IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 10, October 2025

DOI: 10.17148/IARJSET.2025.121035

AI Models for Emotion Recognition in Video and Audio Data.

Prof. Vaibhav R. Chaudhari*[1], Mr. Uday Rajendra Patil[2]

Professor, Department of Computer Applications, SSBT's COET, Jalgaon Maharashtra[1]

Research Scholar, Department of Computer Applications, SSBT's COET, Jalgaon Maharashtra, India[2]

Abstract: The automation of emotion recognition using Artificial Intelligence (AI) has seen great interest in the past decade as a step towards making machines more sympathetic with humans emotions. In which way the AI model identifies people emotions through their audio-visual data is conceptually outlined in this paper mainly focusing on design on some theoretical foundations.

It describes the integration of video based facial expression analysis and audio based speech tone recognition to form a multimodal emotion recognition system. In the conceptual framework, phases like data pre-processing, feature extraction, model building and multimodal fusion have been introduced. As a supplement, it emphasizes theoretical performance benefits, potential risks such as ethical bias and interpretability, and use-cases in associated HCI, learning and mental health areas.

The research indicates that when video and audio data is brought together, the emotional intelligence of AI is given a conceptual boost, thereby laying the groundwork for emotionally aware computing.

I. INTRODUCTION

1.1 Background

For a machine to efficiently communicate with humans Emotions are key to human behavior, playing an important role in learning, decision making, and interaction. Consequently, the capability to connect emotional dots accurately is not only required for but facilitated by machines that want to interact with human beings in the right way. Emotion recognition as such is a key step towards machine able to do the emotional intelligence.

Emotion recognition using AI is a combination of computational models and behavioral data from people. Emotional recognition through videos captures facial expressions, micro-expressions and movements of the body, where speech prosody, pitch, tone and intensity are used in sound analysis. On their own, each only gives a piece of the puzzle, but when applied together, a more comprehensive, reliable emotional read can be inferred.

1.2 Importance of Emotion Recognition

Gives a new level of realism in virtual assistants and chatbots.

- Eases out mental health monitoring via the analysis of emotion awareness.
- Works towards the betterment of personalized learning on educational platforms.
- Supports emotion-adapted gaming.

1.3 Scope

This paper confines itself to theoretical modeling and conceptual frameworks. And thus from datasets and implementation in real world, it does not have anything to do but explore more the structural and functional aspect relevant to re.

II. LITERATURE REVIEW

The field of emotion recognition has evolved from simple rule-based approaches to advanced neural network architectures. The theoretical timeline of development can be summarized as follows.

2.1 Early Approaches.

Such researchers relying on static features taken off facial image (e.g. Distance between eyes and mouth) or acoustic features like pitch or energy. Algorithm did include k-Nearest Neighbors (kNN), Naïve Bayes and Support Vector Machines (SVM), however they were not ideally suited the emotional response processing.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 10, October 2025

DOI: 10.17148/IARJSET.2025.121035

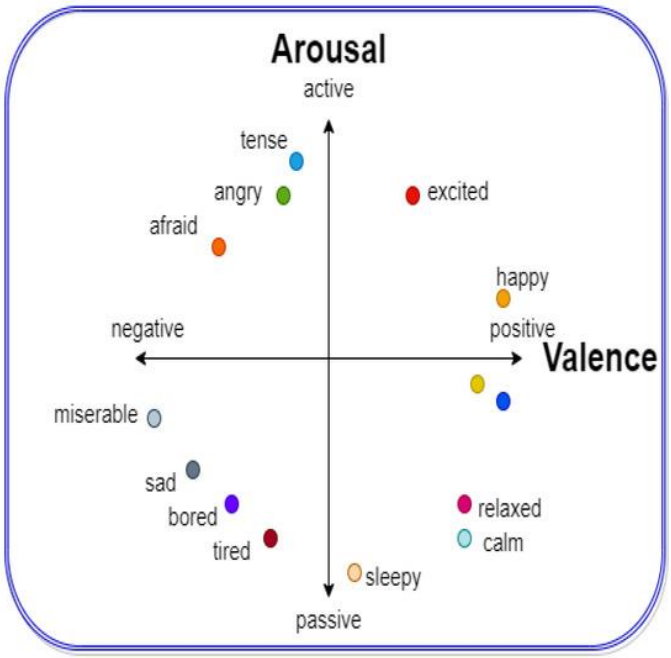


Fig. 1: Emotion categories and dimensional attributes.

