Indian Institute of Information Technology,
Design and Manufacturing, Kancheepuram

CS21B1069
CS21B1072
CS21B1076

Project

MultiModal Emotion detection

1 Contribution

INTRODUCTION

This project aims to create a robust and accurate Emotion Detection System by combining the outputs of three models, each specializing in a different modality:

1. Facial Image Analysis – Handled by Ansh(CS21B1076).
2. Speech Emotion Recognition – Handled by Vaibhav(CS21B1072).
3. Facial Action Unit Detection – Handled by Rahul(CS21B1069).

The outputs of these models are integrated to predict the final emotional state. This multi-modal approach enhances accuracy and reliability, making it applicable in areas like stress detection for mental health and human-computer interaction.

2 Literature review

Research into emotion detection spans various modalities, with key advancements in facial analysis, speech recognition, and feature-based approaches like action unit detection.

1. Facial Action Unit Detection
   - Based on Paul Ekman's Facial Action Coding System (FACS), which maps muscle movements (AUs) to emotions.
   - Recent work focuses on combining deep learning with traditional FACS methods, leveraging models like Vision Transformers and ensemble learning.
   - Challenges include limited labeled datasets and low AU occurrence rates in frames.
2. Speech Emotion Recognition
   - Studies reveal that auditory cues like pitch, intensity, and pauses are strong indicators of emotional states.
   - Modern techniques employ CNNs, RNNs, and transformer-based models (e.g., Wav2Vec2.0) for accurate analysis.
   - Dataset diversity and small size remain key limitations.
3. Facial Image Emotion Recognition
   - Literature highlights the use of facial landmarks and deep learning architectures like ResNet and VGG to extract emotions.
   - A dual focus on categorical emotions (e.g., happiness, fear) and dimensional analysis (valence and arousal) has improved interpretability.
   - Key challenges involve data imbalance, low-resolution inputs, and variations in lighting and pose.

3 Dataset and Modelling

3.1 Facial Action Unit Detection

Overview

- Utilizes the Facial Action Coding System (FACS) to decode emotional states by analyzing facial muscle movements.
- Action Units (AUs), such as brow raising (AU1) or lip tightening (AU23), are mapped to emotions.

Datasets and Preprocessing

- Dataset: DISFA, annotated with AUs and intensity levels.
- Preprocessing: Facial landmarks detected using DLIB; Contrast Limited Adaptive Histogram Equalization (CLAHE) applied for image enhancement.

Model Details

- Multi-label classification with XGBoost ensemble trained on features extracted from a pre-trained Vision Transformer (DinoV2).
- Mapped AUs to 7 basic emotions using traditional FACS rules.
- Achieved an F1-Score of 0.2 and accuracy of 28% on sample data.

3.2 Speech Emotion Recognition

Overview

- Extracts emotional states from speech by analyzing acoustic features.
- Emotions like anger, fear, and happiness signal varying stress levels.

Datasets and Preprocessing

- Dataset: IITKGP-SEHSC, comprising 12,000 utterances across 7 emotions.
- Feature Extraction: Techniques like MFCC, Chroma, and Mel Spectrogram were employed.

Model Details

- Architecture: CNN with Conv1D and dense layers for classification.
- Training: Optimized using Adam optimizer with a learning rate of 0.0005.
- Achieved an accuracy of 79% on the test set.

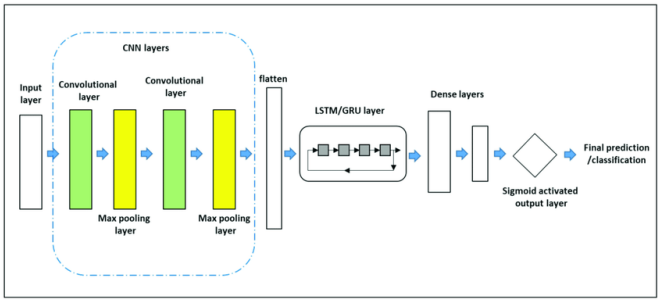

Figure 1: Model Architecture

3.3 Facial Image Emotion Recognition

Overview

- Identifies emotions from facial images using landmark detection.
- Introduced valence (pleasantness) and arousal (emotional intensity) for a continuous range of emotion classification.

