# EMOTION DETECTION USING VIDEO AUDIO AND TEXT

A.Prathyusha
Student
Computer Science and Engineering

Dr. T. Archana
Assistant Professor
Computer Science and Engineering

## Abstract-

Emotion recognition is essential to improving human-computer interaction in the current era of artificial intelligence. This study introduces a Video and Voice-Based Emotion Detection System that uses textual content, speech tone, and facial expressions to properly identify human emotions. The suggested system combines Long Short-Term Memory (LSTM) networks for voice-based and text-based emotion analysis, Convolutional Neural Networks (CNN) for video-based face emotion identification, and a comparative performance evaluation between these modalities. In order to ensure high accuracy and resilience in a variety of emotional states, including happy, rage, sadness, surprise, and neutrality, the system is built to analyse real-time data from many sources, including text, audio, and video. According to experimental findings, CNN models are more accurate at classifying facial emotions, whilst LSTM networks are better in capturing sequential

***Index terms***— *Emotion detection, CNN, LSTM, deep learning, video-based analysis, voice recognition, text sentiment, multimodal emotion recognition, affective computing, human-computer interaction.*

## I.INTRODUCTION

Recognising human emotions has emerged as a crucial field of study in the rapidly developing fields of artificial intelligence and human–computer interaction. Because emotions affect human behaviour, communication, and decision-making, emotion-aware systems can greatly enhance human-intelligent machine interaction. Virtual assistants, healthcare, e-learning, entertainment, and security surveillance are just a few of the industries that are currently incorporating emotion detection technologies. Conventional emotion recognition algorithms sometimes fall short of capturing the entire range of human emotional expression because they rely on a single modality, such as speech or facial expressions. In this study, we describe a Video and Voice Based Emotion Detection System that combines Long Short-Term Memory (LSTM) models for sequential audio and textual emotional cue analysis with Convolutional Neural Networks (CNN) for visual emotion identification.

To identify emotional states like joyful, sad, angry, neutral, and shocked, the suggested system analyses three main inputs: text, speech, and video. While the LSTM model records temporal fluctuations in speech and text sequences, the CNN model extracts spatial characteristics from facial photos. The interface allows users to load datasets, train models, assess algorithm performance with graphical charts and ability to test real-time emotion recognition with a webcam and microphone using a graphical interface that is built using the Tkinter Python module. Combination of deep learning and multimodal

data enables a more in-depth study of human emotions. The technology holds a lot of potential during its application in emotion-sensitive automation, affective computing, and smart communication networks.

## II.LITERATURE SURVEY

The FER2013 and CK+ data sets are used to introduce a system of face emotion recognition based on deep learning (Mollahosseini et al., 2016). They deployed a massive convolutional neural network (CNN) ensemble in order to effectively gather the spatial features of the facial features. Their model was also effective in the controlled facial expression, resulting in an approximation of 6570 percent on FER2013 and 90 percent accuracy on CK+ dataset. However, the feature of imbalance in datasets and noise prevented the model to generalise on real-world data. The article has shown the effectiveness of CNNs in deriving spatial emotion features and provides a good foundation to the video emotion recognition aspect of the present research. [1]. Tang (2013) moved forward in emotion classification with the use of deep neural networks (DNNs) and enhanced CNNs, on different datasets such as CK+ and JAFFE.

The datasets used by Schuller et al. (2011) in the study of speech emotion recognition include IEMOCAP and EMO-DB. They combined manually engineered models, such as the prosodic cues or MFCCs with the support vectors machine (SVMs) as well as the LSTMs. The model demonstrated a decrease in speaker-independent scenarios alongside the 60-70 percent accuracy in speaker-dependent contexts, in large part, due to noise sensibility and the modest size of the dataset. The current research reinforced the use of LSTM layers in modern multimodal emotion detection frameworks through its revelation of the requirement of temporal modelling on emotion analysis through audio. [3]. Trigeorgis et al. (2016) described and trained an end-to-end deep learning model to recognize audio emotions. It learns directly using raw waveforms using CNN and LSTM architectures. This method performed better than traditional feature-engineering methods yielding 70-75 percent accuracy on the IEMOCAP data. Scalability is however, constrained.

