Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

Multimodal Emotion Recognition: A Review

[1] Neha Mathur, [2] Paresh Jain

[1] Research Scholar, Department of Computer Science and Engineering, Suresh Gyan Vihar Unviersity, Jaipur
[2] Professor, Department of Electrical Engineering, Suresh Gyan Vihar Unviersity, Jaipur

Abstract:- In the current era of artificial intelligence, recognition of emotion plays a critical role in health care, mental health, in education system for monitoring students' engagement, in drivers stress monitoring in transportation, in security and surveillance, in marketing for analysing customer satisfaction level and in entertainment and social media areas. The researches mostly focussed on single modality i.e. either facial expression, vocal expressions, psychological signals but these approaches often degrade in performance due to less robustness and more sensitive to noise. So, emotion recognition from more than one sources easily addresses above challenges. This paper provides a comprehensive survey of recent advances in multimodal emotion recognition enlightening the fusion strategies, feature extraction techniques, datasets usage and also enlightening their limitations, which provides a valuable reference for researchers and practitioners working with recognition of emotions.

1. Introduction

Emotion is an essential component of human from which they express their feelings for anything. Whatever the human think, feel and their behaviour all is reflected in the form of emotions. Emotion also reflects a human decision and perception towards anything. The recognition of emotion is the process of identifying the emotions of human. It has its application in many areas like while designing of product all pros and cons are monitored based on the decision taken which can be made by monitoring customer experiences imitated by their emotions [1]. The severity of risk on soldiers as well as on pilots can be detected by monitoring the emotions, which is being a major contribution of emotion recognition towards defence and aerospace areas. In public transportation also emotion recognition has its application as by monitoring the emotions of driver while driving, a safety measures can be suggested and thus taken [2]. The emotion recognition plays a vital role in the interaction of human and computer known as human computer interaction (HCI) system under which computerized system recognizes the emotions of human. Recent HCI systems used in various areas such as e-health, e-learning, recommender systems, smart home, smart city and systems like chat bot. The emotions can be recognized from different forms such as speech, aphorisms, facial expression, video, lengthy text, short messages and emojis.[3]

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

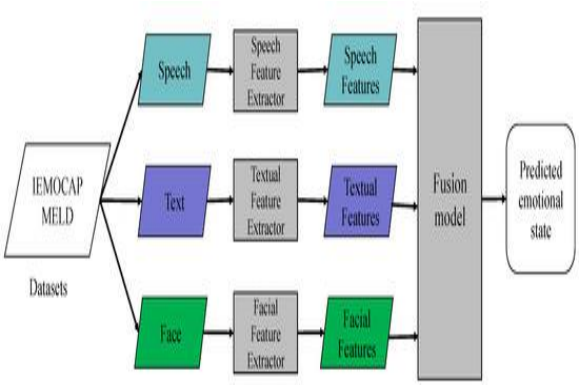

Fig 1. Multimodal emotion recognition framework

The above figure 1 provides the multimodal emotion recognition framework in which it uses a dataset and from a single data set three different modalities can be worked upon in parallel. But each modality is processed separately in different feature extractor. Thereafter the extracted features are fed into a fusion model.

With the technology breakthrough, people are totally relying on smart phones and thus, generally remain active on social media. This generates a huge amount of big data which contains not only the textual information but also videos. The reviews of the consumers can be in any form either text, or recorded videos. The platforms of social media like YouTube, Facebook, Instagram, where users show their reviews, opinions and different ways of using a product from the videos. Many times, such videos also show comparisons between different products of different brands [4]. These videos help buyers to take decisions about a product. Thus, for the recognition of emotions videos are more beneficial to analyse over textual data. As while using videos for analysing, we get multimodal data which provides abundant behavioural cues and both audio and visual data. From the multimodal data, we get clear understanding about the intention of an individual from the facial expression while speaking and audio data

1.1.Scope of this survey

The emotion recognition while using multimodal data provides many cues, which are needed for decision making, thus remain crucial research topic for the research communities related to AI and NLP technology. Thus, there is need for a time-to-time review of literature to define challenges that come across and future work, for the interested researchers. This paper gives a survey of recent advances on multimodal approaches for emotion recognition. Most of multimodal approaches reviewed in this paper uses audio visual data.
The rest of paper is organized as follows section II describes the existing audio-visual multimodal approaches. Finally, section III offers conclusion.

