Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Multimodal Emotion Recognition using Deep Feature Fusion and Attention Mechanism: A Survey

Faisal Majeed
Department of Computer Science and Engineering
Ganga Institute of Technology and Management
kablana, India

Poonam Dhankhar
Department of Computer Science and Engineering
Ganga Institute of Technology and Management
kablana, India

***Abstract* - Multimodal Emotion Recognition (MER) is a central area of affective computing offering a more precise, accurate and robust understanding of human emotional states. It is achieved by the integration of multiple data streams at a time like text data, speech data, and visual data information.[1] This report offers a comprehensive survey of the methodologies that are driving the modern MER systems with a main focus on the integration of deep feature fusion and advanced attention mechanisms. I have proposed an evolutionary framework consisting of four stages such as foundation, optimization, multi scale, and cross fusion of these channels, so as to classify the technological progress that is moving MER toward real time application in human computer interaction (HCI).[1] It undertakes a systematic analysis of various fusion strategies including early, late, hybrid, and intermediate layer fusion and underscores the critical importance of techniques such as cross modal, recursive, and multi-scale attention in integrating the challenges occurring from data heterogeneity. The document outlines that Multimodal Large Language Models (MLLMs) and generative reasoning currently constitute the state-of-the-art system providing potent instruments for explaining the hidden relations within intricate social interactions.[5] Also, the survey strongly recommends integration of Explainable AI (XAI) to ensure requisite transparency and capture user confidence in emotion aware systems.[7] By preparing the performance metrics which are derived from the datasets including IEMOCAP, MELD, and CMU-MOSEI, the report establishes the present performance benchmark and outlines the future research designs for the development of multimodal systems which are characterized by enhanced robustness, adaptability, and inclusivity.[1]**

Keywords - Multimodal Emotion Recognition (MER), Human Computer Interaction (HCI), Emotion aware system.

I. INTRODUCTION

Emotion recognition is a fundamental prerequisite for the development of empathetic and intelligent machines which are capable of smooth interaction with human users. Emotions have different psychological aspects that are demonstrated across diverse biological and behavioral channels including facial expressions, vocal rhythmic patterns, language related choices, and physiological signals like heart rate or skin conductance.[11] The early research focused mostly on unimodal systems and therefore these approaches often suffered from a lack of robustness in real-world environments due to various factors such as variable lighting, acoustic noise, and the inherent ambiguity of single-channel signals.[7] For instance, a textual "I am fine" can be interpreted as either positive or deeply distressed depending on whether it is accompanied by a smile or a trembling voice.[1]

Multimodal Emotion Recognition (MER) addresses these limitations by taking the full advantage of the complementary and redundant nature of different modalities. By integrating information from multiple sources, the MER systems offer a more broader representation of a person's emotional state and thus mimicking the way humans naturally fuse multisensory cues.[11] The core architecture of a modern MER system typically involves three stages such as modality specific feature extraction, feature fusion, and classification or regression.[2]

In recent years, the integration of deep learning has revolutionized MER particularly by the use of deep feature fusion and attention mechanisms.[1] Attention mechanisms allows a model to dynamically prioritize the most informative features among the modalities. This effectively addresses the reality i.e., not all our sensory inputs contribute equally to an emotional judgment in every context.[11] For example, in a loud noisy environment, the visual modality may carry more weight than the sound related modality.[16]

The transition from discriminative tasks to generative reasoning represents the most recent milestone in the field. The advent of Multimodal Large Language Models (MLLMs) has enabled systems to go beyond categorical labeling toward understanding the "why?" behind an emotion and it's a shift facilitated by instruction tuning and large scale multimodal datasets.[5] As these systems move from research labs to sensitive applications such as mental health diagnosis and autonomous driving, the challenges faced are also significant. The Challenges of data synchronization, modality missingness, and explainability have become central themes of investigation for scholars.[8] This survey aims to map this complex landscape by providing a technical and insightful overview of the state-of-the-art in MER.

II. RELATED WORK

The trajectory of multimodal emotion recognition has been marked by a steady progression in the sophistication of how features are extracted and merged. Historically, the field moved from manual feature engineering to deep learning, and finally to the current era of Transformer-based architectures and MLLMs.