Datasets and Preprocessing

- Dataset: FER2013, consisting of 35,887 grayscale images.
- Challenges include class imbalance and low-resolution images.

Model Details

- Initial attempts with simple CNNs and custom architectures showed limited success due to imbalanced data.
- Final Model: Pre-trained ResNet50 with additional layers for classification.
- Techniques like data augmentation and class weighting improved performance, achieving 65.02% accuracy and a mean F1 score of 64.85%.

3.4 Integration of Modalities

- Outputs from all three models are combined using an ensemble approach to predict the final emotion.
- This integration leverages complementary strengths from visual, auditory, and facial muscle-based emotion detection.

TRAINING:

- Trained on Kaggle GPU P100
- Framework used : Tensorflow
- Optimizer : Adam
- Batch size : 128
- Loss Function : Focal Loss (to handle class imbalance)
- Epochs : 80
- Train, validation, Test split : 52:13:35

```
Epoch 80/80
456/456 ━━━━━━━━━━━━━━━━━━━━ 367s 804ms/step - accuracy: 0.4043 - auc: 0.9957 - loss: 0.0011 - val_accuracy: 0.3871 - val_auc: 0.9
925 - val_loss: 0.0016
```

Figure 2: Loss and PR-AUC at the end of 80th epoch for FACS.

PERFORMANCE ON TESTING AND VALIDATION OF FACS

- Used PR-AUC as a metric to assess models training and Focal loss as loss Function to monitor model throughout training.
- Accuracy of each action unit and precision and recall through classification report are also used.

