International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Human Emotion Recognition Using Multi-modal Deep Learning: A Review of Methods, Datasets, and Challenges

Ayushi Parmar, Prof. Chandni Sikarwar

Abstract: Human emotion recognition plays a vital role in the field of affective computing and finds wide-ranging applications in healthcare, education, robotics, and human-computer interaction. Traditional unimodal approaches—based solely on facial expressions, speech, or physiological signals—often face limitations due to varying environmental conditions, individual differences, and signal noise. To address these challenges, the use of multi-modal deep learning has gained significant momentum, as it combines multiple data streams such as visual, auditory, textual, and physiological inputs to enhance the accuracy and robustness of emotion detection.

This review paper presents a detailed examination of recent developments in the area of multi-modal deep learning for human emotion recognition. We explore various deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformer-based architectures, and their effectiveness in processing different modalities. In addition, we study various fusion strategies—early, late, and hybrid fusion—and their respective contributions towards improving recognition performance.

The aim of this paper is to provide researchers and practitioners with valuable insights into the current landscape, ongoing challenges, and future opportunities in this rapidly growing domain.

Keywords: Keywords—Human Emotion Recognition, Multi-modal Deep Learning, Emotion Detection, Deep Learning Architectures, Multi-modal Fusion

1. Introduction

Human Emotion Recognition (HER) has emerged as an important area of study, especially in the context of improving human-computer interaction and aiding mental health interventions. In recent years, there has been a growing interest in using multi-modal deep learning techniques to develop systems that can understand human emotions more accurately and reliably. Unlike traditional methods that rely on a single type of input—such as facial expressions or voice—multi-modal systems combine information from various sources including facial cues, speech patterns, body language, and physiological signals like EEG and ECG, thus offering a more holistic understanding of emotional states.

2. Literature Review

Emotions are integral to human experience and significantly influence our perception, decision-making, and interpersonal interactions [1]. In the context of emotion recognition, understanding how emotions are modelled and classified is essential.

One of the most widely accepted frameworks is proposed by Paul Ekman, who suggested that emotions are universal and can be recognised across cultures. He identified six fundamental emotions, Happiness, Sadness, Anger, Fear, Surprise and Disgust [2].

These emotions are generally expressed through facial movements, tone of voice, and gestures. However, emotional expression can vary subtly based on cultural and individual factors.

Besides categorical models, there are dimensional models of emotion, which provide a more flexible and continuous representation [5]. For instance: Russell's Circumplex Model represents emotions on two axes:

- Valence (positive to negative)
- Arousal (low to high energy)

For example, joy is high-arousal and positive-valence, while sadness is low-arousal and negative-valence.

Plutchik's Wheel of Emotions proposes eight primary emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. These are arranged in a wheel-like structure with varying intensities and combinations.

These models provide a theoretical base for designing emotion recognition systems that can adapt to nuanced emotional states rather than treating them as fixed categories.

2.1 Modalities Used for Emotion Recognition

In any multi-modal system, the type and quality of data—or modalities—used play a crucial role. Combining multiple modalities helps improve accuracy, especially in complex or real-world scenarios [3]. Below, we discuss the most commonly used modalities in emotion recognition.

2.1.1 Facial Expression-Based Emotion Recognition

Facial expressions offer powerful non-verbal cues that are closely linked to emotional states. Certain features are extracted like Eyebrow movement, eye openness, lip curvature, facial muscle dynamics [4]. Some of the common techniques are tradition techniques like Facial Action Coding System (FACS) and deep learning techniques like CNNs such as VGG-16, ResNet, and Vision Transformers (ViT).

2.1.2 Speech-Based Emotion Recognition

Speech carries emotional information through tone, pitch, pace, and intensity. Certain features are extracted like Prosodic which include Pitch, loudness & rhythm, Spectral which includes MFCCs and spectrograms, Linguistic like Emotion-rich words and semantic cues.

To deal with these certain approaches includes traditional approaches like GMM and HMM. Whereas Deep learning approaches like RNNs, LSTMs, Transformers (e.g., BERT for speech-text fusion) [6].

2.1.3 Physiological Signal-Based Emotion Recognition

Emotions trigger physiological changes in the body that can be measured with appropriate sensors.

Signals used:

- EEG (brain activity)
- ECG (heart rate)
- GSR (skin conductance)
- EMG (muscle movement)

Techniques:

- Traditional classifiers: SVM, k-NN
- Deep learning: CNN-LSTM hybrids, Autoencoders

2.1.4 Text-Based Emotion Recognition

Text data - whether from social media, chat logs, or transcripts—can provide insight into emotions through language. Certain features are extracted like Lexical: Emotion words and sentiment scores, Syntactic like Sentence structure, Semantic like Word embeddings (Word2Vec, GloVe, BERT) [7].

