Advances and Challenges in Multi-Modal Emotion Recognition: A Comprehensive Investigation

Kefan Yao

Wuxi United international School, Wuxi, China

daniel20090828@icloud.com

Abstract. Multi-Modal Emotion Recognition (MER) combines information from two or more modalities—such as speech, facial expressions, video, and physiological signals—to more accurately infer people's emotional states. Previous work shows that relying on a single modality often misses important cues: for instance, audio may capture tone but not facial micro-expressions, video may capture expression but not internal arousal. Recent systems using CNNs, Transformers, or hybrid fusion architectures, applied in driving-safety and healthcare contexts, have improved accuracy significantly, especially when handling missing or noisy modalities. This survey reviews such methods, discusses current challenges like interpretability, modality mismatch, and real-time deployment, and suggests future directions including lightweight models, privacy-preserving fusion, and cross-domain generalization. Furthermore, it highlights the growing importance of explainable and adaptive models that can dynamically adjust to changing environments and user contexts. As MER continues to evolve, these innovations will enhance emotion-aware applications in human-computer interaction, mental health, and intelligent systems.

Keywords: Multi-modal emotion recognition, machine learning, deep learning.

1. Introduction

People have feelings, emotions that may indicate what they are thinking about now. There are two easiest emotions, which are positive and negative. To be specific, there are many other complicated emotions such as anger, execitement or sadness. These emotions establish relationships between people, form the diversity of communication, and include a lot of information. Muti-modal emotion recognition, using two or more emotional expressions, for example, words, videos, and images to analyse emotions. Multi-modal emotion recognition aims to recognise people's feeling right now in various and complicated situations. It can use many methods such as multimodal emotion recognition, machine learning to improve the accuracy of identifying the emotions, erecting a more harmonious relationships between people. Accurately grasping the attitudes of relevant parties can significantly enhance the user experience of Artificial Intelligence (AI) products and has important applications in quality inspection, interaction, risk control, public opinion supervision, and other areas. For instance, in the service industry, understanding whether customers are satisfied with the service can help improve the quality of service, increase customer acquisition and satisfaction; in e-commerce, users' preferences for a certain product and its competitors are highly valuable information; in human-computer interaction, knowing the emotional state of the conversation partner can help us adopt appropriate language expressions in a timely manner to show comfort and understanding, thereby enhancing the interaction experience. There are numerous scenarios in real-world applications that require information on emotions or attitudes, and in such cases, sentiment analysis algorithms provide a way to extract these key pieces of information.

The previous studies have contributed to the multi-modal emotion recognition, such as emotion features extraction. Emotional feature extraction attempts to obtain the emotional information contained in different ways of human emotional expression. Different information modalities have different ways of obtaining emotional features, detection and intervention of depression and emotional disorders. Researchers are attempting to use algorithms to extract patients' language expressions, facial expressions and voice features to automatically assess their depressive states etc.

This paper intends to provide a survey of current methods and applications in MER. After this introduction, Section 2 (Method) reviews different methodological paradigms (e.g. CNN-based fusion, Transformer/attention models) and application domains (such as driving/automotive and healthcare/patient monitoring). Section 3 (Discussion) elaborates on persistent challenges (interpretability, modality incompleteness, cross-domain issues, etc.) and future research directions. Finally, Section 4 (Conclusion) wraps up the findings, main contributions, experimental results, limitations, and plans for future work.

2. Method

2.1. Model-Style Approaches

2.1.1 CNN / Early Deep Fusion Methods

In earlier MER work, researchers often used CNNs or other deep nets separately for each modality, then fused later. For instance, Tzirakis et al. built an end-to-end model combining audio and video: a CNN processes audio, a ResNet-50 processes visuals, and then a dual-stream LSTM fuses and models the temporal dynamics. That system outperformed handcrafted-feature baselines on RECOLA / AVEC datasets [1]. Another method is to better align different modalities before fusing. Liu et al. used Deep Canonical Correlation Analysis (DCCA) to map EEG and face features into a shared latent space, so that when fused the representations are more coherent and robust under noise [2]. One more example: Zhao et al. used wav2vec 2.0 for speech embeddings and BERT for text embeddings, then performed multi-level fusion (co-attention early fusion + late fusion) to boost MER. They showed improved performance on IEMOCAP compared to baselines [3].