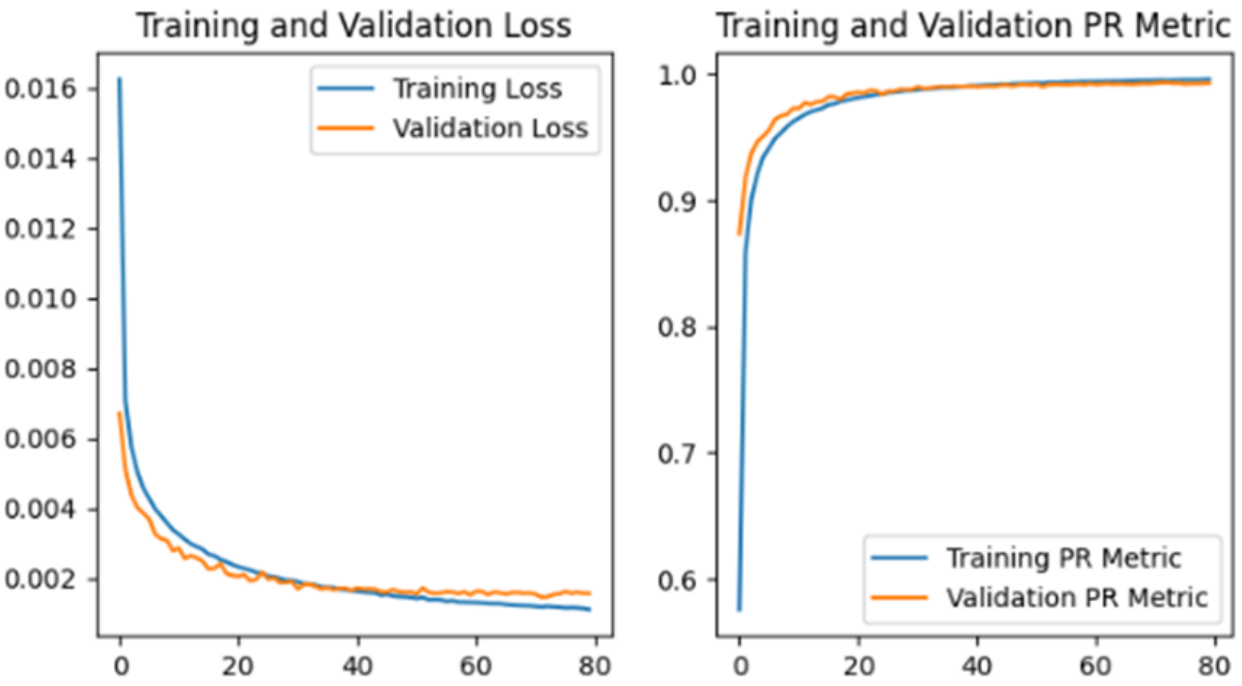

Figure 3: Loss and PR-AUC curves.

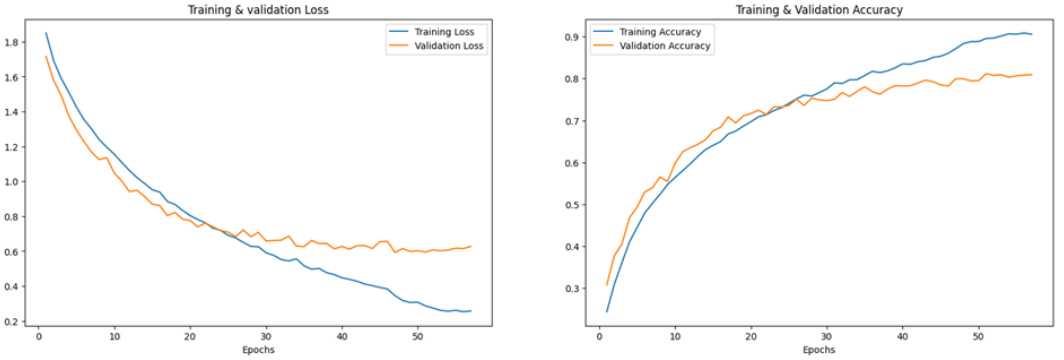

Figure 4: Accuracy and Loss Curves for speech

Classification Report:

	precision	recall	f1-score	support
angry	0.79	0.86	0.82	221
disgust	0.81	0.66	0.73	242
fear	0.80	0.77	0.78	208
happy	0.76	0.78	0.77	239
neutral	0.80	0.86	0.83	209
sad	0.76	0.75	0.75	237
surprise	0.79	0.83	0.81	219
accuracy			0.79	1575
macro avg	0.79	0.79	0.79	1575
weighted avg	0.79	0.79	0.78	1575

Figure 5: Classification Report

4 Challenges

1. Limited publicly available datasets for AUs and FER.
2. Class imbalance in datasets, leading to reduced model generalizability.
3. Ambiguity in expressions across individuals and modalities.

5 Conclusions

- The proposed multi-modal Emotion Detection System combines advanced techniques in speech, facial images, and action unit detection to achieve robust emotion recognition. By addressing challenges like dataset limitations and class imbalances, the system promises high accuracy and wide applicability. Future work includes refining integration methods and exploring larger, more diverse datasets.

5.1 Applications

1. Healthcare: Detecting stress and emotional disorders in patients.
2. Human-Computer Interaction: Enabling robots to interpret emotions for better interaction.
3. Workplace Monitoring: Understanding employee stress and engagement levels.

6 References

FER:

1. Roy, A. K., Kathania, H. K., Sharma, A., Dey, A., and Ansari, M. S. A. (2024). Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition. (ResEmoteNet - Rank1 - FER2013, Rank1 - RAF-DB)

2. Pham, L., Vu, T. H., and Tran, T. A. (2021). Facial expression recognition using residual masking network. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 4513–4519. (https://paperswithcode.com/paper/facial-expression-recognition-using-residualEnsemble ResMaskingNet with 6 other CNNs - Rank2 - FER2013)

3. El Boudouri, Y. and Bohi, A. (2023). Emonext: an adapted convnext for facial emotion recognition. In 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), pages 1–6. (https://paperswithcode.com/paper/emonext-an-adapted-convnext-for-facialEmoNeXt - Rank3 - FER2013)

