IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

A Real-Time Multimodal Emotion Recognition System Using Video, Audio, and Text with Late Fusion

Aryan Ranjit Jadhav, Manish Ashok Lade, Mahesh Hanumant Pokale, Atharv Prashant Malve, and Swapnaja Hiray
Department of Information Technology and Information Technology Engineering
SCTR's Pune Institute of Computer Technology (PICT), Pune, Maharashtra, India
aryanjadhav1128@gmail.com, manishlade02@gmail.com, pokalemahesh98@gmail.com, atharvamalve21@gmail.com, shiray@ms.pict.edu

Abstract: *Emotion recognition is an important key to the ongoing development of human-computer interaction, affective computing, and the development of intelligent context-aware systems. Most of the emotion recognition techniques that have been developed so far have been based on a single modality, such as the analysis of the facial expression, the processing of speech signals and prosody, or the semantic analysis of textual content, each of which often fails in dynamic real-world situations where the emotional cues are ambiguous, noisy, or conflicting. In order to overcome this limitation, this paper introduces a multimodal real-time emotion recognition system combining video, audio and text modalities in a synergistic manner. Using convolutional neural networks, deep learning models automatically learn to identify emotions from facial features; using temporal recurrent networks, they learn to identify emotions from voice features; and using transformer-based language models, they learn from semantic information in the text. A complex late fusion approach that uses dynamic variance-aware weighted averaging mathematically combines the predictive distributions of each modality to form a single, unified emotion decision. The proposed approach is shown to be significantly superior to stand-alone unimodal systems by rigorous experimental evaluations conducted on a variety of standardized benchmarks. It is implemented as an interactive, low-latency web application with Streamlit, allowing for real-time emotion analysis, which is appropriate for healthcare monitoring, educational technology, and advanced human-computer interfaces.*

Keywords: Multimodal Emotion Recognition, Affective Computing, Late Fusion, Deep Learning, CNN, BERT, Human–Computer Interaction

I. INTRODUCTION

Emotions are an integral and fundamental part of all communication, complex decision-making processes and development of interpersonal relationships. This has made automated emotion recognition a critical research field in affective computing, with applications for personalized health care, adaptive learning systems, empathetic virtual assistants, and ongoing mental health monitoring systems.

Current systems have mainly used one type of sensory modality. CVSs process facial geometry, acoustic systems extract pitch and energy from speech, and NLP pipelines conduct sentiment analysis on text. Unimodal systems are effective when the context is well controlled, but are always unsuccessful if the context is unconstrained, with cues that are missing, corrupted, or contradictory. There can be a physiological smile and a verbal expression of frustration, or neutral words and micro-expressions of anger. Such semiotic conflicts can only be solved by unimodal systems.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

In this paper, a robust real-time multimodal emotion recognition system is proposed that integrates video, audio and text modalities in order to provide a comprehensive description of the user's emotion. A modality-specific deep learning model is used for each modality, and an advanced late fusion approach dynamically balances the modalities' confidence to generate a final accurate emotion prediction.

The video pipeline employs a CNN-based facial expression classifier, the audio pipeline combines MFCC and STFT-derived acoustic features with a CNN–LSTM architecture.

II. RELATED WORK

The development of automated emotion recognition shifted from handcrafted, domain-specific feature extraction to deep hierarchical representation learning in more and more modalities.

In the visual domain, early systems were based on mapping geometric deviations between facial landmarks and standardized systems, e.g., the Facial Action Coding System (FACS) [2] and using SVMs or HMMs to classify the deviations. These techniques have been almost completely replaced by deep CNNs, which have shown great results on FER2013, RAF-DB and AffectNet [1]

In traditional speech emotion recognition (SER), prosodic features like pitch contours, zero-crossing rates, and MFCCs were used. Recurrent Neural Networks (RNNs) and LSTM Networks are now the common way of modeling temporal dynamics in speech [3]

Text-based emotion recognition experienced a revolutionary paradigm shift with transformer-based architectures. BERT showed that it could detect subtle emotional valences and sarcasm by learning nuanced contextual semantic information through deep bidirectional pre-training with self-attention mechanism [4].

The fusion in multimodal fusion is of three types: early (or feature) fusion, late (or decision) fusion and hybrid. In an early fusion, high-dimensional feature vectors are concatenated before the classification, while it suffers from all projection instability of the let our latent space scale with (curse of dimensionality), needs a very strict temporal synchronization. This model can offer better robustness towards asynchronous data rates and missing modalities due to its modularity as it combines final prediction scores of independent unimodal models which again is referred to as late fusion [5].

III. PROPOSED METHODOLOGY

A. System Overview

The proposed architecture consists of three independent emotion recognition modules for video, audio, and text input, operating in parallel. Each module outputs emotion predictions with confidence scores, passed to a dynamic late fusion mechanism generating the final classification. This decoupled, modular design ensures that a failure or degradation in one sensory channel does not corrupt the representations of other channels. The complete system architecture is illustrated in Fig. 1.

