JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Multimodal Emotion Recognition: A Comprehensive Review

Rushal Chauhan, Rajeev Mathur

Computer Engineering Dept., Research Scholar, Indus University, Ahmedabad, chauhanrushal.24.rs@indusuni.ac.in
Information & Communication Technology Dept. , Professor, IICT- Indus University Ahmedabad, director@iict.indusuni.ac.in

***ABSTRACT*— Multimodal Emotion Recognition (MER) is one of the newest merger technologies to date and can be used in affective computing, human-computer interaction, and sentiment analysis. This review examines state-of-the-art deep learning-based approaches to fusion in MER systems, focusing on early and late fusion strategies. This paper attends to the capture of inter-module interactions in text-audio-video recognition systems and its impact on emotion recognition. The need for explainable models, generalization beyond the original corpus, and additional data or modality integration are some of the critical gaps addressed. The review outlines unresolved issues, focusing on few-shot learning and large language models as a means of increasing flexibility and generalization.**

Keywords*— *Multimodal Emotion Recognition (MER), Early Fusion, Late Fusion, Hybrid Fusion

I. INTRODUCTION

Understanding emotions is important for human-computer interactions as it helps artificial intelligence (AI) systems understand and respond to a user's emotions. Older strategies primarily focused on capturing facial expressions, speech, or even text to analyze feelings without considering the output from other channels. These methods tend to get lost in the confusion of dealing with several unknowns because emotions are constructed and influenced by several different aspects at the same time [8].

The fusion of text, audio, and visual information for emotion recognition is Multimodal Emotion Recognition (MER), and it has proven to be quite effective. Incorporating various modalities allows MER to overcome the limitations of single modality approaches. For instance, facial expressions are important for emotion recognition, but they are not very helpful in cases of emotion suppression; however, tone of voice and linguistic information are very helpful [9]. Advances in deep learning and other algorithms have automated and optimized many processes, which has increased research aimed at enhancing MER's functionality.

The arrival of deep learning has substantially advanced MER capabilities. Models that have been trained previously, such as BERT for text, Wav2Vec for speech, and transformer for video, have achieved the state-of-the-art by learning valuable feature representations directly from raw data [25]. In addition, attention mechanisms have contributed to improving learning via the ability to vary the importance given to different modalities over time, and in turn, the accuracy of more context-aware emotion recognition [30].

However, challenges with MER exist. Significant among them is the need for large annotated multimodal datasets, as deep learning models rely on good quality training data. Also, it is a still relevant concern related to generalization across different datasets and real-world situations. Models trained on one dataset are often poor at generalizing to a different as well as unseen environment[32].In addition, because deep learning-based multimodal architecture is computationally complex, they may not be feasible for use in real-time applications, or on resource-limited devices [21] .

A significant second issue relates to the interpretability of MER models. Many deep-learning architectures have been designed as "black boxes", which can complicate understanding of their reasoning and decisions. Such ambiguity can be especially troublesome in sensitive applications like healthcare and affective computing where interpretability and reliability are critical. To combat this issue, researchers are beginning to utilize explainable AI (XAI) frameworks and attention visualization techniques for MER to improve transparency in the model and enhance trust in automated emotion recognition [22]. This review provides an overview of current MER techniques, with a focus on early and late fusion, attention mechanisms, and challenges of the field. We also highlight the upcoming directions in research: adding physiological signals, utilizing few-shot learning, and examining large language models to better develop generalization and adaptability in MER.

Figure 1 summarizes research methodologies generally used by researchers for the emotion recognitions. Incorporation of different source of data in multimodal emotion recognition improves emotion recognitions.

