Driver Emotion Recognition Using a Hybrid Convolutional and Long Short-Term Memory Neural Network

Vaishnavi W
Indian Institute of Technology
Hyderabad
ai20btech11025@iith.ac.in

Nishant
Indian Institute of Technology
Hyderabad
bm22btech11013@iith.ac.in

Abstract

Real-time emotion recognition during driving has significant potential of adding more to the level of safety that comes with driving. The technology has come to a level where the processing of physiological and visual data can occur concurrently to give a conclusion of the emotion of the driver. This paper introduces a framework for the analysis of processing and interpretation of such multimodal data. We use electrocardiogram (ECG) readings and video recordings of ten male subjects undergoing simulated driving conditions for the study. We describe here how we process these inputs to extract features that we consider to be critical. Highly inspired from [1], our framework combines these capabilities using an integrated ML model (Convolutional Bi-LSTM) signal processing and computer vision techniques incorporated. Towards this, we aim to make available a framework that identifies four different emotional states, i.e., Baseline, Cognitive, Emotional, and Normal driving states, which could aid strongly for developing adaptive safety mechanisms in vehicles. The preliminary results may suggest a promising line of implication: the recognition of emotional state with reasonable accuracy by our method and offer great efficacy of our approach in the same, paving the way for real-time systems of the detection of emotions in dynamic environments.

1. Introduction

Over the past few years, there has been a growing interest in being able to detect and track human emotions precisely in real-time, to be used in the effort of enhancing safety and well-being not only for the driver but the entire human race. [2, 4, 3] Sensor technologies and machine learning algorithms have advanced to the point that it is now technically feasible and inexpensive for psychological states of drivers to be continually monitored and for the automobile to intervene in predicting risky behaviors, avoiding them culminating in critical incidents. This is one study that takes advantage of the convergence of technologies for mobile communication with advanced computational methodologies, thus realizing real-time processing of voluminous datasets capturing the complex emotional responses of drivers.

The present research work reports on the development of a robust system for emotion recognition that involves multimodal data integration. It mainly covers video and electrocardiogram (ECG) signals captured during the execution of simulated driving scenarios. A comprehensive dataset of 10 male subjects was used in exploring the intricate dynamics of emotional shifts across varied driving conditions: baseline (non-driving), cognitive, emotional, and normal driving states. The use of simultaneous video and ECG data provides a platform of rich input that permits one to have a rich basis for the detection of the emotional states based on physiological and behavioral indicators.

This paper proposes a new methodology of preprocessing and analyzing these sources of data. It has been mainly achieved by using digital signal processing and machine learning techniques for improved accuracy in classifying emotions. ECG data fused with video data are done in our system, not only for its increased detection power but to provide insights into the interplay of different modalities of emotional expression. Knowing all of the above and the technology in this field, we believe that we should be able to make good contributions to the development of the mentioned intelligent systems that will ensure the safety of driving environments.

2. Methodology

2.1. Dataset

The dataset utilized in this study comprises recordings obtained from 10 male subjects, aged approximately between 20 to 35 years, under simulated driving conditions. Each subject participated in sessions where they experienced four distinct emotional states: Baseline (non-driving), Cognitive, Emotional, and Normal driving. For each emotional state, 2-minute video clips were captured, resulting in a total of 40 video clips (10 subjects × 4 emotions). Additionally, simultaneous electrocardiogram (ECG) readings were collected corresponding to each 2-minute video segment.

2.2. Data Preprocessing

Each of the sample video and ECG datasets were divided into four 30-second segments. This was done to expand the sample size to 160.

2.2.1 Electrocardiogram Data

The ECG signals were preprocessed to extract relevant features for emotion detection. Initially, raw ECG signals were processed by applying a bandpass filter to eliminate noise and artifacts, to enhance signal quality. This was done using 'ecg.ecg' function from the biosppy library. While the filtered signal provides time-domain features, Welch's method was employed to compute the power spectral density of the ECG signals, facilitating the extraction of frequency-domain features. This technique divides the signal into overlapping segments, computes the fourier transform for each segment and then averages the resulting power spectrum. These features were then concatenated into a single feature array.

