Fusion of Modalities for Emotion Recognition with Deep Learning

Shourya Gyanvarsha
Department of Electrical Engineering,
NIT Rourkela
Rourkela, Odisha, India
shourya.pizzy54@gmail.com

Swayamjit Mohanty
Department of Electrical Engineering,
NIT Rourkela
Rourkela, Odisha, India
swayamjit.mohanty2805@gmail.com

Himanshu Sekhar Sahoo
Department of Electrical Engineering,
NIT Rourkela
Rourkela, Odisha, India
himanshuss2001@gmail.com

Prof. Dipti Patra
Department of Electrical Engineering,
NIT Rourkela
Rourkela, Odisha, India
dpatra@nitrkl.ac.in

***Abstract*—The paper delineates a pioneering advancement in emotion recognition technology, showcasing a sophisticated multimodal system adept at discerning human emotions through the fusion of video, audio, and facial features, facilitated by state-of-the-art deep learning methodologies. By amalgamating information from diverse sensory modalities, the system attains remarkable precision in classifying emotions, surpassing the efficacy of unimodal approaches.**

Through a comprehensive array of experiments and evaluations, the study substantiates the system's prowess in accurately deciphering emotional states. Notably, the fusion of multimodal cues enables nuanced insights into human affective responses, transcending the limitations of individual modalities and furnishing a holistic understanding of emotional dynamics.

***Index Terms*—audio-visual emotion recognition; xlsr-Wav2Vec2.0 transformer; transfer learning; Action Units; RAVDESS; speech emotion recognition; facial emotion recognition**

I. Introduction

Recent strides in affective computing aim to equip machines with human-like perceptual abilities, with emotion recognition standing as a critical challenge essential for applications in human-computer interaction and mental health monitoring. Multimodal emotion recognition systems have gained prominence for their capacity to capture nuanced emotional expressions by integrating data from diverse sensory modalities such as facial expressions, vocal intonations, and body language.

Deep learning, particularly Convolutional Neural Networks (CNNs) like ResNet architectures and recurrent neural networks (RNNs) with LSTM units, has revolutionized emotion recognition. ResNet variants have demonstrated remarkable efficacy in processing visual data, while MFCC and LSTM models excel in capturing emotional cues from audio signals. Benchmark datasets like RAVDESS have played a crucial role in advancing these systems, providing a robust foundation for development and evaluation. This paper presents a comprehensive exploration of multimodal emotion recognition, aiming to develop robust systems capable of discerning subtle nuances in human affective states, with implications spanning human-computer interaction and mental health monitoring.

II. Motivation and Objective

The demand for emotion recognition from video is driven by evolving societal needs and accelerated by the post-COVID-19 era's embrace of contactless technologies. This transformative tool finds applications in aiding individuals with autism spectrum disorders and ensuring road safety in the automotive industry by detecting driver stress and fatigue. Additionally, in human-machine interactions, it serves as a proactive monitor of emotional states, with implications for healthcare and overall well-being. Inspired by these factors, our project focuses on 'Multimodal Emotion Recognition using Deep Learning,' aiming to integrate facial expression and speech models for improved accuracy and performance.

A. Research Problem

Emotion recognition systems are gaining attention for their potential in healthcare, but single-modal approaches often struggle to capture the complexity of human emotions. This limitation hampers accurate assessment, highlighting the need for improved multimodal systems. Integrating facial expressions, speech patterns, and physiological signals could offer a more comprehensive understanding of emotions. However, a lack of comprehensive reviews and analyses in healthcare contexts impedes the development and adoption of effective solutions.

B. Research Objectives

- Applicability of multimodal emotion recognition systems for mental health :
  - Explore multimodal emotion recognition systems in healthcare, focusing on mental health monitoring and intervention.

- Evaluate existing approaches, identify areas for enhancement and innovation.

- Develop and evaluate a novel multimodal emotion recognition system:
  - Develop and assess a new multimodal emotion recognition system for healthcare.
  - Emphasize accuracy, reliability, and real-time performance to meet healthcare demands.

