset consists of driver static images in naturalistic driving scenarios in varying illumination conditions, different persons driving, different zoom levels, different placements of camera, etc. Among these 459 detected facial regions, 38 of them are either with no face, small fraction of face or face with too much occlusion which are of no use for expression classification. Hence, a total of 421 face regions and facial landmark points are extracted from the dataset.The face regions detected in the data are manually categorized into seven emotion categories (*surprise, happy, neutral, sad, angry, fear, disgust*), by a group of people. Since VIVA and DriveFace dataset have lot of different expressions, we only take the basic emotion (majority vote) to manually classify these for validation of the emotion recognition stage using the fine-tuned VGG16 networks. These ROI and landmark points are input to the emotion recognition stage using the fine-tuned VGG16 network. Extracted features from both the networks are integrated using the weighted summation given in Equation 5. Then soft-max layer is applied to classify the input data into corresponding classes and the accuracies of the corresponding models are reported in Figure 2. For comparative analysis we check the accuracy of our proposed method on fine-tuned AlexNet and VGG16 networks using (a) only the detected Face ROI on both AlexNet and VGG16 networks, (b) only the detected landmark points using both AlexNet and VGG16 networks and (c) our proposed approach of integrating using weighted summation of both appearance features extracted from the detected face ROI and geometric features from the detected landmark points from two separate networks (AlexNet for both and VGG16 for both) for emotion recognition. In this case, we perform experiments on both AlexNet and VGG16 for comparison. In Figure 2, we can see that AlexNet and VGG16 give 86.4% and 90.2% accuracy on ROI image whereas on facial landmark points both networks give 47.7% and 49.3% accuracy. The accuracy of emotion

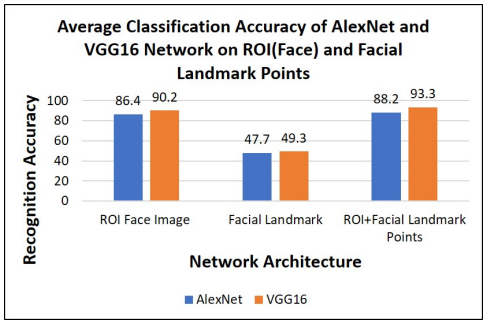

Fig. 2. The average recognition accuracy of experiments carried out using ROI after face detection, facial landmark and both facial +landmark. We got the highest accuracy of 93.3% in integrated feature classification (ROI+Facial Landmark Points)

Fig. 3. Some example images from VIVA face dataset [18] where the drivers face is occluded. Driver's face is not detected in these images in Stage 1. Occlusion generally occurs because the driver is performing some activity, which may distract him/her from the road.

detection achieved by extracting appearance and geometric features using two AlexNets and integrating these features is 88.2%. As VGG16 networks gives better accuracy in facial landmark points and ROI face data, we use VGG16 for our framework and integrating the appearance (ROI face data) and geometrical features (facial landmark) obtained from two VGG16 networks using weighted summation achieves 93.3% accuracy. Using the knowledge of the placement of camera and average face position of the driver for a car, a pre-defined window can be specified where the drivers face should be continuously visible in that region. This helps in faster detection of the face because of a reduced search area in the image. This step serves two purposes. Firstly, reducing the computation time for face detection and eliminating the chances of face detection of persons who are not driving (i.e. passengers), noise and reducing the chances of false detections. Secondly, as the pre-specified window is fixed such that an attentive driver's face always lies within the frame of the pre-set window, any non-detection of face in the pre-set window implies the driver bending away or being inattentive/distracted. This non-detection of face may also be because of occlusion, which also sometimes means that the driver is performing some other task which is distracting his attention from the road (as shown in Figure 3).