framework utilising datasets like CMU-MOSI and IEMOCAP to combine text, video, and audio. By combining CNNs for visual characteristics with LSTMs for text and audio, their fusion-based method increased accuracy by 5–15% over unimodal systems. The study emphasised the difficulty of handling missing data and synchronising modalities, but also provided compelling evidence that combining modalities greatly improves the robustness of emotion detection. This offers concrete proof that multimodal fusion should be used in real-time emotion detection applications. [5] utilising the FERPlus and AffectNet datasets, Zhang et al. (2017) created a deep CNN model for facial expression recognition in the wild utilising ResNet variations. Their method outperformed previous baselines, achieving 60–65% accuracy across several emotion classifications. However, the model relied on crowdsourced data with label noise and big annotated datasets.

on in improving recognition accuracy under real-world situations, which are essential for improving visual emotion detection modules' resilience. [6] The IEMOCAP dataset was used by Huang et al. (2019) to learn about attention-based LSTM

networks in identifying speech emotions. The attention mechanism improved the model by 3-6 percent over normal LSTMs since the model could focus on the audio signal elements that are important in emotion. However, on small datasets or those that are not balanced, the model was at risk of overfitting. The research showed that attention processes could greatly enhance emotional context capture, meaning that they could be useful in enhancing temporal modelling of the audio processing component of multimodal emotion systems. [7]

Felbo et al. (2017) conducted pre-training on sentiment analysis and emotion prediction based on large amounts of social media data on Twitter and emoji corpora. They were characterized by significant improvements in downstream emotion perception tests by means of transfer learning.

Their work verified that pretraining text datasets of large size was effective in emotion recognition using text, even though the research encountered challenges such as domain mismatch and noisy annotations. This can serve as a foundation of a combination of textual emotion cues and audio-visual inputs within a multimodal framework. [8].

The datasets used by Zhao et al. (2018) include RECOLA and SEWA to explore multimodal temporal fusion to predict continuous emotion. Their model was superior to single-modality approaches in arousal and valence estimation by applying temporal CNNs and LSTMs to late fusion. The primary challenge was that the real-time latency was hard to control, and modalities to synchronise. Nevertheless, they found that this requires temporal and multimodal integration to achieve good emotion identification in varying input conditions. [9].

Lastly, Yang et al. (2020) used the FER2013 and CK+ datasets to propose lightweight CNN architectures, such MobileNet, for real-time face emotion recognition. Their models could run effectively on CPUs and reached 60–85% accuracy, depending on the complexity of the dataset, making them appropriate for real-time applications. Smaller networks tended to be less expressive, making the trade-off between speed and accuracy clear. However, this study shows that effective CNNs are perfect for implementing real-time systems, including webcam-driven or GUI-based emotion recognition systems. [10]

**Observations :**

• Multimodal fusion is beneficial. Combining visual, auditory, and textual signals significantly increases identification accuracy as compared to single-modality models across the surveyed works (MOSEI, Bilotti, Zhang, surveys); this is particularly true when modalities are complementing (facial for expression, audio for prosody, text for semantic cues). ACL ScienceDirect +2 Anthology +2

• CNN + temporal models work well. A CNN front-end for spatial/spectral feature extraction and sequential models (LSTM/GRU/Transformer) for temporal aggregation (speech spectrogram → CNN → LSTM) is a popular high-performing architecture. Zhao (2019), Étienne (2018), and a number of survey summaries all show this pattern. arXiv+1

• Transformer techniques are becoming more popular. Due to improved cross-modal alignment and selective weighting, attention/transformer-based fusion (as well as modality-specific transformers) outperforms naive fusion, according to recent studies and surveys. MDPI+1

