Third RBCDSAI Conference on Deployable AI (DAI 2023)

Interpretable Multimodal Emotion Recognition using Facial Features and Physiological Signals

Puneet Kumar[†] and Xiaobai Li[*], Centre for Machine Vision and Signal Analysis, University of Oulu, Finland

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 information from visual facial features and remote photoplethysmography (rPPG) signals extracted from videos. A permutation feature importance-based interpretability technique 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 Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset demonstrate the improvement in the emotion classification performance on combining the complementary information from multiple modalities.

Introduction

- Emotion Recognition: unimodal vs multimodal
- Real-life applications ⇒ deployable
- Prominent modalities: visual and physiological
- rPPG: Non-invasive, additional dynamic information
- Interpretability techniques in literature for emotion recognition: visual modality ✓ multimodal ✗

Contributions:

- A multimodal emotion recognition framework
- A permutation feature importance (PFI) based interpretability technique
- Experiments on IEMOCAP dataset, quantitative & qualitative results and modality-wise contribution scores

Conclusion

- Demonstrated the importance and feasibility of multimodal emotion recognition using physiological and visual information
- Head-start for real-world applications of interpretable emotion analysis using aforementioned modalities

Proposed Method

The proposed method is illustrated in Figure 1 and consists of following four phases:

1) Feature Extraction

rPPG Signals: Input video → region of interest (ROI) → Haar cascades → mean intensity → rPPG signal

Facial Features: For each face in every frame, compute 68 landmarks using Dlib shape predictor and extract facial features from them

2) Multimodal Fusion: early and late fusion of the extracted rPPG and visual features

3) Emotion Classification: Deep ResNet based convolutional networks for rPPG and visual models

4) Interpretability

- Permute the values of each feature
- Measure the resulting impact on model's performance
- Estimate the feature importance from the difference among model performance scores
- Find rPPG features' importance scores, visual features' importance scores and overall importance scores
- Compute individual modality's contribution

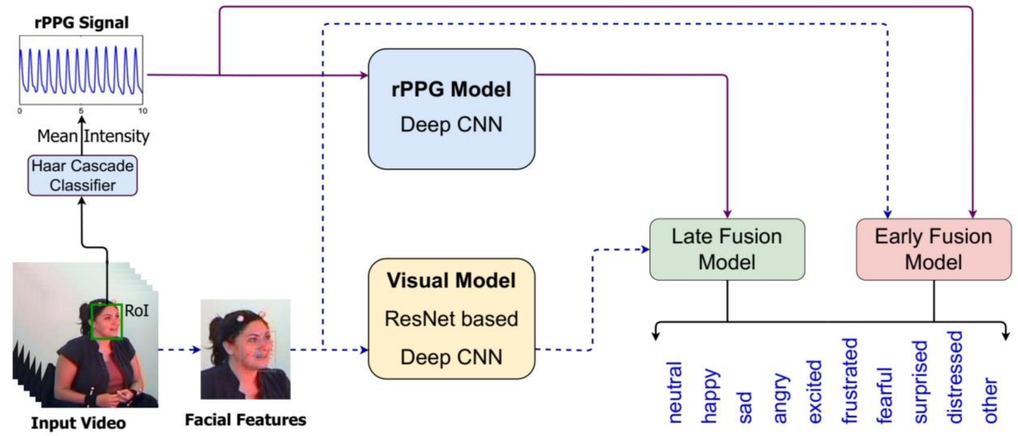


Figure 1: Schematic illustration of the proposed framework

Experiments & Results

Experimental Setup

- IEMOCAP dataset with 10,039 video samples
- 10 discrete emotion labels (*neutral*, *happy*, *sad*, *angry*, *excited*, *frustrated*, *fearful*, *surprised*, *distressed* and *other*)
- Model training – epochs: 50, batch size: 32, learning rate: 0.001
- Model evaluation using accuracy, precision, recall & F1 score metrics and modality-wise contribution scores

Results

- Tables 1 & 2: performance for individual & fusion models
- Better emotion recognition accuracy for fusion models than the models using individual modalities
- Late fusion underperforms compared to early fusion

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%

[†] *denotes the presenting author and* [*] *denotes the corresponding author*

UNIVERSITY OF OULU

RBC DSAI
ROBERT BOSCH CENTRE FOR DATA SCIENCE AND ARTIFICIAL INTELLIGENCE IIT MADRAS

Third RBCDSAI Conference on Deployable AI (DAI 2023)

Interpretable Multimodal Emotion Recognition using Facial Features and Physiological Signals