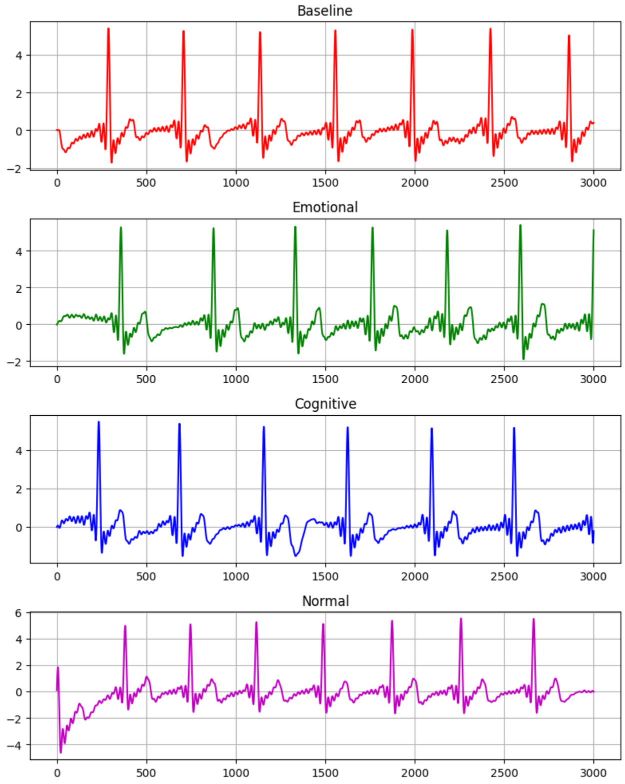

Figure 1: Filtered input ECG signal for a subject

2.2.2 Video Data

Initially, each 30-second video segment was decomposed to approximately 60 frames. These frames were then converted to grayscale, to reduce computational complexity. Following this, they were cropped and resized to a size of 240x240 pixels.

2.3. Architecture

2.3.1 ECG-Based Emotion Detection

The processed ECG signals are input into a Bidirectional Long Short-Term Memory (BiLSTM) network, leveraging its ability to capture long-term dependencies. This is followed by a dense layer and a softmax activation for emotion classification.

2.3.2 Video-Based Emotion Detection

From the image frames, facial features are extracted using Gabor wavelet transform. Gabor is a convolution filter representing a combination of gaussian and sinusodial term, known to be commonly used in facial recognition.

$$g(x,y) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\psi\right) \quad (1)$$

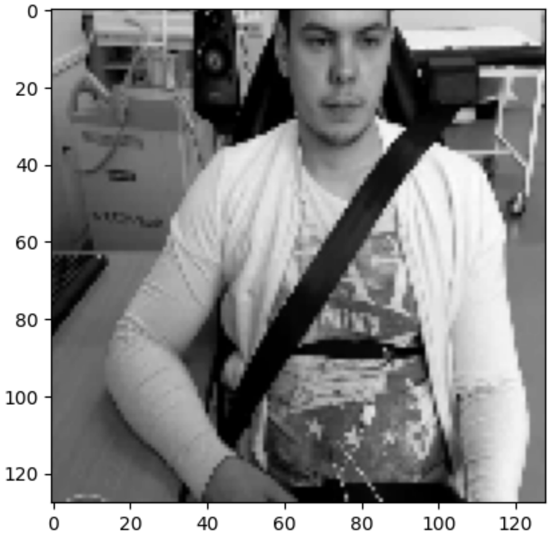

Figure 2: A frame from the video

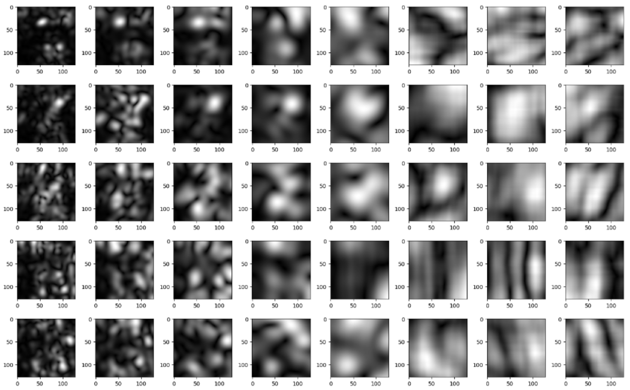

Figure 3: Extracted Gabor Wavelet Features

The resulting facial features undergo gaussian smoothing. To reduce the dimension of our features, Principal Component Analysis (PCA) is applied. The feature vectors are then fed into a feedforward layer, followed by softmax activation for classification.

