Interpretable Multimodal Emotion Recognition using Facial Features and Physiological Signals

Puneet Kumar and Xiaobai Li*
CMVS, University of Oulu, Finland.
{puneet.kumar, xiaobai.li}@oulu.fi

Abstract

This paper aims to demonstrate the importance and feasibility of fusing multimodal information for emotion recognition. It introduces a multimodal framework for emotion understanding by fusing the information from visual facial features and rPPG signals extracted from the input videos. An interpretability technique based on permutation feature importance analysis has also been implemented to compute the contributions of rPPG and visual modalities toward classifying a given input video into a particular emotion class. The experiments on IEMOCAP dataset demonstrate that the emotion classification performance improves by combining the complementary information from multiple modalities.

Keywords: Affective Computing, Interpretable & Deployable AI, Multimodal Analysis, rPPG, Facial Features.

1 Introduction

Emotions, characterized by a rich and complex mix of physiological and cognitive states, hold significant importance across multiple fields such as psychology, human-computer interaction, affective computing, and even extending to broader domains such as virtual reality, user experience design, healthcare, and education [1]. Understanding and accurately interpreting emotions is essential in human communication and social interactions [2]. With the surge in the development and accessibility of multimodal sensing technologies, researchers can explore multiple modalities to enhance the accuracy and robustness of emotion recognition systems [3]. The current research trend focuses on building Artificial Intelligence (AI) systems that can be deployed for real-life applications [4].

Two such modalities, facial expressions and physiological signals, have garnered significant attention due to the rich information they offer and their non-invasive nature [5]. Facial expressions, direct and non-invasive indicators of emotion, have been thoroughly investigated [6]. Various techniques involving the extraction of facial landmarks, local descriptors, or holistic representations have been proposed to capture nuanced variations in facial muscle movements that reflect different emotional states [7]. Physiological signals, such as remote photoplethysmography (rPPG) signals, provide another layer of emotional cues. These signals, obtained through non-contact video-based techniques, offer insights into physiological changes associated with emotional responses [5]. The interplay of these two modalities offers a more holistic understanding of emotions, thus enhancing the robustness of emotion recognition systems [8].

Emotion classification through audio-visual information is a well-established research task [9, 10, 11]. However, recognizing emotion using the physiological context along with the audio-visual information score for further exploration [5]. Furthermore, despite the significant advancements, many multimodal emotion recognition models do not provide meaningful interpretations for their predictions [12, 13]. Most existing interpretability techniques have been implemented for visual modality and have yet to be fully explored for multimodal analysis [14, 15, 6].

This paper proposes an interpretable multimodal emotion recognition framework that extracts rPPG signals and facial features from the input videos and uses their combined context for emotion detection. The Haar cascades classifier [16] has been implemented to extract the rPPG signals, whereas a pre-trained ResNet-34-based network extracts the visual features. Further, early and late fusion approaches that integrate the static facial expression features and dynamic rPPG signals to capture both spatial and temporal aspects of emotions have been incorporated.

An interpretability technique based on permutation feature importance (PFI) [17] has also been incorporated that computes the contribution of rPPG and visual modality towards classifying a given input video into a particular emotion class. The experiments performed on Interactive Emotional Dyadic Motion Capture (IEMOCAP)

---

*Corresponding Author: xiaobai.li@oulu.fi

dataset [18] have resulted in an accuracy of 54.61% while classifying the input videos into ten emotion classes ('neutral,' 'happy,' 'sad,' 'angry,' 'excited,' 'frustrated,' 'fearful,' 'surprised,' 'distressed' and 'other'). The increased performance on using the multimodal context than the individual accuracies on using rPPG or visual modality alone advocates the importance of leveraging the multimodal context for emotion understanding. The average contributions of rPPG and visual modalities towards emotion recognition have been computed as 37.67% and 62.33%, respectively.