B. Experiments on DriveFace dataset

DriveFace dataset [19] to generalize our proposed approach, we perform experiments on DriveFace dataset that contains images of different drivers in real scenarios. There are a total of 606 samples having size 640×480 pixels each, acquired by 2 female and 2 male drivers. In this case also, we take a survey of many people and categorize the data

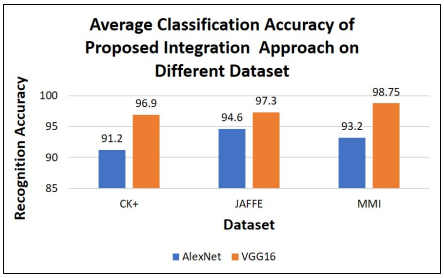

Fig. 4. Average recognition accuracy of proposed approach on different facial expression datasets. The recognition accuracy of our proposed framework is 96.9% on CK+ dataset, 97.3% on JAFFE dataset and 98.7% MMI dataset

into different emotions such as 'happy', 'surprise', 'talking' and 'neutral' faces. The confusion matrices of the proposed

TABLE I
CONFUSION MATRIX (IN %) OF 4 FACIAL EXPRESSION OF DRIVEFACE DATASET [19].

	Happy	Talking	Surprise	Neutral
Happy	**96.0**	0	2.0	2.0
Talking	1.7	**96.3**	1.0	1.0
Surprise	1.4	0	**97.0**	1.6
Neutral	2.2	0	2.0	**95.8**

approach are shown in Table I. We got an average recognition accuracy of 96.27%. We notice that in each expression we got more than 95% accuracy. 'Surprise' is highly confused with 'neutral' due to intra-class variations and labelling of the expression. For some person, a surprised face may appear as neutral, and for some others, 'neutral' may be a smiling face.

C. Experiments on Facial Expression Dataset

To generalize our proposed approach we perform experiments on some more datasets.

JAFFE dataset [20] have 7 facial expression 'Happy', 'Surprise', 'Sad', 'Disgust', 'Fear', 'Anger', and 'Neutral' dataset performed by 10 Japanese females.

MMI dataset [21] contains samples of images displaying various emotions similar to JAFFE. This dataset has images captured in three views. One is the static frontal-view, the second being the profile-face view, and a third dual-view (side mirror view).

Cohn-Kanade Dataset (CK+) dataset [22] contains 5870 images (640×490) of various expressions 'Happy', 'Surprise', 'Sad', 'Disgust', 'Fear', 'Anger', and 'Neutral'.

We observe a very good recognition accuracy as reported in Fig 4. On each of these datasets, we perform cross-subject validation, such that in each iteration $2/3$ data is used for training and $1/3$ data for testing. Figure 4 shows that the recognition accuracy of our proposed approach is more than 95% on all the facial expression recognition datasets. This shows that our proposed framework is capable of robustly and accurately recognizing the drivers emotions, if the camera is optimally mounted to be able to see the drivers face continuously.

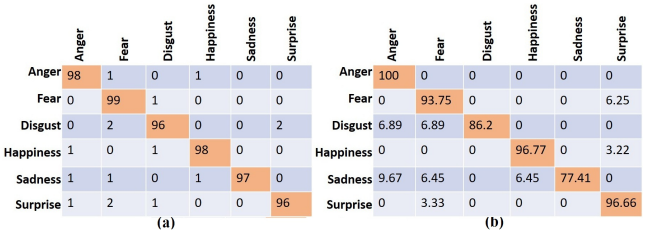

Fig. 5. (a) Confusion matrix (in %) of our proposed approach on JAFFE dataset (b) Confusion matrix of paper [23] on JAFFE dataset.

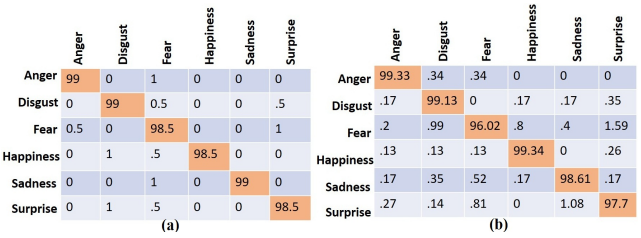

Fig. 6. (a) Confusion matrix (in %) of our proposed approach on MMI dataset (b) Confusion matrix of paper [24].

D. Comparison of proposed approach with state-of-the-art methods

For the comparative analysis we have shown the confusion matrix of our proposed approach in Fig 5 with [23]. In this paper, the authors classify the expression using appearance features of active facial patches. They use SVM to classify the extracted features into expression categories and achieve 91.8% accuracy while we achieve 97.3% accuracy. From the confusion matrix in Fig 5(a) we can see that in all expressions, we achieved more than 95% accuracy while in paper [23] there is low accuracy in case of sadness and disgust in Fig 5(b). In all expression cases our method outperforms except in case of anger we get an accuracy of 98% while they get 100%.