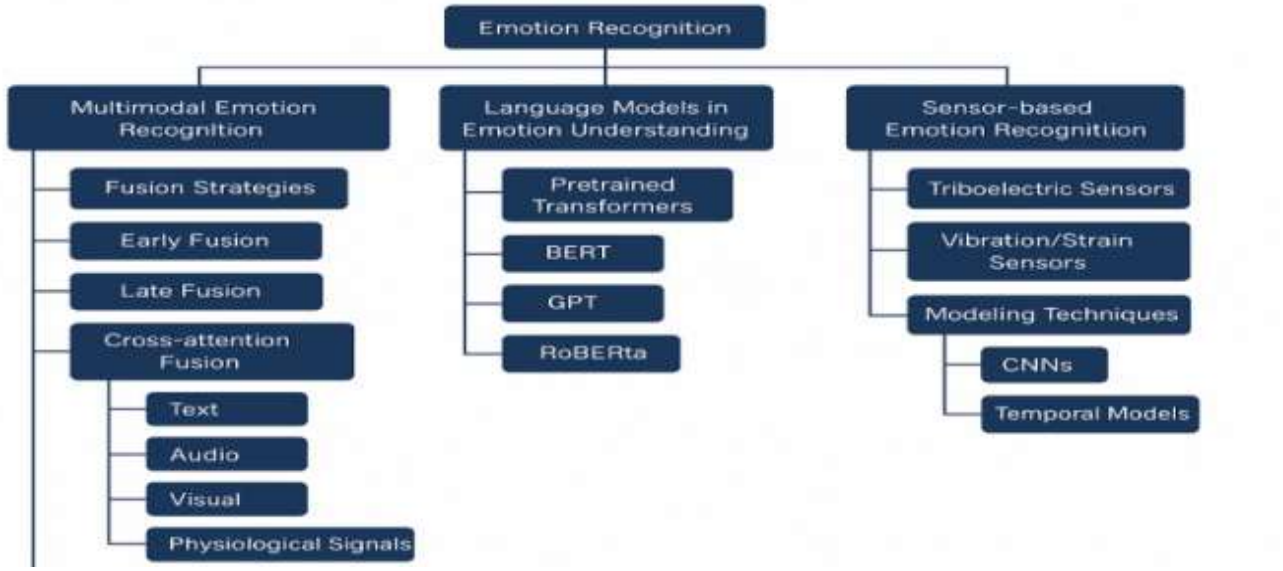

Figure 1 :Summarization of Research Methodologies for Emotion Recognition.

II.MULTIMODAL EMOTION RRCOGNITION (MER) FRAMEWORK

The major components and workflow is described as under:

Input Modalities: The system takes three types of input: Video, Audio, Text

a. Cross Attention Mechanism: In aggregate, the features extracted could be leveraged for cross-modality attention in order to encode relationships between the modalities -

Each of the attention points uses the query-key-value framework with attention to dynamically weight information exchanged between the three modalities.

b. Hierarchical Attention: After completing the cross-attention processing, two hierarchical attention layers are implemented to revise representations of the features and recognize higher order multimodal interaction ties.

c.Classification Layers: Those revised features are then passed through the novel fully-

connected, layer, which takes all input tensor of all features and applies a Softmax activation layer for the classification of the emotional

d. Predicted Emotion Output: The estimate from the model is the predicted emotion from the multimodal fused observation.

The Figure 2 represents a deep learning-based Multimodal Emotion Recognition (MER) framework that integrates text, audio,