The contributions of this paper can be summarized as follows:

- A multimodal emotion recognition framework has been proposed to classify a given video into discrete emotion classes. It extracts the dynamic rPPG signals from the input videos and combines them with static facial expressions using early and late fusion approaches.
- An interpretability technique has been incorporated that computes the contribution of rPPG and visual modalities towards emotion classification using the PFI algorithm.
- Extensive experiments have been performed on the IEMOCAP dataset, and the results have been presented in terms of accuracy, precision, recall, F1 score, and modality-wise contributions toward emotion classification.

2 Proposed Method

The proposed framework has been diagrammatically depicted in Figure 1 and described in the following sections.
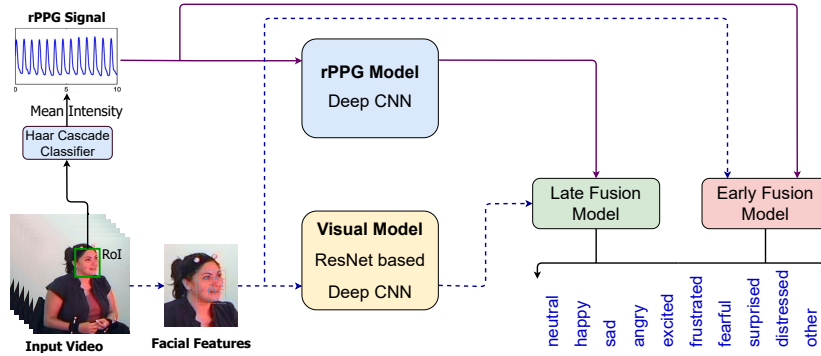

Figure 1: Schematic illustration of the proposed framework.

2.1 Preprocessing and Feature Extraction

The video files are loaded and processed frame by frame using OpenCV (cv2) library [1] and processed to extract rPPG signals and facial features.

i) rPPG Signals Extraction: Face detection within each video frame during the rPPG signal extraction process is accomplished using Haar cascades [16]. The region of interest (ROI), predominantly the facial region, is isolated from each frame, after which the mean intensity is computed to generate the rPPG signal for each video. The calculation of the mean intensity within the ROI ($\bar{I}c$) is represented in Eq. 1.

$$\bar{I}c = \frac{1}{N} \sum_{x=1}^{W} \sum_{y=1}^{H} I_{x,y,c} \tag{1}$$

Where $I_{x,y,c}$ is the intensity of the pixel at location (x, y) for color channel c in the ROI, and N is the total number of pixels in the ROI, whereas W and H represent the width and height of the ROI, respectively, and $c \in R, G, B$.

ii) Facial Features Extraction: Facial feature extraction employs Dlib's shape predictor [19], which is a version of the ResNet-34 trained on Face Scrub dataset[20] to identify the facial landmarks in a given image of a face. As per Eq. 2, it identifies 68 facial landmarks for each detected face within every frame, distinguishing unique facial characteristics.

$$P = D(F, \{L_i\})$$
$$F = [f_1, f_2, \ldots, f_n] \tag{2}$$

---

[1]https://opencv.org/

Where F represents the face detected in a frame, P represents the predicted points on the face, $D(F, \{L_i\})$ is the function for predicting points on the face, and L_i is the set of landmark points for the ith point. As signals from different videos might differ in length, it becomes crucial to standardize the input for the neural network model. This standardization is achieved by zero-padding $\bar{I}$ and P to match the maximum signal length.

2.2 Multimodal Feature Fusion

Early fusion and late fusion approaches are used to combine the rPPG signals and facial features.

i) Early Fusion: In the early fusion approach, the rPPG signals and facial features are concatenated before being fed into the model. The fused data are then passed through a neural network comprising a flatten layer, followed by CNN layers of dimensions 512 and 256, and the final layer of size equal to the number of classes. The flatten layer transforms the 3D input tensor into a 1D tensor, and the subsequent CNN layers functions perform the classification task. The model structure is represented as per Eq. 3.