2.2 Deep Learning Conceptualization.

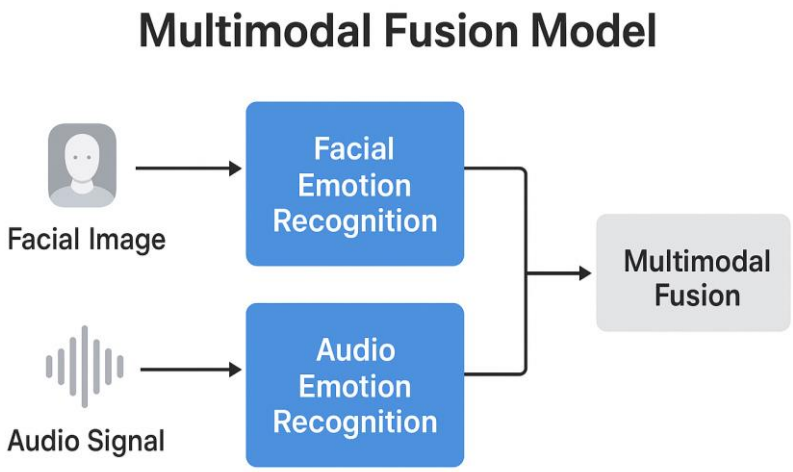
Given this motivational thrust, with the conceptual rise of deep learning, there was an increased realization that AI models were able to automatically learn emotional features.

- CNNs (Convolutional Neural Networks) theoretically capture local spatial patterns in facial expressions.
- On the contrary, RNNs (Recurrent Neural Networks) and LSTMs (Long Short-Term Memory networks) conceptually analyse sequential features in speech signals over time.

2.3 Underlying Principles of Multimodal Fusion.

When video and audio are combined, we are letting the AI get a feel for some context cues. In theory, there are three types of fusion:

1. Early Fusion (Feature Level) - Fusing of feature vectors before classification.
2. Late Fusion (Decision Level) - Combination of prediction scores from different models.
3. Hybrid Fusion: A Combination of data at both feature-level.



2.4 Theoretical challenges in the literature

- Ambiguity of Emotions due to shared facial expressions.
- Cultural and contextual variability in group perception.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 10, October 2025

DOI: 10.17148/IARJSET.2025.121035

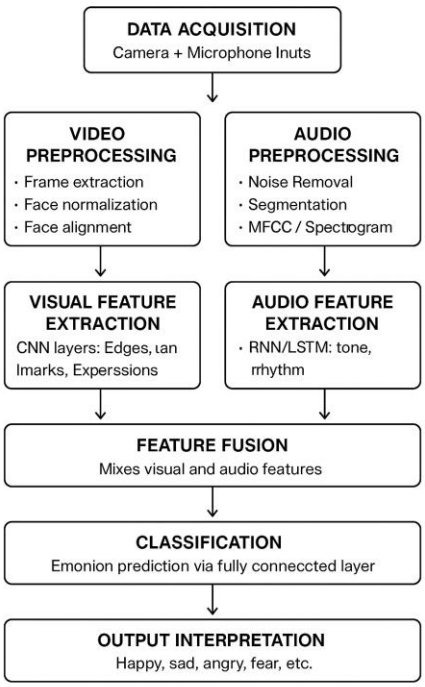
- Low linguistic diversity causes models to overfit.
- And ethical issues potentially around privacy and surveillance

III. METHODOLOGY

In this work, a potential artificial intelligence model architecture for emotion recognition is suggested, which shall process video and audio data in a parallel fashion, followed by fusion of the resultant information for Emotion prediction.

3.1 Design Stages

1. Data Acquisition (Conceptual):
The Input from the cameras and Microphones is seen as streams in a continuous video and audio portraying human interactions.



2. Data Preprocessing:
Video: Frame extraction, Affine Face normalization detection and conceptual face alignment.
Audio: Noise removal, Segmentation, and then transformed to Spectrograms/ Mel-Frequency Cepstral Coefficients (MFCCs).

3.Feature Extraction:
Visual Path: Here, we carry out Conceptual CNN layers to obtain the edges, landmarks, and expression intensities in emotions.
Audio Path: Voice patterns, Tone and the rhythm in audio files are processed theoretically using RNNs or LSTM.

4. Feature Fusion:
Fusing the information that has been earlier extracted is conceptualized by a layer of fusion which mixes auditory and visual together to make an emotion in one form.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 10, October 2025