III. Background of Emotion recognition

- Kaiming He, et al., pioneered residual learning for more effective training of deep neural networks. Their 152-layer residual nets achieved a remarkable 3.57% error on ImageNet, winning 1st place at ILSVRC 2015. Additionally, they improved COCO object detection by 28% due to their deep representations. [5]
- Jitesh Kumar Bhatia, Juginder Pal Singh, et al., introduced a flexible facial recognition model for emotion recognition, considering real-life complexities in emotions. Their approach incorporates behavioral features and biometric visual characteristics, aiming to address the challenge of handling continuous and varied facial expressions. [4]
- A. A. A. Zamil, et al., explored emotion detection using speech signals, employing Mel Frequency Cepstrum Coefficient (MFCC) features and Logistic Model Tree (LMT) classifiers. Their approach demonstrated effective emotion classification on Emo-DB and RAVDESS datasets, achieving a maximum accuracy of 70% for 7 different emotions. [1]
- Y. Wang, et al., conducted a study on emotion recognition based on prosodic parameters across multiple languages. Their analysis of time, energy, pitch, and formant parameters revealed the potential for recognizing basic emotion states using simple prosodic parameters across languages. [8]
- A. Sharma, et al., conducted a comprehensive review of recent trends in human emotion detection, particularly focusing on facial expression recognition. Their literature review explores various machine learning techniques and applications in the field, addressing challenges and comparing methods' advantages, disadvantages, and accuracy. [2]
- M. Kumar and S. Srivastava proposed utilizing artificial neural networks for emotion detection through facial expressions. Their study focuses on six predefined facial expressions to determine behavior and mood, highlighting the significance and broad applications of emotion detection. [6]
- V. Hosur and A. Desai advocate face reading for understanding human behavior. Their method, Facial Emotion Recognition using CNNs, analyzes images, extracting and classifying facial expressions into emotions. It involves background removal and facial feature extraction. They highlight applications in medicine, education, investigations, and human-robot interaction. [3]
- Sinoara, R, et al., emphasize text semantics in text mining but find a lack of integrated research. Their systematic mapping study, based on 1693 selected works from 3984 identified in five digital libraries, provides an overview and identifies gaps, guiding researchers. Despite numerous studies, processing semantic aspects in text mining remains challenging. [7]

IV. Methods

The methodology for exploring multimodal emotion detection with ResNet began by assembling diverse datasets containing facial images, audio recordings, and video clips to capture a wide range of emotional expressions. Each modality underwent preprocessing to extract pertinent features crucial for emotion recognition. Subsequently, ResNet models were trained independently on these modalities, leveraging hierarchical representation learning to capture intricate patterns from raw data. Further refinement of the models was conducted specifically for emotion recognition tasks, optimizing their architectures and parameters to enhance accuracy.

Following individual training, model predictions from each modality were merged using various fusion techniques to exploit the complementary information across modalities. These fusion strategies aimed to improve the overall emotion detection performance by integrating the strengths of different data sources. Performance evaluation metrics such as accuracy and precision were then employed to gauge the effectiveness of ResNet for multimodal emotion detection, providing valuable insights into the system's capabilities and guiding potential enhancements for future research endeavors.

A. Speech Emotion Recognition

- **ResNet-18:** ResNet-18 is a convolutional neural network architecture renowned for its effectiveness in training deep networks. It consists of 18 layers and is part of the ResNet (Residual Network) family introduced by Microsoft Research. The key innovation of ResNet-18 lies in its residual blocks, which utilize skip connections to alleviate the vanishing gradient problem. These blocks enable smoother gradient flow during training by adding the original input to the output of convolutional layers. The architecture includes four stages, each containing several residual blocks. Downsampling is achieved through strided convolutions or max-pooling layers to reduce spatial dimensions and broaden receptive fields. At the end of the network, global average pooling layers are used instead of fully connected layers, reducing parameters and computational complexity while maintaining performance in image classification tasks.