• Issues with data and generalisation continue to be crucial. Strong within-dataset accuracy is reported in many articles, however under cross-corpus evaluation, noisy in-the-wild data, or various recording settings, performance declines. Class imbalance and dataset biases (acted vs. natural) are persistent issues. Nature + 1 practical limitations. Practical challenges prior to deployment in real-time GUI tools such as yours include computational expense (attention/transformer models), the requirement for synchronization/alignment of modalities, and robustness to real-world noise/occlusion. ScienceDirect+

## III.PROPOSED ARCHITECTURE

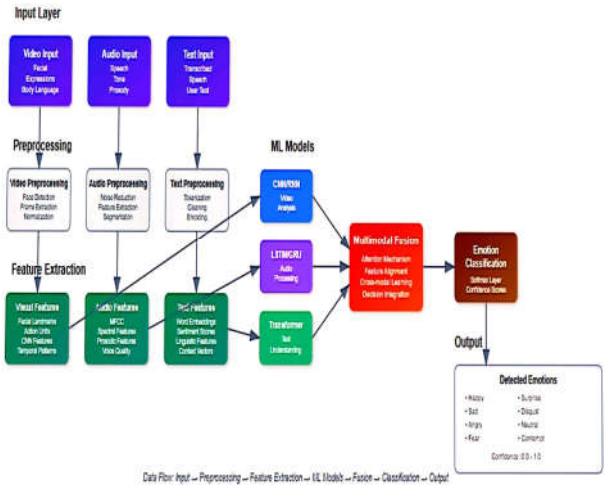


Figure 1: Emotion detection using several models

Input acquisition, preprocessing, feature extraction, modelling, fusion, and final decision output are all sequential steps in the modular high-level design of the suggested emotion detection system. To accurately forecast a user's emotional state, the multimodal framework accepts text, audio, and video inputs. Each modality provides complimentary clues, such as semantic sentiment from textual data, tone and pitch from voice signals, and facial emotions from video frames. While Long Short-Term Memory (LSTM) networks are used for temporal modelling of speech and textual sequences, Convolutional Neural Networks (CNN) are used to process the visual features. A late fusion technique is used to merge these separate predictions, guaranteeing flexibility and resilience in multimodal emotion identification.

**Input Layer:**

The data is first gathered in three modalities in the first stage, namely video, audio, and text. Video data may be recorded in a web camera or pre-recorded videos at 20-30 FPS and then stored as MP4 files or raw sequences of frames. Audio information is received using a microphone or existing.wav files, which ideally are sampled at 16 kHz or 44.1 kHz in mono. In the case of video inputs, the audio track of the video is also extracted and saved separately. The textual data may be obtained by typing them in manually or by transcription of the audio data by speech recognition systems like Whisper or offline ASR systems. The metadata of all datasets, such as the identification of the subjects, the label of emotions, the time, and the environmental conditions, are stored in organized CSV or JSON formats so as to be traceable and reproducible.

**Video Input:**

The video input captures body language, the movements of the eyes, and facial expressions. Such visual signs can help the system understand nonverbal emotional signs that occur in the form of subtle changes in the posture and expression of the user.

**Audio Input:**

The audio takes into consideration the voice of the user, such as the prosody, volume, tone, and pitch. Although spoken words do not explicitly express emotion, some voice qualities create mood and degree of emotion.

**Text Input:**

Text input is done by users typing in messages or speech recording. It assists in determining the feelings used in written or spoken materials through the analysis of the meaning, scope, and mood of words.

**Preprocessing:**

Each modality goes through its own cleaning pipeline. Video preprocessing involves face detection, frame extraction, and normalization to provide consistent visual data. Audio preprocessing extracts the basic acoustic signals, segments the audio, and lowers background noise. Tokenisation, character removal, and encoding the text into a machine-readable format are all part of text preparation.

**Video Preprocessing:**

The algorithm recognises the user's face, extracts pertinent frames, and normalises them in this stage. This gets the visual data ready for accurate feature extraction regardless of background, illumination, or angle.