A. Evolution of Attention-Based Models (2020-2025)

The integration of attention mechanisms has redefined how systems handle the "modality gap"the semantic and structural differences between heterogeneous data streams.[1] A four stage evolutionary model characterizes the research context of the last half-decade:

1. **Foundation Stage**: Early attention models focused on basic weighted representations within a single modality to improve feature relevance.
2. **Optimization Stage**: Models began to employ simple cross-modal interactions which often focuses on temporal attention to capture the dynamic changes in emotion from the beginning of data processing.
3. **Multi-Scale Stage**: In this stage, research shifted toward extracting features at different granularities. For instance, ScaleVLAD (2023) demonstrated the value of multi-scale feature extraction in capturing both local micro-expressions and global emotional context.
4. **Cross-Fusion Stage**: Cutting edge models like RJCMA (2024) and MemoCMT (2025) represent the current trend of deep, recursive cross-modal interaction, where all the modalities are iteratively aligned and refined to achieve a deeper aggregation of information.

B. Key Architectural Breakthroughs

Specific models have played a pivotal role in establishing the benchmarks for conversational and non-conversational MER. Explained in Table I.

TABLE I

Model Name	**Year**	**Primary Innovation**	**Key Performance Metric**
AM-RNN	2021	Speaker state GRU + pair-wise attention for conversational context.[1]	81.29% Accuracy (IEMOCAP).[1]
TEMMA	2022	Early fusion combined with temporal attention.[1]	Improved dynamic emotion tracking.[1]
RJCMA	2024	Recursive joint cross-modal attention for deeper feature alignment.[1]	Advanced cross-modal interaction.[1]
MemoCMT	2025	Bidirectional cross-attention for deep aggregation.[1]	State-of-the-art feature refinement.[1]
CENet	2025	Feature transformation strategy for input refinement.[1]	High robustness to input noise.[1]

C. Convergence with Large Language Models

The shift toward MLLMs has transformed MER into a reasoning task. The models like **Emotion-LLaMA** utilize emotion specific encoders such as HuBERT for audio and MAE/VideoMAE for visual features which are aligned into a shared space with a language model backbone.[6] This allows the system to leverage the vast world knowledge of LLMs to interpret complex social scenarios.[6] Similarly, the **BeMERC** framework incorporates speaker behaviors, also including facial micro expressions and body postures into a vanilla MLLM to model emotional dynamics during a conversation.[26]

The emergence of these models is supported by new datasets like **MER-Caption**, which contains over 115,000 video description pairs which lead to shifting the focus from simple labels to descriptive emotion understanding.[21] This reflects a broader trend toward open vocabulary emotion recognition where the system is not limited to a predefined set of categories like "happy" or "sad" but can describe nuanced emotional states in natural language.[21]

III. CHALLENGES, PROBLEMS AND METHODOLOGIES COMPARISON

Despite the rapid advancement of MER, the field faces several persistent challenges related to the nature of multimodal data and the complexity of human affect.

A. Technical Challenges in Multimodal Data

1) Modality Heterogeneity and Alignment

The fundamental differences between the different data such as textual, acoustic, and visual data leads to a problem which is known as the "modality gap." The textual data operates on symbolic discrete units, while acoustic data is a continuous frequency signal and the video/visual data consists of a spatial-temporal sequence.[18] This successfully brings these varied sensory inputs together into a single channel and makes their cohesive representation a significant technical challenge. Furthermore, temporal misalignment i.e, mismatch in timing of the various signals is a common issue.[29] Emotional cues in different modalities rarely occur at the exact same moment. This makes necessary the use of advanced mechanisms, such as the Multimodal Transformer (MulT) which are designed to learn these underlying latent alignments.[2]

2) Missing Modalities and Data Sparsity

In real world applications, it is common for one or more modalities to be missing or corrupted due to problems such as hardware failure, environmental noise or occlusion.[14] Traditional fusion methods often assume complete data which leads to a significant degradation in performance particularly when inputs are incomplete.[22] To address this, researchers have proposed methods like **Modality Invariant Feature Acquisition and Retrieval Augmented MER (RAMER)**. This RAMER uses contrastive learning (bringing similar things closer and different apart) to acquire features that are consistent across modalities or retrieve related data to compensate for missing signals.[22]

3) The Modality Balance Dilemma

A common problem in MER is the tendency of models to over relying on a dominant modality typically text and ignoring the small signals in other data modalities.[30] This is called "modality balance dilemma" and it prevents the system from fully utilizing the complementary information provided by audio and vision.[30] Also, the standard decoupling methods used in practice might cause "modality specialization disappearance" where the unique predictive capability of unimodal data is lost during fusion.[30] The **EMOE (Modality-Specific Enhanced Dynamic Emotion Experts)** model addresses this by using a mixture of experts to dynamically adjust modality weights based on sample features.[30]