B. Face Emotion Recognition

- **ResNet-34:** ResNet-34 is an extension of the ResNet architecture, featuring 34 layers. Like other ResNet variants, it utilizes residual connections to enable effective training of deep neural networks. The architecture consists of several stages, each containing multiple residual

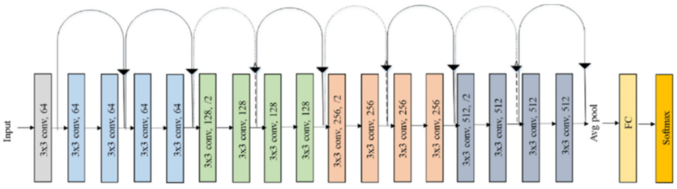


Fig. 1. ResNet-18 Architechture

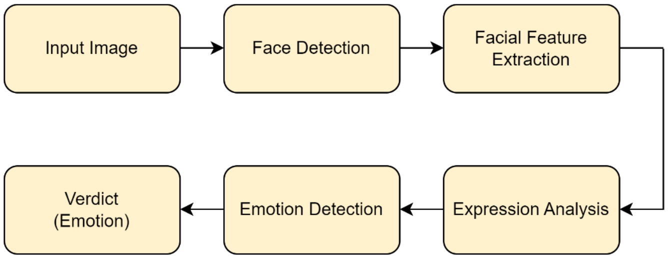


Fig. 2. Block diagram for the implemented systems.

blocks. These blocks include two convolutional layers with skip connections that facilitate smoother gradient flow during training. Downsampling is achieved through strided convolutions or max-pooling layers to reduce spatial dimensions and increase receptive fields. At the end of the network, global average pooling layers are employed instead of fully connected layers, reducing parameters and computational complexity while maintaining high performance in image classification tasks. Overall, ResNet-34 is celebrated for its ability to train deep networks effectively and achieve state-of-the-art results in various computer vision tasks.

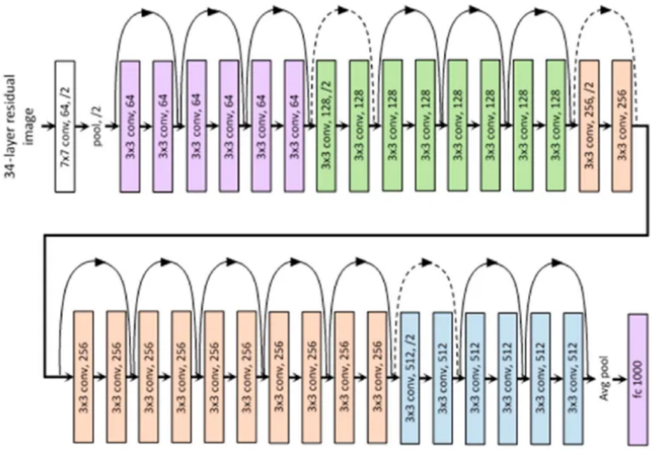


Fig. 3. ResNet-34 Architecture

C. Proposed Model

Our framework comprises three distinct systems: the speech emotion recognizer, facial emotion recognizer, and text recognizer. These systems operate independently, analyzing speech, facial expressions, and text inputs, respectively. To generate the final emotion prediction, we integrate the outcomes of these three systems.