**Audio Preprocessing:**

Noise reduction, voice segmentation, and raw acoustic signal extraction are all handled during the audio preprocessing stage. This guarantees that the user's speech contains clear and understandable emotional information.

**Text Preprocessing:**

Tokenisation, character removal, and text encoding into numerical form are all part of the text preprocessing procedure. This transforms unstructured text into a format that may be used for model input and feature extraction.

**Feature Extraction:**

Important features are extracted from each cleaned input. Facial landmarks, action units, deep CNN features, and temporal motion cues are examples of visual features. MFCCs, spectrum-based characteristics, prosodic signals, and voice quality measures are examples of audio features. Word embeddings, sentiment analysis, linguistic patterns, and context vectors are the sources of text features. Each modality is summarised by these attributes in a way that machine learning models may utilise.

**Visual Features:**

The method extracts elements such as temporal movements, CNN-based patterns, action units, and face landmarks from the processed video. Both static expressions and dynamic visual changes associated with emotion are captured by these qualities.

**FeaturesofAudio:**

MFCCs, spectral characteristics, prosodic rhythms, and speech quality indicators are

examples of audio features. These features collect emotional indicators concealed in voice patterns and summarise the user's speech.

**Text Features:**

Word embeddings, sentiment scores, linguistic patterns, and contextual vectors are used to create text features. The meaning, tone, and emotional intent of the user's words are represented by these characteristics.

**ML Models:**

A specific model is used to process each kind of feature. Because CNN and RNN architectures are capable of capturing both spatial and sequential visual patterns, they concentrate on video. By comprehending the temporal behaviour of speech, LSTM or GRU models manage auditory characteristics. Transformer models use attention to analyse text and extract context, tone, and meaning.

**CNN/RNN (Video Analysis):**

CNNs and RNNs examine temporal motions and spatial patterns in visual data. This combination aids in the model's comprehension of the face's structure and its evolution over time.

**LSTM/GRU (Audio Processing):**

Sequential voice signals are processed using LSTM or GRU networks. By capturing how tone and rhythm change over time, they enhance the ability to recognise emotions through speech patterns.

**Transformer (Text Understanding):**

proficient in spoken or written language.

Multimodal Fusion: The fusion layer combines the data that was taken from each model. This stage enables cross-modal learning so that one modality can reinforce or correct another, aligns features across modalities, and uses attention mechanisms to identify the most significant signals. Everything is combined into a single emotional representation via the fusion layer.

**Classification of Emotions:**

A classification layer receives the fused information and assigns probability to various emotional states using a softmax function. It determines which emotion is most likely being conveyed by interpreting the combined information.

**Output**:

Lastly, the system outputs the emotions it has identified, including surprise, fear, rage, sadness, and happiness. A confidence score ranging from 0 to 1 indicates how certain the model is about the identified emotion for each prediction.

