Gated Cross-Modal Fusion Mechanism for Audio-Video-based Emotion Recognition

Himanshu Kumar
Department of Computer Science, Central University of Tamil Nadu, Thiruvarur, Tamil Nadu, India
himanshukphd20@students.cutn.ac.in (corresponding author)

Martin Aruldoss
Department of Computer Science, Central University of Tamil Nadu, Thiruvarur, Tamil Nadu, India
martin@cutn.ac.in

Received: 28 October 2024 | Revised: 22 November 2024 | Accepted: 27 November 2024

ABSTRACT

Due to its potential uses in security, surveillance, mental health monitoring, and human-computer interaction, artificial emotion recognition employing video and audio modalities has attracted a lot of attention. This study focuses on optimal cross-modal fusion techniques to enhance the precision and robustness of multimodal audio-video-based emotion recognition. Specifically, this study introduces a gated cross-modal fusion mechanism in audio-video-based emotion recognition, known as Compact Bilinear Gated Pooling (CBGP). The novelty of this work is that CBGP fusion is being applied to the emotion recognition task for the first time to integrate the extracted features and reduce the dimensionality of the audio and video modalities using 1DCNN and 3DCNN deep neural architectures, respectively. This novel approach was tested and verified on three benchmark datasets: CMU-MOSEI, RAVDESS, and IEMOCAP, each containing multimodal data representing a range of emotions, including happy, sad, fear, anger, neutral, and disgust. Experimental results show that CBGP consistently outperformed state-of-the-art fusion techniques, such as early fusion, late fusion, hybrid fusion, and others. CBGP extracts the relevant features, leading to higher accuracy and F1 scores due to its dynamic gating mechanism that selectively emphasizes relevant feature interactions. This study suggests that the integration of gating mechanisms within fusion processes is vital to improve emotion recognition. Future work will focus on extending these findings to real-time applications, exploring multitask learning frameworks, and enhancing the interpretability of multimodal emotion recognition systems.

Keywords-attention mechanism; bilinear pooling; emotion recognition; feature extraction; fusion strategies

I. INTRODUCTION

Emotion recognition utilizing multimodal data has evolved as an important area of research, with applications ranging from human-computer interaction to mental health monitoring and security. Researchers aim to develop systems that can properly discern human emotions by combining audio and video inputs. This study addresses the challenge of improving fusion strategies to enhance the performance of audio-video-based emotion recognition systems. Existing traditional approaches, such as intramodal, intermodal, and hybrid fusion techniques, have shown nominal efficiency and limited improvement due to difficulties in capturing complex correlations among cross-model features. Intramodal fusion involves integrating and exploring correlations of features within a single modality, such as combining different audio features or various video descriptors. Some attention-based techniques, such as relation-attention, self-attention, and transformer-attention, achieve effective intramodal fusion. In contrast, intermodal fusion combines features across different modalities, integrating cross-model audio and video features to create a cohesive representation.

This study introduces a gated cross-modal fusion-based approach to address the complex challenge of integrating features from audio and video modalities, which are inherently different in nature. Single modality emotion recognition relies on a single data source, making it simpler but limited in scope. In contrast, multimodal-based emotion recognition is more complex and challenging, requiring the synchronization of diverse data and the resolution of conflicting signals. The gated mechanism in the fusion approach is designed to capture the intricate relationships between audio and video features, enhancing the system's ability to recognize emotions with greater accuracy. The contributions of this study are as follows:

- Applies Compact Bilinear Gated Pooling (CBGP) to the multimodal emotion recognition task for the first time.
- Tests and validates the gated cross-modal fusion approach experimentally in audio-video-based emotion recognition.

- Adopts the optimal cross-model fusion mechanisms to fuse audio and video features.
- Implements, trains, and validates the model across three different datasets.
- The proposed CBGP cross-model fusion outperforms the state-of-the-art fusion models, such as early fusion, late fusion, and hybrid fusion, in addressing audio-video-based emotion recognition challenges.
- Highlights the importance of the gating fusion mechanism and provides a robust fusion method with good accuracy and versatility.

Experiments were conducted on three well-known datasets: IEMOCAP [1], RAVDESS [2], and CMU-MOSEI [3]. The IEMOCAP dataset provides a diverse set of emotions captured through audio-visual recordings, while RAVDESS offers a controlled environment with professional actors portraying various emotional states. CMU-MOSEI is a large-scale multimodal sentiment analysis dataset that offers a rich source for evaluating the effectiveness of the proposed fusion strategy.