- **Facial Emotion Recognizer:**
  1) **Video Segmentation:** The initial step divides the main video into smaller segments based on factors like time intervals or scene changes.
  2) **CNN Model (ResNet-34):** Each segment is passed through ResNet-34, known for its effectiveness in image classification. It extracts features from each frame, treating them as individual images.
  3) **Residual Networks (ResNet):** ResNet utilizes residual blocks, easing the training of deep networks by learning residual mappings.
  4) **Predictions:** ResNet-34 generates predicted values for each segment, representing aspects like object recognition or emotional content.
  5) **Concatenation:** Predictions from all segments are combined into a single sequence, maintaining temporal order.
  6) **LSTM Model (Sequential Model):** The concatenated sequence is inputted into an LSTM model, capturing dependencies and patterns in sequential data.
  7) **Sequential Emotion Analysis:** The LSTM analyzes emotional content sequentially, considering how emotions evolve.
  8) **Output:** The LSTM's output provides the final classification of emotions for the entire video, accounting for both spatial and temporal aspects.
- **Speech Emotion Recognizer:**
  1) **MFCC Extraction:** Mel-Frequency Cepstral Coefficients (MFCC) are derived from the audio signal to capture its characteristics. Frames of the audio signal undergo processing involving windowing, FFT, Mel filtering, logarithm computation, and DCT to produce MFCC coefficients, representing spectral features.
  2) **ResNet-18 Model:** ResNet-18, renowned for image classification, is adapted to process MFCC features for emotion prediction.
  3) **Input Representation:** The MFCC coefficients extracted from each audio frame form a sequence of feature vectors. These feature vectors can be arranged as a 2D matrix, where the rows represent time frames and the columns represent the MFCC coefficients.
  4) **Convolutional Layers:** Multiple convolutional layers extract spatial patterns from MFCC data treated as an image-like input.
  5) **Residual Blocks:** The output of the convolutional layers is flattened and passed through one or more fully connected layers. These layers perform the final classification based on the learned features, ultimately predicting the emotion label associated

with the input audio.

6) **Pooling Layers:** Pooling layers reduce spatial dimensions of feature maps while retaining important information.
7) **Fully Connected Layers:** The output of the convolutional layers is flattened and passed through one or more fully connected layers. These layers perform the final classification based on the learned features, ultimately predicting the emotion label associated with the input audio.
8) **Emotion Prediction:** ResNet-18's output provides a probability distribution over emotion classes, selecting the class with the highest probability as the predicted emotion for the audio segment.

In the final stage of the emotion prediction process, the feature representations obtained from the text, audio, and video inputs are combined into a single representation. This combined representation, often referred to as a feature map, captures the relevant information from each modality (text, audio, video) that is indicative of the underlying emotional content.

The feature maps from text, audio, and video inputs are concatenated to form a unified feature map, preserving individual characteristics while creating a comprehensive representation. This concatenated map is then fed into the final classification layer, which predicts the associated emotion. Typically consisting of fully connected layers and softmax activation, this layer produces a probability distribution over emotion categories. By combining information from multiple modalities, the model enhances emotion prediction accuracy, leveraging their complementary nature for robustness and effectiveness.

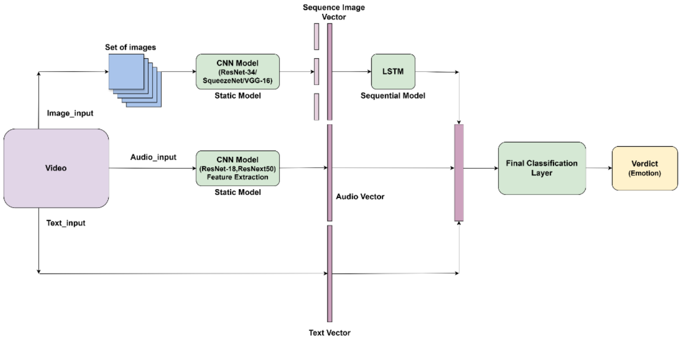


Fig. 4. Architecture of the Proposed Model

V. Dataset

The **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** serves as a comprehensive and multifaceted resource for investigating emotional expressions within the realms of speech and song. The dataset comprises 7356 files featuring contributions from twenty-four professional actors—twelve women and twelve men—ensuring gender balance. Each actor delivers lexically matched phrases in a neutral North American accent, spanning emotions like calmness, happiness, sadness, anger, fear, surprise, and disgust in both speech and song. To enrich emotion representation, expressions are crafted at two intensity levels—normal and strong—alongside a neutral expression, fostering a diverse portrayal for robust emotion recognition capabilities.

In the specific context of our research project, the focus is laser-sharp on audio-visual emotion recognition within the domain of speech. This necessitates a meticulous selection process, directing our analysis exclusively towards the full Audio-Video material and the speech channel, with songs excluded from consideration. This deliberate refinement streamlines the dataset to 1440 videos, each meticulously curated with a maximum duration of 5.31 seconds and a minimum duration of 2.99 seconds, aligning with the project's emphasis on speech-centric emotion recognition.