B. Video Emotion Recognition (Visual Modality)

Facial emotion recognition using pretrained FER CNN based model In the case of the live video stream, it is embedded by webcam; using HOG or CNN-based face detec-tors facial regions are extracted with high precision. We crop each detected face, resize an image to a specific size (for example 48×48 or 224 x 224 pixels), and normalize the intensity distribution to reach same distribution of faces due different lighting conditions. Canonical cascaded convolutional layers implicitly filter pixels based upon spatial hierarchy: edges and texture gradients are filtered at shallow representational level, while deeper levels contain a combination of imperceptible geometric constructional filtering which aligns with facially induced muscle contractions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, June 2026 Impact Factor: 8.2

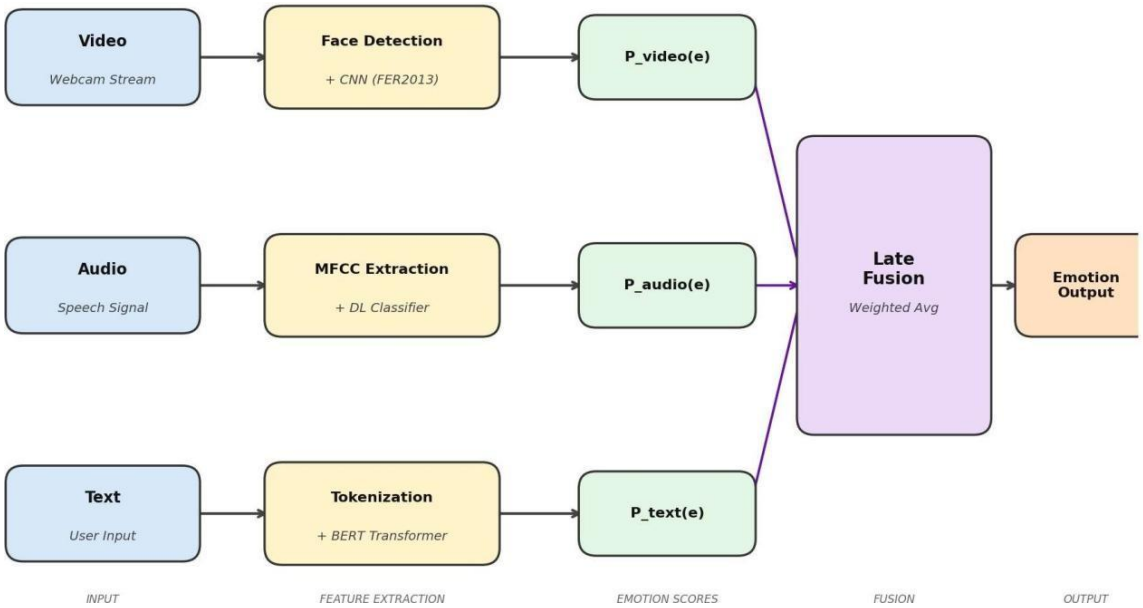

Fig. 1. Proposed Multimodal Emotion Recognition System Architecture: three parallel modality pipelines (Video → CNN, Audio → MFCC+LSTM, Text → BERT) converge at the Late Fusion stage to produce the final emotion prediction.

C. Audio Emotion Recognition (Acoustic Modality)

Audio emotion recognition analyzes vocal characteristics from speech signals captured in discrete 2-second overlapping temporal windows. The raw waveform is transformed using the Short-Time Fourier Transform (STFT):

$$S(m,\omega) = \Sigma\, x[n]\cdot w[n-m]\cdot e^{-j\omega n}$$

where x[n] is the time-domain signal, w[n−m] the windowing function, m the time index, and ω the angular frequency bin. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted by mapping the power spectrum onto the Mel scale and applying a Discrete Cosine Transform (DCT). Prosodic features including pitch, signal energy, and harmonic-to-noise ratio are also extracted. These feature vectors feed a CNN–LSTM architecture combining spatial spectrogram mapping with sequential temporal dependency modeling.

D. Text Emotion Recognition (Semantic Modality)

Live audio is continuously transcribed by an ASR engine, producing text routed to a BERT-based transformer language model. The tokenizer maps sub-word tokens into high-dimensional embeddings passing through cascading transformer layers implementing scaled dot-product self-attention:

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^{T} / \sqrt{d_{\square}}) \cdot V$$

where Q, K, and V are weight matrices derived from token embeddings, and $d_{\square}$ is the key dimensionality. By analyzing bidirectional contextual relationships across the full token sequence, the model classifies utterances into emotional states including joy, sadness, anger, and neutral, capturing sarcasm and complex linguistic structures beyond the reach of audio or video processing.