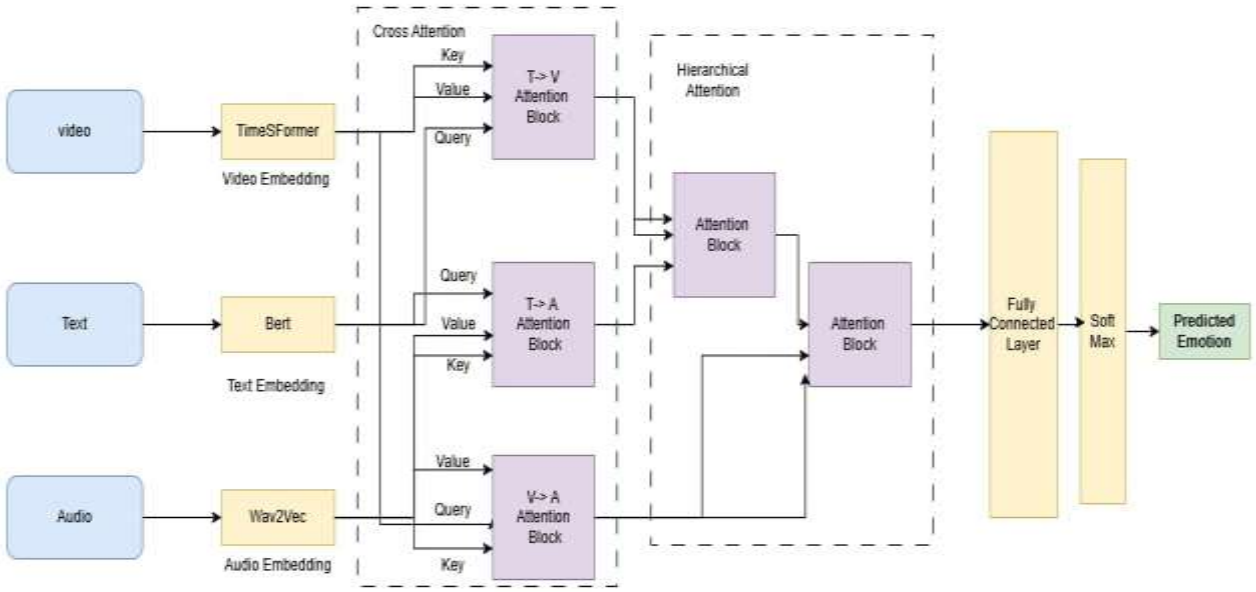

Figure 2: Multimodal Emotion Recognition (MER) framework

and video modalities.

The model establishes essential dependencies within text, audio, and video sections. The attention mechanisms will recognize dependencies within the feature objective, establishing the reduction of any noise, subsequently improving feature representation. The hierarchical attention architecture will validate the comprehensive fusing of multimodal feature data. This model also provides an approach to recognize selective dependence, ultimately balancing prediction accuracy and computational efficiency.

2.1 Fusion Techniques in MER

MER systems primarily employ early or late fusion strategies to integrate information from multiple modalities effectively. Early fusion (Feature-level fusion) is considered the most straightforward among fusion techniques. It involves combining features extracted from multiple data sources or modalities into a single unified feature vector. This early fusion method allows for capturing fundamental correlations between modalities but may struggle with modeling more complex relationships. This approach can yield good performance when the relationships between modalities are direct and linear. Late fusion (Decision-level fusion) operates at a later stage, where separate models are trained on each modality, and their outputs—such as predicted labels or confidence scores—are combined using strategies like majority voting or weighted averaging. This method has a moderate level of complexity and is particularly useful when feature integration is difficult or impractical. Hybrid methods can enhance system robustness by diversifying the sources of information and balancing their influence. Interpretability also varies: simpler hybrids may remain understandable, while complex ones might resemble black-box systems. When well-designed, hybrid fusion can achieve high performance by effectively blending complementary techniques. Early, Late and Hybrid fusion strategies play a critical role in determining the performance and robustness of the recognition system and are further explored s below -

2.1.1 Early Fusion (Feature-level fusion)

Early fusion consists of combining features from diverse modalities before entering a neural network. This process allows for joint feature learning, capturing complex interactions between modalities at an early phase. General approaches include:

- **Feature concatenation:** A naive method which fuses raw features together across multiple modalities prior to model input. It is simplistic but may not fully leverage cross-modal relationships [23].

• **Attention based fusion:** Attention methods such as self-attention and cross-modal attention and multi-layer attention have been shown to enhance feature interactions via weighting modalities to maximize attention of the model for particular time intervals, specifically when identifying emotions from unimodal inputs [2].