2.3.3 Integration of ECG and Video Data for Emotion Detection

Equal sized output features from the dense layers of both ECG (BiLSTM) and video data networks (CNN) are concatenated. These concatenated vectors are subsequently fed into a neural network for comprehensive emotion classification.

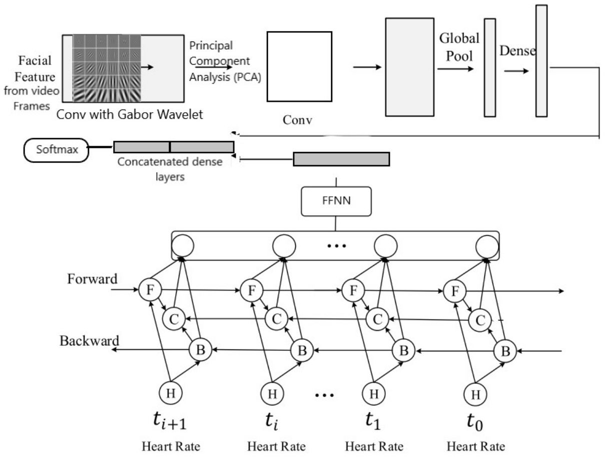

Figure 4: Process chart of the proposed model

3. Results

3.1. ECG based Emotion Detection

The performance of the emotion detection model is evaluated through various metrics and visualizations as described below. The training and validation accuracy plots over epochs are depicted in Figure 5. The model achieved a training accuracy of 61% and a test accuracy of 59%.

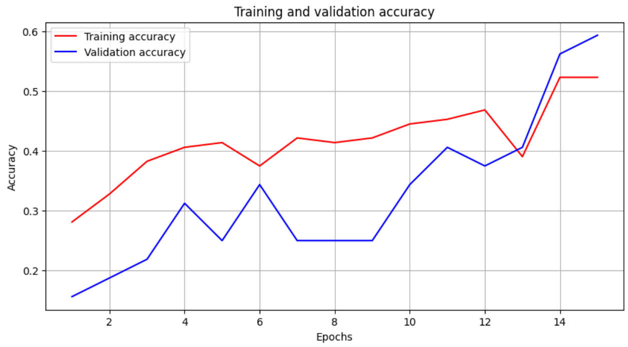

Figure 5: Training and Validation Accuracy Plot

Table 1: Class wise classification Report

Classes	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	30
1	0.52	0.77	0.62	35
2	0.42	0.42	0.42	33
3	0.69	0.30	0.42	30

Table 2: Confusion Matrix of Recognition Results

	Baseline	Cognitive	Emotional	Normal
Baseline	**100**	0	0	0
Cognitive	0	**60**	20	20
Emotional	0	14.29	**71.43**	14.29
Normal	0	10	80	**10**

3.2. Video-Based Emotion Detection

Figure 6 illustrates how our model's accuracy evolves with each training epoch. During training, it reached an impressive accuracy of 97%, while maintaining a 78% accuracy rate on unseen data.

Table 3: Class wise classification Report

Classes	Precision	Recall	F1-score	Support
0	0.91	1.00	0.95	10
1	0.56	1.00	0.71	5
2	0.80	0.57	0.67	7
3	0.86	0.60	0.71	10