To deal with these methods include Traditional: Naïve Bayes, SVM and Deep learning: RNNs, BERT, GPT

3. Solution Domain: Deep Learning Techniques for Multi-modal Emotion Recognition

Deep learning has brought a paradigm shift in the domain of human emotion recognition, especially when dealing with diverse and complex data from multiple modalities. The ability of deep learning models to automatically extract and learn hierarchical representations from raw data has proven beneficial for emotion analysis [6].

In recent years, various architectures—such as CNNs, RNNs, LSTMs, and Transformers—have been effectively employed in processing facial expressions, speech signals, physiological data, and textual information. Moreover, the integration of these modalities using fusion strategies has further improved the system's performance in real-world scenarios [6].

3.1 Deep Learning Architectures for Multi-modal Emotion Recognition

Convolutional Neural Networks (CNNs):

CNNs are widely used in image and spatial data analysis. They have been particularly effective in extracting features from facial images and spectrograms derived from speech signals.

Recurrent Neural Networks (RNNs):

RNNs are designed for handling time-series or sequential data. They are commonly used in speech-based emotion recognition and text analysis [7].

LSTM networks are a refined version of RNNs designed to capture long-term dependencies. They use memory cells and gate mechanisms to control the flow of information.

Transformer Models

Transformers have emerged as powerful alternatives to RNNs and LSTMs, especially in the domain of Natural Language Processing and, more recently, in multi-modal fusion tasks.

Popular Models:

- **BERT**: For text-based emotion recognition.
- **Vision Transformer (ViT)**: For facial emotion analysis.
- **Multimodal-BERT and Wav2Vec**: For combining audio, visual, and textual features.

3.2 Generative Adversarial Networks (GANs) for Data Augmentation

GANs are primarily used to address the problem of limited training data. They generate synthetic samples that mimic real data, improving model generalisation.

3.3 Multi-modal Fusion Strategies in Deep Learning

The effectiveness of multi-modal emotion recognition largely depends on how different data streams are combined. Fusion strategies can be broadly classified into the following:

Early Fusion (Feature-Level Fusion)

Combines features from all modalities at the input stage. Example: Concatenating MFCC features from speech with facial embeddings from a CNN.

Late Fusion (Decision-Level Fusion)

Each modality is processed independently, and final predictions are merged. Example: Output from a CNN (facial expression) and an LSTM (speech) combined at the decision level.

Hybrid Fusion

Combines both early and late fusion techniques, often incorporating attention mechanisms. Example: Fuse features from CNN and LSTM at a hidden layer, and refine the decision using a transformer or attention model.

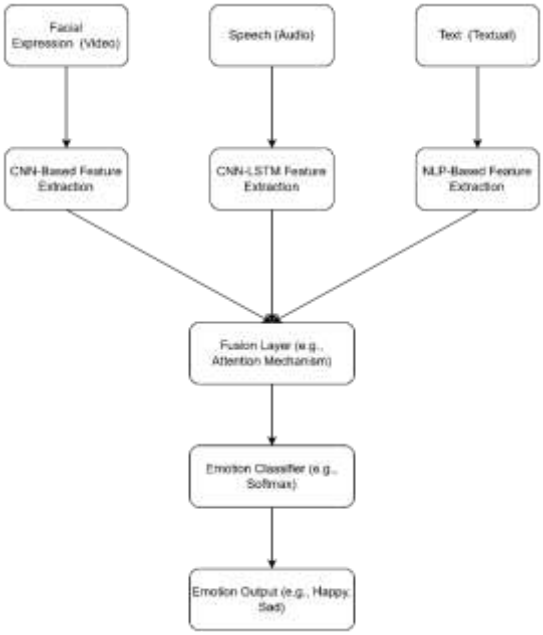

Figure 1: Architecture of the Multimodal Emotion Recognition System

4. Benchmark Datasets

For any deep learning-based system, especially in the area of human emotion recognition, the availability of high-quality and diverse datasets is of utmost importance. In multi-modal emotion recognition, datasets must include multiple types of data such as facial expressions, speech, physiological signals, and text, ideally recorded under real-life conditions. Well-annotated and representative datasets enable researchers to train, validate, and benchmark their models reliably.

4.1 Facial Expression-Based Emotion Datasets

Facial expressions are among the most expressive and commonly studied modalities for emotion recognition. The following datasets are extensively used in training and evaluating models based on facial features.

(i) FER-2013 (Facial Expression Recognition 2013)

Contains over 35,000 grayscale facial images classified into seven categories—Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise[8].

