Journal of HyoJeong Academia

Article

Emotion Recognition Technique: A Comprehensive Review

Qurat Ul Ain Aisha[1], Ahhyeon Lee[1], Byung-Gyu Kim[1,2,*] and Jiwoo Kang[1,2]

[1] Department of IT Engineering, Sookmyung Women's University, Seoul, 04310, Korea
[2] Artificial Intelligence Innovation Research Center, Sookmyung Women's University
*Correspondence: bg.kim@sookmyung.ac.kr

Abstract: Emotion recognition is rapidly advancing within the realm of artificial intelligence (AI), spurred by progress in deep learning, multimodal data handling, and broader availability of large datasets. This paper offers an in-depth overview of emotion recognition strategies, focusing on biometric signals (e.g., audio, visual, physiological, and brain signals), conventional machine learning methods, and the latest deep learning architectures, including Transformers, two-dimensional Convolution Neural Networks (CNNs) 2D CNNs, and three-dimensional CNNs (3D CNNs). We further examine multimodal systems and the role of Large Language Models (LLMs) in merging textual, audio, and video information for more precise emotion assessments. Key challenges such as real-time implementation, data biases, and cultural diversity are also highlighted. Ethical issues related to privacy and the potential misuse of emotion recognition technologies receive attention, as does a discussion of emerging applications in healthcare, human-computer interaction (HCI), mental health monitoring, and education. In sum, this review aims to contribute to the scholarly conversation around evolving emotion recognition methodologies and their applications in practical systems.

Keywords: Emotion recognition, Transformers, Biometric signals, Multimodal approach, Large Language Models, Sentiment Analysis, Human-Computer Interaction

Citation: Qurat Ul Ain Aisha, Ahhyeon Lee, Byung-Gyu Kim, Jiwoo Kang. Emotion Recognition Technique: A Comprehensive Review. *JHJA 3(1)*, **2025**.

Received: February 17, 2025
Accepted: March 17, 2025
Published: March 29, 2025

Publisher's Note: JHJA stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Emotions are central to human interaction, influencing our thought processes, choices, and social connections. Automated emotion recognition, which classifies or detects emotional states algorithmically, has garnered growing interest in fields like psychology, AI, computer vision, speech processing, and computational neuroscience. Its adoption has the potential to transform multiple areas, such as HCI, mental health surveillance, e-learning, robotics, and entertainment.

Earlier research in emotion recognition typically relied on facial expressions, voice signals, and physiological measurements through handcrafted features (e.g., facial action units, Mel-Frequency Cepstral Coefficients (MFCCs)). However, the emergence of deep learning—incorporating CNNs, Recurrent Neural Networks (RNNs), and Transformer-based methods—has greatly increased accuracy and facilitated modeling of complex emotional signals. Moreover, the incorporation of multiple modalities (e.g., text, video, audio, EEG) has delivered a more comprehensive analysis of affective states, significantly boosting performance.

This paper surveys emotion recognition technology, tracing developments from feature-based systems to modern deep learning and multimodal solutions. In particular, it underscores the role of LLMs in refining text-based and multimodal emotion recognition. Key issues like data imbalance, cross-cultural differences, user privacy, and real-time.

2. Emotion Recognition Techniques

Emotion recognition systems can be distinguished by input modality: audio, visual, physiological (EEG, GSR, ECG, etc.), or textual data. Increasingly, these modalities are integrated for enhanced performance. Below, we provide a detailed look at different modalities and the impact of novel deep learning methods.

2.1. Audio-Based Emotion Recognition

2.1.1. Traditional Methods

Handcrafted Features: Classic strategies for speech emotion recognition (SER) used features like MFCCs, pitch, and energy [1]. Classifiers such as SVMs and HMMs were often chosen to capture patterns in these traditional features [1].

Prosodic Features: Rhythm, stress, and intonation all reflect emotional states (e.g., intense emotional states may manifest as higher pitch or amplitude).

2.1.2. Deep Learning Advancements

1D CNNs and RNNs: Li et al. [2] employed RNN-based SER, illustrating advantages in modeling temporal variations in speech. In general, CNNs and RNNs (e.g., LSTMs or GRUs) have outperformed hand-engineered methods [3].

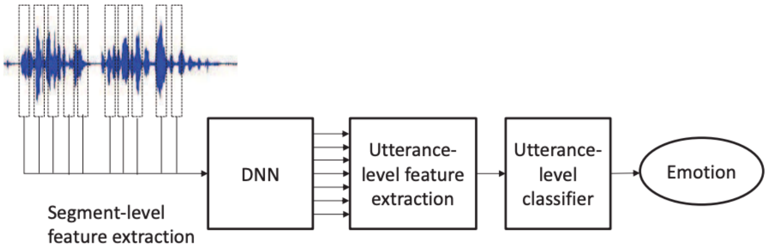

DNN
Utterance-level feature extraction
Utterance-level classifier
Emotion
Segment-level feature extraction

Figure 1. Deep Neural Network for speech emotion recognition [3].

Figure 1 illustrates the architecture of a deep neural network designed for speech emotion recognition. It highlights the process of segmenting audio signals, extracting spectral features, and employing layers of convolutional and recurrent neural networks to capture temporal dependencies, ultimately leading to emotion classification.