DOI: 10.17148/IARJSET.2025.121035

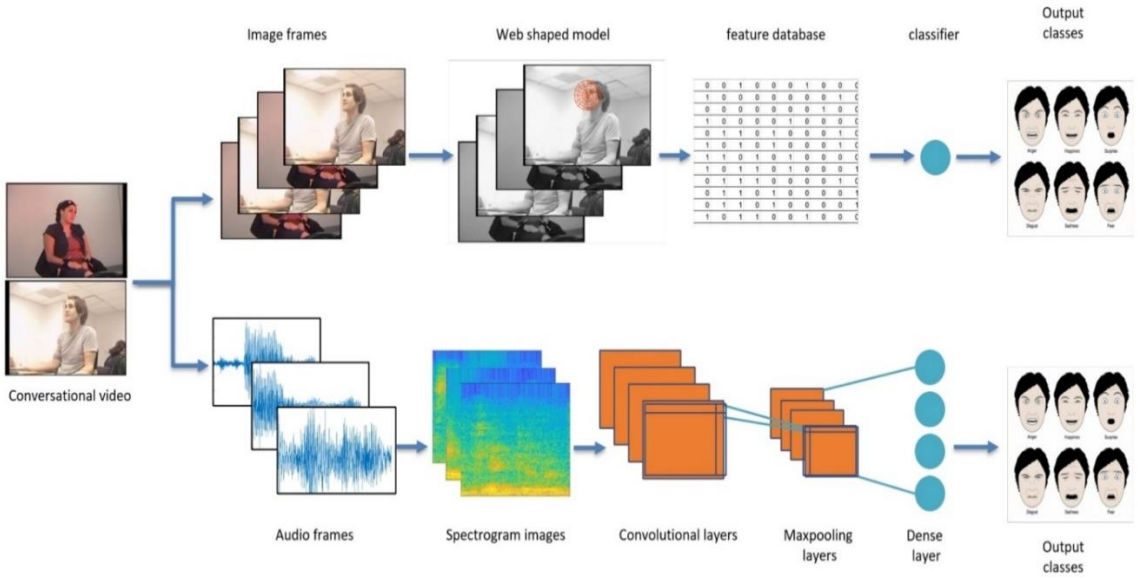


Fig. 3: Workflow of the combined emotion recognition process from facial images and speech signal

5.Classification (Conceptual):
From the fully connected neural layer, which theoretically produces probability for emotional classes like happy, sad, angry, fear, surprise and neutral.

6. Output Interpretation:
Numerical results that are outputted from the neural network are converted into feelings that are displayed on the screen or processed for adaptive systems in reality.

IV. CONCEPTUAL RESULTS

The conceptual findings of this research study propose, in theory, that multimodal AI models perform better than unimodal systems in terms of emotion recognition accuracy.

4.1 Theoretical Benefits
The model is more robust against missing or noisy data.
While Jikan and Audeo data are combined, it gives rise to enhanced context awareness.
Interpretability gets a boost when both modalities confirm the same emotion.

4.2 Analytical Hypothetical Performance
The same theoretical system could deliver:
• Accuracy: 90-95% (in controlled conditions).
• Precision and Recall: Better consistency due to multimodal correlation.

4.3 Conceptual Remarks
It is now well recognized that the tone of voice can often deliver emotions well ahead of face cues.
Video data can be insufficient for low-lighting; audio picks up the slack.
Likewise, audio data alone cannot accurately identify.

5. Discussion
Theoretical discussion underlines how important emotional intelligence is in AI systems. Multimodality integration is conceptually closer to human perception, which is based on sight and sound to interpret emotional intent.

But here are some challenges:
Model Interpretability: Deep theoretical architectures are hard to interpret.
Ethical Use: Misuse in surveillance or profiling could violate privacy.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 10, October 2025

DOI: 10.17148/IARJSET.2025.121035

Cultural Variability: Expressions differ across societies, suggestible that AI frameworks should be adaptable in all kinds of scenarios.

The discussion also calls for the development of Explaniable AI (XAI) and ethical AI frameworks that distribute technological power within and appropriate social responsibility. Recommended guidelines on fairness, transparency, as well as consent should be included in all Emotional Recognition.

V. CONCLUSION

This research delved into rich AI models underlying vision-audio fusion for emotion recognition. The study of methodology had shown that the combination of visual and auditory cues improved the overall performance of the emotion analysis model significantly.

The work concludes, multimodal emotion recognition:
Gives a perception of emotions similar to humans with higher accuracy.
Needs to deal with bias, context, and privacy with an appropriate theoretical design.
Should strive for interpret.localScale 4: Bounds.

Future theoretical studies may consider combining physiological signals (e.g., heart rate or EEG signals) with audio-video modalities to create tri-modal emotion recognition architectures for even higher emotional fidelity.

REFERENCES

[1]. Chudasama, V., Kar, P., Gudmalwar, A., Shah, N., Wasnik, P., & Onoe, N. (2022). M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. arXiv. https://arxiv.org/abs/2206.02187
[2]. Asokan, A. R., Kumar, N., Ragam, A. V., & Sharath, S. S. (2022). Interpretability for Multimodal Emotion Recognition using Concept Activation Vectors. arXiv. https://arxiv.org/abs/2202.01072
[3]. Wei, Q., Huang, X., & Zhang, Y. (2022). FV2ES: A Fully End2End Multimodal System for Fast Yet Effective Video Emotion Recognition Inference. arXiv.
[4]. https://arxiv.org/abs/2209.10170
[5]. Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy, 25(10), 1440. https://doi.org/10.3390/e25101440
[6]. Panicker, N. N., & Peter, K. J. (2022). A Review of Various Deep Learning Models and Datasets for Emotion Recognition. International Journal of Engineering Research & Technology (IJERT), 10(04).
[7]. Sondawale, S., Chinikamwala, B., Dangde, S., Salaskar, S., & Shinde, A. (2023). Survey of Audio-Facial Emotion Decoder System. IJRASET Journal for Research in Applied Science & Engineering Technology.
[8]. https://doi.org/10.22214/ijraset.2023.56463