• **Cross-modal embeddings:** In this approach, multimodal data is aligned in the similar embedding space to facilitate generalizable features across datasets while limiting the impact of modality-specific biases [1].

The research findings suggest that attention based fusion methods outperform feature concatenation approaches due to their ability to adequately encode interdependencies between modalities [20]. However, early fusion methods can have a higher computational cost due to high-dimensional feature spaces.

2.1.2 Late Fusion (decision-level fusion)

Late fusion involves processing each modality in isolation and combining the predictions from multiple modalities at the decision level, which provides the modularity and flexibility. Several techniques have been proposed for late fusion, and these techniques include:

• **Weighted averaging:** This technique involves assigning weights to each individual modality prediction based on reliability so that the more informative modalities have greater contributions to the final prediction [28].

• **Meta-learning-based approach:** This method involves optimizing fusion with another learning model that sequentially selects the most reliable modality given an input, thus improving adaptability across datasets [28].

• **Ensemble learning:** This technique consists of combining the predictions of multiple unimodal classifiers through majority voting or stacking [11].

Late fusion is especially useful when the different modalities feature independent noise characteristics, making it more robust for real-world applications [18]. In comparison to early fusion, late fusion also allows the use of specialized models for each modality, improving interpretability and reducing overhead complexity. However, it may hold less capability than early fusion in capturing deep interactions between modalities.

2.1.3 Hybrid Fusion

Hybrid fusion methods, which refer to approaches that make use of both early and late fusion, are an emerging research area with the objective of optimizing multimodal integration and performance [19].

Here is a list of common **Hybrid Fusion Techniques** used in Multimodal Emotion Recognition (MER):

Common Hybrid Fusion Techniques:

• **Cross-modal Attention + Ensemble Classification**
  ○ Combines early-stage feature alignment using attention mechanisms with late-stage decision fusion (e.g., majority voting or stacking).
  ○ Example: Cross-modal Transformer + Softmax Ensembles [7].
• **Joint Embedding + Weighted Decision Fusion**
  ○ Learns a shared representation space from multiple modalities and then applies weighted prediction fusion [10].
  ○ Balances fine-grained interactions and robust decision-making.
• **Graph-based Hybrid Fusion**
  ○ Uses Graph Neural Networks (GNNs) to model relationships between modalities, followed by separate decision layers.
  ○ Improves modality interaction modeling and interpretability [14].
• **Attention-Gated Late Fusion**
  ○ Applies attention mechanisms to control the contribution of each modality before decision fusion [27].
  ○ Helps adaptively weigh modalities based on context [27].

2.2. CHALLENGES IN MER

2.2 Challenges in MER

Despite advancements in MER, several challenges persist:

1. Lightweight and Explainable Models

Deep learning models employed in MER tend to be computationally costly and hard to interpret. Solutions to overcome these include: Model compression methods like pruning and quantization to minimize computational overhead [11] and explain-ability techniques like Grad-CAM to visualize attention regions in neural networks [18].

2. Cross-Corpus Generalization

Models learned on one dataset tend not to generalize well to new datasets. Solution of this is domain adaptation methods to transfer knowledge between datasets [19]. Another method is adversarial learning methods to enhance robustness against distributional shifts [3].

3. Multimodal Fusion Optimization

Existing fusion methods are unable to model subtle cross-modal dependencies. Future work needs to concentrate on adaptive weighting methods that dynamically assign significance to various modalities [31] and graph-based fusion methods to better model interdependencies [12].

4. Expanding Modalities

Adding other modalities like physiological signals (EEG, heart rate) can better support emotion recognition. Future efforts need to concentrate on building large-scale databases combining physiological and behavioral information [24] and examining the significance of micro-expressions and gestures in Multimodal emotion recognition (MER) [17].

II. RELATED WORK

Multimodal emotion recognition has been intensively studied using a variety of fusion approaches, deep learning architectures, and interpretability methods.