B. Comparison of Fusion Methodologies

Table II

Fusion Strategy	**Timing**	**Strengths**	**Weaknesses**
Early (Feature-Level)	Before classification training.[2]	Captures early correlations; low computational complexity.[2]	Mere concatenation may obscure unique modality features; risk of feature degradation.[2]

Late (Decision-Level)	After individual modality predictions .[2]	Flexible; does not require temporal synchronization; allows local optimization.[2]	Ignores inter-modal interactions; final accuracy may be limited.[2]
Hybrid	Integrates early and late stages.[2]	Maximizes emotional information utilization; captures interplay and modality expertise.[2]	High design complexity; escalated computational demands.[2]
Intermediate (Deep)	Within network layers of deep models.[2]	Leverages deep learning representation; models intra- and inter-modality interactions.[2]	Requires complex architectural design (e.g., Aligned Word-level vs. Unaligned features).[2]

Fusion strategies are generally categorized by the stage at which information integration occurs. Each approach has distinct strengths and weaknesses that influence its suitability for specific applications discussed in Table II.

1) Intermediate Fusion Sub-Architectures

Intermediate fusion has evolved into highly specialized sub-types:

- **Simple Concatenation Fusion:** The features are processed independently and then joined as input for the next deep layer to learn interaction information.
- **Utterance-Level Interaction Fusion:** This layer focuses on modeling interactions at the level of the entire spoken statement.
- **Fine-Grained Interaction Fusion:** This layer is important for capturing nuanced interactions across distinct time steps. The models like **MARN (Multi-Attention Recurrent Network)** use Multi-Attention Blocks to discover cross view dynamics at each time step while **MulT** adapts streams from one modality to another to handle asynchrony.[2]

C) Attention Mechanisms and Their Mathematical Foundations

Attention mechanisms have become the "glue" that binds multimodal features together. The core principle involves weighting input elements by their importance in a given context.[20]

1) Scaled Dot-Product and Cross-Modal Attention

The standard attention algorithm which is used in Transformers calculates an attention matrix by multiplying a Query (Q) matrix with a Key (K) matrix then is followed by a softmax and scaling operation. This is then multiplied by a Value (V) matrix.[20]

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In self-attention, Q, K, and V all originate from the same modality. In **cross-modal attention**, Q comes from one modality (e.g., text) while K and V come from another (e.g., audio), allowing the text to "attend" to relevant acoustic cues.[20]

2) Advanced Attention Mechanisms

The newer architectures uses specialized attention mechanisms to enhance feature representation:

- **Parallel Cross-Modal Attention (PCMA):** It ensures consistency and alignment by combining cross-modal attention in parallel rather than sequentially.
- **Semantic Attention (SEMA):** It is used to ensure the visual modality prioritizes information aligned with the audio modality. It utilizes a Hadamard element wise product to fuse features and a sigmoid activation to generate a semantic map.
- **Cross-Modality Relation Attention (CMRA):** It captures temporal dynamics while integrating complementary information, improving robustness to noise.[16]

D. Higher-Order Fusion: Tensor and Bilinear Methods

Since simple addition or concatenation may not fully capture complex interactions, researchers use techniques like bilinear pooling and tensor decomposition to model them better.

1) Tensor Decomposition Fusion

Tensor fusion involves creating an outer product of modality features, but this leads to an explosion in dimensionality. To mitigate this, **Tucker Decomposition** is used to reduce the model's parameter count and lower the risk of overfitting by approximating the full tensor with a core tensor and factor matrices.[32]

2) Bilinear Pooling and Semi-Tensor Product

Bilinear pooling mechanisms like **Factorized Bilinear Pooling (FBP)** and **Compact Bilinear Gated Pooling (CBGP)** allow for a more expressive fusion of features.[33] The **Semi-Tensor Product (STP)** method has emerged as a promising alternative to traditional tensor methods. This offers greater flexibility and reduced computational demands by allowing matrix multiplication between dimensions that do not strictly match.[34]

E) Comparative Benchmark Results

The efficacy of these methodologies is rigorously evaluated on public datasets. Recent state-of-the-art results demonstrate a consistent improvement in weighted F1-scores and accuracy, discussed in Table III.