Puneet Kumar

Postdoctoral Researcher

Centre for Machine Vision & Signal Analysis

University of Oulu, Finland

Puneet.Kumar@oulu.fi

www.puneetkumar.com

FICORE
Finnish Indian Consortia for Research and Education

UNIVERSITY OF JYVÄSKYLÄ

UNIVERSITY OF HELSINKI

CeRAI
Centre for Responsible AI

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

Introduction: Background

- Emotion Recognition: unimodal vs multimodal [1]
- Real-life applications ⇒ deployable [2,3]
  - Healthcare,
  - Education,
  - Human computer interaction,
  - User experience design, etc.
- Prominent modalities [4,5,6]
  - Visual facial features
  - Physiological signals
- Why remote photoplethysmography (rPPG)?
- Emotion understanding interpretability
  - Visual [7,8] ✓
  - Multimodal [9,10] ✗

Video 1: Need for multimodal processing

Introduction: Contributions

- A multimodal emotion recognition framework
  - Extract static facial expressions
  - Extract dynamic rPPG signals
  - Compute multimodal context using early and late fusion approaches
  - Classify a given video into discrete emotion classes
- An interpretability technique
  - Incorporates permutation feature importance (PFI) algorithm
  - Computes the contribution of rPPG and visual modalities towards emotion classification
- Extensive experiments
  - Dataset: Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [11]
  - Quantitative results: accuracy, precision, recall, and F1 score
  - Qualitative results: modality-wise contributions toward emotion classification

Proposed Method

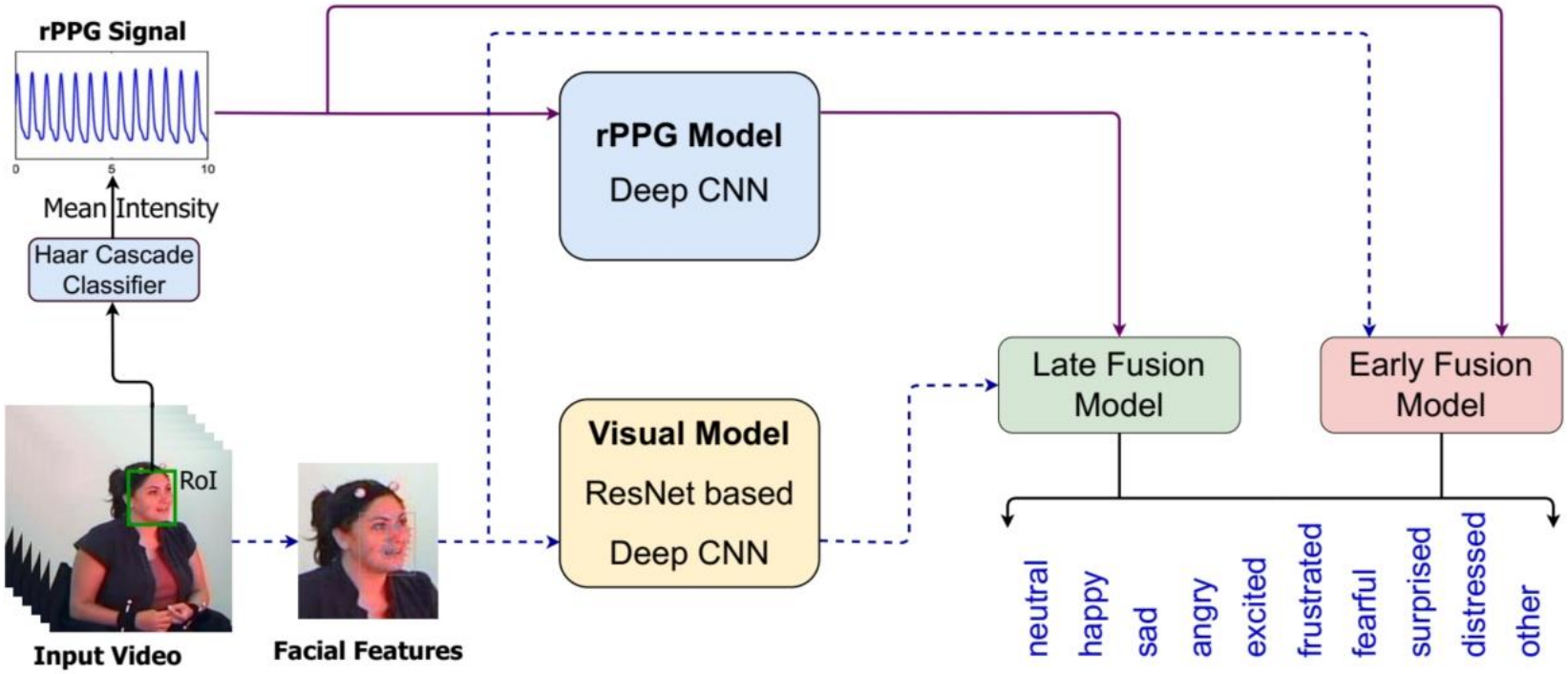


Figure 1: Schematic illustration of the proposed framework

Proposed Method: Feature Extraction

rPPG Signals Extraction:

Input Video:

→ Region of interest (ROI)

→ Haar cascades [12]

→ mean intensity (Eq 1)

→ rPPG Signal

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

Where,

N: total number of pixels in the ROI

W & **H**: width and height of the ROI

c: color channel, **c** ∈ {R, G, B}

$\bar{I}_c$: Mean pixel intensity

$I_{x, y, c}$: Pixel intensity at location (**x**, **y**) for color channel **c** in the ROI

Facial Features Extraction:

- Dlib shape predictor [13]
- For each face in every frame:
  Compute 68 facial landmarks as per Eq. 2.
- Landmarks: facial characteristics

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

Where,

P: the predicted points on the face

D(**F**, L_i): function for predicting points on the face

L_i: set of landmark points for the i^{th} point

F: face detected in a frame

Proposed Method: Multimodal Fusion & Emotion Classification

Multimodal Fusion:

Early Fusion:

$$I' = \text{concatenate}(\bar{I}c, P)$$
$$I'' = \text{flatten}(I') \quad (3)$$
$$F_{early} = \text{NNet}(I'', C)$$

Where,

I: input shape
C: number of classes
$\bar{i}_c$: mean intensity within the ROI from rPPG signals
P: facial features
NNet: the early fusion network
F_{early}: output of the early fusion

Late Fusion:

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

Where,

$M_{rPPG}(\bar{i}_c)$: output of the rPPG model
$M_{facial}(P)$: output of the visual model
w_1 and w_2 are the weights of rPPG and visual models

Emotion Classification:

rPPG Model:

- Input: rPPG signals
- Output: discrete emotion classes
- Deep Convolutional Neural Network (CNN)
- Activation function: Rectified Linear Unit (ReLU)
- Optimizer: Adam

Visual Model:

- Input: facial features
- Output: discrete emotion classes
- ResNet-based Deep CNN
- Activation function: ReLU
- Optimizer: Adam

Proposed Method: Interpretability

- Permutation feature importance (PFI) [14]:
  - Permute the values of each feature
  - Measure the resulting impact on model performance
  - Estimate the feature importance from the difference of model performance scores
- PFI of feature **j**: difference in the model score on permuting **j**

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

Where,

PFI(j): permutation feature importance of feature **j**

$\mathbf{E_{\pi}[f(X^{(i)})]}$: expected value of the model score over all samples in the dataset

$\mathbf{E_{\pi}[f(X^{(i)}_{\pi_j})]}$: expected value of the model score when the values of feature **j** are permuted according to some permutation **π**

$\mathbf{X^{(i)}\pi_j}$: dataset **X(i)** with the values of feature **j** permuted according to **π**

- Find rPPG features' importance scores, visual features' importance scores and overall importance scores
- Compute individual modality's contribution

Experiments

- Dataset: Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [11]
  - 10,039 video samples
  - Ten discrete emotion labels (neutral, happy, sad, angry, excited, frustrated, fearful, surprised, distressed and other.

- Model training
  - NVIDIA RTX 4090 GPU
  - 50 epochs
  - batch size: 32
  - learning rate: 0.001.

- Model evaluation
  - Metrics: accuracy, precision, recall, and F1 score.

Results

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%

Conclusion & Future Scope

- Conclusion
  - Emotion recognition: accuracy (individual modalities) > accuracy (multimodal fusion)
  - Late fusion underperforms compared to early fusion
  - Importance & feasibility of multimodal emotion recognition
  - Head-start for the real-world applications with interpretable emotion understanding

- Future Scope
  - Cross-dataset experiments on larger and more diverse datasets
  - Incorporation of more modalities such as audio, text, and other physiological signals
  - Development of more in-depth interpretability mechanisms to explain the role of individual features

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] Andŕe Altmann, Laura Tolo¸si, Oliver Sander, and Thomas Lengauer. Permutation Importance: A Corrected Feature Importance Measure. Bioinformatics, 26(10):1340–1347, 2010.