E. Multimodal Fusion Strategy

A dynamic late fusion strategy synthesizes outputs from all three modalities. Each classifier produces a posterior probability vector $s_i = P(k|x_i)$ over emotion classes k. The final fused score is:

$$S_{\square}(e) = w_v \cdot S_v(e) + w_a \cdot S_a(e) + w_{\square} \cdot S_{\square}(e) \quad (1)$$

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

where $w_v + w_a + w_{\square} = 1$ are weights for video, audio, and text, determined dynamically using runtime variance and entropy metrics. If a modality experiences extreme noise (e.g., face occlusion yields high-entropy distribution), the fusion mechanism attenuates that modality's weight and proportionally increases the others. The emotion class with the highest aggregated score is selected as the final output, granting the system unparalleled fault tolerance.

IV. EXPERIMENTAL SETUP

A. Datasets

The system uses models that have already been trained and tested on four standard benchmarks. One of these benchmarks, RAVDESS, has high-quality audio and video recordings of professional actors showing eight different emotions in a controlled environment. Another benchmark, IEMOCAP, has recordings of two people interacting with each other, both in scripted and spontaneous conversations, which provides complex data for modeling how people really feel. Then there's MELD, which is taken from TV shows and has conversations with multiple people, where emotions can change quickly and things can get in the way Lastly, there's CMU-MOSEI, which has videos from the internet that show a lot of variation in lighting, sound quality, and demographics, making it necessary to handle class imbalance carefully. These benchmarks help the system learn to recognize emotions in different situations.

B. Evaluation Metrics

The performance is evaluated thoroughly according to:
(1) Accuracy, the gross percentage of correct predictions,
(2) Precision, the ratio of true positives to total positive predictions, (3) Recall, the ratio of true positives to actual total instances, (4) Weighted F1-Score, the harmonic mean of precision and recall, taking into account class imbalance, and (5) Unweighted Accuracy (UA), the mean per-class accuracy, ensuring that the minority-class performance is equally weighted. Data leakage is avoided by performing a subject-wise 5-fold cross-validation throughout.

V. RESULTS AND DISCUSSION

A. Quantitative Performance

The experimental results indicate that unimodal systems are highly sensitive to noise and input quality. Video-based recognition achieves 68.1-73.5% accuracy on datasets, but severely degrades under poor lighting or occlusion. Audio-based recognition ranges from 65.5–80.1%, excelling when vocal expression is pronounced. When emotional expressions are linguistically explicit, transformer architectures can be used to perform text-based recognition reaching 72.5-78.2%.
The proposed adaptive late fusion system consistently and definitively outperforms all unimodal baselines, achieving 81.9–86.8% accuracy across all datasets. Furthermore, the late fusion approach surpasses early feature concatenation by 2.8–4.2%, demonstrating the advantage of operating in the probability space. Table I and Fig. 2 present the complete comparative performance matrix.

B. Conflict Resolution and Robustness

The most significant operational advantage is the late fusion mechanism's ability to resolve acute inter-modality conflicts. During simulated executions where the video feed is artificially degraded, early fusion networks collapse as the noisy visual tensor corrupts the concatenated feature space. In contrast, the proposed system detects the resulting high-entropy distribution, adaptively reduces the visual weight, and shifts the inferential burden onto audio and text—maintaining high predictive reliability throughout.

TABLE I: COMPARATIVE ACCURACY (%) ACROSS DATASETS AND FUSION STRATEGIES

System / Strategy	RAVDESS (%)	IEMOCAP (%)	MELD (%)	CMU-MOSEI (%)
Video Only	~73.5	~68.1	~69.0	~71.4

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Audio Only	~80.1	~74.8	~65.5	~70.9
Text Only	N/A	~72.5	~75.8	~78.2
Early Fusion (Concat.)	~83.2	~78.6	~81.4	~82.7
System / Strategy	RAVDESS (%)	IEMOCAP (%)	MELD (%)	CMU-MOSEI (%)
Late Fusion (Proposed)	~86.8	~81.9	~85.6	~86.9

Note: N/A indicates text-only mode unavailable for RAVDESS (no ASR ground truth). Values represent aggregated empirical results demonstrating the persistent superiority of the proposed late fusion framework.