Table III

Dataset	**Year**	**Model /Method**	**Weighted (%)**	**Modalities Used**
IEMOCAP	2025	Domain Gen. + GNN [9]	+0.97% gain	A, V, T
IEMOCAP	2025	Emotion -LLaMA [25]	89.10	A, V, T
MELD	2025	Dialogue MLLM [35]	68.57	A, V, T
MELD	2025	Domain Gen. + GNN [9]	+0.65% gain	A, V, T
CMU-MOSEI	2025	Domain Gen. + GNN [9]	+1.09% gain	A, V, T
CMU-MOSEI	2025	AHOT [36]	88.30	A, V, T

Research suggests that models incorporating **Graph Neural Networks (GNN)** perform better in conversational settings because they can model both global dialogue context and local speaker-specific emotional changes.[9] For example, when the GNN module was removed in an ablation study, accuracy on IEMOCAP decreased by 3.69%.[9]

IV. FUTURE DIRECTION

The future of Multimodal Emotion Recognition lies in moving beyond simple classification toward systems that are interpretable, reasoning capable, and universally applicable.

A. Generative Reasoning and Emotion Understanding

The transition from static classification to generative reasoning is a primary trend. The future systems will likely leverage the "emergent capabilities" of MLLMs to perform Multimodal Emotion Reasoning (MER) which involves synthesizing fine grained signals like prosodic features (pitch, energy) and facial micro expressions to decode latent causality. This allows the system to understand why an emotion is being expressed, rather than just what the label is.[37]

B. Explainable Artificial Intelligence (XAI)

A critical requirement for the deployment of MER in sensitive domains (e.g., healthcare, legal) is transparency. Future research will prioritize **Multimodal XAI** which provides "natural explanations" that are human-interpretable and adaptable to user preferences.[38] The techniques like **SHAP (Shapley Additive Explanations)** and **Grad-CAM** are being integrated to identify which modality or specific feature (e.g., a specific word or facial movement) contributed most to a prediction. This approach not only increases trust but also helps in identifying and mitigating model biases.[40]

C. Cross-Domain and Cross-Lingual Generalization

As the global demand for MER increases, systems must become capable of generalizing across different cultures and languages. Recent work on the MERC-PLTAF method has shown promise in cross-lingual settings, specifically between English and Mandarin by using refined feature extraction and sophisticated fusion to overcome language barriers.[41] Future systems will need to address the cultural variance in emotional expression to ensure fairness and accuracy across diverse populations.[23]

D. Real-Time and Edge Computing

The practical deployment of deep learning models is often hampered by high computational overhead. Innovations aimed at reducing real-time response latency (with some frameworks already reporting a 70% reduction) will be essential for integration into smart environments and social robotics.[7] Edge computing applications, which process data locally to preserve privacy and reduce latency, represent a significant growth area for the MER market.[42]

E. Meta-Learning and Parameter-Efficient Tuning

To solve the data scarcity problem, the use of Meta-Learning and parameter efficient tuning (e.g., **LoRA** - Low-Rank Adaptation) will allow models to adapt to new tasks or emotional categories with minimal labeled data.1 This will enable the development of inclusive and adaptive systems that can "learn to learn" about individual emotional quirks or specialized clinical contexts.[1]

V. CONCLUSION AND FUTURE WORK

The development of Multimodal Emotion Recognition systems has undergone a profound transformation. It is evolving into a sophisticated sensing paradigm that emulates the human capacity for multisensory affect perception. The integration of deep feature fusion and attention mechanisms has been the primary factor of this progress, enabling systems to overcome the limitations of unimodal approaches and the challenges of heterogeneous data. By 2025, the field has reached a state of maturity where models can not only recognize emotions with high accuracy on benchmark datasets but also begin to reason about the underlying context using large scale multimodal backbones.

The four-stage evolutionary model which were identified in this survey which begin from the foundation of feature extraction to the current state-of-the-art in cross fusion and generative reasoning shows a clear pathway toward deeper modality interaction and more refined attention strategies.1 The challenges such as modality misalignment, missing data, and the black-box nature of deep models still persist. The emergence of XAI, contrastive learning, and MLLMs provide the solution for addressing these issues.[5]

Moreover, the future of MER lies in the creation of intelligent, inclusive, and adaptive systems that can navigate the complexities of real world human interaction. There is a need of transparency and reasoning along with the predictive accuracy so that the next generation of emotion aware systems will play a crucial role in enhancing the quality of human machine communication in the fields of healthcare, education, and beyond

REFERENCES

[1] L Han, L., et al., Attention Mechanism in Multimodal Emotion Recognition From 2020 to 2025: Technological Evolution, Challenges, and Future Prospects, February 23, 2025

[2] Gu, H., et al., A comprehensive survey of deep learning-based multimodal emotion recognition, 2023

[3] A Survey on Multimodal Emotion Recognition: Methods, Datasets, and Future Directions, 2026