$$
\begin{aligned}
I' &= \text{concatenate}(\bar{I}c, P) \\
I'' &= \text{flatten}(I') \\
F_{early} &= \text{NNet}(I'', C)
\end{aligned}
\tag{3}
$$

Where I is the input shape, C denotes the number of classes, $\bar{I}c$ is the mean intensity within the ROI from the rPPG signals, P represents the facial features, $NNet$ represents the early fusion network and F_{early} is the output of the early fusion.

ii) Late Fusion: In the late fusion approach, the rPPG and visual models are trained separately, and their outputs are combined using a weighted average. Eq. 4 represents a late fusion approach where the models are trained separately, and their outputs are combined in the final output F_{late}.

$$
F_{late} = w_1 \cdot M_{\text{rPPG}}(\bar{I}c) + w_2 \cdot M_{\text{facial}}(P)
\tag{4}
$$

Where $M_{\text{rPPG}}(\bar{I}c)$ and $M_{\text{facial}}(P)$ represent the outputs of the rPPG model and the visual model, respectively, and w_1 and w_2 are the weights assigned to each model's output in the final fusion.

2.3 Emotion Classification

This study employs three separate models for emotion classification. Two of these models operate independently, utilizing rPPG signals and facial features. The third model operates via 'early fusion,' exploiting the combined context of data from the rPPG and visual models. The outputs of these individual models are then collaboratively integrated through a 'late fusion' approach that uses a weighted addition technique. The individual models, based on rPPG signals and facial features, are constructed as follows.

i) rPPG Model: This model utilizes a Deep Convolutional Neural Network (CNN) with two hidden layers. It incorporates Rectified Linear Unit (ReLU) activation functions for emotion classification derived from rPPG signals.

ii) Visual Model: This model, built on facial features, employs a ResNet-based Deep CNN with two hidden layers and ReLU activation functions.

2.4 Interpretability

An explainability method based on permutation feature importance (PFI) [17] is implemented, which is used to estimate the importance of features by permuting the values of each feature and measuring the resulting impact on model performance. The PFI of feature j is the decrease in the model score when values of feature j are randomly permuted. PFI for a feature j is the difference in the model score when the values of feature j are randomly permuted. Eq. 5 mathematically represents the concept of permutation feature importance.

$$
PFI(j) = E_\pi[f(X^{(i)})] - E_\pi[f(X^{(i)}_{\pi_j})]
\tag{5}
$$

Where $PFI(j)$ is the permutation feature importance of feature j, $E_\pi[f(X^{(i)})]$ is the expected value of the model score over all samples in the dataset when the model is scored normally, $E_\pi[f(X^{(i)}_{\pi_j})]$ is the expected value of the model score when the values of feature j are permuted according to some permutation π, and $X^{(i)}_{\pi_j}$ denotes the dataset $X^{(i)}$ with the values of feature j permuted according to π.

3 Results and Discussion

3.1 Experimental Setup

The emotion classification experiments have been performed on the IEMOCAP dataset [18] consisting of 10,039 videos labeled with ten discrete emotion labels ('neutral,'' happy,' 'sad,'' angry,' 'excited,' 'frustrated,' 'fearful,' 'surprised,' 'distressed' and'other'). The model training has been trained on NVIDIA RTX 4090 GPU for 50 epochs with a batch size of 32 and a learning rate of 0.001. The performance has been evaluated using accuracy, precision, recall, and F1 score metrics.

3.2 Results

Table 1 summarizes the accuracy of the individual and fusion models, whereas the average contributions of rPPG and visual modalities towards emotion recognition in the early fusion setup are presented in Table 2. The proposed framework has demonstrated an emotion classification accuracy of 54.61%, and the average contributions of rPPG and visual modalities towards emotion recognition have been computed as 37.67% and 62.33%, respectively.

Table 1: Detailed performance of the individual and fusion models.