2. Literature Review

There are several ways of expressing emotions by human behaviours such as through facial expression, speech, text, gesture, etc. the psychologists use above ways in judging a human and for making an opinion, psychologists generally use expressions obtained from modalities such as face and voice and sometimes messages. The unimodal system of emotion recognition has limited performance as it is based on either facial modality or vocal modality. In order to improve the performance of systems, multi-modality in expressions is used.

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

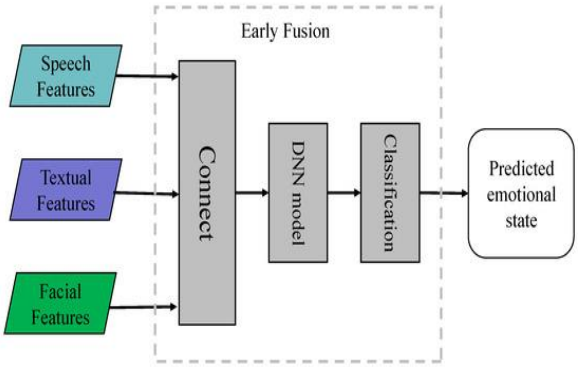

Fig 2. Multimodal emotion recognition framework based on early fusions [4]

The above figure 2 represents a framework of multimodal emotion recognition based on early fusion means initially the features from different modalities are extracted and thereafter fed into the different neural network.

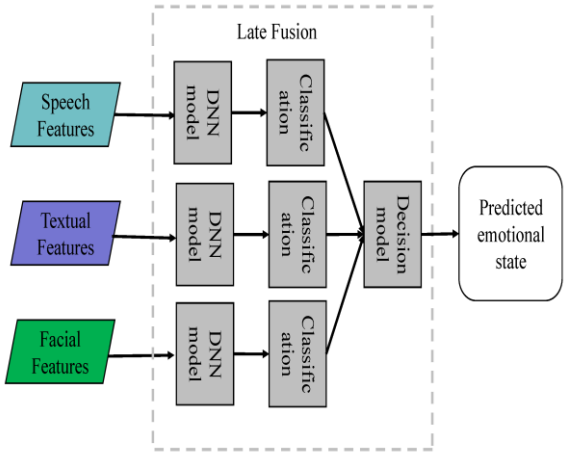

Fig 3. Multimodal emotion recognition framework based on late fusions [4]

The above figure 3 provide the framework for multimodal emotion recognition based on late fusion in which feature extraction and training of feature are done separately for each modality and based on which predictions are made about the features. These predicted results are combined later on to obtain the final emotion category.

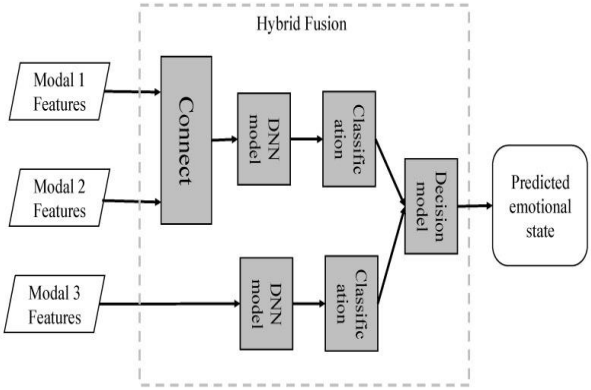

Fig 4. Multimodal emotion recognition framework based on hybrid fusions [4]

The above figure 4 provide the multimodal emotion recognition framework. The development of different strategies involving fusion of modalities is a hot research area [4]. This paper surveys only audio-visual modalities and includes only recent advances instead of all the studies. The following table 1 lists the existing popular data fusion strategies for facial–vocal expression-based emotion recognition with respect to the utilized database, type of emotion categorization, audio, and facial features, recognition methods (i.e. audio classifier (A), visual classifier (V), and audio-visual bimodal fusion approach (AV)), classifier fusion modality, recognition performance, and publication year. The following table 1 provides some approaches for multimodal emotion recognition techniques.