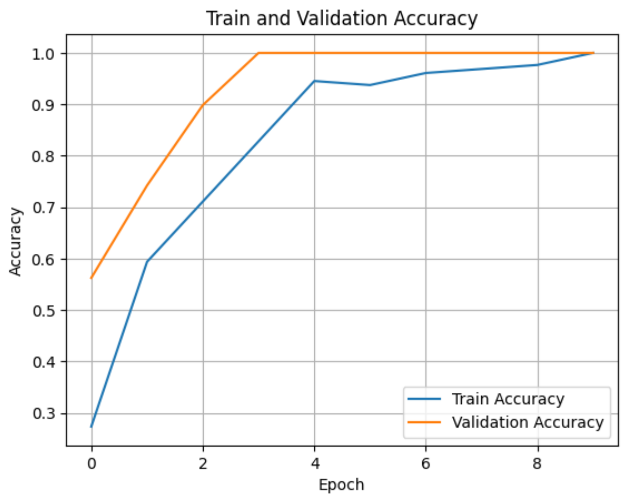

Figure 6: Training and Validation Accuracy Plot

Table 4: Confusion Matrix of Recognition Results

	Baseline	Cognitive	Emotional	Normal
Baseline	**100**	0	0	0
Cognitive	0	**100**	0	0
Emotional	14.28	14.28	**57.14**	14.28
Normal	0	30	10	**60**

3.3. Integration of ECG and Video Data for Emotion Detection

Figure 7 depicts the fusion model's accuracy with each training epoch. During training, it reached an accuracy of 98.4%, while having an impressive 84.3% accuracy rate on test data.

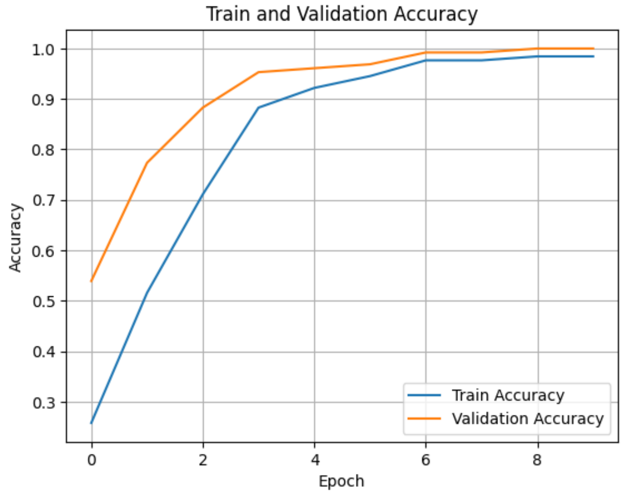

Figure 7: Training and Validation Accuracy Plot

Table 5: Class wise classification Report

Classes	Precision	Recall	F1-score	Support
0	0.91	1.00	0.95	10
1	0.83	1.00	0.91	5
2	0.83	0.71	0.77	7
3	0.78	0.70	0.74	10

Table 6: Confusion Matrix of Recognition Results

	Baseline	Cognitive	Emotional	Normal
Baseline	**100**	0	0	0
Cognitive	0	**100**	0	0
Emotional	0	0	**71.43**	28.57
Normal	10	10	10	**70**

4. Conclusion

Our experiments show that relying solely on either video or ECG data for emotion detection yielded limited accuracy. While the ECG-based model performed well for baseline and emotional states, it struggled with distinguishing between other emotions, possibly due to overlapping signal patterns. Similarly, the video-based model faced difficulty detecting certain emotions due to subtle face expressions. However, integrating video data with ECG signals through multimodal fusion led to enhanced performance. It underscores the effectiveness of combining modalities for more robust emotion detection.

References

[1] Guanglong Du, Zhiyao Wang, Boyu Gao, Shahid Mumtaz, Khamael M. Abualnaja, and Cuifeng Du. A convolution bidirectional long short-term memory neural network for driver emotion recognition. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4570–4578, 2021. 1

[2] Min Hu, Haowen Wang, Xiaohua Wang, Juan Yang, and Ronggui Wang. Video facial emotion recognition based on local enhanced motion history image and cnn-ctslstm networks. *Journal of Visual Communication and Image Representation*, 59, 12 2018. 1

[3] Dazhi Jiang, Kaichao Wu, Dicheng Chen, Geng Tu, Teng Zhou, Akhil Garg, and Liang Gao. A probability and integrated learning based classification algorithm for high-level human emotion recognition problems. *Measurement*, page 107049, 09 2019. 1

[4] suowei wu, Zhengyin Du, weixin li, and di Huang. Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze. pages 40–48, 10 2019. 1