2.1.2 Transformer / Attention-Based Methods

Transformer models are gaining lots of popularity in MER because they can model interactions between modalities via cross-attention, and capture long dependencies. Hazmoune et al. provide a taxonomy of Transformer-based MER methods, showing how different attention fusion layers work in MER pipelines [4]. One concrete model is Joyful. It mixes global contextual embeddings with each modality's specific features and applies graph contrastive learning so that the fused representations are more separable. It achieves state-of-the-art performance on several conversational MER benchmarks [5]. Another example is MemoCMT, which uses cross-modal Transformer blocks to strengthen interactions between modalities. It showed strong results on multimodal emotion datasets across multiple emotion categories [6]. In 2024, professor Shayaninasab and Babaali used pre-trained Transformers on each modality (text, audio, video), then try different fusion (feature-level, decision-level) and get a solid accuracy (around 75.4%) on IEMOCAP [7].

2.2. Driving / Automotive Applications

In vehicles, recognizing driver emotions like fatigue, anger, or stress can improve safety systems. Espino-Salinas et al. fused driver motion data (steering, acceleration) with facial geometry images. They used CNNs to get visual embeddings, classical machine learning on motion features, and then fused with a simple network. In simulations, they got 96% accuracy [8]. Xiang et al. constructed a dataset combining in-car video, physiological signals, and CAN bus data. Their experiments demonstrated that fusing all modalities outperforms any single modality in classifying driver emotions [9]. Luan et al. proposed MHLT (Multi-Head Layer Transformer) for real-time driving. It adaptively weights audio and video via attention, and outputs both emotion categories and intensity scores [10]. A recent preprint introduced a federated learning framework for in-vehicle MER, where each car trains locally and just shares model updates (no raw data). This helps preserve privacy while improving system performance across cars [11].

2.3. Healthcare / Patient Monitoring Application

In clinical or hospital settings, MER can help monitor patient stress, mood changes, pain, or emotional states. Mutawa et al. designed a real-time MER system combining EEG signals and facial optical flow. They fused these heterogeneous signals via machine learning classifiers to continuously monitor patient emotions [12]. He et al. surveyed BCI-based MER combining EEG, electrodermal activity (EDA), and audio-visual data. They discuss how noise, modality mismatch, and missing data pose challenges, but the potential for patient monitoring is strong [13]. Chen et al. built a deep multimodal fusion framework that includes audio, video, and text modalities. They applied it to clinician–patient interactions and showed that the model can detect subtle emotional shifts during therapy sessions [14].

3. Discussion

3.1. Persistent Challenges

Interpretability. Many MER systems, especially Transformer or contrastive learning models, act like black boxes. It's hard to know why a model gave a certain emotion prediction, or which modality contributed most. In areas like medicine or autonomous driving, this lack of transparency is a big barrier to trust and acceptance.

Modality Incompleteness and Synchronization. In real life, modalities often drop out or degrade: facial video might be blurred, audio might be noisy, sensors might fail. Aligning modalities in time is also hard. Some works try conditional attention or modality dropout to deal with missing data, but they often fail under severe loss.

Cross-Domain Generalization. Models trained in controlled lab settings often don't generalize well to real-world deployments. Differences in environment, devices, culture, lighting, noise, and participants cause domain shift. A model might work well on one dataset but fail in a new scenario.

Resource Constraints & Real-Time Needs. Many MER methods (especially Transformer-based fusion) are heavy in computation and memory. Deploying them on vehicles, wearable devices, or hospital monitoring equipment—where latency and hardware are constrained—is hard. You need lightweight, efficient models.

Privacy & Security. MER deals with extremely sensitive data: facial images, voice, EEG, etc. Transmitting or storing raw data can cause privacy leaks. Techniques like differential privacy, encryption, or federated learning could help, but integrating them without hurting MER performance is not easy.

3.2. Future Research Direction

Adaptive & Explainable Fusion. Future models should dynamically adjust how they fuse modalities depending on signal quality, and provide human-readable explanations (e.g. attention heatmaps, modality weights). That would help in trusting the system and diagnosing errors.