Model	Accuracy	Precision	Recall	F1 Score
rPPG	37.45%	0.37	0.38	0.38
Facial Features	46.42%	0.49	0.49	0.49
Late Fusion	41.17%	0.43	0.42	0.42
Early Fusion	54.61%	0.56	0.58	0.57

Table 2: Average contribution of each modality towards emotion recognition.

Modality	Contribution
rPPG	37.67%
Visual	62.33%

Table 1 shows that both the individual models performed reasonably well. However, the fusion model outperformed the individual models, demonstrating the advantage of combining rPPG signals and facial feature information for emotion recognition.

3.3 Discussion

This paper presents a compelling case for including multimodal context in emotion recognition. While the models trained on individual modalities show moderate performance, their fusion significantly improves emotion recognition accuracy. It emphasizes the complementarity of these modalities in capturing emotional states. However, the late fusion of modalities underperforms compared to the early fusion approach, indicating that integrating modalities at an earlier stage allows for more effective learning of emotional states.

However, this study has a few limitations of the proposed work. The IEMOCAP dataset, while widely used, may limit the generalizability of the findings. Cross-dataset experiments on larger and more diverse datasets could further strengthen the results. Moreover, more modalities such as audio, text, and other physiological signals can also be incorporated for emotion recognition. Finally, a more in-depth interpretability mechanism can be developed to explain the role of individual features in emotion detection.

4 Conclusion

This work presents a multimodal emotion recognition framework using rPPG signals and facial features. It paves the way for practical applications where transparent and interpretable emotion understanding is important. The results highlight the benefits of integrating multiple modalities for emotion recognition, with an early fusion approach yielding the highest accuracy. While there are limitations and potential improvements, our study provides a promising direction for future research in emotion recognition, emphasizing the importance of multimodal data and fusion techniques.

References

[1] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Elsevier Information Fusion Journal*, 37:98–125, 2017.

[2] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. Cross Subject Multimodal Emotion Recognition Based on Hybrid Fusion. *IEEE Access*, 8:168865–168878, 2020.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(2):423–443, 2018.

[4] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. Challenges in Deploying Machine Learning: A Survey of Case Studies. *ACM Computing Surveys*, 55(6):1–29, 2022.

[5] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial Video-based Physiological Signal Measurement: Recent Advances and Affective Applications. *Signal Processing Magazine*, 38(6):50–58, 2021.

[6] Sarthak Malik, Puneet Kumar, and Balasubramanian Raman. Towards Interpretable Facial Emotion Recognition. In *The 12th Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021.

[7] Nannan Wang, Xinbo Gao, Dacheng Tao, Heng Yang, and Xuelong Li. Facial Feature Point Detection: A Comprehensive Survey. *Neurocomputing*, 275:50–65, 2018.

[8] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31(1):39–58, 2009.

[9] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning Multi-level Deep Representations for Image Emotion Classification. *Neural Processing Letters*, pp. 1–19, 2019.

[10] M Xu, F Zhang, and S Khan. Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion. In *IEEE Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1058–1064, 2020.

[11] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *The 31st AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 6818–6825, 2019.

[12] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[13] Luca Longo et al. Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In *The Springer International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)*, pages 1–16, 2020.

[14] Marco Tulio Ribeiro, S. Singh, and C. Guestrin. Why Should I Trust You? Explaining Predictions of Any Classifier. In *International Conference on Knowledge Discovery & Data mining (KDD)*, pages 1135–1144, 2016.

[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *The IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[16] Sander Soo. Object Detection using Haar Cascade Classifier. *Institute of Computer Science, University of Tartu*, 2(3):1–12, 2014.

[17] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive Emotional dyadic MOtion CAPture data. *Language Resources and Evaluation*, 42(4), 2008.

[19] Davis E. King. DLIB Models. https://github.com/davisking/dlib-models, 2016. Accessed on 21.05.2023.

[20] Hong-Wei Ng and Stefan Winkler. A Data Driven Approach to Cleaning Large Face Datasets. In *IEEE International Conference on Image Processing (ICIP)*, pages 343–347. IEEE, 2014.