Transformer models have greatly improved MER. Methods that utilize BERT for text, Wav2Vec for speech, and Transformer for video analysis have shown better performance by identifying deeper contextual relationships within each modality [3]. These architectures enhance feature extraction and representation learning, resulting in more efficient emotion recognition.

Cross-modal attention models have further elaborated MER with dynamically balancing different modalities' information. Hierarchy attention networks, especially, have been very effective in aggregating text, acoustic, and visual features into models, making it possible to highlight the most salient emotional cues [31].

Explaining MER models is another prime area of research. Attention visualization, explainable AI (XAI) methods, and interpretable deep learning models are being designed in order to increase model transparency. These developments are especially important for use cases in sensitive fields like health care and mental health analysis, where interpretability is mandatory [12].

Dataset generalization continues to be an issue of importance in MER studies. Most models generalize well across certain datasets but do poorly in the case of unseen data. In an attempt to tackle this, researchers have been looking at domain adaptation strategies like adversarial training and contrastive learning to enhance model robustness across different datasets [24].

Upcoming research directions are to integrate physiological signals (heart rate, EEG) into MER models, promoting self-supervised learning methods, and using large-scale pre-trained models to provide greater adaptability and performance. Recent research also investigated the merging of reinforcement learning and adversarial networks to advance the robustness and real-world usability of MER systems [17]. These developments are likely to define the future of multimodal emotion recognition architectures. Table 1 shows the Key Findings of Fusion Technique based Existing Research Studies.

Table 1: Techniques and Key findings

Model/ Method	***Year***	***Fusion Technique***	***Key Findings***
HuBERT + Wav2Vec2 [16]	2024	Early Fusion	Improved accuracy in speech-text emotion recognition
CNN + LSTM [6]	2023	Early Fusion	Robust to noise and adaptable across datasets
MemoCMT - Transformer-based Cross-modal Attention [7]	2025	Hybrid Fusion	Outperformed traditional models in MER tasks

Model/ Method	*Year*	*Fusion Technique*	*Key Findings*
ResNet + BiLSTM [13]	2021	Early Fusion	Enhanced feature extraction for visual-audio MER
Multimodal GAN [26]	2022	Late Fusion	Improved domain adaptation and generalization
Capsule Graph Networks [4]	2021	Hybrid Fusion	Better accuracy than traditional fusion methods.
Graph Neural Networks [15]	2023	Hybrid Fusion	Improved relationship modeling between modalities

Table 2: Performance Comparisons of recent state-of-the-art models

Model	*Year*	*Ref*	*Technique*	*Modality*	*Dataset*	*Accuracy (%)*
LMR-CBT	2024	[10]	Self-supervised transformer	Video, Audio	IEMOCAP, CMU-MOSI, CMU-MOSEL	81
Parallel-Net	2024	[13]	Cross-attention multimodal fusion	Text, Audio, Visual	IIT-R SIER	89.68
BERT-ViT	2024	[26]	Tensor Product	Visual + Text	WeChat, China	93.65
Emotion-LLaMA	2024	[4]	Instruction-tuned LLM with reasoning	Multimodal (V + T + A)	MERR	69.61
PSiFI-CNN	2024	[10]	Sensor fusion with CNN	Physiological (strain, vocal)	PSiFI Dataset	93.3

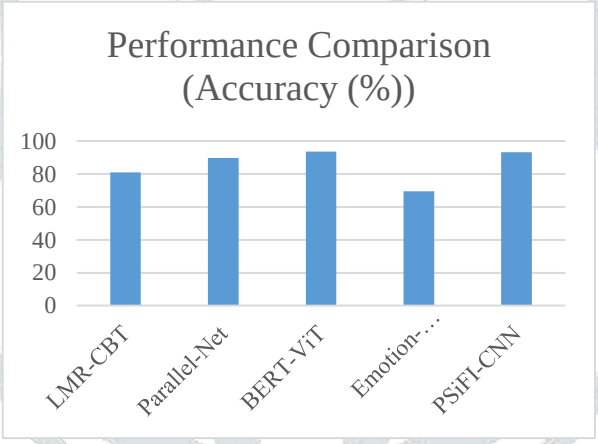

Figure 3: Performance Comparison of Recent Research Methods

Figure 4 shows comparison of model with datasets used for the results generation. This comparison shows that use of LLM model such as BERT and deep learning model CNN with visual and text data gives best results.

