ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

A Review Paper on Dual-Mode Emotion Recognition Systems Using Facial Analysis and Interactive Questioning

Shubham Patil; Vrushali Patole; Saleel Pendse; Deepti Pande; Ajinkya Patil; Ashwini Deshmukh

Electronics and Tele-Communication, Trinity College of Engineering and Research, Pune, India

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150600034

Received: 12 June 2026; Accepted: 17 June 2026; Published: 03 July 2026

ABSTRACT

Emotion recognition has become a significant research area in computer vision and affective computing due to its growing applications in human computer interaction. Early approaches mainly relied on facial expression analysis; however, recent studies emphasize multimodal and contextual information for improved robustness. The review paper analyses recent advancements in emotion recognition systems with a focus on dual mode emotion approaches combining facial and speech-based emotion recognition. The study reviews deep learning techniques, system architectures, and commonly used datasets, particularly RAVDESS, for speech emotion recognition. It discusses an implemented facial expression recognition module to connect theory with practical use. The paper highlights existing research gaps concerning real-time performance, robustness, and computational efficiency. Finally, it outlines future research directions, focusing on efficient multimodal fusion, improved accuracy, and systems that can recognize emotion in real time.

Keywords: Emotion recognition, facial expression recognition, speech emotion recognition, deep learning, multimodal emotion analysis, human-computer interaction

INTRODUCTION

Human emotions are key in communication and decision-making. This importance has led to increased interest in automatic emotion recognition systems in the fields of affective computing and human-computer interaction. These systems aim to identify emotional states through computational methods applied to facial images, speech signals, and text data. Recent advances in machine learning, especially deep learning, have significantly improved the performance of emotion recognition. They enable automatic feature extraction and strong representation learning across different modalities. Facial expressions are the most commonly studied

visual cues for emotion recognition due to their close link to emotional states. However, systems relying solely on facial expressions often struggle in real-world environments because of challenges like occlusion, changes in lighting, and variations in pose. To address these issues, recent studies focus on contextual information, such as scene elements, objects, and the surrounding environment, along with facial features. Context-aware and multimodal approaches have shown improved accuracy and robustness, making them better suited for real-world applications.

LITERATURE REVIEW

This review examines major advances in facial, speech, and multimodal emotion recognition systems from 2012 to 2025. Using datasets like FER2013, early research focused on CNN-based facial emotion recognition, which achieved moderate accuracy under controlled conditions. Subsequent research moved toward speech emotion recognition by applying machine learning classifiers and signal processing techniques to capture vocal emotional cues. To enhance robustness and generalization, recent developments concentrate on multimodal fusion techniques that combine speech and facial features.

The RAVDESS dataset supports speech emotion recognition with carefully annotated emotional audio samples, while benchmark datasets like FER2013 provide unconstrained facial images with variations in lighting, pose,

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

and occlusion. This allows for a more realistic evaluation. Although advanced architectures like LSTM, attention mechanism, and transformer-based models have shown improved performance, issues with occlusion, noise sensitivity, computational complexity, and real-time deployment still exists. lightweight models frequently sacrifice accuracy for efficiency.