4. Vignesh, S., Mahadevan, S., Muthukumarasamy, S., and Sridhar, R. (2023). A novel facial emotion recognition model using segmentation vgg-19 architecture. International Journal of Information Technology, 15. (Segmentation VGG-19 - Rank4 - FER2013)

5. Georgescu, M.-I., Ionescu, R. T., and Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. IEEE Access, 7:64827–64836. (https://paperswithcode.com/paper/local-learning-with-deep-and-handcraftedLocal Learning Deep+BOW - Rank5 - FER2013)

Speech Emotion Recognition:

1. ireless Pers Commun 130,515–525 (2023).
2. Dillon R, Teoh AN, Dillon D. Voice analysis for stress detection and application in virtual reality to improve public speaking in real-time: A review. arXiv preprint arXiv:2208.01041. 2022 Aug 1.
3. Schmidt, P., Reiss, A., & Van Laerhoven, K. "Wearable-Based Stress and Affect Detection during Real-Life Driving." IEEE Transactions on Affective Computing, vol. 9, no. 3, pp. 286-298, July-Sept. 2018.

4. Pradhan, N., & Sethi, R. "Stress Detection Using Speech Features and Machine Learning Techniques." 2020 IEEE Bombay Section Signature Conference (IBSSC), pp. 29-34, 2020.

5. Zhao, L., Li, Y., & Tao, J. "Stress Detection from Speech with Acoustic and Lexical Features." IEEE Access, vol. 8, pp. 189232-189241, 2020.

6. Fayek, H.M., Lech, M., & Cavedon, L. "On the Use of Convolutional Neural Networks for Speech Emotion Recognition." 2017 IEEE International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-5, 2017.

7. Tariq, M. U., & Rehman, A. "Speech Emotion Recognition Using Acoustic Feature Visualization and Convolutional Neural Networks." 2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3), pp. 48-52, 2019.

FACS:

1. AUFormer: Vision Transformers are Parameter-Efficient Facial Action Unit Detectors Kaishen Yuan, Zitong Yu, Xin Liu, Weicheng Xie, Huanjing Yue, Jingyu Yang

2.Deep Adaptive Attention for Joint Facial Action Unit Detection and Face Alignment Zhiwen Shao, Zhilei Liu, Jianfei Cai, Lizhuang Ma

3.K. Zhao, W. -S. Chu and H. Zhang, "Deep Region and Multi-label Learning for Facial Action Unit Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 3391-3399, doi: 10.1109/CVPR.2016.369.

4.FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features Yufeng Yin, Di Chang, Guoxian Song, Shen Sang, Tiancheng Zhi, Jing Liu, Linjie Luo, Mohammad Soleymani

5.Emotion Recognition Using Transformers with Masked Learning Seongjae Min, Junseok Yang, Sangjun Lim, Junyong Lee, Sangwon Lee, Sejoon Lim

6. J. Yang, F. Zhang, B. Chen and S. U. Khan, "Facial Expression Recognition Based on Facial Action Unit," 2019 Tenth International Green and Sustainable Computing Conference (IGSC), Alexandria, VA, USA, 2019, pp. 1-6, doi: 10.1109/IGSC48788.2019.8957163.

7. DISFA: A Spontaneous Facial Action Intensity Database S.Mohammad Mavadati, Student Member, IEEE, Mohammad H. Mahoor, Member, IEEE, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn

8. Towards Independent Stress Detection: A Dependent Model Using Facial Action Units Carla Viegas; Shing-Hon Lau; Roy Maxion; Alexander Hauptmann

9. Automatically Recognizing Facial Indicators of Frustration: A Learning-centric Analysis Joseph F. Grafsgaard; Joseph B. Wiggins; Kristy Elizabeth Boyer; Eric N. Wiebe; James C. Lester

10. A real-time robust facial expression recognition system using HOG features Pranav Kumar; S L Happy; Aurobinda Routray