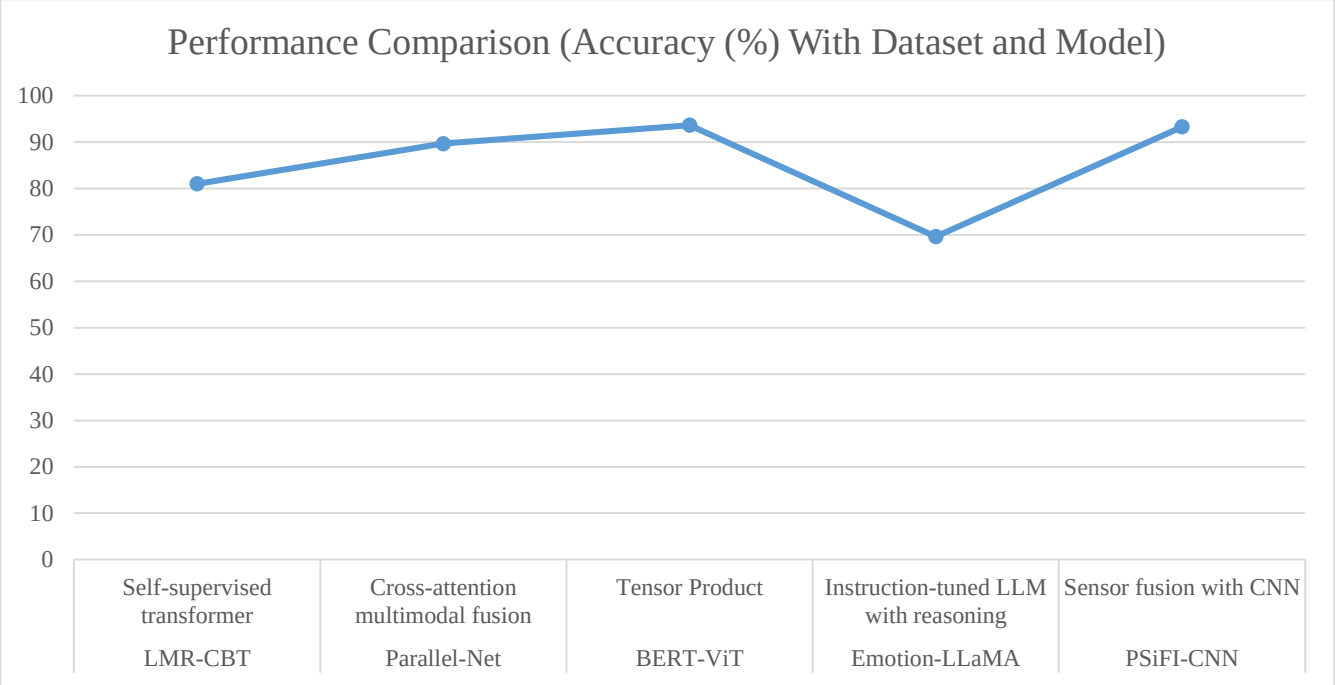

Figure 4 : Performance Comparison of Different Models with Datasets

III. CONCLUSION

In conclusion, our work emphasizes the strong influence of attention mechanisms in enhancing multimodal emotion recognition in an early fusion setting. Through efficient extraction of intricate interactions among text, audio, and video features, the attention-based model outperforms the non-attention method on all major performance metrics consistently, proving its capacity to better understand emotional signals and handle class imbalance. Additionally, research has indicated that attention-based methods like cross-modal embeddings and transformer-based networks provide better generalization across datasets. Nonetheless, the non-attention model is still a viable option in low-computational-resource scenarios, providing competitive performance with lower complexity. Late fusion methods, such as weighted averaging and ensemble approaches, also offer strong alternatives when computational efficiency and interpretability are more important.