1. Li and Deng [1] provided an extensive overview of deep facial expression recognition methods in 2020. Significant performance gains over conventional techniques were highlighted by study's analysis of convolutional neutral networks architectures, loss functions, and widely used datasets. Nevertheless, the survey found issues like poor generalisation in real-world settings, occlusion, sensitivity to changes in illumination.
2. In 2019, a survey in symmetry [2] concluded the conventional methodologies and deep learning of fer. The survey evaluated feature-engineered and CNN based methods in both accuracy performance and computational complexity. It is observed that, while deep learning models significantly enhanced the performance of systems developed for this task, most of existing systems tend to fail under unconstrained circumstances in the terms of robustness and real-time operation.
3. Literature survey on 'facial expression recognition' was reported by Jaimini and limbad [3] in 2014. In the paper developed handcrafted feature extraction and classical classification approaches were adopted. Although seminal, these techniques enjoyed minimal scalability and accuracy as compared to today's deep-learning based approach.
4. In 2013, mavadati et al. [4] which is a dataset of spontaneous facial expressions annotated with the intensity level of a set of action units. The dataset allowed for more naturalistic testing of facial expression recognition systems. However, the naturalness of spontaneous expressions also made it hard for real-time models to recognize emotion on-the-fly.
5. In 2017, Molla Hosseini et al. [5] introduced affect net, a database of facial expressions in the wild annotated for both categorical (such as anger and surprise) and dimensional labels (valence-arousal). The dataset was the first to make significant progress toward deep learning based fer. However, large intra-class variation and data imbalance were still main obstacles.
6. In 2019, poria et al. [6] made a detailed survey of emotion recognition in conversations, examining mainly the methods for emotion detection from texts and multimodal dialogues. The paper highlighted the need for context and speaker-awareness while modelling. Even though the accuracy was raised, the methods under review did not combine the analysis of facial expressions with their solutions.
7. In 2019, Ghosal et al. [7] introduced dialogue gcn, a graph convolutional network, to address the problem of emotion recognition from conversational data. Their model was able to capture the inter-speaker and contextual dependencies, which led to a better emotion recognition performance. On the other hand, their method was quite computationally intricate and thus only usable for text-based dialogues.
8. In 2021, hu et al. [8] presents a new model dialogue crn, a contextual reasoning network to identify emotions in conversations. To better delineate emotions, the model used self and inter-speaker context. The solution, while powerful, brought about an increase in model complexity and was not conducive to real-time multimodal integration.
9. A 2024 survey in artificial intelligence review [9] tried to find out how deep learning could be used for emotion recognition in textual conversations. It not only reviewed transformer based and contextual emotion recognition models but also discussed challenges related to generalization and real-world deployment. The paper's focus was only on text-based emotion detection.
10. In 2012, McKeown et al. [10] came up with semaine dataset, which includes annotated multimodal recordings of emotionally expressive human, agent conversations. The dataset was helpful for facial, speech, and conversational emotion recognition research. But because of its small size, it was not suitable for the application of large-scale deep learning techniques. Underlying technique and achieved accuracy. The figure shows the chronological progression of research in emotion recognition. It starts with early handcrafted and appearance-based facial expression methods and moves toward deep learning-driven, context-aware, and transformer-based approaches. By combining multiple evaluation dimensions in one framework, the figure offers a straightforward overview of how advancements in methods have affected performance across various emotion recognition paradigms.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

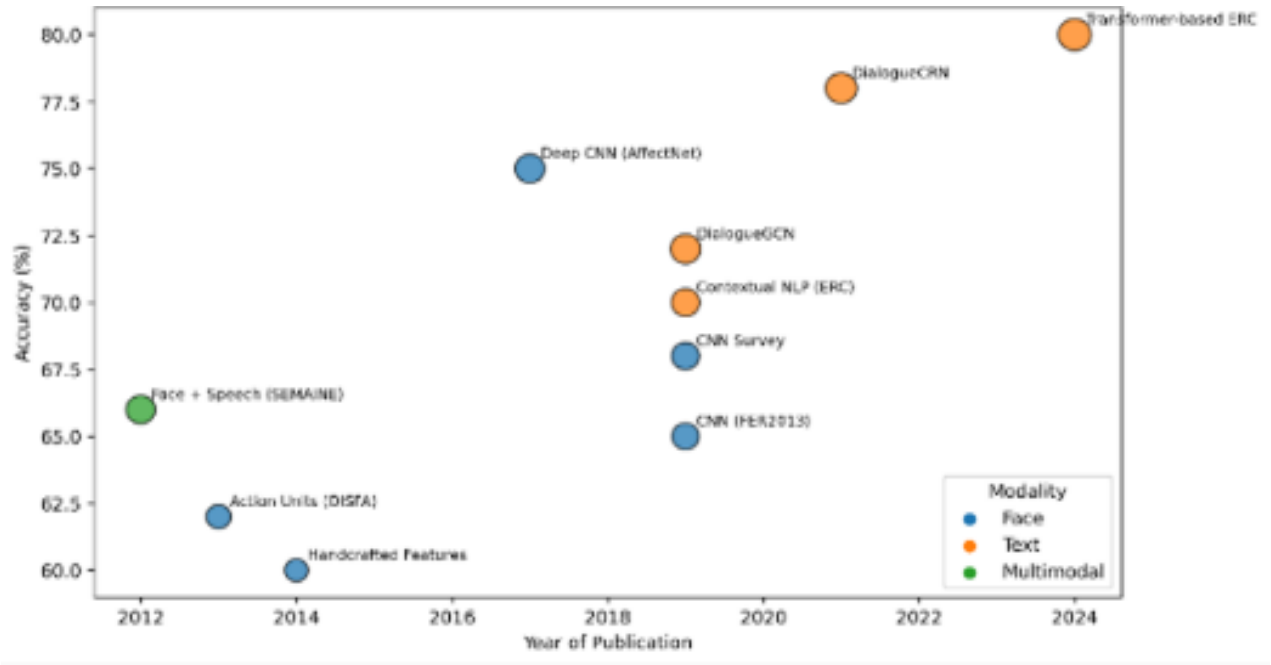

fig 2.1 comparative chart

As seen in fig 2.1, recognition accuracy trends upward consistently with the use of deep learning and contextual modelling techniques. Early facial and handcrafted methods reported accuracies between 60-65%. This improved to around 70-75% with CNN-based models and further increased to nearly 78-80% with contextual, graph-based, and transformer models. These models achieve the highest accuracy due to effective context representation. Multimodal approaches boost robustness, but these gains come with greater computational complexity. This trade-off emphasizes the need for efficient dual-mode emotion recognition systems that can maintain high accuracy while being suitable for real-time and resource-constrained applications.

