WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 7, Number 3, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v7n3.02
Journal homepage: https://wepub.org/index.php/IJCSIT/index

# A Survey on Multimodal Emotion Recognition: Integrating Cues for a Deeper Understanding of Affect

Zhanpeng Li [1, 2, 3], Yuming Qi [1, 2, 3, *], Sanpeng Deng [1, 2, 3], Xiumin Shi [1, 2, 3]

[1] Tianjin University of Technology and Education, Tianjin 300222, China

[2] Tianjin Key Laboratory of Intelligent Robot Technology and Application, Tianjin 300350, China

[3] Tianjin Bonus Robotics Technology Co., Ltd, Tianjin 300350, China

*Corresponding Author

## ABSTRACT

Multimodal Emotion Recognition (MER) has emerged as a crucial area of research in artificial intelligence and human-computer interaction, aiming to build systems that can understand human affective states by integrating information from various modalities. This review provides a comprehensive overview of the MER landscape, synthesizing insights from foundational and recent literature. We delve into the primary modalities utilized—including visual (facial expressions), acoustic (speech prosody), textual (language content), and physiological signals—and discuss the state-of-the-art deep learning techniques for feature extraction within each. A central focus is placed on multimodal fusion strategies, from early (feature-level) and late (decision-level) fusion to more sophisticated Transformer-based and attention mechanisms that capture complex inter-modal dynamics. We also examine the role of advanced architectures like Multimodal Large Language Models (MLLMs) and techniques such as knowledge distillation for handling real-world challenges like modality missingness. Key benchmark datasets that have propelled the field forward are described. Finally, we outline the persistent challenges, including data scarcity, modality misalignment, and real-world robustness, and propose promising future research directions to advance the development of more accurate, robust, and context-aware affective computing systems.

## KEYWORDS

Multimodal Emotion Recognition; Affective Computing; Deep Learning; Feature Fusion; Sentiment Analysis; Transformer Models

## 1. INTRODUCTION

Emotion is a fundamental aspect of human experience, influencing cognition, behavior, and social interaction [1, 2]. Enabling machines to recognize and appropriately respond to human emotions is a cornerstone of advanced human-computer interaction (HCI) and artificial intelligence. As comprehensive surveys have shown, the field has evolved from unimodal analysis to complex multimodal integration to better capture the nuances of affect [3, 4, 22]. Multimodal Emotion Recognition (MER) addresses this challenge by emulating the human ability to perceive affect through a combination of cues, such as facial expressions, tone of voice, spoken words, and even physiological responses.

Unimodal systems often struggle with ambiguity [5, 6]. For instance, the phrase "That's great" can be sincere or sarcastic depending on the vocal prosody and facial expression. By integrating information from multiple modalities, MER systems can overcome these limitations, creating a more holistic and accurate understanding of an individual's emotional state [7, 8]. This capability has significant

applications in diverse domains, including healthcare, education, intelligent driving systems, and social robotics.

The rapid advancement of deep learning has revolutionized the field [9]. Architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and especially Transformers have provided powerful tools for extracting salient features and modeling the complex dependencies both within and between modalities [10, 12, 16]. This paper presents a comprehensive survey of the current state of MER, synthesizing key findings from the literature. We explore the foundational concepts, review the key modalities and feature extraction techniques, provide an in-depth analysis of multimodal fusion strategies, and discuss recent trends. We conclude by identifying the major challenges and outlining future research directions.

## 2. MODALITIES AND FEATURE EXTRACTION

An effective MER system begins with robust feature extraction from individual modalities. The most commonly used modalities are visual, acoustic, textual, and physiological.

### 2.1. Visual Modality

The face is a primary channel for emotional expression. While early methods used hand-crafted features, deep learning, particularly CNNs, now dominates. Models like ResNet, pre-trained on large image datasets [12], are widely used to extract deep spatial features. For video data, which captures the temporal evolution of expressions, architectures such as 3D-CNNs or a combination of CNNs and RNNs are employed to learn robust spatio-temporal representations [8, 10].

### 2.2. Acoustic Modality

Speech contains both linguistic content and acoustic cues. MER focuses on the latter, extracting prosodic features from the audio signal. Standard features include Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy, often extracted using toolkits like openSMILE [13]. Deep learning approaches often apply CNNs to spectrograms [5, 20, 21], followed by RNNs to model the temporal sequence of these features.

### 2.3. Textual Modality

With the prevalence of social media, text is a vital modality. The paradigm has shifted to contextualized representations using pre-trained language models like BERT and its variants. These models generate dense vector embeddings that capture rich semantic information [16], which can be further enhanced with external knowledge for improved sentiment analysis [7, 8].

### 2.4. Physiological Modality

Physiological signals like electroencephalography (EEG) and electrocardiography (ECG) offer an objective measure of emotional arousal [15, 18, 19]. As recent surveys highlight, their integration provides a robust channel for MER. Deep learning models, including 1D-CNNs and attention mechanisms, are increasingly used to learn discriminative patterns directly from these time-series signals, revealing correlations between physiological responses and emotional states.