2.1.3. Self-Supervised Transformer Models:

Models like Wav2Vec 2.0 [4] and HuBERT [5] have set new benchmarks in speech representations, particularly in scenarios with limited labeled data.

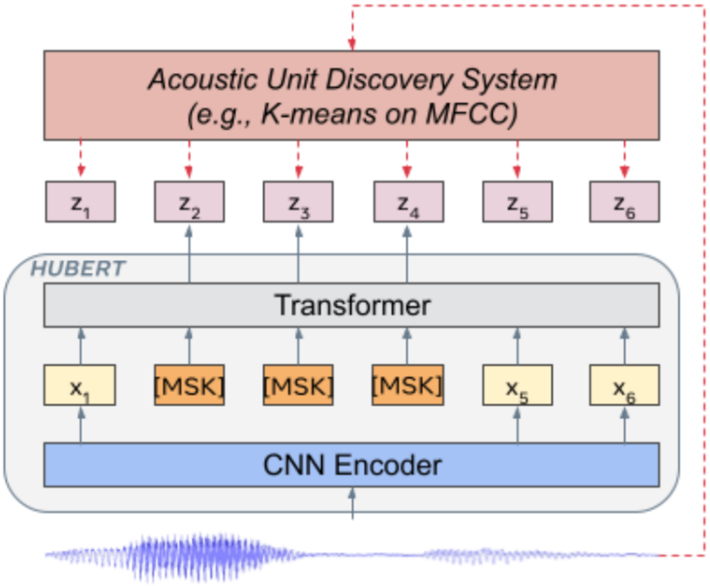


Figure 2. The HuBERT approach [5].

Figure 2 presents the HuBERT pipeline for self-supervised speech representation learning. It shows how raw audio is first encoded via CNN, followed by a Transformer network that performs masked prediction using pseudo-labels generated through clustering. This iterative process refines the model's understanding of speech features, even in low-resource settings.

2.1.4. Datasets

Popular datasets in speech emotion recognition include (Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS**)**, Interactive Emotional Dyadic Motion Capture (IEMOCAP**)**, and (Berlin Database of Emotional Speech (EmoDB**)**, each containing diverse emotional categories and varying levels of speaker diversity [6].

2.2. Visual-Based Emotion Recognition

Facial expression recognition (FER) remains a core topic due to the significance of facial cues. Major breakthroughs in computer vision have pushed FER systems to higher accuracy levels.

2.2.1. CNN-Based Approaches

2D CNNs: Used to learn spatial characteristics from static images (e.g., photographs or single video frames) [7].
3D CNNs: Fuse spatial and temporal dimensions by mapping multiple consecutive frames, enabling detection of micro-expressions over short durations [8].

2.2.2. Vision Transformers (ViTs)

Self-Attention Mechanisms: ViTs directly apply Transformer architectures to image patches, capturing long-range relations [9].

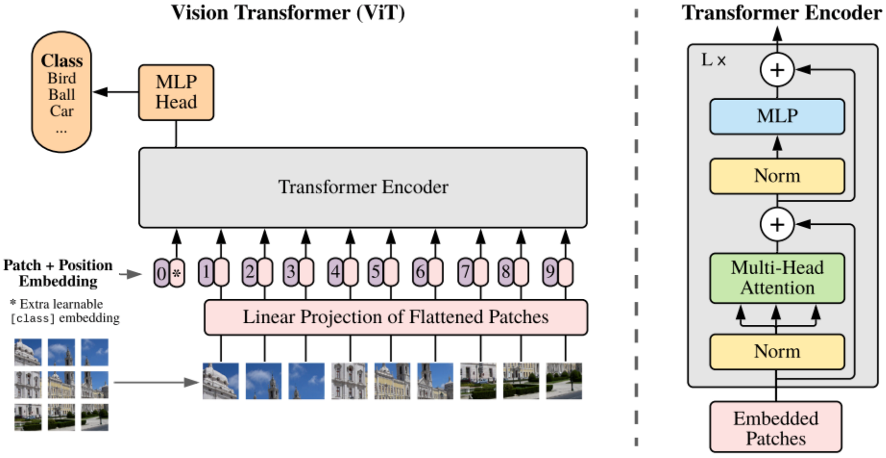


Figure 3. Model overview for Self-Attention Vision Transformer Mechanism [9].

This figure details the Vision Transformer (ViT) architecture. It depicts the process of splitting an image into patches, embedding these patches with positional information, and processing them through self-attention layers. The model's ability to capture long-range dependencies is emphasized, making it effective for tasks such as facial emotion recognition.

Data Requirements and Regularization: Large datasets or self-supervised pretraining (Self Distillation with no Labels (DINO [10]), and Masked Auto Encoder (MAE [11]) can mitigate overfitting by generating generalized representations.

2.2.3. GANs for Data Augmentation

Data Synthesis**:** GANs can generate diverse facial images under different emotional states [12]. Conditional GANs are especially helpful for balancing underrepresented categories.

2.2.4. Datasets and Benchmarks

Well-known visual emotion datasets include Extended Cohn-Kanade Dataset (CK+), AffectNet, Facial Expression Recognition 2013 (FER2013), and Real World Affective Database (RAF-DB). These datasets vary in resolution, complexity, and the number of annotated emotion classes (e.g., basic emotions vs. compound emotional states) [13].