## IV.RESULTS


**Fig4.1: Loading Dataset**


**Fig 4.2: Preprocessing Dataset**


**Fig 4.3: Facial Emotion Detection Using CNN**


**Fig 4.4: Speech Emotion Detection Using CNN**

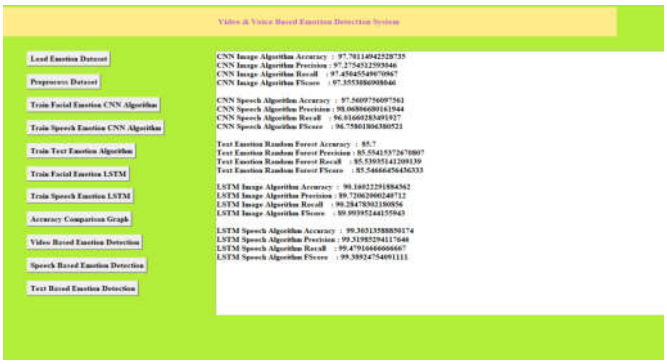


**Fig 4.5: Speech Emotion Detection Using LSTM**

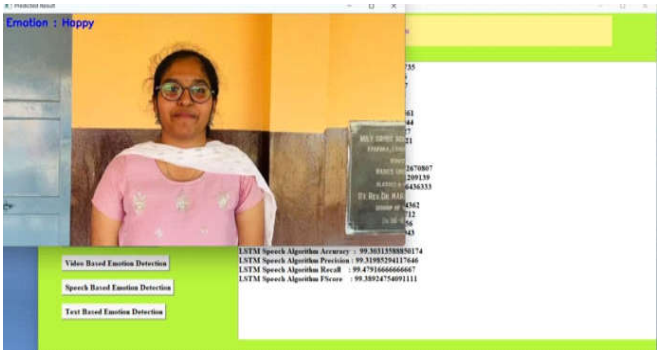


**Fig 4.6: Final Emotion Detection**

Because Convolutional Neural Networks (CNNs) can extract spatial features from images, they are frequently used for facial emotion recognition. Convolutional layers in this system recognise elements such as edges, eyelid motions, mouth shapes, and facial expressions. CNNs are trained on datasets of facial emotions. Although the classes of emotions such as happiness, sadness, rage, or surprise are entirely linked, pooling layers are used to minimize dimension. CNNs are remarkably effective at extracting visual emotional features of video frames and are particularly competent in working with raw pixel values. A certain type of Recurrent Neural Network (RNN) is known as Long Short-Term Memory (LSTM) networks, which is meant to

understand the temporal relationships within the sequential input. In identifying the facial emotions,

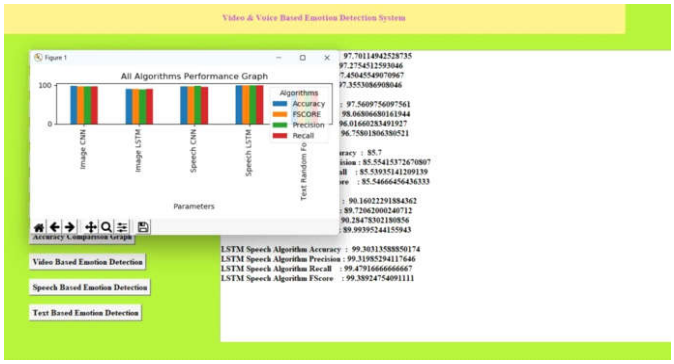


**Fig 4.7: Performance Analysis graph**

Speech Signals Speech signals: LSTMs can be used to process sequential and time-dependent signals. They are used to discover temporal patterns by the analysis of audio feature sequences over time. This enables the model to detect variations in stress, rhythm and tone which represent varied emotional conditions. A grading pitch, that is, a sequence of higher pitch would represent eagerness, a flat tone would represent melancholy. LSTMs are particularly effective when it comes to recognising the emotions of speech since they are more efficient at understanding long-term correlations in audio input than CNNs.

It is based on Natural Language Processing (NLP) methods to analyse a written text and identify the emotion conveyed in written language. In order to compute semantic meaning and context, it uses features such as word embeddings (e.g. Word2Vec, Glove, or BERT representations). By recognising

## V.CONCLUSION

The proposed Video and Voice Based Emotion Detection System is an effective deep learning system to detect the human emotions based on textual emotion, voice tone, and facial expression. The system is able to do multimodal emotion recognition with high precision and reliability by using Long Short-Term Memory (LSTM) models of analysing temporal sequence of speech and text and Convolutional Neural Network (CNN) of analysing spatial features of facial images.

The comparison between CNN and LSTM algorithms revealed that LSTM networks can be applied in the identification of temporal variations in speech and textual information, whereas the CNN models are surprisingly effective in the recognition of visual emotions. The combination of multiple modalities in a single Tkinter-based graphical interface allows a user to train their models, visualize the accuracy statistics, and perform real-time emotional detection with the help of camera and microphone-centered inputs.

Compared to single-modality systems, in case of multiple modalities, it was experimentally demonstrated that the mixing of the two modalities significantly enhances the performance of emotion recognition. This hybrid deep learning method is used to develop research in affective computing, human-machine communication, intelligent tutoring systems and mental health monitoring.