(ii) AffectNet

One of the largest facial emotion datasets, consisting of over 1 million facial images scraped from the internet and labeled into eight emotional categories [9].

(iii) CK+ (Cohn-Kanade Extended Dataset)

Contains 593 sequences of facial expressions from 123 participants, recorded in controlled environments [10].

4.2 Speech-Based Emotion Datasets

Speech signals convey rich emotional content through variations in tone, pitch, rhythm, and speed. The datasets mentioned below are commonly used in speech emotion recognition systems.

(i) RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)

Consists of 1,440 recordings by professional actors speaking in various emotional tones like happiness, sadness, anger, fear, and surprise [11].

(ii) IEMOCAP (Interactive Emotional Dyadic Motion Capture Database)

Offers over 12 hours of audio-visual data recorded from dyadic sessions, annotated for multiple emotional states.

(iii) CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)

Comprises over 7,000 audio clips recorded by 91 actors, covering six basic emotions.

5. State-of-the-Art Approaches and Comparisons

In recent years, research in multi-modal emotion recognition has seen remarkable progress owing to the evolution of deep learning architectures, advanced fusion techniques, and the availability of diverse datasets. Modern systems now combine facial expressions, voice characteristics, text content, and physiological signals to gain a more complete and accurate understanding of human emotions.

This section presents an overview of leading deep learning-based techniques, highlighting how each approach performs across different modalities and benchmark datasets.

5.1 State-of-the-Art Approaches for Multi-modal Emotion Recognition

Various deep learning models have been proposed for emotion recognition, and their effectiveness depends largely on the modality in question and the type of fusion used.

5.1.1 Convolutional Neural Networks (CNNs)

CNNs are predominantly used for analysing spatial features in visual and audio data.

Applications:

- **Facial Expression Recognition**:
  - VGG16/VGG19: Perform well on datasets like FER-2013 and AffectNet.
  - ResNet-50/101: Deep residual networks that provide higher accuracy through skip connections.
  - DeepFace/FaceNet: Face recognition models adapted for emotion detection.
- **Speech Emotion Recognition**:
  - Speech signals are converted into spectrograms (visual representations of sound), which are then fed into CNNs to classify emotions.

5.1.2 Recurrent Neural Networks (RNNs) and LSTM

For sequential and time-series data, RNNs and LSTM models are particularly effective.

Applications:

- **Speech**: Capture emotional variations in tone, rhythm, and pace.
- **Text**: Understand emotional flow across sentences.
- **EEG/ECG**: Identify emotional responses through physiological signal patterns.
- Bi-directional LSTM (Bi-LSTM): Reads sequences from both directions, offering a more holistic view of temporal dependencies.
- GRU (Gated Recurrent Unit): A lightweight alternative to LSTM, suitable for real-time systems.

5.1.3 Transformer-Based Models

Transformer models have revolutionised emotion recognition, especially in multi-modal settings, due to their attention-based mechanisms and parallel processing capabilities.

5.2 Comparative Analysis of State-of-the-Art Models

The following table presents a comparative analysis of SOTA models based on performance across different datasets.

Table 1:

Model	Modalities Used	Dataset(s)	Strengths	Limitations
CNN (VGG-16)	Facial expressions	FER-2013	Strong spatial feature extraction	Lacks temporal processing
Bi-LSTM	Speech signals	RAVDESS	Captures temporal dependencies	High training time
BERT	Text-based	EmoContext	Pre-trained NLP model for text emotions	Requires large labeled text data
Vision Transformer	Facial expressions	AffectNet	Self-attention improves facial emotion classification	Requires high computational power

6. Applications of Multi-modal Emotion Recognition

The potential of multi-modal emotion recognition extends across a wide spectrum of real-world domains. By intelligently combining facial expressions, voice modulation, textual content, and physiological signals, emotion-aware systems can provide highly contextual and human-like responses in various settings. This section highlights some of the most promising application areas.

6.1 Healthcare and Mental Health Monitoring

Emotion recognition has emerged as a valuable tool in the healthcare sector, particularly for psychological assessment and emotional well-being monitoring.

6.2 Human-Computer Interaction (HCI) and Intelligent Assistants

In the domain of HCI, emotion-aware systems are being developed to create machines that can empathise and respond in a more human-centric manner.

6.3 Education and E-Learning Platforms

In smart classrooms and online learning platforms, emotion recognition can be used to measure student engagement and emotional state during lessons.

Systems can adjust the pace, content, or difficulty level depending on whether the learner appears confused, bored, or attentive, thus promoting personalised learning.

6.4 Marketing and Customer Experience

Marketers use emotion recognition systems to gauge consumer reactions to advertisements, products, or services.