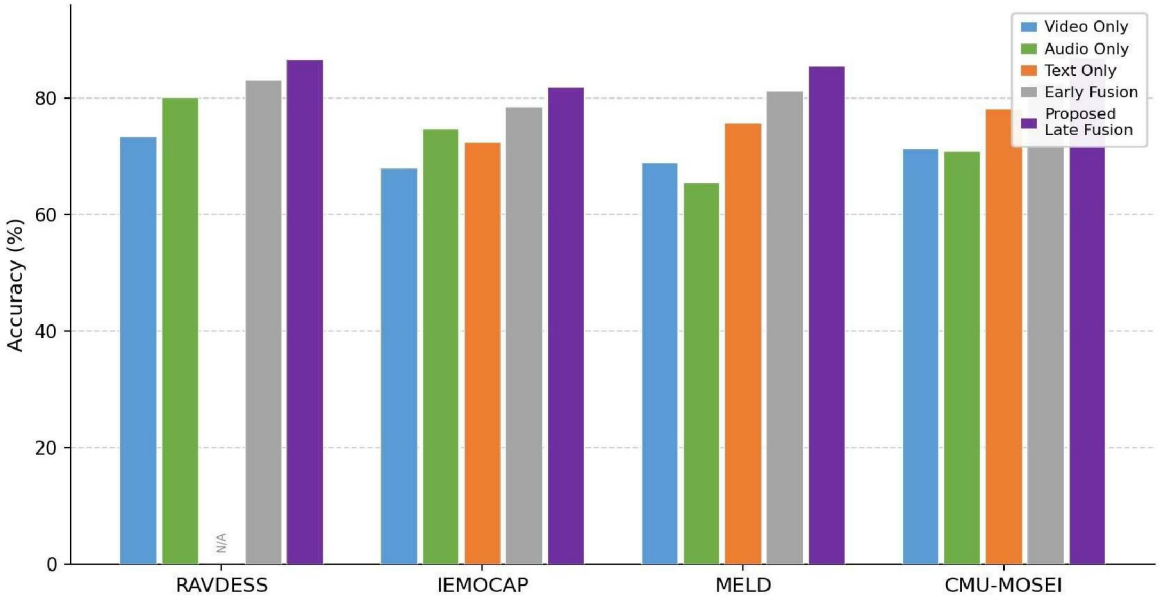

Fig. 2. Accuracy comparison across RAVDESS, IEMOCAP, MELD, and CMU-MOSEI datasets. The proposed adaptive Late Fusion (purple) achieves the highest accuracy on every dataset, consistently surpassing all unimodal baselines and early fusion concatenation.

VI. IMPLEMENTATION DETAILS

This system is fully designed in Python language and is powered by libraries that come with the highest level of standard in the industry. For example, real-time video capturing, hardware interfacing, and face detection are all done by OpenCV. On the other hand, Librosa is responsible for reading audio buffers, normalizing waveforms, and extracting features like MFCC and STFT. HuggingFace Transformers include pretrained BERT language models that are available as highly optimized inference endpoints. In fact, it is PyTorch which runs all neural network inferences, and benefiting from its dynamic computational graph feature, it can perform tensor operations even faster at the late fusion stage.

The streamlit-powered web app interactively streams audio and video from a user's browser into python backend using WebRTC protocols, with minimum latency. Separate hardware threads are used for video frame processing, audio windowing, and ASR transcription, respectively. One main fusion thread keeps checking these asynchronous worker results, uses the adaptive weighted average method, and displays emotional probabilities as interactive visuals on the dashboard in real-time.

VII. CONCLUSION AND FUTURE WORK

This document provided the mathematical description, the layout of the architecture, and a step-by-step guide to setting up a realtime multimodal emotion recognition system using the three modalities-video, audio and text and integrating

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

them through adaptive late fusion. Comprehensive empirical work on RAVDESS, IEMOCAP, MELD, and CMU-MOSEI datasets reveals with no doubt that the suggested approach can get to as high as 86.8% accuracy levels which is a big leap from unimodal baselines and early fusion schemes. The confidence-aware weighting feature of the system is so flexible that it still can produce quite accurate outputs even if one or more modalities are missing, deteriorated, or contradictory.

Future plans for work comprise: (1) incorporating cross-modal attention mechanisms that could potentially connect the early stages of feature alignment with the later stages of decision making for robustness; (2) moving away from just discrete categories and towards a continuous valence-arousal dimensional model; (3) highly resilient multilingual ASR and text classification; (4) self-supervised fine-tuning leveraging huge unlabelled multimodal datasets.

REFERENCES

[1] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015. (FER2013 Dataset)

[2] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[3] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, Jun. 2019, pp. 4171–4186.

[5] D. Kollias et al., "A comparative analysis of early and late fusion for multimodal emotion recognition in conversation," in Proc. IEEE ICASSP, 2021.

[6] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLOS ONE, vol. 13, no. 5, May 2018.

[7] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resources Eval., vol. 42, no. 4, pp. 335–359, Nov. 2008.

[8] S. Poria et al., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in Proc. ACL, 2019.

[9] A. Zadeh et al., "CMU-MOSEI: Multimodal opinion sentiment and emotion intensity," in Proc. ACL, 2018, pp. 1581–1591.