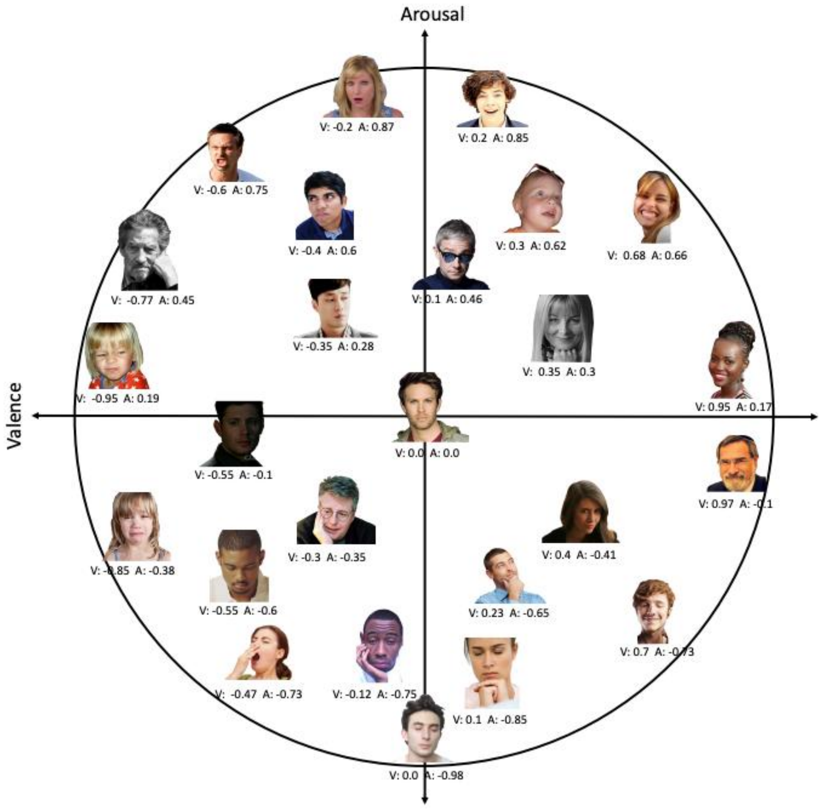


Figure 4. Human emotion states [13].

Figure 4 maps human emotions on a valence-arousal plane, providing a visual representation of how different emotions are distributed based on their positivity/negativity and intensity. It serves as a useful reference for understanding the underlying emotional dimensions used in many recognition models.

2.3. Physiological and Brain Signal-Based Emotion Recognition

Physiological signals, including Electroencephalography (EEG), Electrocardiography (ECG), Galvanic Skin Response (GSR), and Electromyography (EMG), often reflect genuine affective states, as they are more difficult to consciously control than facial or vocal cues.

2.3.1. EEG-Based Recognition

Signal Processing and Feature Extraction: Time-frequency transforms or spectrograms are widely used for EEG data [14].

CNNs and RNNs: CNNs capture spatial-spectral information, and RNNs model temporal behavior [15]. In addition, Transformers have been tested on EEG to identify correlations across electrodes [16].

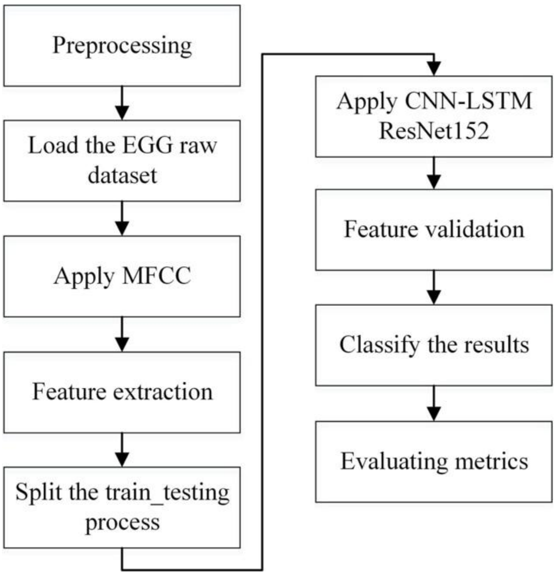


Figure 5. Overview of the CNN_LSTM Hybrid methodology [15].

Figure 5 demonstrates a hybrid architecture that combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks for processing EEG signals. It outlines the stages from data preprocessing and feature extraction to sequential modeling and final classification, highlighting the synergy between spatial and temporal analysis.

2.3.2. Transformer Architectures

Recent studies leverage transformer models for EEG time series, using self-attention to capture relationships across electrodes and time steps [16]. This can improve classification accuracy in both offline and real-time systems.

2.3.3. Datasets

Common EEG-based emotion datasets include Database for Emotion Analysis using Physiological Signals (DEAP), SJTU Emotion EEG Dataset (SEED), Database for Emotion Recognition Through EEG and ECG(DREAMER), and *Dataset for Affect, Personality, Mood, and Engagement* (AMIGOS), featuring recordings of participants exposed to emotional stimuli such as music videos or film clips [17].