V. FUTURE DIRECTIONS

In the future, continued optimization of these architectures, incorporation of other modalities like physiological signals, and use of cutting-edge methods like few-shot learning and large language models will be crucial to increasing generalization and making broader real-world applicability possible. Also, advances in explainability, cross-corpus adaptation, and the creation of lightweight models will be necessary to close the gap between research and real-world use. In order to progress MER, upcoming research must focus on Integrating Large Language Models (LLMs): Utilizing LLMs for multimodal comprehension. Additional research emphasis is required for real-time applicability enhancement by Optimizing lightweight architectures for mobile device deployment and improving transfer learning methods by allowing models to generalize across various datasets.

REFERENCES

[1] Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis. In Proceedings of the 25th International FLAIRS Conference (pp. 202–207). AAAI Press.

[2] Cambria, E., Livingstone, A., & Hussain, A. (2012). The hourglass of emotions. In Lecture Notes in Computer Science (Vol. 7403, pp. 144–157). Springer. https://doi.org/10.1007/978-3-642-35139-6_13

[3] Chakhtouna, A., et al. (2024). Unveiling embedded feature in Wav2Vec2 and HuBERT models for speech emotion recognition. In 5th International Conference on Industry 4.0 & Smart Manufacturing (pp. 2560–2569). https://doi.org/10.1016/j.procs.2024.02.074

[4] Cheng, Z., Cheng, Z.-Q., He, J.-Y., et al. (2024). Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. NeurIPS 2024.

[5] Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. (2024). Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In Proceedings of the 2024 Conference on Neural Information Processing Systems (NeurIPS).

[6] Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., & Zitnik, M. (2023). Multimodal learning with graphs. Nature Machine Intelligence, 5(4), 340–350. https://doi.org/10.1038/s42256-023-00624-6

[7] Fu, Z., Liu, F., Xu, Q., et al. (2024). LMR-CBT: Learning modality-fused representations with CB-Transformer for multimodal emotion recognition from unaligned multimodal sequences. Frontiers of Computer Science, 18, 184314. https://doi.org/10.1007/s11704-023-2444-y

[8] Geetha, A., Vinayakumar, R., Soman, K. P., & Hussain, A. (2024). Multimodal emotion recognition with deep learning. Information Fusion, 105. https://doi.org/10.1016/j.inffus.2024.101000

[9] Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition. Engineering Applications of Artificial Intelligence, 133. https://doi.org/10.1016/j.engappai.2024.106839

[10] He, Q., Li, X., Kim, D. W. N., Jia, X., Gu, X., Zhen, X., & Zhou, L. (2020). Feasibility study of a multi-criteria decision-making based hierarchical model for multi-modality feature and multi-classifier fusion: Applications in medical prognosis prediction. Information Fusion, 55, 207–219. https://doi.org/10.1016/j.inffus.2019.09.001

[11] Ieracitano, C., Adeel, A., Morabito, F. C., & Hussain, A. (2019). A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. Neurocomputing. https://doi.org/10.1016/j.neucom.2019.01.008

[12] Khan, M., Tran, P. N., Pham, N. T., et al. (2025). MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion. Scientific Reports, 15, 5473. https://doi.org/10.1038/s41598-025-89202-x

[13] Kumar, P., Malik, S., & Raman, B. (2024). Interpretable multimodal emotion recognition using hybrid fusion of speech and image data. Multimedia Tools and Applications, 83, 28373–28394. https://doi.org/10.1007/s11042-023-16443-1

[14] Lee, J. P., Jang, H., Jang, Y., et al. (2024). Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface. Nature Communications, 15, 530. https://doi.org/10.1038/s41467-023-44673-2