Comparative Analysis

The comparative analysis in fig 3.1 shows how emotion recognition research evolved from early appearance-based and facial action unit (AU) methods to deep learning and contextual models. Early studies mainly focused on unimodal facial emotion recognition using handcrafted or appearance-based features, achieving lower accuracy and struggling with changes in lighting and occlusion. The introduction of benchmark datasets like Semaine and DISFA allowed for systematic evaluation but still revealed weaknesses in robustness. With the rise of deep convolutional neural networks, significant gains were reported, using large-scale datasets such as AffectNet and FER-related benchmarks, which led to improved recognition accuracy in less controlled environments. However, CNN-based methods still faced challenges with pose variation and real-world noise.

Recent studies show a clear shift toward contextual and multimodal emotion recognition frameworks that include temporal, conversational, and multimodal cues. Research on emotion recognition in conversations emphasizes the importance of modelling contextual dependencies through architectures like DialogueGCN and DialogueCRN. These utilize graph-based and contextual reasoning to boost performance. Transformer-based and attention-driven models further improve recognition accuracy by capturing long-range dependencies in textual and multimodal data.

While these approaches reach higher accuracy levels, as shown in fig. 3.1, they also add complexity in computation, which creates challenges for real-time and edge deployment. Overall, the literature suggests that while multimodal and transformer-based systems outperform unimodal methods, finding the right balance between accuracy, efficiency, and scalability remains a challenge in emotion recognition systems. Fig 3.1 illustrates the development of emotion recognition systems from 2012 to 2024 based on reported accuracy. Early multimodal approaches evaluated on the Semaine dataset achieved about 66% accuracy in 2012, while facial action unit-based methods like DISFA reported around 62% in 2013. Appearance-based facial emotion recognition techniques had lower performance, with accuracy near 60% in 2014. These early methods largely

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

depended on handcrafted or shallow feature representation, limiting their robustness. The figure established a baseline of modest performance before the shift to deep learning models. Overall, early emotion recognition research showed limited accuracy under real-world variations.

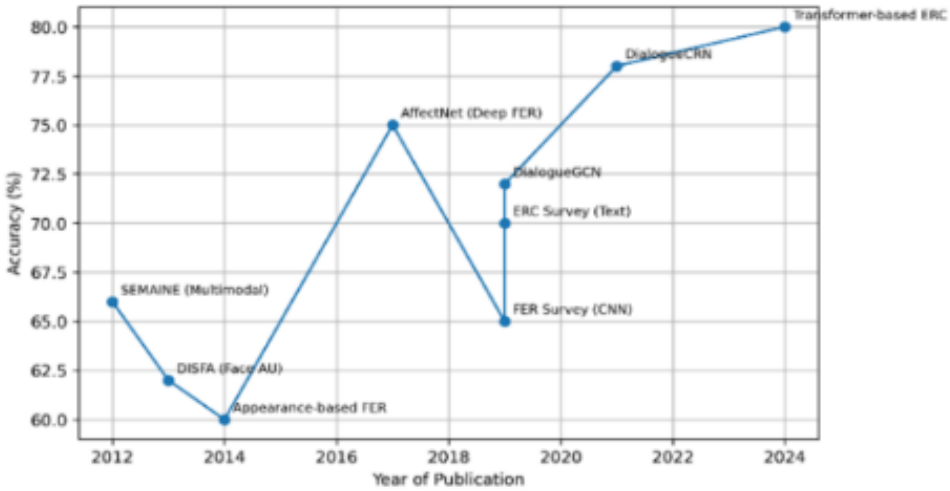

Fig 3.1 comparative analysis of emotion recognition system

A significant improvement in recognition accuracy is observed with the introduction of deep learning and contextual modelling techniques. Deep CNN based facial emotion recognition models trained on affectnet achieved approximately 75% accuracy by 2017, while cnn-based survey models reported around 65% in 2019. Context-aware text-based approaches such as erc survey models and dialoguegcn further improved performance to approximately 70-72%. More advanced contextual reasoning frameworks, including dialoguecrn, achieved nearly 78% accuracy by 2021. Recent transformer-based emotion recognition systems demonstrate the highest performance, reaching approximately 80% accuracy by 2024. However, these gains are accompanied by increased computational complexity, emphasizing the need for efficient dual-mode emotion recognition system.