# 3. MULTIMODAL FUSION STRATEGIES

The core challenge in MER is to effectively fuse the information from different modalities. The advancement of Multimodal Emotion Recognition (MER) hinges on effective fusion of multi-source data, with diverse fusion strategies exhibiting distinct performance trade-offs. Advantages, and disadvantages of mainstream fusion methods are systematically compared in Table 1.

## 3.1. Early, Late, and Hybrid Fusion

Early (feature-level) fusion [22] concatenates feature vectors at the input stage, allowing the model to learn low-level interactions. Late (decision-level) fusion combines the outputs of separate unimodal models, offering flexibility but missing low-level cues. Hybrid strategies [24] seek to combine the best of both worlds, often fusing features at intermediate layers of a deep network.

## 3.2. Attention-Based and Transformer-Based Fusion

The attention mechanism has become the dominant paradigm for modern fusion techniques.

Attention Mechanisms: These allow the model to dynamically weigh the importance of different features, both within a modality (self-attention) and between modalities (cross-modal attention). This enables the model to focus on the most salient information for the emotion recognition task.

Transformer-Based Fusion: The Multimodal Transformer (MulT), proposed by Tsai et al. (2019) [10], revolutionized the field by using a series of cross-modal attention blocks to let each modality attend to the others. This allows for deep, iterative fusion of information over time. Subsequent works like DialogueRNN [23] have adapted this for conversational contexts by modeling speakers and context, while recent models like M2Fnet [24] continue to refine these fusion networks.

**Table 1.** Comparison of Advantages, Disadvantages of Multimodal Fusion Methods

| Fusion Method | Advantages | Disadvantages |
|---|---|---|
| Early Fusion (Feature-level) | Captures low-level cross-modal correlations; retains raw data info; simple training. | High-dimensional feature space (increased complexity/computation); strict data alignment requirements; may miss complex interactions. |
| Late Fusion (Decision-level) | High flexibility (independent modal optimization); strong robustness; easy debugging. | Ignores fine-grained interactions; loses detailed info; low adaptability to new needs. |
| Hybrid Fusion | Balances strengths of early/late fusion; enhances global understanding. | Complex implementation; risk of overfitting (more parameters). |
| Attention-Based Fusion | Dynamically weights feature importance (intra/inter-modal); focuses on salient info for MER. | High computational complexity (weight calculation/updating). |
| Transformer-Based Fusion | Enables deep, iterative cross-modal fusion; handles sequential data/long-range dependencies well. | Complex structure (high training/inference cost); requires large datasets. |

# 4. RECENT TRENDS AND ADVANCED TOPICS

## 4.1. Multimodal Large Language Models (MLLMs)

The emergence of powerful MLLMs [17], such as GPT-4V [27], represents a new frontier. These models, pre-trained on vast web-scale multimodal data, demonstrate a remarkable ability to perform MER with zero-shot or few-shot capabilities, leveraging their immense world knowledge to understand context and nuance.

## 4.2. Handling Missing Modalities with Knowledge Distillation

In real-world scenarios, data from one or more modalities can be noisy or missing. Knowledge distillation has been proposed as an effective technique to address this. The approach involves training a powerful multimodal "teacher" model on complete data and then using its knowledge to train a unimodal "student" model, significantly improving its robustness when operating alone.

## 4.3. Self-Supervised Learning

Models like CLIP have pioneered learning transferable visual models from natural language supervision on a massive scale [28]. While not designed specifically for MER, this pre-training paradigm creates powerful encoders that can be fine-tuned for emotion recognition, mitigating the need for large, manually annotated datasets.

# 5. DATASETS AND BENCHMARKS

The advancement of MER relies heavily on public benchmark datasets, as detailed information on the core datasets that support its development is shown in Table 2.

IEMOCAP [26]: A foundational dyadic conversational dataset with audio, video, and text, annotated with categorical and dimensional emotions.

CMU-MOSI & CMU-MOSEI [29, 30]: Large-scale datasets of online opinion videos annotated with sentiment intensity, serving as standards for multimodal sentiment analysis.

MELD [31]: A multi-party conversational emotion dataset extracted from the TV series Friends.

DEAP [32]: A widely used dataset for emotion analysis using physiological signals (EEG, etc.) elicited by music videos

**Table 2.** Datasets

| Dataset | Text | Audio | Video | Physiological |
|---|---|---|---|---|
| IEMOCAP | √ | √ | √ | × |
| CMU-MOSI & CMU-MOSEI | √ | √ | √ | |
| MELD | √ | √ | × | × |
| DEAP | × | × | × | √ |

# 6. CHALLENGES AND FUTURE DIRECTIONS

Despite significant progress, MER still faces key challenges.

Data Scarcity and Annotation: High-quality annotated multimodal datasets are expensive to create [25]. Future work should focus on self-supervised learning and few-shot learning to reduce reliance on labeled data [28].

Modality Alignment and Temporal Dynamics [10, 23]: Aligning asynchronous data streams remains a hurdle. More sophisticated temporal modeling is needed to capture long-range dependencies.