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

Table 1. Multimodal Emotion Recognition Techniques with limitations

Purpose	Methodology	Modality	Database	Limitations	Year
To associate face with voice in a video	Joint audio-visual analysis of videos [5]	A+V	set of 400 randomly sam-pled YouTube videos.	When multiple speakers speaking one after the other without a break accuracy decreases	2017
To associate an audio clip with face image	CNN architecture [6]	A+V	VoxCeleb and VGGFace	Performance decreases for same gender, age and nationality	2018
Audio visual matching	A distance learning method using ELM and a loss function [7]	A+V	VoxCeleb and VGGFace	Does not consider interaction between the different modalities	2019
Biometric matching	Disjoint mapping network (DIMNet) [8]	A+V	VoxCeleb and VGGFace	Performance decreases rapidly as number of values for nationality increases	2019
to model the correlation between features modalities	multimodal attention network (MMAN) [9]	**V+T**	IEMOCAP	This approach does not show good performance for sad emotion category	2020
Lack of labelled data during fusion of modalities	BERT-like SSL models (Deep Bidirectional Transformers) [10]	S+T	IEMOCAP, CMU-MOSEI, and CMU-MOSI	This fusion technique only fuses modalities of text and speech signals and thus not used for videos	2020
To classify text and acoustic data	DCNN and a Bi-direction RNN [11]	**S+T**	IEMOCAP	Problem of overfitting	2020
Cross modal heterogeneous issue between audio and video data	Adversarial metric learning method [12]	A+V	VoxCeleb and VGGFace	Across gender/nationality/age mistakes also occur In matching and retrieval tasks. the AML may produce apparently incorrect results.	2021

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

To extract audio and visual features and integrate them	MoBEL(Mixture of Brain Emotional Learning) [13]	A+V	eNTERFACE'05	Does not give good performance for long videos	2021
To study combined effects of vision and hearing loss while human interaction	Freiburg Activity and Visual Contrast Test	A+V	Geneva Multimodal Emotion Portrayals (GEMEP) [14]	Study only allows measuring the possible acute effects of sensory impairments and thus disregards any long-term adaptation	2021
To deal with late fusion and end-to-end fusion strategy	end-to-end Deep Neural Network (DNN)	A+V	Ryerson Audio–Visual Database of Emotional Speech and Song dataset and Crowd-Sourced Emotional Multi Modal Actors dataset.[15]	the inferior case where uni-modal solution better than multi-modal solution still exists, which suggests data augmentation cannot generalize multi-modal features. However, the ratio of mismatched learned and target patterns are ranging along with the shuffling of the sub-datasets.	2021
to learn interactive information between audio and text modality for multimodal emotion recognition	cross-modal attention [16]	**A+T**	IEMOCAP	It does not classify all the emotions classes angry, happy, sad and neutral	2021
To predict valence of the RECOLA dataset [17]	visual facial expression embedding network	**A+V**	RECOLA	High computational cost and datasets should be labelled	2021

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

Audio-visual emotion recognition with cross-modal attention	Improved GhostNet for facial features; LFCNN for audio feature extraction; tree-like LSTM (tLSTM) for multi-stage fusion; decision-level fusion strategy [18]	**A+VV**	CK+, EMO-DB, MAHNOB-HCI (EEG, speech, faces); IEMOCAP, RAVDESS	Less robust on real-world environment changes which also makes computation more complex	2023
Multimodal sentiment analysis with text, audio, visual fusion	Unified Feature Extraction Network (UFEN); Multi-Task Fusion Network (MTFN); Transformer and attention mechanisms; multi-layer extraction with self-attention	**A+V+T**	MOSI, MOSEI, SIMS	In multimodal features, at some places data integration get missing and also suffers from semantic inconsistencies	2023
Fusion of audio-visual modalities	Deep Canonical Correlation Analysis (DCCA) for feature correlation analysis; separate CNN and 3D CNN extractors for audio and visual features [20]	**A+V**	eNTERFACE05	Not able to capture subtle emotion differences	2024
Dimensional audio-visual emotion recognition	CCC metric optimization [21]	**A+V**	AVEC 2017, AFEW (EmotiW2019)	Small dataset limitations and challenges in handling multimodal asynchrony and	2024