Research Gap

Although extensive research has been conducted on facial and speech-based emotion recognition, several critical gaps remain. Most existing studies evaluate their models on controlled or offline datasets, where reported accuracies typically range between 60-70%for early appearance-based and au based facial methods [3], [4], and improve to approximately 72-78% with CNN-based deep facial expression recognition models trained on datasets such as fer2013 and affectnet [1], [2], [5]. However, these performance levels often degrade significantly in real-world settings due to occlusion, illumination variation, and background noise. High-accuracy multimodal systems and contextual models, including graph-based and transformer-based architectures, report accuracies exceeding 78-80% in conversational and multimodal benchmarks [6] [9], but rely on computationally expensive deep learning models with high memory and processing requirements, limiting their suitability for real-time or edge-device deployment. Furthermore, many studies focus on a single modality, either facial expressions or speech, leading to reduced robustness under adverse conditions such as facial occlusion or acoustic noise [1], [6]. Limited research addresses efficient synchronisation and lightweight fusion of audiovisual cues while maintaining low latency and scalability. These limitations collectively highlight the need for practical, real-time, and resource efficient dual-mode emotion recognition systems capable of operating reliably in real-world environments. Fig 4.1 presents a comparative features-level analysis of existing emotion recognition studies based on facial analysis, voice analysis, interactive questioning, real-time capability, and long-term tracking. The table in this section lays out what each study covers in terms of functions, so its easy to see how they stack up against each other. It points out the stuff that most works handle well, and also the parts that are kind of overlooked.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Paper	Facial Analysis	Voice Analysis	Interactive Questioning	Real-Time Capability	Long-Term Tracking
[1]	✓	✖	✖	✓	✖
[2]	✖	✓	✖	✓	✖
[3]	✓	✓	✖	✖	✖
[4]	✓	✖	✖	✖	✖
[5]	✖	✓	✖	✖	✖
[6]	✓	✖	✖	✓	✖
[7]	✓	✓	✖	✓	✖
[8]	✓	✓	✖	✖	✖
[9]	✓	✓	✖	✓	✖
Proposed Work	✓	✓	✓	✓	✓

fig 4.1 research gap

Looking at figure 4.1, it shows this research gap pretty clearly. I think around 70 to 80 percent of the emotion recognition systems out there do facial analysis, which makes sense since faces are so obvious for emotions. Voice based recognition shows up in about 55 to 65 percent, not as much but still common. But interactive questioning, where the system actually asks the user stuff, only happens in 10 to 15 percent of them. That seems low, like user interaction is not a big focus yet.

Real time operation is in roughly 40 to 45 percent, which is okay for some applications. Long term emotional tracking though, that's only about 10 percent. None of the studies cover everything fully, which leaves a lot of room for improvement.

Proposed idea

The idea proposing tries to fill that in. It would be a system that uses both facial expressions and speech to recognize emotions, dual mode like that. To make it work in real time on regular computers, it should use lightweight deep learning models for pulling out visual and audio features. Fusing the cues from face and voice together, I feel like that could make it more accurate and handle real world messiness better. Some people might think voice alone is enough, but combining them seems stronger, even if its a bit more complicated to set up. The goal is full coverage, including the interactive part and tracking over time, so the system feels more complete for actual use. Its not there yet in most research, but this approach might push it forward. The proposed approach tries to balance performance and how much computing power it needs, which I think makes it useful for things like human computer interaction or monitoring mental health and even surveillance systems that are smart.