The fusion mechanism of audio and video features is crucial in multimodal systems, especially those incorporating humanoid robots [4, 5] or other intelligent agents [6], as it allows realistic and successful human-machine interaction. This integration of features enables the system to detect, evaluate, and respond to a variety of human behaviors and emotions. To achieve this, feature extraction is a critical process that ensures that the system focuses on the most appealing and pertinent areas of object information that it receives while minimizing processing complexity. Feature extraction is the process of transforming raw audio and video input into measurable and relevant qualities for machine learning algorithms. The features extracted from the audio and video modalities are fused with a simple concatenation of feature vectors [7].

In multimodal systems, good feature extraction is crucial for accurately expressing the multimodal input while minimizing data dimensions. Feature selection is the process of determining and selecting the most informative features of an extracted collection. This phase is critical for lowering the dimensionality in the feature space, increasing computing efficiency, and improving the performance of machine learning models. In recent years, deep learning has led to advances in numerous sectors, including healthcare [8], speech recognition [9], music generation [10], electroencephalogram signals [11], and e-learning [12]. Deep neural networks have the advantage of training models efficiently with multimodal data, which also deal with autoencoders [13], intermodal [14], intramodal [15], and generalization techniques using Restricted Boltzmann Machines (RBM) [16].

Deep learning models outperform traditional hand-crafted traditional features [17, 18], and multimodal approaches outperform unimodal-based emotion recognition [19-21]. Attention models have achieved significant improvements when combined with deep neural networks, such as pre-trained networks, including AlexNet [22], VGG16 [23-25], VGGFACE [26], and ResNet [27-29], to automatically extract and fuse features. In the fusion mechanism, previous studies have highlighted the importance of combining the most likely features to achieve notable results. Some of the most widely used fusion mechanisms are feature-level fusion [30-32], decision-level fusion [33-34], hybrid fusion[35-36], attention-based fusion [37-38], and rule-based fusion [39-41].

This study fills the research gap on cross-modal fusion, highlighting the significance of this approach in capturing complex audio-video features. It also provides a comparative experimental performance analysis on benchmark datasets, demonstrating the effectiveness of integrating gating mechanisms in cross-modal fusion.

II. MATERIALS AND METHODS

A. Audio-Video Features and Feature Extraction

The process starts with the video input, extracting both the audio signals and image frames. A 1DCNN [42] was used to extract the audio features, while a 3DCNN [43] was used to extract facial features from the visual modality. The extracted features from both modalities are then fed into a gated cross-modal fusion mechanism layer (CBGP) to fuse the extracted features. Figure 1 illustrates the process from feature extraction to emotion classification, with specific subdivisions in the cross-modal fusion mechanism, including Factorized Bilinear Pooling (FBP), Compact Bilinear Pooling (CBP), and CBGP.

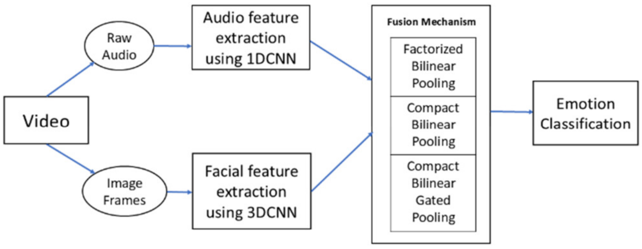
Video
Raw Audio
Image Frames
Audio feature extraction using 1DCNN
Facial feature extraction using 3DCNN
Fusion Mechanism
Factorized Bilinear Pooling
Compact Bilinear Pooling
Compact Bilinear Gated Pooling
Emotion Classification

Fig. 1. Basic architecture of the audio-video-based cross-modal fusion mechanism.

B. Factorized Bilinear Pooling (FBP)

FBP improves traditional bilinear pooling by factorizing the interaction into a lower-rank approximation, reducing overfitting and computational burdens. This allows FBP to capture the more relevant cross-modal interactions with fewer parameters while preserving the expressive power.

$$Z = \sum_{i=1}^{m} (M^T A) . (N^T V) \quad (1)$$

where Z is the pooled feature vector, M and N are bilinear interaction matrices, and A and V are feature vectors from audio and video, respectively.

C. Compact Bilinear Pooling (CBP)

CBP refines bilinear pooling by using a compact representation of bilinear interactions through the tensor sketch technique. It approximates the outer product of feature vectors with random projections, reducing dimensionality while preserving major attentive cross-model interactions.

$$Z = \sum_{i=1}^{m} (proj_a(A)_i) . (proj_v(V_i)) \quad (2)$$

where Z is the pooled feature vector, A and V are feature vectors from audio and video, respectively, and $proj_a$ and $proj_v$ are projection matrices of audio and video features.