VI. Simulation Results

A. Performance comparison analysis of various models for audio and image

- Hyperparameters such as dropout rate are tuned to optimize the performance of deep neural network architectures. Dropout regularization is applied to prevent overfitting by randomly dropping out a fraction of units during training.
- ResNet34, a deep residual network architecture, achieves higher accuracy compared to SqueezeNet and VGG16 for video classification tasks. This superior performance is observed when using a dropout rate of 0.5.
- In addition to higher accuracy, ResNet34 also offers reduced computation time compared to SqueezeNet and VGG16. This reduction in computation time can be attributed to the efficient architecture of ResNet34, which enables faster inference without compromising accuracy.
- The combination of higher accuracy and reduced computation time makes ResNet34 a more efficient choice for video classification tasks.

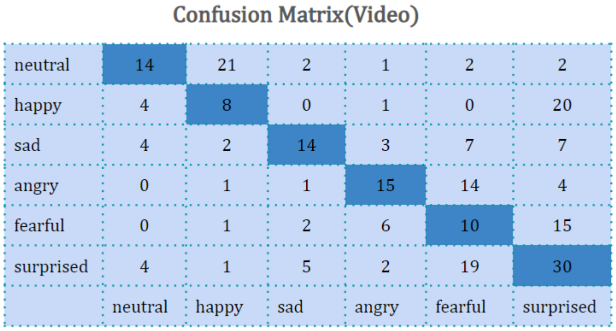
Model	Hyper-parameters	Accuracy	Computation Time
ResNet18+VGG16	Drop-out rate: 0.3	70.83%	73.04 seconds
	Drop-out rate: 0.5	71.98%	49.08 seconds
ResNet18+Squeezenet	Drop-out rate: 0.3	70.73%	36.52 seconds
	Drop-out rate: 0.5	70.80%	35.94 seconds
Proposed Model	Drop-out rate: 0.3	72.17%	32.44 seconds
	Drop-out rate: 0.5	77.99%	28.54 seconds

TABLE I. Table demonstrating the performance analysis of multimodal emotion recognition

B. Facial Emotion Recognition:

- In the realm of facial emotion recognition, the choice of ResNet34 over VGG16 and Squeezenet yielded an accuracy of 45.6%. This comparative evaluation underscores the pivotal role of architectural selection in deep learning tasks, especially in contexts where intricate features from audio data are pivotal for accurate emotion recognition

- Despite the achieved accuracy not reaching desired levels, the utilization of ResNet34 signifies a promising avenue for further investigation and refinement in the field of speech emotion recognition.
- Future research endeavors could focus on nuanced parameter tuning, expansion of training datasets, or exploration of alternative neural network architectures to enhance the accuracy and robustness of emotion recognition from speech signals



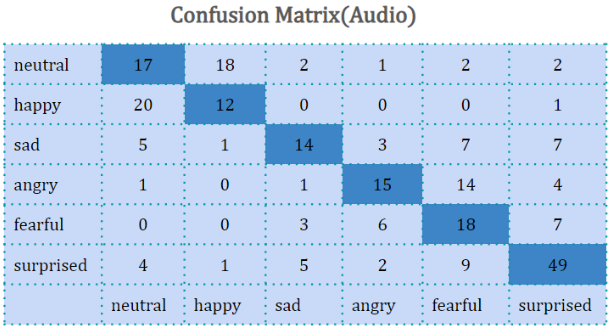
Confusion Matrix(Video)

neutral	14	21	2	1	2	2
happy	4	8	0	1	0	20
sad	4	2	14	3	7	7
angry	0	1	1	15	14	4
fearful	0	1	2	6	10	15
surprised	4	1	5	2	19	30
	neutral	happy	sad	angry	fearful	surprised

Fig. 5. Average Confusion Matrix of Video trained with ResNet34 with validated with an accuracy of 73.84%