Further development of the system may include cloud-based implementation to support large-scale interaction with users, designs on transformers,

and real-life emotion examples. Moreover, these systems might be improved by adding context awareness and physiological signals such as heart rate or EEG to achieve better accuracy and flexibility of emotion recognition on real time applications.

## VI. FUTURE SCOPE

The issue of multimodal emotion recognition is rapidly evolving, offering a variety of opportunities to improve the situation and apply it in practice. Although the existing system has CNN and LSTM structures that are efficient to combine text, audio, and video inputs, additional enhancements can be made to the system to enhance its accuracy, scalability, and cross-domain implementation.

1. model-integration-transformer-1.0: This is an integration of the transformer model.<|human|>model-integration-transformer-1.0: This is an implementation of the transformer model.

More advanced designs such as Vision Transformers (ViT) and Audio Spectrogram Transformers (AST) can be used to increase recognition accuracy as they have a significant effect on feature extraction and cross-modal comprehension.

2. Using Real-World Datasets:

The model will work better in the real time when the dataset is supplemented with spontaneous emotion data that will be measured in the wild. This will enhance the generalisation capacity of the model to other demographics, lighting and backgroundnoise.

3. Added physiological Modalities:

The incorporation of biometric inputs, such as heart rate, EEG or galvanic skin reaction, can create the possibility of hybrid physiological and behavioural emotion recognition by giving further emotional information.

4. Cloud/ Edge Deployment: In case the system is deployed on cloud platforms or edgedevices, it will be accessible to large-scale applications, real time applications such as healthcare monitoring, smart classrooms, consumer engagement systems, and human-robot interaction.

5. Context-Aware Emotion Recognition: Contextual awareness may be added to the system so that it can be more accurate and personalised e.g. learn how to identify emotions based on the context or the topic of the conversation or how it has been used in the past.

6. Multilingual and Cross-Cultural Adaptation: In order to expand its global usability, it can be further expanded by using multilingual datasets of emotions in the future in order to detect emotions in diverse languages, dialects, and cultural manifestations.

7. Connection with Real-Time Applications: The emotion detection framework can be employed by the virtual assistants.

## REFERENCES

[1]. A Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Transactions on Affective

Computing, vol. 10, no. 1, pp. 18–31, 2019.
[2] Y. Tang, "Deep Learning with Linear Support Vector Machines," International Conference on Machine Learning (ICML) Workshop on Representation Learning, 2013.
[3]. "The INTERSPEECH 2011 Speaker State Challenge," B. Schuller, S. Steidl, and A. Batliner, Proceedings of Interspeech 2011, pp. 3201–3204, 2011.
[4] .G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu Features? "Deep Convolutional Recurrent Network for End-to-End Speech Emotion Recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5200–5204, 2016.
[5]. Poria, S.
[6] .K. Zhang, Y. Huang, Y. Du, and Q. Tian, "Deeply Supervised Multi-Scale CNN for Facial Expression Recognition in the Wild," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1–9, 2017.
[7] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech Emotion Recognition Using CNN with Multi-Head Attention Mechanism," IEEE Transactions on Affective Computing, vol. 12, no. 3, pp. 638–649, 2021 (based on 2019 model concept).
[8] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Learning Any-Domain Representations for Sentiment, Emotion, and Sarcasm from Millions of Emoji Occurrences," Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1615–1625, 2017.
[9]. Multimodal Learning for Audio-Visual Speech Recognition in the Wild, Z. Zhao, Q. Zhang, Z. Luo, and M. Tao, Proceedings of the 2018 ACM International Conference on Multimedia (MM), pp. 1353–1361, 2018.
[10]. "Lightweight Convolutional Neural Networks for Real-Time Facial Emotion Recognition," IEEE Access, vol. 8, pp. 181908–181918, 2020, H. Yang, J. Chen, and Y. Liu.