Real-World Robustness [5, 6, 7, 8, 20, 21]: Models trained on clean data often fail in "in-the-wild" scenarios. Developing models robust to noise and missing modalities is critical.

Context and Personalization [4, 24]: Emotion is highly context-dependent. Future systems should incorporate conversational context and user-specific traits for personalized recognition.

Explainability and Ethics [18, 19]: As models become more complex, ensuring their decisions are transparent is vital. Moreover, the use of sensitive data raises significant privacy and ethical concerns that must be addressed.

## 7. CONCLUSION

Multimodal Emotion Recognition is a vibrant and rapidly advancing field. By integrating cues from multiple modalities, MER systems are achieving an increasingly nuanced understanding of human affect. Deep learning, particularly Transformer-based architectures with cross-modal attention, has become the cornerstone of modern MER. However, significant challenges remain. Future research focused on leveraging large pre-trained models, advancing self-supervised and robust learning techniques, and addressing ethical considerations will be key to unlocking the full potential of emotionally intelligent machines.

## REFERENCES


[1] Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, 89(4), 344-350..

[2] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124.

[3] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: from unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125.

[4] Gideon, J., McInroe, A., Brophy, C., Wang, Z., & Fitter, N. T. (2023). A Survey of Affective-Computing-Based Multimodal Emotion Recognition. IEEE Transactions on Affective Computing.

[5] Wu, X., Mou, X., Liu, Y., & Liu, X. (2024). A multimodal emotion recognition algorithm based on speech, text and facial expression. Journal of Northwest University (Natural Science Edition), 54(2), 178-187.

[6] Liu, Z., & Lei, Y. (2024). Design and Experiment of Multi-Modal Sentiment Analysis Model by Fusing Multi-scale Features. Research and Exploration in Laboratory, 43(9), 78-83.

[7] Qiang, Y., Chu, S., & Hu, Y. (2025). MSD-Net: Multimodal Soft Knowledge Distillation for Sentiment Analysis in Real-World Modality Missing Scenarios. Journal of Taiyuan University of Technology.

[8] Yang, R., & Ma, J. (2023). A Feature-Enhanced Multi-modal Emotion Recognition Model Integrating Knowledge and Res-ViT. Data Analysis and Knowledge Discovery, 7(11), 14-25.

[9] Ye, J., Zheng, W., Li, Y., Cai, Y., & Cui, Z. (2017). Multimodal emotion recognition based on deep neural network. Journal of Southeast University (English Edition), 33(4), 444-447.

[10] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. Proceedings of the ACL.

[11] Liu, J., Zhang, P., Liu, Y., Zhang, W., & Fang, J. (2021). Summary of Multi-modal Sentiment Analysis Technology. Journal of Frontiers of Computer Science and Technology, 15(7), 1165-1184

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[13] Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The Munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia.

[14] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124.

[15] Zheng, W. L., & Lu, B. L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development, 7(3), 162-175.

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[17] Li, Y., Wang, X., & Wang, Q. (2025). Research on Multimodal Sentiment Analysis Models for Social Media Based on Multimodal Large Language Models. Information Theory and Practice.

[18] Liu, Y., Yuan, L., Zu, S., Fan, Y., Xie, N., & Yang, Y. (2024). Emotion Recognition Based on Multimodal Physiological Data: A Survey. Journal of University of Electronic Science and Technology of China, 53(5), 720-730.

[19] Ramirez, J. R., Parra, M., & Castellanos, G. (2022). A Survey on Multimodal Fusion for Emotion Recognition Using Physiological Signals. Sensors.

[20] Chen, T., Cai, C., Yuan, X., & Luo, B. (2024). Multimodal emotion recognition method based on multiscale convolution and self-attention feature fusion. Journal of Computer Applications, 44(2), 369-376.

[21] Miao, B., Xu, Y., Zhao, S., & Wang, J. (2024). C-BGA: Multimodal Speech Emotion Recognition Network Combining Contrastive Learning. Computer Engineering and Applications, 60(16), 168-176.

[22] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

[23] Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., & Cambria, E. (2019). DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Proceedings of the AAAI Conference on Artificial Intelligence.

[24] Han, J., Zhang, Z., Poria, S., & Schuller, B. W. (2022). M2FNet: Multi-modal fusion network for emotion recognition in conversation. IEEE Transactions on Affective Computing.

[25] Ma, M., Wu, Z., Wang, S., Zhang, S., & Huang, Y. (2022). Modality-missing knowledge distillation for emotion recognition in conversations. Proceedings of the ACM International Conference on Multimedia.

[26] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., & Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335-359.

[27] Lian, Z., Sun, L., Sun, H., & Lian, Z. (2024). GPT-4v with emotion: a zero-shot benchmark for generalized emotion recognition. Information Fusion.

[28] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., & Sutskever, R. (2021). Learning transferable visual models from natural language supervision. Proceedings of ICML.

[29] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259.

[30] Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. Proceedings of ACL

[31] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Proceedings of ACL.

[32] Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., & Patras, I. (2012). DEAP: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing, 3(1), 18-31.