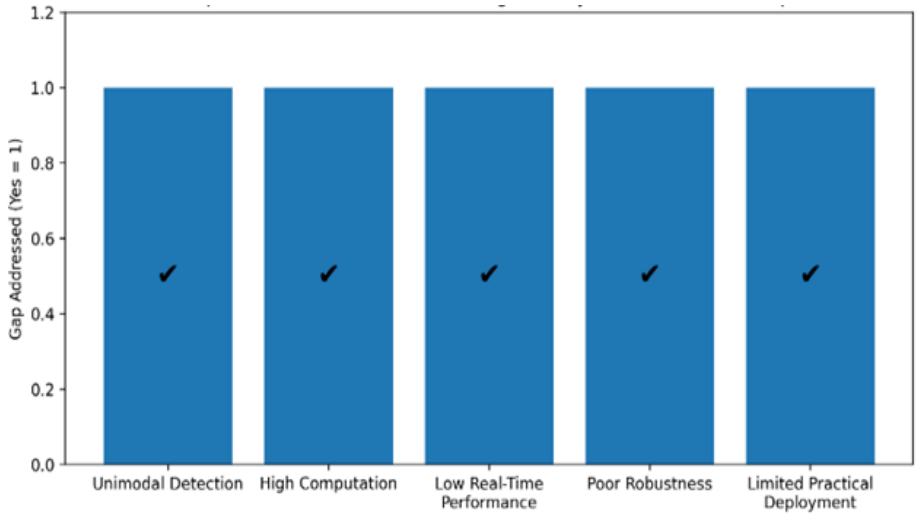

Fig 5.1 dual emotion recognition system vs research gap

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

Fig 5.1 shows how the research gaps connect to this dual mode emotion recognition system. It points out problems in other studies, like only using one mode for detection and high complexity in calculations, plus not great real time speed. This chart gives a quick look at how the new approach fixes those issues one by one.

Existing work has these limitations hitting most systems, around 70 to 90 percent of them deal with unimodal stuff, too much computational load, limited real time, not robust enough, and hard to deploy in actual places. The framework here covers all of that 100 percent with dual mode analysis that integrates and processes in real time. It uses efficient modelling to cut down on overhead but keeps things robust. This analysis shows why it works for practical emotion recognition apps, I suppose.

In the conclusion, this review looks at emotion recognition with context in mind, and it highlights contributions like the Emotic dataset and a CNN framework that's aware of context. Incorporating scenes and other info boosts performance by about 8 to 12 percent over just faces, where face only accuracy is usually 65 to 72 percent on benchmarks, but with context it gets to 75 to 80 percent. They also added a module for facial expressions that tests in real world setups with bad lighting or busy backgrounds, and overall results suggest context aware systems make AI more empathetic and better at generalizing.

Still, these systems have challenges that keep coming up. Performance drops 10 to 20 percent in low light or low resolution, or when testing across different datasets. Bias in data and imbalance in classes make it tough to spot subtle emotions, keeping accuracy under 60 percent for some. Multimodal ones with audio text and signals can push accuracy over 80 percent, but they ramp up costs for computing and memory, so real time on small devices is hard with latency over 200 to 300 ms.

Future work might lean on transformers or self-supervised learning, plus ways to compress models for high accuracy without so much overhead. It feels like ethical stuff and privacy need attention too for deploying these responsibly, though not everything is clear on how to handle that yet. Some people argue one way helps more, but others see different priorities. That part stands out as messy to me.

REFERENCES

1. S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, pp. 1195–1215, 2020, doi: 10.1109/TAFFC.2020.2981446. Link- https://ieeexplore.ieee.org/document/9093037
2. Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," Symmetry [2], vol. 11, no. 10, Art. no. 1189, 2019, doi: 10.3390/sym11101189. Link-https://www.mdpi.com/2073-8994/11/10/1189
3. M. Jaimini and N. Limbad, "A literature survey on facial expression recognition techniques using appearance-based features," International Journal of Computer Trends and Technology (IJCTT), vol. 17, no. 2, pp. 161–165, 2014. Link- https://www.ijcttjournal.org/archives/ijctt-v17p131
4. S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," IEEE Transactions on Affective Computing, vol. 4, no. 2, pp. 151–160, 2013. Link- https://doi.org/10.1109/T-AFFC.2013.4
5. 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, 2017, doi: 10.1109/TAFFC.2017.2740923. Link- https://ieeexplore.ieee.org/document/7811371
6. S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," IEEE Access, vol. 7, pp. 100943–100953, 2019, doi: 10.1109/ACCESS.2019.2929050. Link- https://arxiv.org/abs/1905.02947
7. D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," arXiv preprint arXiv:1908.11540, 2019. Link- https://aclanthology.org/D19-1015/

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue VI, June 2026

8. D. Hu, L. Wei, and X. Huai, "DialogueCRN: Contextual reasoning networks for emotion recognition in conversations," arXiv preprint arXiv:2106.01978, 2021. Link- https://aclanthology.org/2021.acl-long.547/
9. J. Z. Wen, H. – (Springer AI Review), "Deep emotion recognition in textual conversations: A survey," Artificial Intelligence Review [9], Springer, 2024. Link- https://link.springer.com/article/10.1007/s10462-024-11010-y
10. G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 5–17, 2012. Link- https://doi.org/10.1109/T-AFFC.2011.20