C. Speech Emotion Recognition

- ResNet18 outperformed alternative architectures such as ResNext51 and ResNet101, achieving an accuracy of 64.6%. This highlights the importance of architecture selection in achieving optimal performance in emotion recognition tasks.
- Despite the achieved accuracy not reaching desired levels, the utilization of ResNet34 signifies a promising avenue for further investigation and refinement in the field of speech emotion recognition.
- Future research endeavors could focus on nuanced parameter tuning, expansion of training datasets, or exploration of alternative neural network architectures to enhance the accuracy and robustness of emotion recognition from speech signals

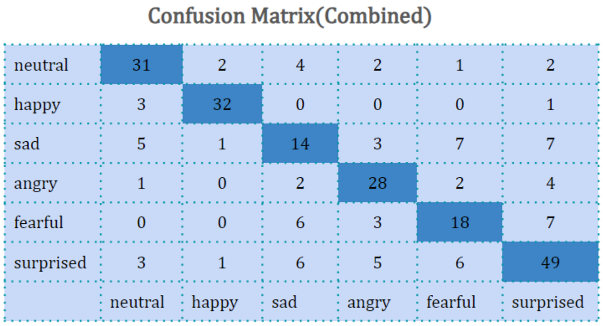


Confusion Matrix(Audio)

neutral	17	18	2	1	2	2
happy	20	12	0	0	0	1
sad	5	1	14	3	7	7
angry	1	0	1	15	14	4
fearful	0	0	3	6	18	7
surprised	4	1	5	2	9	49
	neutral	happy	sad	angry	fearful	surprised

Fig. 6. Average Confusion Matrix of Audio trained with ResNet18 validated with an accuracy of 79.76%

D. Emotion Recognition using Video and Audio:



Confusion Matrix(Combined)

neutral	31	2	4	2	1	2
happy	3	32	0	0	0	1
sad	5	1	14	3	7	7
angry	1	0	2	28	2	4
fearful	0	0	6	3	18	7
surprised	3	1	6	5	6	49
	neutral	happy	sad	angry	fearful	surprised

Fig. 7. Average Confusion Matrix of Audio-Video combined using our proposed model obtains an accuracy of 85.02%

VII. Conclusion

Accurately identifying emotions presents a formidable challenge due to the myriad ways individuals express and interpret them, even within the same linguistic and cultural context. In response, our approach combines two Convolutional Neural Network (CNN) models with Long Short-Term Memory (LSTM) networks, outperforming single-modality models. Despite this progress, further research is crucial to capture the dynamic nature of emotions, as our model still falls short of replicating human-level accuracy. Additionally, certain video frames contain more crucial emotional cues than others, posing a challenge for existing models. However, by integrating visual data with other modalities such as text and audio, our model achieves a remarkable 77.99% accuracy in emotion classification, highlighting the importance of leveraging multiple sources for enhanced performance.

References

[1] A. A. A. Zamil, S. Hasan, S. M. J. B. J. M. A., and Zaman, I. Emotion detection from speech signals using voting mechanism on classified frames. In *International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)* (2019).

[2] A. Sharma, V. B., and Arora, J. Machine learning techniques for real-time emotion detection from facial expressions. In *2nd Edition of IEEE Delhi Section Flagship Conference (DELCON)* (2023).

[3] Hosur, V., and Desai, A. Facial emotion detection using convolutional neural networks. In *IEEE 2nd Mysore Sub Section International Conference (MysuruCon)* (2022).

[4] J. K. Bhatia, J. P. Singh, P. K. S., and Chauhan, V. K. Emotion detection using facial expressions. In *International Conference on System Modeling Advancement in Research Trends* (2022).

[5] Kaiming He, Xiangyu Zhang, S. R. J. S. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016).

[6] Kumar, M., and Srivastava, S. Emotion detection through facial expression using deeplearning. In *International Conference on Information Systems and Computer Networks* (2021).

[7] Sinoara, R. A., A. J. . R. S. O. Text mining and semantics: a systematic mapping study. In *Journal of the Brazilian Computer Society* (2017).

[8] Y. Fan, M. Xu, Z. W., and Cai, L. Automatic emotion variation detection in continuous speech. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (2014).