Similarly, we perform experiments on MMI dataset [21] and compare with [24]. As can be seen from the confusion matrices shown in Fig 6, our proposed approach and DeXpression [24] both achieve a very close accuracy of 98.7% and 98.3% respectively.

We have shown comparison with [14], [16], [17], [24]–[27] state-of-the-art methods on the datasets JAFFE dataset [20], MMI dataset [21] and CK+ dataset [22] in Table II. We find that using network integration of both appearance and geometrical features for expression recognition helps in increasing the recognition accuracy, which enables the framework to better explain novel instances during the test phase. Moreover, using deep network for feature extraction helps in overcoming the shortcomings of hand-crafted feature extraction methods. As in Table II we can see that using LBP+SVM [25] achieved 81.0%, 86.9%, 89.6% accuracy on JAFFE, MMI and CK+ dataset respectively. While using deep network techniques [14], [24], [16] and our proposed framework achieved more than 95% accuracy on these dataset.

TABLE II
COMPARISON OF OUR PROPOSED APPROACH WITH STATE-OF-THE-ART-METHODS ON DIFFERENT DATASET

Author	Method	JAFFE	MMI	CK+
Ali Mollahosseini et al. [14]	DCNN	-	77.9	93.2
Peter Burkert et al. [24]	CNN	-	98.3	**99.6**
Andr Teixeira Lopes et al. [16]	DCNN	86.1	-	96.7
Caifeng Shan et.al. [25]	LBP+SVM	81.0	86.9	89.6
Jun Wang et.al [26]	TC+LDA	-	93.3	82.6
Evangelos Sariyanidi et al. [27]	F-Bases	-	75.12	96.02
Heechul Jung et al. [17]	DTAGN	-	70.24	97.24
Our proposed approach	**VGG16+ROI+FLP**	**97.3**	**98.7**	96.9

E. Time Analysis

Our proposed work is divided into two modules (i) face detection and ROI, Facial landmark points (FLP) extraction(ii) Feature extraction using both data separately, then feature integration and expression classification. In the first step, the images in the dataset are of resolution 968×544. It took approximately 12.1 seconds without pre-specified window for each face region and landmark points to be extracted from each image of the dataset and with pre-specified window the runtime went down to 2.2 seconds. We prepare a fine-tuned model of VGG16 network on all the datasets in separate steps. In the second step, we integrate extracted features on ROI and FLP data and classify the expression. This step takes $.018$ seconds for each frame. Thus our proposed approach take total 2.38 seconds to detect face and FLP points and classify the expression.

V. CONCLUSION

In this paper, we have proposed a framework for driver emotion monitoring based on facial expression recognition, since the driver's emotions play an important role in attentiveness during driving. We use a face detection model based on mixture of trees with shared pool of parts to robustly detect the drivers face in the wild within a pre-specified window in the image. The output of the face detection module is the ROI containing the face region and the landmark points on the face. We give these as input to two separate fine-tuned VGG16 networks and integrate the extracted features using weighted summation to classify the facial expression and discover the emotional state of the driver "in the wild". Experimental results show that our system is accurate, robust and works in various challenging situations.

REFERENCES

[1] E. Roidl, B. Frehse and R. Höger, Emotional states of drivers and the impact on speed, acceleration and traffic violationsa simulator study, *Accident Analysis & Prevention*, Vol. 70, pp. 282–292, 2014.

[2] M. A. Trógolo, F. Melchior, and L. A. Medrano, The role of difficulties in emotion regulation on driving behavior, *Journal of Behavior, Health & Social Issues*, Vol. 6, no. 1, pp. 107–117, 2014.

[3] C. E. Izard, Emotion theory and research: Highlights, unanswered questions, and emerging issues, *Annual review of psychology*, Vol. 60, pp. 1–25, 2009.

[4] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien, Emotion on the roadnecessity, acceptance, and feasibility of affective computing in the car, *Advances in human-computer interaction*, Vol. 2010, 2010.