Author response to reviewers' comments

Official Review of Paper19 by Reviewer kBjF

RBCDSAI DAI 2023 Conference Paper19 Reviewer kBjF

30 May 2023 (modified: 31 May 2023) RBCDSAI DAI 2023 Conference Paper19 Official Review Readers: Program Chairs, Reviewers, Paper19 Authors

Summary:
The authors present their early results on fusing multimodal information for emotion recognition.

Strengths:
- As the authors mention, this article summarizes some promising results for including multimodal context in emotion recognition.
- The authors also identify the importance of early fusion.

Weaknesses:
The article could benefit by clarifying the following points.

- The usage of variable x in eq (1) is ambiguous. Is the x in $\sum x$ the x coordinate of the pixel.
- Is it not clear what 1^W represents in eq(1).
- The authors have not described how the landmark points L_i are obtained.
- In eq (5), it is not clear why the permutation π is required for computing the expected value of the model score. Should it be $E_\pi[f(X(i))]$ or simply $E[f(X(i))]$

Rating: 4: Accept (candidate for best article)
Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Author response to reviewer's comments

RBCDSAI DAI 2023 Conference Paper19 Authors *Puneet Kumar (privately revealed to you)*

05 Jun 2023 (modified: 05 Jun 2023) RBCDSAI DAI 2023 Conference Paper19 Official Comment Chairs, Paper19 Authors

Comment:
Dear reviewer, Thank you for providing your valuable comments. Please find the author response explaining the rebuttal or changes incorporated in the manuscript.

• Equation (1) was typeset incorrectly. We have corrected it now.

• The Dlib's shape predictor is a ResNet-34-based model trained to identify the facial landmarks in a given image of a face. We have elaborated on this explanation in Section 2.1 (ii).

• The subscript π denotes the expectation over all possible permutations of a particular feature. We are not interested in any specific permutation but the average effect of all possible permutations; hence the expectation over π is used.

Thanks & Regards, Authors, DAI2023 Paper#19

Official Review of Paper19 by Program Chairs

RBCDSAI DAI 2023 Conference Program Chairs

01 Jun 2023 RBCDSAI DAI 2023 Conference Paper19 Official Review Readers: Program Chairs, Reviewers, Paper19 Authors

Summary:
The paper presents a simple method to integrate two different types of features for the task of emotion detection in human faces.

Strengths:
-> The idea proposed is simple and sound -> The model has been validated on a real world dataset

Weaknesses:
-> I think equation (1) is typeset incorrectly. The summation I assume runs from 1 to W -> this is simply the average of the pixel intensities in the region of interest. This can be corrected in the final draft.

-> The authors mention using deep networks but only use 2 hidden layers, which is perhaps not deep. Have they tried out increasing the number of layers to observe the performance?

Rating: 3: Accept
Confidence: 3: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

Author response to reviewer's comments

RBCDSAI DAI 2023 Conference Paper19 Authors *Puneet Kumar (privately revealed to you)*

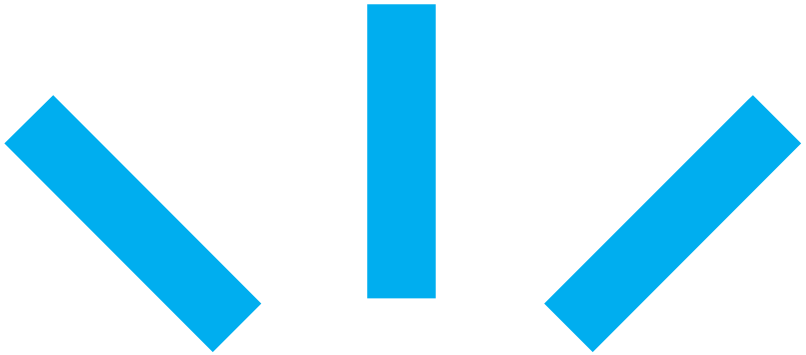
05 Jun 2023 RBCDSAI DAI 2023 Conference Paper19 Official Comment Authors

Comment:
Dear reviewer, Thank you for providing your valuable comments. Please find the author response explaining the rebuttal or changes incorporated in the manuscript.

• Equation (1) was typeset incorrectly. Thank you for pointing it out. We have corrected it now.

• We experimented with 2, 3, and 4 hidden layers. The network with 2 layers performed better than the rest of the choices; hence, it was implemented. It should be noted that this was a relatively simpler use-case and our aim in this work was to develop a baseline model to utilize visual and rPPG features for emotion recognition from facial videos. In the future, we intend to include more modalities, experiment on larger and more diverse datasets, and implement more complex networks.

Thanks & Regards, Authors, DAI2023 Paper#19



Thank you!

Puneet Kumar

www.puneetkumar.com

Puneet.Kumar@oulu.fi