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

				temporal misalignment	
EEG and audio-visual emotion recognition	Modality specific encoders with comparative representation learning [22]	**EEG+A+V**	DFEW and MAFW	Real-time computation constraints; difficulty in EEG data samples	2024
Multimodal emotion recognition with DCNN fusion	Empirical-based fusion (EBF) [23] with three separate DCNN models	**A+V+T**	IEMOCAP, RAVDESS, MELD	Duplicity in features, complex computation, overfitting due to small datasets	2024
Audio-visual emotion recognition with attention mechanisms	Multi-scale channel attention (MCA) module; global interactive fusion (GIF) [24] with attention	A+V	RAVDESS and SAVEE [25]	Model is less robust; found difficulty in handling noise in datasets and efficiently allocation of parameter	2025
Hybrid multi-attention network for audiovisual fusion	Hybrid Attention of Single and Parallel Cross-Modal (HASPCM) [26]	**A+V**	AffWild2 and AFEW-VA	increased computation due to attention mechanisms; less robust towards missing modalities	2025
Multimodal emotion recognition	Language-supervised foundation models with LoRA adaptation [27]	**A+V**	DFEW and MAFW datasets	Heavily dependent on pre-trained models for transfer learning	2025

4. Conclusions and Discussions

Earlier audio-visual emotion recognition models, including CNN- and RNN-based fusion frameworks, primarily rely on early or late fusion strategies. While these methods demonstrate reasonable performance in controlled environments, they often struggle to model fine-grained temporal dependencies and inter-modal correlations. Compared to such approaches, the proposed method consistently achieves superior performance across evaluation metrics, indicating more effective exploitation of complementary audio and visual cues. In particular, conventional methods treat audio and visual modalities either independently or with limited interaction, leading to suboptimal fusion. The improved results of the proposed approach suggest that explicitly modelling cross-modal dependencies is critical for emotion recognition, especially in scenarios involving subtle affective expressions. Recent attention-based methods have shown notable

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

improvements by selectively focusing on emotionally salient features. However, many of these models employ single-level or modality-specific attention, which restricts their ability to capture hierarchical and cross-modal emotional dynamics. Hybrid fusion strategies combining early and late fusion have been proposed to address modality imbalance and noise sensitivity. While these methods improve robustness, they often introduce increased computational complexity without proportional performance gains. Recent foundation-model-based approaches leverage large-scale pretraining and language supervision to enhance emotion understanding. While such models achieve strong results, they typically require extensive computational resources and large annotated datasets.

References:

[1] Andrew, B., Adam, B., Damien, D., Gary, M., Gawain, M. (2017). Dynamic Analysis of Automatic Emotion Recognition Using Generalized Additive Mixed Models. View Source

[2] Imbert, L., Moirand, R., Bediou, B., Koenig, O., Chesnoy, G., Fakra, E., Brunelin, J. (2022). A Single Session of Bifrontal tDCS Can Improve Facial Emotion Recognition in Major Depressive Disorder: An Exploratory Pilot Study. Biomedicines, 10(10), 2397. 10.3390/biomedicines10102397

[3] Agustinus Suradi, , Md Safat Hossain, (2025). Development of an Emotion Recognition System Based on Deep Learning for Human-Computer Interaction Applications. Journal of Electrical, Computer and Informatics, 1(2), 01-10. 10.70062/jeci.v1i2.198

[4] Shou, Yuntao, Tao Meng, Wei Ai, Fangze Fu, Nan Yin, and Keqin Li. "A comprehensive survey on multi-modal conversational emotion recognition with deep learning." *ACM Transactions on Information Systems* (2023).

[5] K. Hoover, S. Chaudhuri, C. Pantofaru, M. Slaney, I. Sturdy. Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers. [Online], Available: https://arxiv.org/abs/1706.00079, 2017.

[6] A. Nagrani, S. Albanie, A. Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, USA, pp. 8427−8436, 2018. DOI:10.1109/CVPR.2018.00879c.

[7] R. Wang, H. B. Huang, X. F. Zhang, J. X. Ma, A. H. Zheng. A novel distance learning for elastic cross-modal audio-visual matching. In Proceedings of IEEE International Conference on Multimedia & Expo Workshops, IEEE, Shanghai, China, pp. 300−305, 2019. DOI: 10.1109/ICMEW.2019.00-70

[8] Y. D. Wen, M. Al Ismail, W. Y. Liu, B. Raj, R. Singh. Disjoint mapping network for cross-modal matching of voices and faces. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, USA, 2019.

[9] Pan, Zexu, Zhaojie Luo, Jichen Yang, and Haizhou Li. "Multi-modal attention for speech emotion recognition." *arXiv preprint arXiv:2009.04107* (2020)

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

[10] Siriwardhana, Shamane, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. "Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition." *arXiv preprint arXiv:2008.06682* (2020).