[5] A. Allamehzadeh and C. O. Monreal, Automatic and manual driving paradigms: Cost-efficient mobile application for the assessment of driver inattentiveness and detection of road conditions, *Intelligent Vehicles Symposium (IV)*, 2016.

[6] A. Tawari and M. M. Trivedi. Robust and continuous estimation of driver gaze zone by dynamic analysis of multiple face videos, *IEEE Intelligent Vehicles Symposium Proceedings*, 2014.

[7] C.S. Dula, and E. S. Geller. Risky, aggressive, or emotional driving: Addressing the need for consistent communication in research, *Journal of safety research* 34.5 (2003): 559-566.

[8] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, Emotion recognition using facial expressions, *Procedia Computer Science*, Vol. 108, pp. 1175–1184, 2017.

[9] X. Zhu and D. Ramanan, Face detection, pose estimation, and landmark localization in the wild,*In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879-2886. IEEE, 2012.

[10] R. O. Mbouna, S. G. Kong, and M. G. Chun, Visual analysis of eye state and head pose for driver alertness monitoring, *IEEE transactions on intelligent transportation systems* 14.3: 1462-1469,2013.

[11] J. M. Wang, H. P. Chou, C. F. Hsu, S. W. Chen, and C. S. Fuh, Extracting driver's facial features during driving, *Intelligent Transportation Systems (ITSC), 14th International IEEE Conference on*, pp. 1972–1977, 2011.

[12] H. Gao, A. Yüce, and J.P. Thiran, Detecting emotional stress from facial expressions for driving safety,*Image Processing (ICIP), 2014 IEEE International Conference on*, pp. 5961–5965, 2014.

[13] Y. Zhang and C. Hua, Driver fatigue recognition based on facial expression analysis using local binary patterns, *Optik-International Journal for Light and Electron Optics*, Vol. 126, no. 23, pp. 4501–4505, 2015.

[14] A. Mollahosseini, D. Chan, and M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, *WACV, 2016 IEEE Winter Conference on*, pp. 1–10, 2016.

[15] H. W. Ng, V.D. Nguyen, V.Vonikakis, and S.Winkler, Deep learning for emotion recognition on small datasets using transfer learning, *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443–449, 2015.

[16] A. T. Lopes, E.Aguiar, A.F. DeSouza, and T.Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, *Pattern Recognition*, Vol. 61, pp. 610–628, 2017.

[17] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, *Computer Vision (ICCV), 2015 IEEE International Conference on*, pp. 2983–2991, 2015.

[18] S. Vora, A. Rangesh, and M. M. Trivedi, On generalizing driver gaze zone estimation using convolutional neural networks, *Intelligent Vehicles Symposium (IV), 2017 IEEE*, pp. 849–854, 2017.

[19] K. Diaz-Chito, A. Hernández-Sabaté, and A. M. López, A reduced feature set for driver head pose estimation, *Applied Soft Computing*, Vol. 45, pp. 98–107, 2016.

[20] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, Coding facial expressions with gabor wavelets, *AFGR, Proceedings. Third IEEE International Conference on*, Vol. 3, pp. 200–205, 1998.

[21] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, Web-based database for facial expression analysis, *Multimedia and Expo, ICME 2005. IEEE International Conference on*, pp. 5–pp, 2005.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 94–101, 2010.

[23] S. Happy and A. Routray, Automatic facial expression recognition using features of salient facial patches, *IEEE transactions on Affective Computing*, Vol. 6, no. 1, pp. 1–12, 2015.

[24] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki, Dexpression: Deep convolutional neural network for expression recognition, *arXiv preprint arXiv:1509.05371*, 2015.

[25] C. Shan, S. Gong, and P. W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, *Image and Vision Computing*, Vol. 27, no. 6, pp. 803–816, 2009.

[26] J. Wang and L. Yin, Static topographic modeling for facial expression recognition and analysis, *Computer Vision and Image Understanding*, Vol. 108, no. 1-2, pp. 19–34, 2007.

[27] E. Sariyanidi, H. Gunes, and A. Cavallaro, Learning bases of activity for facial expression recognition, *IEEE Transactions on Image Processing*, Vol. 26, no. 4, pp. 1965–1978, 2017.