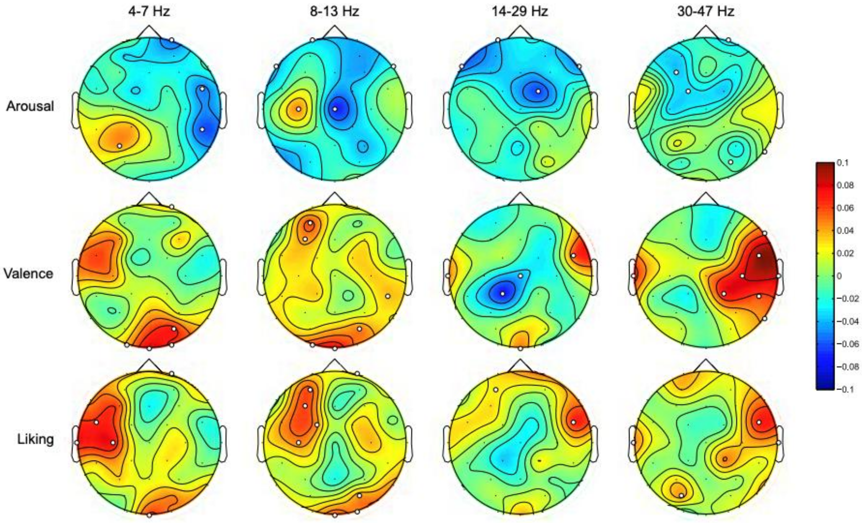


Figure 6. EEG Topographic Maps Showing Brain Frequency [17].

This figure presents topographic maps of the brain, illustrating the spatial distribution of EEG frequency bands across the scalp. It visually correlates specific frequency responses with different emotional states, offering insights into the neural dynamics that underpin affective processing.

2.3.4. Other Physiological Signals

ECG and GSR: Heart rate variability and skin conductance can correlate with emotional intensity [18].

EMG: Facial electromyography (fEMG) can tracks subtle facial muscle movement that may elude standard visual analysis [19].

2.4. Text-Based Emotion Recognition

Text-based approaches detect emotional or affective content from written or transcribed speech. Modern large-scale language models have dramatically boosted accuracy.

2.4.1. Classical NLP Methods

Lexicon-Tools: Simple features such as Bag-of-Words, TF-IDF, or lexicons were common historically but can fail with sarcasm or sophisticated language structures [20].

2.4.2. Transformer-Based Models

BERT, GPT, and Variants: Fine-tuning these large pre-trained networks on emotion-tagged data (e.g., MELD, DailyDialog) has yielded substantial improvements [21, 22].

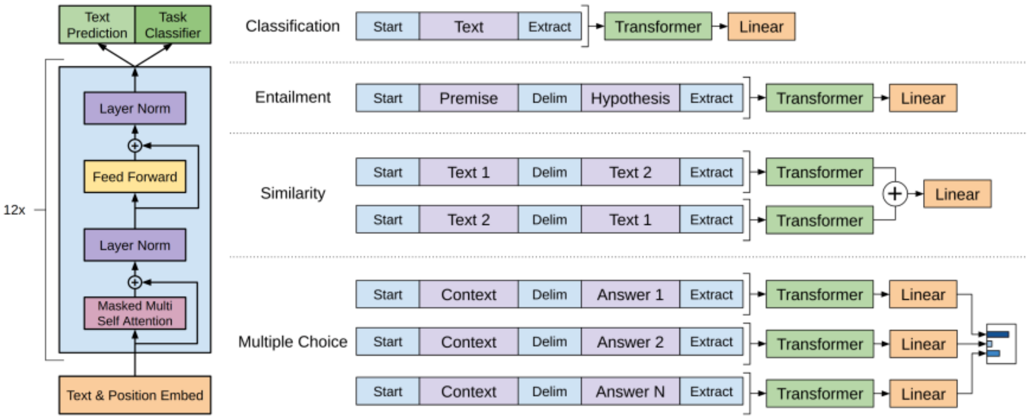


Figure 7. BERT architecture for emotion recognition [21].

This diagram outlines the BERT model's structure as applied to emotion recognition. It shows how text inputs are tokenized and embedded, then processed through multiple Transformer layers with self-attention mechanisms, culminating in a classification head that assigns emotion labels based on contextual cues.

2.4.3. Contextual Understanding

Self-attention enables capturing more context, which is crucial for irony, slang, and hidden emotional nuances [23].

3. Multimodal Approaches Using Large Language Models (LLMs) and Video Data

The multimodal fusion of textual, audio, visual, and physiological signals has gained momentum due to its ability to model emotional states more comprehensively and robustly. This section focuses on how Large Language Models (LLMs) and advanced video-based architectures contribute to enhanced multimodal emotion recognition.

3.1. Text and Video-Based Emotion Recognition

3.1.1. CLIP and Video Vision Transformers (ViViT)

CLIP (Contrastive Language-Image Pretraining): Proposed by Radford et al. [24], CLIP learns joint representations of images and text via a contrastive objective on large-scale web data. While initially designed for zero-shot image classification, CLIP embeddings have proven valuable in emotion recognition tasks where textual and visual cues must be aligned.

ViViT (Video Vision Transformer): ViViT extends transformer-based approaches to video by modeling spatiotemporal features. By leveraging self-attention across video frames, it can capture changes in facial expressions, body language, and scene context [25].