[15] Lee, J. P., Jang, H., Jang, Y., Kim, H., Choi, S., Park, H., & others. (2024). Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface. Nature Communications, 15, 530. https://doi.org/10.1038/s41467-023-44673-2

[16] Liu, J., et al. (2021). Multimodal emotion recognition with capsule graph convolutional based representation fusion. In ICASSP 2021 (pp. 6339–6343). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9413608

[17] Ma, F., Li, Y., Ni, S., Huang, S.-L., & Zhang, L. (2022). Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN. Applied Sciences, 12(1), 527. https://doi.org/10.3390/app12010527

[18] Mahmud, M., Kaiser, M. S., Hussain, A., & Vassanelli, S. (2018). Applications of deep learning and reinforcement learning to biological data. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2018.2878899

[19] Ouifak, H., & Idri, A. (2023). Application of neuro-fuzzy ensembles across domains: A systematic review of the two last decades (2000–2022). Engineering Applications of Artificial Intelligence, 124. https://doi.org/10.1016/j.engappai.2023.106582

[20] Poria, S., Agarwal, B., Gelbukh, A., Hussain, A., & Howard, N. (2014). Dependency-based semantic parsing for concept-level text analysis. In Lecture Notes in Computer Science (Vol. 8403, pp. 113–127). Springer. https://doi.org/10.1007/978-3-319-04921-2_10

[21] 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. https://doi.org/10.1016/j.inffus.2017.02.003

[22] Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing, 174, 50–59. https://doi.org/10.1016/j.neucom.2015.01.095

[23] Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D., & Bandyopadhyay, S. (2013). Enhanced SenticNet with affective labels for concept-based opinion mining. IEEE Intelligent Systems, 28(2), 31–38. https://doi.org/10.1109/MIS.2013.10

[24] Wang, S.-W., & Yu, S.-N. (2021). Emotion recognition based on photoplethysmography using ResNet and BiLSTM networks. In 2021 International Conference on e-Health and Bioengineering (EHB) (pp. 1–4). IEEE. https://doi.org/10.1109/EHB52898.2021.9657742

[25] Shen, J., & Hussain, A. (2024). AIMDiT: Modality augmentation for emotion recognition in conversations. arXiv. https://arxiv.org/abs/2402.00001

[26] Xiang, A., Qi, Z., Wang, H., Yang, Q., & Ma, D. (2024). A multimodal fusion network for student emotion recognition based on transformer and tensor product. In ICSECE 2024 (pp. 1–4). IEEE. https://doi.org/10.1109/ICSECE61636.2024.10729485

[27] Xiang, A., Qi, Z., Wang, H., Yang, Q., & Ma, D. (2024). A multimodal fusion network for student emotion recognition based on transformer and tensor product. In Proceedings of ICSECE 2024 (pp. 1–4). IEEE. https://doi.org/10.1109/ICSECE61636.2024.10729485

[28] Xiong, F., Sun, B., Yang, X., Qiao, H., Huang, K., Hussain, A., & Liu, Z. (2019). Guided policy search for sequential multitask learning. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(1), 216–226. https://doi.org/10.1109/TSMC.2018.2803220

[29] Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K., & Hussain, A. (2019). Cross-modality interactive attention network for multispectral pedestrian detection. Information Fusion, 50, 20–29. https://doi.org/10.1016/j.inffus.2018.09.008

[30] Zhang, S., Wang, Y., & Hussain, A. (2024). Deep learning-based multimodal emotion recognition. Expert Systems with Applications, 237. https://doi.org/10.1016/j.eswa.2023.120267

[31] Zhou, H., Zhao, Y., Liu, Y., Lu, S., An, X., & Liu, Q. (2023). Multi-sensor data fusion and CNN-LSTM model for human activity recognition system. Sensors, 23(10), 4750. https://doi.org/10.3390/s23104750

[32] Zou, S., Chen, L., & Hussain, A. (2022). Improving multimodal fusion with main modal transformer. Knowledge-Based Systems, 258. https://doi.org/10.1016/j.knosys.2022.109999