Facial analysis, voice tone during feedback, and text reviews are analysed to determine customer satisfaction and emotional appeal, helping businesses improve customer engagement strategies.

6.5 Security and Surveillance

Emotion recognition systems are being integrated into surveillance cameras to enhance security measures.

These systems can identify potentially suspicious or aggressive behaviour by analysing micro-expressions, body language, or vocal stress patterns, thus acting as a preventive tool in sensitive environments.

6.6 Entertainment and Media

The media industry is utilising emotion-aware AI to deliver customised content.

Streaming platforms can suggest content based on the viewer's emotional state, while virtual characters in films or games can respond in an emotionally intelligent manner.

7. Challenges and Knowledge Gaps

Despite the progress made, multi-modal emotion recognition still faces several notable challenges:

- **Data Scarcity and Imbalance**: There is a lack of large-scale, balanced datasets that include diverse emotions and demographics, especially from underrepresented regions and cultures.
- **Integration Complexity**: Fusing modalities such as EEG, ECG, and gesture inputs remains a technically challenging task.
- **Cross-modal Understanding**: Although multi-modal systems are effective, the deeper mechanisms of how different modalities interact are not yet fully understood.
- **Lack of Real-world Adaptability**: Many emotion recognition systems are trained in controlled settings and fail to generalise to dynamic, real-world environments.

8. Conclusion

Multi-modal deep learning has opened new frontiers in the field of human emotion recognition. By incorporating various data sources such as facial expressions, voice, text, and physiological signals, researchers can now build more accurate, adaptable, and human-centric systems. While the field has witnessed substantial advancements, several challenges still remain—particularly regarding data diversity, real-time processing, and understanding the interplay between different modalities.

References:

1. Zhang, Jianhua., Yin, Zhong., Chen, Peng., & Nichele, S.. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Inf. Fusion, 59, 103-126. http://doi.org/10.1016/j.inffus.2020.01.011
2. Abdullah, S.., Ameen, S.., Sadeeq, M.., & Zeebaree, Subhi R. M.. (2021). Multimodal Emotion Recognition using Deep Learning. Journal of Applied Science and Technology Trends. http://doi.org/10.38094/JASTT20291
3. Li, Shan., & Deng, Weihong. (2019). Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. IEEE Transactions on Image Processing, 28, 356-370. http://doi.org/10.1109/TIP.2018.2868382
4. Wang, Junke., Wu, Zuxuan., Chen, Jingjing., & Jiang, Yu-Gang. (2021). M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection. Proceedings of the 2022 International Conference on Multimedia Retrieval. http://doi.org/10.1145/3512527.3531415
5. Ng, Hongwei., Nguyen, Viet Dung., Vonikakis, Vassilios., & Winkler, Stefan. (2015). Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. http://doi.org/10.1145/2818346.2830593
6. Kahou, Samira Ebrahimi., Bouthillier, Xavier., Lamblin, Pascal., Gülçehre, Çaglar., Michalski, Vincent., Konda, K.., Jean, Sébastien., Froumenty, Pierre., Dauphin, Yann., Boulanger-Lewandowski, Nicolas., Ferrari, Raul Chandias., Mirza, Mehdi., Warde-Farley, David., Courville, Aaron C.., Vincent, Pascal., Memisevic, R.., Pal, C.., & Bengio, Yoshua. (2015). EmoNets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User

Interfaces, 10, 99 - 111. http://doi.org/10.1007/s12193-015-0195-2
7. Khalil, R. A.., Jones, E.., Babar, M. I.., Jan, T.., Zafar, M. H.., & Alhussain, T.. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access 7, 117327-117345.
http://doi.org/10.1109/ACCESS.2019.2936124
8. Cao, Houwei., Cooper, David G.., Keutmann, M. K.., Gur, R.., Nenkova, A.., & Verma, R.. (2014). CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing, 5, 377-390. http://doi.org/10.1109/TAFFC.2014.2336244
9. Noroozi, F.., Corneanu, C.., Kamińska, D.., Sapinski, T.., Escalera, Sergio., & Anbarjafari, G.. (2018). Survey on Emotional Body Gesture Recognition. IEEE Transactions on Affective Computing, 12, 505-523. http://doi.org/10.1109/TAFFC.2018.2874986
10. Liu, Wei., Qiu, Jielin., Zheng, Wei-Long., & Lu, Bao-Liang. (2021). Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition. IEEE Transactions on Cognitive and Developmental Systems, 14, 715-729. http://doi.org/10.1109/TCDS.2021.3071170
11. Chowdary, M. K.., Nguyen, Tu N.., & Hemanth, D.. (2021). Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications, 35, 23311-23328. http://doi.org/10.1007/s00521-021-06012-8