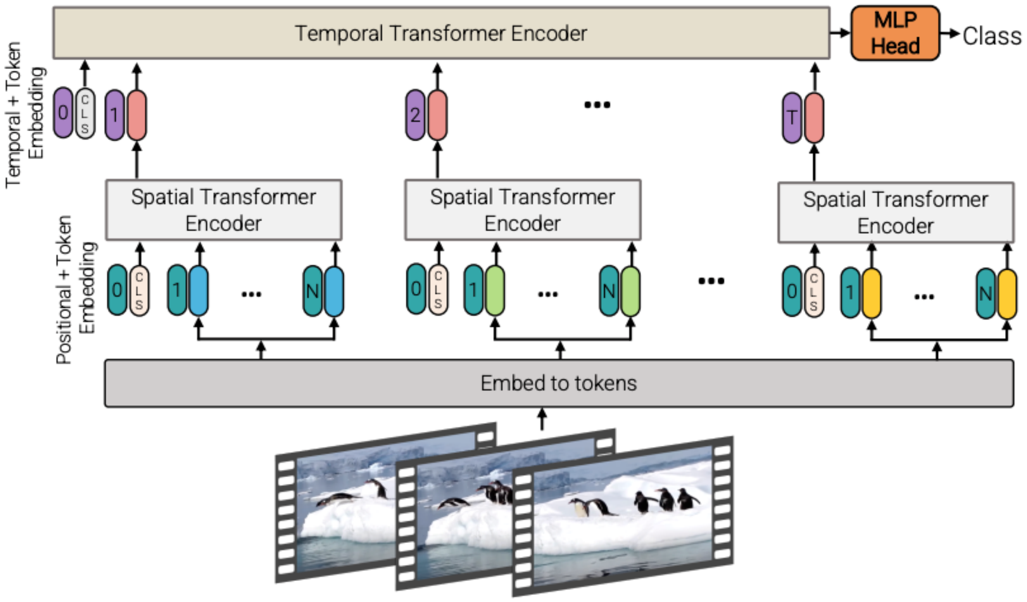

Temporal Transformer Encoder
MLP Head
Class
Temporal + Token Embedding
0
CLS
1
2
T
Spatial Transformer Encoder
Spatial Transformer Encoder
Spatial Transformer Encoder
Positional + Token Embedding
0
CLS
1
N
Embed to tokens

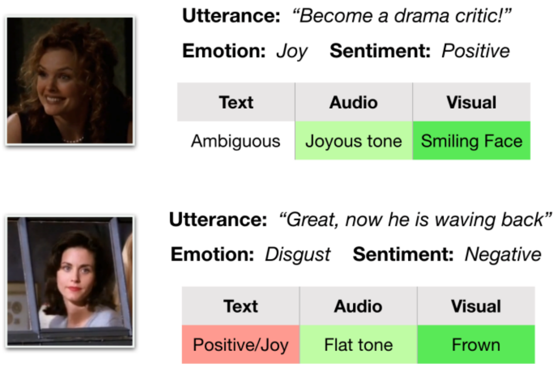
Figure 8. ViViT Factorised Encoder (Model 2) [25].

Figure 8 details the ViViT model for video-based emotion recognition. It highlights how the model factorizes video input into spatial and temporal components, processing each through specialized Transformer encoders to capture dynamic facial expressions and body language over time.

Multimodal Alignment: By combining CLIP and ViViT, researchers can exploit large-scale pretrained representations of language and video, mapping textual emotional cues (e.g., transcripts or textual descriptions) to visual patterns of expression for improved accuracy.

3.1.2. Applications and Datasets

Multimodal Emotion Lines Dataset (MELD): Originally created for multimodal sentiment and emotion analysis in multi-speaker dialogues, MELD includes synchronized audio, video, and transcripts from TV show episodes [26] as shown in Figure 9.



Utterance: *"Become a drama critic!"*

Emotion: *Joy* **Sentiment:** *Positive*

Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

Utterance: *"Great, now he is waving back"*

Emotion: *Disgust* **Sentiment:** *Negative*

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

Figure 9. Example from MELD [26].

This example from the MELD dataset shows a multimodal dialogue scenario, where synchronized audio, video, and textual data are aligned. It exemplifies how different modalities are integrated to improve the accuracy and contextual understanding of emotion recognition in conversational settings.

Interactive Scenarios: In social robotics and VR-based training, the fusion of textual and video inputs helps systems respond more empathetically to user emotions [27].

3.2. Combining Audio, Visual, and Textual Inputs

3.2.1. Transformer Fusion Networks

Transformer fusion networks employ cross-attention or co-attention mechanisms to merge embeddings from different modalities, such as speech spectrograms, video frames, and text transcripts [28]. By jointly learning these modalities, the system can capture correlated features (e.g., tone of voice matching facial expressions) more effectively.