[4] Boulahia, S. Y., et al., Early, intermediate and late fusion strategies for robust deep learning-based multimodal emotion recognition, September 30, 2021

[5] Shou, Y., et al., Multimodal Large Language Models Meet Multimodal Emotion Recognition and Reasoning: A Survey, September 29, 2025

[6] Cheng, Z., et al., Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning, 2024

[7] Multimodal Emotion Recognition with Explainable AI for Cognitive Human-Computer Interaction in Smart Environments, August 4-6, 2025

[8] Enhancing Emotion Detection Accuracy and Transparency through Multimodal Fusion and Explainable AI, March 6-7, 2025

[9] Xie, J., et al., Multimodal Emotion Recognition Method Based on Domain Generalization and Graph Neural Networks, February 23, 2025

[10] Lu, Y. & Feng, H., Multimodal emotion recognition method based on an Adaptive High-order Transformer Network, October 27, 2025

[11] Zhang, J. et al., Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, 2023

[12] Parmar, A., Human Emotion Recognition Using Multi-modal Deep Learning: A Review of Methods, Datasets, and Challenges, 2024

[13] Shou, Y., et al., Multimodal large language models (MLLMs) for emotion recognition and reasoning, 2025

[14] Zhu, X. et al., A Review of Key Technologies for Emotion Analysis Using Multimodal Information, 2024

[15] Multimodal Emotion Recognition: Methods, Datasets, and Future Directions, 2026

[16] Hybrid Multi-ATtention Network (HMATN) for Audio-Visual Emotion Recognition, April 2025

[17] A unique deep multimodal emotion model based on capsule graph transformer networks, 2025

[18] Tsai, Y. H. H., et al., Multimodal Transformer (MulT) for Unaligned Multimodal Language Sequences, 2023

[19] Emotion recognition from facial and speech features using attention mechanisms, 2025

[20] Vaswani, A. et al., Attention Is All You Need, 2017

[21] Lian, Z. et al., AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models, January 27, 2025

[22] Liu, R. et al., Contrastive Learning Based Modality-Invariant Feature Acquisition for Robust Multimodal Emotion Recognition With Missing Modalities, 2025

[23] Geetha, A. V. et al., Cultural and individual differences in emotional expression: Challenges and future directions, 2025

[24] Cheng, Z. et al., Emotion-LLaMA: sensitivity to language and audio encoders, 2024

[25] Cheng, Z. et al., Emotion-LLaMA F1 Score benchmarks on

MER2023 Challenge, 2024

[26] Fu, P. et al., BeMERC: Behavior-aware MLLM-based framework for Multimodal Emotion Recognition in Conversations, 2025

[27] Lian, Z. et al., MER 2025: When Affective Computing Meets Large Language Models, April 28, 2025

[28] Wu, C., et al., Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects, November 2025

[29] Advanced cross-modal fusion mechanisms: FBP, CBP, and CBGP, 2023

[30] Fang, Y., et al., Emoe: Modality-Specific Enhanced Dynamic Emotion Experts, 2025

[31] A systematic summary of multimodal fusion methods: early, late, hybrid, and intermediate, 2023

[32] Wang, R., et al., Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking, 2025

[33] Liu, F. et al., A Parallel Multi-Modal Factorized Bilinear Pooling Fusion Method Based on the Semi-Tensor Product for Emotion Recognition, 2022

[34] Liu, F. et al., Semi-Tensor Product based Multi-modal Fusion Method for Emotion Recognition, 2022

[35] Zhang, J., et al., DialogueMLLM: A Generative MERC framework based on Multimodal Large Language Models, 2025

[36] Lu, Y. & Feng, H., Experimental results on IEMOCAP and CMU-MOSEI using AHOT, 2025

[37] Cheng, Z. et al., Multimodal emotion analysis shifting from static classification to generative reasoning, 2024

[38] Multimodal, Affective and Interactive eXplainable Artificial Intelligence (MAI-XAI 25), October 2025

[39] Gradient SHAP XAI method for identifying significant features in MER, 2025

[40] Mitigating model biases using XAI in Multimodal Emotion Recognition, 2025

[41] MERC-PLTAF: Multimodal Emotion Recognition in Conversations with prompt learning and temporal attention fusion, 2025

[42] Global Multimodal Emotion Recognition Analysis market size and projections, 2024

[43] Xie, J. et al., Feature extraction using RoBERTa, OpenSmile, and DenseNet, 2025

[44] ICASSP 2025 Multimodal Emotion and Intent Recognition Challenge (MEIJU), 2025