D. Compact Bilinear Gated Pooling (CBGP)

CBGP enhances and builds on CBP by adding a gating mechanism that adjusts or selectively emphasizes features based on their relevance, using a learned softmax function to modulate feature interactions before pooling. In CBGP, the feature vectors A and V undergo compact bilinear pooling as described in CBP, but before the final summation, the resulting interaction vector is element-wise multiplied by a gating vector $G' \in R^d$, where d is the dimensionality of the compact representation. The gating vector is computed as:

$$G' = \sigma(W_G(A', V') + b_G) \quad (3)$$

where σ is the softmax function, W_G is the weight matrix, b_G is the bias vector, and A', V' are audio and video feature vectors.

1) Training Method

Let A' and V' represent the audio and video modalities. The feature extraction functions f_a and f_v are applied to the audio and video modalities, generating the feature vectors:

$$F_A = f_a(A')$$

$$F_v = f_v(V')$$

CBP uses random projections to project the high-dimensional feature vectors into a lower-dimensional space before combining them.

- Random projection for audio features: $Z_A = P_A F_A$
- Random projection for video Features: $Z_V = P_V F_V$

where Z_A and Z_V are the projections of audio and video, and P_A and P_V are the projection matrices of audio and video features. Then, an element-wise multiplication of the projected vectors is calculated:

$$Z' = Z_A \circ Z_V \quad (4)$$

For gated pooling, the gating vector $G' \in R^d$ is calculated, where, d is the dimensionality of the compact representation. The gating vector is calculated by:

$$G' = \sigma(W_G(A', V') + b_G) \quad (5)$$

where σ is the softmax function. Then, the gating mechanism is applied to the element-wise multiplied vector:

$$Z'' = G' \circ Z' \quad (6)$$

Finally, the elements of the gated interaction vector are summed to obtain the final pooled vector using:

$$Z = Sum(Z'') \quad (7)$$

This entire mechanism can be summarized by:

$$Z = \sum_{i=1}^{m} (\sigma(W_G\,(A, V) + b_G))_i . (Z_A)_i) . (Z_V(V_i) \quad (8)$$

where Z is the pooled feature vector, and Z_A and Z_V are the projections of audio and video.

This mathematical analysis shows that CBGP can identify the optimal fusion approaches that can be applied to audio-video-based emotion recognition systems, ultimately contributing to the development of more robust and accurate emotion recognition technologies. The gating mechanism allows controlling the flow of information between the layers while selecting and rejecting the relevant or non-relevant (based on correlation feature score) inputs. As not all features are equally important at every step or time frame, the gating mechanism dynamically assigns weights to features to capture complex regions more effectively.

2) How the Gating Mechanism Works

A gating unit is applied to control the flow of information, acting as a filter and deciding which features from a modality should be passed on and which should be suppressed. The gating vector is calculated using (3). In this approach, features are audio and video modalities that interact through the outer product. The outer product allows the 1DCNN and 3DCNN to capture the interactions between every feature of one modality and every feature of the other modality in a compact manner. However, the full outer product is computationally expensive due to the high dimensionality it generates, but compact bilinear pooling reduces the high dimensionality. This process ensures that the combined feature representation is both expressive and compact. The gating mechanism is integrated into the pooling process to selectively emphasize the relevant features and suppress irrelevant ones. The gating vector is learned alongside the fused features and the gate assigns weights to different features dynamically. Extracted features are fed as input to the CBGP layer, which combines them into a unified representation. This fused representation is passed to the softmax function and the subsequent (output) layer for classification.

III. RESULTS AND DISCUSSION

The performance of CBGP was evaluated and compared with the state-of-the-art mechanisms, including early fusion, late fusion and hybrid fusion, on the IEMOCAP, RAVDESS, and CMU-MOSEI datasets.

A. Datasets

- IEMOCAP [1]: This dataset contains 12 hours of audiovisual data from 10 actors engaging in scripted and improvised dialogues, annotated with emotions such as happiness, sadness, anger, and neutrality.
- RAVDESS [2]: This dataset contains 24 professional actors performing 7350 clips of emotional speech and singing, annotated for emotions such as happiness, sadness, anger, and fear.
- CMU-MOSEI [3] comprises 23,453 video segments from over 1000 speakers, annotated with sentiment and emotion labels, showcasing a broad spectrum of emotions in diverse contexts.

The results are summarized in the tables below, highlighting the contributions of each fusion method to the overall system performance.