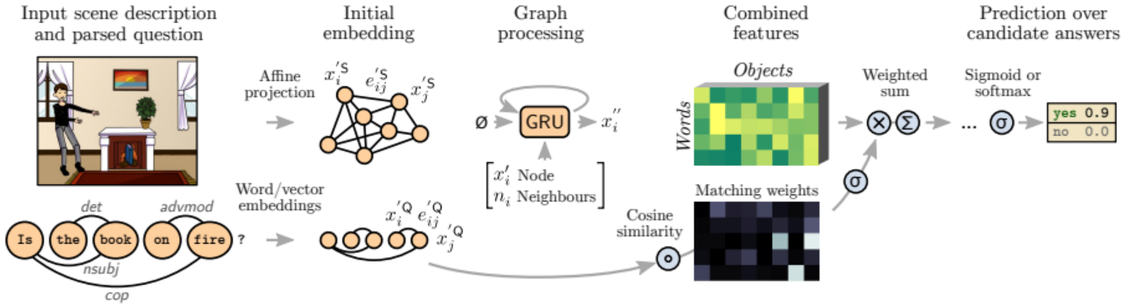


Figure 10. Graph Neural Network for emotion recognition [28].

In Figure 10, this diagram illustrates a Graph Neural Network (GNN) architecture tailored for multimodal emotion recognition. It demonstrates how visual and textual features are represented as nodes and edges within a graph, enabling the model to capture complex interdependencies between modalities for more robust emotion inference.

3.2.2. Multimodal Large Language Models (MLLMs)

GPT-4 and Beyond: Recent large-scale models, such as GPT-4 [29], incorporate additional capabilities to handle diverse input forms, including images and structured data. When paired with specialized encoders for audio or video, GPT-4 can produce holistic emotional inferences that account for both context and nuance.

Benefits of LLMs are as follows.

Context-Rich Understanding: LLMs excel at capturing contextual information from text, thus bridging semantic gaps between language and other modalities.

Transfer Learning Potential: With instruction-tuning or few-shot learning, MLLMs can adapt to new emotion recognition tasks with minimal labeled data.

4. Key Challenges in Emotion Recognition

Despite notable progress, several significant challenges must be addressed to achieve robust, scalable, and ethically sound emotion recognition systems.

4.1. Dataset Biases

4.1.1. Demographic Bias

Variations in emotional expression across age, gender, and ethnicity can lead to dataset imbalances. For instance, many facial expression datasets contain predominantly younger subjects of certain ethnicities [30].

4.1.2. Contextual Bias

Certain datasets capture emotions in controlled environments that do not necessarily generalize to real-life settings (e.g., acted expressions vs. spontaneous expressions) [31].

4.1.3. Mitigation Strategies

Methods such as data augmentation, domain adaptation, and bias-aware training (e.g., adversarial debiasing) can help reduce performance gaps [32].

4.2. Cross-Cultural Variations

4.2.1. Universal vs. Culture-Specific Expressions

While certain facial expressions may be universal (e.g., happiness, sadness), subtle cues can differ widely across cultures. Models trained on a single cultural context may not generalize to others [33].

4.2.2. Linguistic Nuances

In text-based analysis, idiomatic expressions, emoticons, or culturally specific references can confound emotion classifiers.

4.2.3. Potential Approaches

Cross-cultural dataset collection (e.g., from multilingual or multicultural sources), multi-language pretrained LLMs, and culture-adaptive models can address these variations [34].

4.3. Real-Time Performance

4.3.1. Computation and Latency

Emotion recognition in real-time systems, such as social robotics or wearable devices, demands computationally efficient models.

4.3.2. Lightweight Architectures

Model compression techniques (e.g., quantization, knowledge distillation) and real-time capable neural architectures are key to deploying emotion recognition on resource-constrained devices [35].

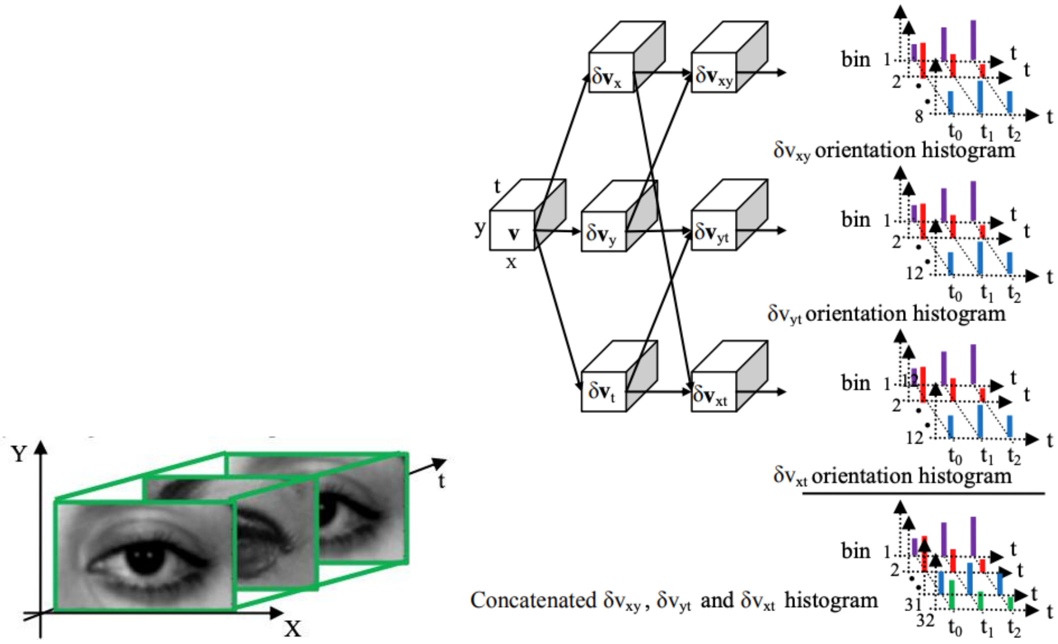

Y
X
t
y
t
x
v
δv_x
δv_y
δv_t
δv_xy
δv_yt
δv_xt
bin
1
2
8
12
31
32
t
t0 t1 t2
δv_xy orientation histogram
δv_yt orientation histogram
δv_xt orientation histogram
Concatenated δv_xy, δv_yt and δv_xt histogram

Figure 11. Facial cube [35].