[11] Priyasad, Darshana, Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. "Attention driven fusion for multi-modal emotion recognition." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3227-3231. IEEE, 2020

[12] A. H. Zheng, M. L. Hu, B. Jiang, Y. Huang, Y. Yan, B. Luo. Adversarial-metric learning for audio-visual cross-modal matching. IEEE Transactions on Multimedia, 2021. DOI: 10.1109/TMM.2021.3050089.

[13] Farhoudi, Z., Setayeshi, S., Razazi, f., Rabiee, A., Razazi, F., Iran, P. o. D. o. E. a. C. E. S. a. R. B. I. A. U. T., Rabiee, A., Iran, P. o. D. o. C. S. D. B. I. A. U. I. (2020). Emotion recognition based on multimodal fusion using mixture of brain emotional learning. 10.30699/icss.21.4.113

[14] Yan, Jingjie, Peiyuan Li, Chengkun Du, Kang Zhu, Xiaoyang Zhou, Ying Liu, and Jinsheng Wei. "Multimodal Emotion Recognition Based on Facial Expressions, Speech, and Body Gestures." *Electronics (2079-9292)* 13, no. 18 (2024).

[15] Luna-Jiménez, Cristina, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M. Montero, and Fernando Fernández-Martínez. "Multimodal emotion recognition on RAVDESS dataset using transfer learning." *Sensors* 21, no. 22 (2021): 7665.

[16] Krishna, D. N. "Using large pre-trained models with cross-modal attention for multi-modal emotion recognition." *arXiv preprint arXiv:2108.09669* 2 (2021).

[17] Schoneveld, Liam, Alice Othmani, and Hazem Abdelkawy. "Leveraging recent advances in deep learning for audio-visual emotion recognition." *Pattern Recognition Letters* 146 (2021): 1-7

[18] Pan, Jiahui, Weijie Fang, Zhihang Zhang, Bingzhi Chen, Zheng Zhang, and Shuihua Wang. "Multimodal emotion recognition based on facial expressions, speech, and EEG." *IEEE Open Journal of Engineering in Medicine and Biology* 5 (2023): 396-403.

[19] Cheng, Hongju, Zizhen Yang, Xiaoqi Zhang, and Yang Yang. "Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion." *IEEE transactions on affective computing* 14, no. 4 (2023): 3149-3163.

[20] Tuerhong, Gulanbaier, Yelei Jin, and MAIRIDAN WUSHOUER. "Emotion Analysis in Speech Based on Audio-Visual Fusion." *Available at SSRN 4857339*.

[21] Praveen, R. Gnana, and Jahangir Alam. "Cross-attention is not always needed: Dynamic cross-attention for audio-visual dimensional emotion recognition." In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6. IEEE, 2024.

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 391-400

[22] Lee, Ju-Hwan, Jin-Young Kim, and Hyoung-Gook Kim. "Emotion recognition using eeg signals and audiovisual features with contrastive learning." *Bioengineering* 11, no. 10 (2024): 997.

[23] Kulkarni, Shailesh, S. S. Khot, and Yogesh Angal. "Deep Multimodal Fusion Convolutional Neural Network for Emotion Recognition." *Library of Progress-Library Science, Information Technology & Computer* 44, no. 3 (2024).

[24] Fu, Ziwang, Feng Liu, Hanyang Wang, Jiayin Qi, Xiangling Fu, Aimin Zhou, and Zhibin Li. "A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition." *arXiv preprint arXiv:2111.02172* (2021).

[25] Zhang, Peng, Hui Zhao, Meijuan Li, Yida Chen, Jianqiang Zhang, Fuqiang Wang, and Xiaoming Wu. "Audio-Visual Emotion Recognition Based on Multi-Scale Channel Attention and Global Interactive Fusion." In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2144-2150. IEEE, 2023.

[26] Moorthy, Sathishkumar, and Yeon-Kug Moon. "Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion." *Mathematics* 13, no. 7 (2025): 1100.

[27] Zhao, Jiaxing, Xihan Wei, and Liefeng Bo. "R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning." *arXiv preprint arXiv:2503.05379* (2025).

Correspondence to: Neha Mathur Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur, India
Corresponding author. E-mail addresses: nmdoll@gmail.com