Lightweight, Edge-Deployable Models. Using techniques like model distillation, quantization, or conditional computation (skipping parts of the network dynamically) can yield smaller models suitable for real-time use on devices or vehicles.

Self-Supervised & Contrastive Pretraining. Since emotion-labelled data is scarce, using large amounts of unlabeled multimodal data with self-supervised or contrastive learning can help build robust features. The Joyful model is a good example of combining contrastive learning with modality fusion [5].

Federated & Privacy-Preserving Learning. To protect privacy and handle distributed data (e.g. across vehicles or hospitals), federated MER training is promising. Each node trains locally and only shares model updates. This reduces risk of raw data leaks.

Generalization vs Personalization. A model needs to generalize to new settings (cross-domain), but also adapt to individual users. Future systems may strike a balance: a base general model plus an online personalization module that fine-tunes per user.

More Modalities & Contextual Inputs. Beyond audio, visual, and physiological signals, future MER systems might incorporate environment sensors (light, noise, temperature), task contexts, behavior logs, and time-series context. Also, moving from static recognition to modeling emotional change dynamics, transitions, or causality would be a next step.

4. Conclusion

To close the loop, recall that in the introduction this paper emphasized how accurate detection of emotions across modalities is key for improving human-computer interaction, healthcare, and safety. This survey's main contribution is systematically comparing how MER systems deploy CNN-based fusion, Transformer-based attention, and hybrid models across driving and clinical monitoring scenarios. From experiments reported in the reviewed papers, systems that combine modalities generally outperform single-modality models—for example, decision-level fusion in Almulla achieved ~80 % accuracy versus much lower rates in individual modalities, and He et al. showed that aBCI systems with EEG + behavior or peripheral signals yield better recognition than EEG alone. However, there are limitations: existing studies often use constrained datasets, have insufficient real-world deployment tests, and struggle with missing or noisy modality data. As a future plan, more work is needed on model robustness under challenging data conditions, lightweight / deployable architectures, and privacy-preserving methods that do not compromise performance.

References

[1] Tzirakis P, Trigeorgis G, Nicolaou M A, et al. End-to-End Multimodal Emotion Recognition using Deep Neural Networks. arXiv preprint arXiv:1704.08619, 2017.

[2] Liu W, Qiu J-L, Zheng W-L, Lu B-L. Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis. arXiv preprint arXiv:1908.05349, 2019.

[3] Hazmoune S, et al. Using Transformers for Multimodal Emotion Recognition. Engineering Applications of Artificial Intelligence, 2024.

[4] Li D, Wang Y, Funakoshi K, Okumura M. Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimodal Emotion Recognition. arXiv preprint arXiv:2311.11009, 2023.

[5] Nguyen T, et al. MemoCMT: Cross-modal Transformer Fusion for Emotion Recognition. Micromachines, 2024.

[6] Espino-Salinas C H, et al. Multimodal Driver Emotion Recognition Using Motor Activity and Facial Geometry. Frontiers in Artificial Intelligence, 2024.

[7] Xiang Z, et al. A Multimodal Driver Emotion Dataset and Baseline Models. Engineering Applications of Artificial Intelligence, 2024.

[8] Luan X, Wen Q, Hang B. Intelligent Driver Emotion Recognition with Model-Level Multimodal Fusion. Frontiers in Physics, 2025.

[9] Zhang K, et al. Real-Time Emotion Recognition via Multimodal Federated Learning. arXiv preprint, 2025.

[10] Mutawa A M, et al. Real-Time Multimodal Patient Emotion Recognition System. Biomedical Signal Processing and Control, 2024.

[11] He Z, et al. Advances in Multimodal Emotion Recognition Based on Brain–Computer Interfaces. Frontiers in Neuroscience, 2020.

[12] Chen L, et al. A Multimodal Emotion Recognition System for Patient–Clinician Interactions. Journal of Emotion Recognition, 2024.

[13] He Z, Li Z, Yang F, Wang L, Li J, Zhou C, Pan J. Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain Sci.* 2020;10(10):687.

[14] Almulla MA. A multimodal emotion recognition system using deep convolution neural networks. *J Eng Res.* 2024;13(4).