They introduce the concept of constructing a "facial cube" by stacking sequential video frames to capture subtle, micro-expressions as shown in Figure 11. It shows how spatial and temporal information is combined to detect fleeting emotional cues that may not be visible in isolated images.

4.4. Privacy and Ethical Concerns

4.4.1. Informed Consent

Recording and analyzing individuals' emotional states can be intrusive.

4.4.2. Misuse and Surveillance

Unchecked use of emotion recognition in surveillance or advertising raises ethical questions regarding autonomy and potential misuse.

4.4.3. Regulatory Frameworks

Governments and organizations must develop guidelines to ensure responsible deployment of affective computing systems [36].

5. Applications and Future Directions

5.1. Applications

5.1.1. Healthcare and Mental Health

Depression Detection: Continuous monitoring of emotional cues in speech, facial expressions, or physiological signals can aid early detection of depression and other mental health disorders [37].

Telemedicine: Emotion-aware virtual assistants can enhance remote patient consultations, providing clinicians with additional diagnostic information.

5.1.2. Human-Computer Interaction (HCI)

Adaptive Interfaces: Emotion recognition can enable adaptive user interfaces that respond to the user's emotional state, improving usability and user satisfaction.

Virtual Assistants: Personal assistants (e.g., Alexa, Siri) may use emotional cues to respond more empathetically.

5.1.3. Education

E-Learning: Tracking student engagement and frustration can help tailor teaching materials and provide real-time feedback to educators [38].

Intelligent Tutoring Systems: Emotion-sensitive tutoring systems can adapt lesson difficulty and provide motivational support.

5.1.4. Entertainment and Gaming

Personalized Content: Streaming platforms can recommend content aligned with a user's emotional state, potentially increasing user engagement.

Interactive Gaming: Emotionally responsive games that adapt narratives or difficulty based on player affect.

5.2. Future Research Directions

5.2.1. Explainable Emotion Recognition

Interpretability methods (e.g., attention visualization, saliency maps) can increase trust in deep learning-based systems, particularly in high-stakes applications such as healthcare.

5.2.2. Long-Term Emotion Tracking

Most current work focuses on short-term or discrete emotion recognition. Future systems should model affective trajectories over extended periods to capture more complex emotional phenomena [39].

5.2.3. Fusion of Wearable Sensors

Integrating additional signals from wearables (e.g., smartwatches, electrodermal activity sensors) could enhance real-time emotion monitoring in daily life.

Multicultural and Multilingual Models: Broadening dataset diversity and building culture-aware models remain critical for global deployment.

LLMs for Multi-Modal Conversational Analysis: Future LLMs may incorporate advanced fusion mechanisms, seamlessly integrating multiple modalities (video, audio, text, EEG) in more human-like conversational interfaces [40].

6. Conclusion

Over the years, emotion recognition has progressed from simple feature extraction to sophisticated deep learning and multimodal fusion approaches. Transformers, including Vision Transformers and self-supervised models such as Wav2Vec 2.0, have expanded capabilities to unify audio, text, and visual information. However, practical applications still face challenges regarding fairness, cross-cultural adaptability, latency constraints, and ethical dilemmas. Achieving transparent, bias-aware, and privacy-respecting solutions is critical for future expansion. With continued investigation into interpretable AI methods, model efficiency, and culturally balanced datasets, emotion recognition systems will likely become increasingly robust and influential. By bridging fundamental research and real-world implementation, these systems hold immense promises for healthcare, HCI, education, and beyond.

References

1. Schuller, B. et al. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans. Affective Comput. 1*(2), 119–131 (2010).
2. Li, Z., Chen, Z. and Han, J. Speech emotion recognition using recurrent neural networks. *Neurocomputing 36*(9), 2207–2215 (2011).
3. Han, K., Yu, D. and Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proc. of Interspeech, MAX Atria, Singapore (1-18 Sep.2014).
4. Baevski, A., Zhou, H., Mohamed, A. and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst. (NeurIPS) 33*, 12449–12460 (2020).
5. Hsu, W. et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Lang. Process 29*, 3451–3460 (2021).
6. Livingstone, S. and Russo, F. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE 13*(5), e0196391 (2018).
7. Li, S. and Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affective Comput. 1*(1), 1–1 (2020). doi: 10.1109/TAFFC.2020.2981446.
8. Fan, Y., Lu, X., Li, D. and Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proc. of 18th ACM Int. Conf. Multimodal Interact (ICMI), Boulder, CO, USA (16-20 Oct. 2018).
9. Dosovitskiy, A. et al. An Image is Worth 16×16 Words: Transformers for image recognition at scale. In Proc. of Int. Conf. Learn. Representations (ICLR), Vienna, Austria (4 May 2021).
10. Caron, M. et al. Emerging properties in self-supervised vision transformers. In Proc. of IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Nashville, TN, USA (20-25 June 2021).
11. He, K. et al. Masked autoencoders are scalable vision learners. In Proc. of IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA (18-24 June 2022).
12. Goodfellow, I. et al. Generative adversarial nets. *Adv. Neural Inf. Process. Syst. (NeurIPS) 27*, 2672–2680 (2014).
13. Mollahosseini, A., Hasani, B. and Mahoor, M. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affective Comput.10*(1), 18–31 (2017).
14. Zheng, W., Zhu, J. and Lu, B. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Trans. Affective Comput.10*(3), 417–429 (2019).
15. Zhang, L., Yin, Y. and Zhang, D. EEG-based emotion recognition using hybrid CNN and LSTM models. *Neurocomputing 415*, 316–325 (2021).
16. Roy, S., Saffary, R., Ahmed, M. and Chen, F. Transformer-based architectures in EEG emotion recognition: A comprehensive review. *Front. Neurosci.16*, 1016862 (2022).
17. Koelstra, S. et al. DEAP: A database for emotion analysis using physiological signals. *IEEE Trans. Affective Comput. 3*(1), 18–31 (2012).
18. Greco, A., Faes, L. and Nollo, G. The importance of electrodermal activity in emotion-based affective computing. In Proc. of 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Orlando, FL, USA (16-20 Aug. 2016).
19. Nielsen, K., Henriksen, T. and Staum, M. Facial electromyography for analyzing emotional responses to film scenes. *PLoS ONE 15*(10), e0239982 (2020).
20. Cambria, E., Schuller, B., Yunqing, X. and Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst. 28*(2), 15–21 (2013).
21. Devlin, J., Chang, M., Lee, K. and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. of NAACL-HLT, Minneapolis, Minn, USA (4-7 June 2019).
22. Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. Improving language understanding by generative pre-training. OpenAI Rep. (2018).
23. Zhang, S., O'Connell, T., Danilevsky, M., Araki, J. and Aharoni, E. A contextualized emotion recognition system: Using text as an anchor modality. In Proc. of AAAI Conf. Artif. Intell., Virtual (2-9 Feb. 2021).
24. Radford, A. et al. Learning transferable visual models from natural language supervision (CLIP). In Proc. of 38th Int. Conf. Mach. Learn. (ICML), Virtual (18-24 July 2021).
25. Arnab, A. et al. ViViT: A video vision transformer. In Proc. of IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, QC, Canada (10-17 Oct. 2021).
26. Poria, S., Majumder, N., Hazarika, D. and Mihalcea, R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proc. of 57th Annu. Meet. Assoc. Comput. Linguist. (ACL), Florence, Italy (28 July-2 Aug. 2019).
27. Zheng, W., Wang, H. and Lu, B. Emotion recognition from multimodal data. *IEEE Trans. Cogn. Dev. Syst. 12*(2), 165–174 (2020).
28. Teney, D., Liu, L. and Van Den Hengel, A. Graph-structured representations for visual question answering. In Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA (21-26 July 2017).
29. OpenAI, GPT-4 technical report (OpenAI, 2023).
30. Howard, A. et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704,* 04861 (2017).
31. Zeng, Z., Pantic, M., Roisman, G. and Huang, T. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell. 31*(1), 39–58 (2009).

32. Wang, A., Narayanan, A. and Russakovsky, O. REVISE: A tool for measuring and mitigating bias in visual datasets. In Proc. of Eur. Conf. Comput. Vis. (ECCV), Online (23-28 Aug. 2020).
33. Jack, R., Garrod, O. and Schyns, P. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biol. 24*(2), 187–192 (2014).
34. Li, Y., Zhou, Y., Ji, F. and Chen, W. Towards cross-cultural emotion recognition: A multilingual and multi-contextual perspective. *IEEE Trans. Affective Comput. 1*(1), 1–1 (2022). doi: 10.1109/TAFFC.2022.3150990.
35. Polikovsky, S., Kameda, Y. and Ohta, Y. Facial micro-expressions recognition using high-speed camera and 3D-gradient descriptor. In Proc. of 3rd Int. Conf. Crime Detect. Prev. (ICDP), London, UK (3 Dec. 2009).
36. Vander, W. Ethical frameworks for emotion recognition in AI. *AI Ethics 1*(1), 1–13 (2021).
37. Cummins, N., Scherer, S., Krajewski, J. and Schuller, B. A review of depression detection in speech. *IEEE Trans. Affective Comput. 9*(1), 1–19 (2017).
38. Bosch, N., D'Mello, S. and Mills, C. What emotions do students experience during science learning? An affective computing approach. *J. Educ. Data Mining 5*(3), 117–151 (2013).
39. Barrett, L. How Emotions Are Made: The Secret Life of the Brain (Houghton Mifflin Harcourt, 2017).
40. Yang, Z. et al. XLNet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst. (NeurIPS) 32*, 5753–5763 (2019).