TABLE I. PERFORMANCE COMPARISON ON IEMOCAP [1]

Fusion mechanism	Accuracy (%)	F1-score (%)	Remarks
Early fusion	55.4	54.9	Limited features interacted
Late fusion	59.02	58.1	Lacks fine-grained feature interaction
Hybrid fusion	62.5	61.8	Improved over early and late fusion
CBGP	79.2	78.3	Best performance, particularly for nuanced emotions

Table I shows that in the IEMOCAP dataset, CBGP outperforms the state-of-the-art fusion mechanisms, achieving the highest accuracy and F1 score. The gating mechanism in CBGP effectively captures subtle emotional cues, leading to improved classification, especially in cases of complex emotions such as anger and sadness.

TABLE II. PERFORMANCE COMPARISON ON RAVDESS [2]

Fusion mechanism	Accuracy (%)	F1-score (%)	Remarks
Early fusion	70.5	69.4	Limited features interacted
Late fusion	72.6	71.2	Lacks fine-grained feature interaction
Hybrid fusion	75.8	74.3	Moderate improvement
CBGP	85.1	84.6	Excels in both speech and song emotions

Table II shows that in the RAVDESS dataset, CBGP again leads, demonstrating its ability to manage the diverse range of emotional expressions in both speech and songs. The gating mechanism proves beneficial in distinguishing emotions that are expressed differently in spoken versus sung formats.

TABLE III. PERFORMANCE COMPARISON ON CMU-MOSEI

Fusion mechanism	Accuracy (%)	F1-score (%)	Remarks
Early fusion	67.3	65.4	Limited features interacted
Late fusion	70.4	69.2	Lacks fine-grained feature interaction
Hybrid fusion	72.6	71.4	Moderate improvement
CBGP	80.3	79.2	Best for fine-grained emotion detection

Table III shows that in the CMU-MOSEI dataset, CBGP achieved the highest scores, particularly excelling at recognizing fine-grained emotions. Its ability to dynamically adjust the importance of different feature interactions allows it to handle nuanced and varied expressions found in the dataset.

B. Ablation Study

To evaluate the performance of CBGP, an ablation study was carried out comparing CBGP to classify and evaluate its effectiveness with state-of-the-art fusion methods, including early fusion, late fusion, and hybrid fusion. The aim was to assess the impact of the CBGP mechanism on the fusion of features and dimensionality reduction when applied to multimodal data, using 1DCNN for audio and 3DCNN for video. Figure 2 illustrates the performance comparison of the baseline fusion models and gated cross-modal fusion (early, late, hybrid, and CBGP) in terms of accuracy and F1 scores across the IEMOCAP, RAVDESS, and CMU-MOSEI datasets.

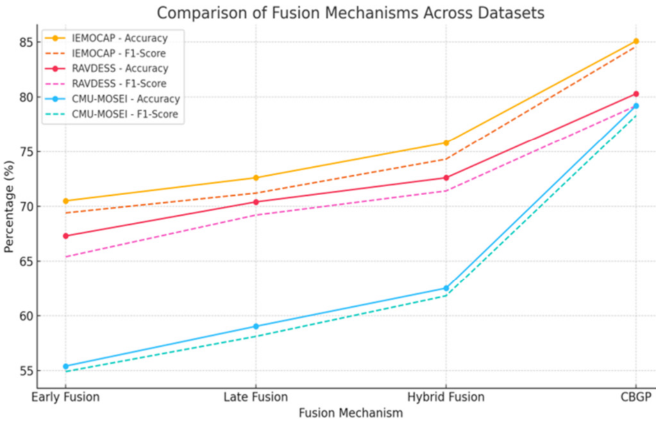

Fig. 2. Performance comparison of the baseline fusion model and gated cross-modal fusion (early, late, hybrid, and CBGP) in terms of accuracy and F1 scores across the IEMOCAP, RAVDESS, and CMU-MOSEI datasets.

C. Feature Extraction

The audio and video modalities undergo independent feature extraction processes using specialized neural networks. For audio, a 1DCNN is used to capture temporal and spectral patterns effectively without losing resolution. For video, a 3DCNN is used to extract facial features by simultaneously processing the spatial and temporal dimensions of image frames, allowing the network to understand motion and spatial structures.

D. Model Training

The 1DCNN architecture for audio data processes MFCC features extracted from spectrograms, effectively capturing temporal patterns. It consists of three Conv1D layers (filters: 64, 128, 256) with ReLU activation, batch normalization, max pooling, and dropout (20-40%) to reduce overfitting. These layers are followed by a flatten layer and two fully connected dense layers (units: 128, 64) with additional dropout (50%). The model employs L2 regularization, the Adam optimizer with a learning rate of 0.001, and is trained with categorical cross-entropy loss.

The 3DCNN extracts spatial and temporal features from video frames using three 3D convolutional layers (filters: 32, 64, 128; kernel size: 3×3×3), 3D max-pooling layers, a dropout rate of 0.3, and two fully connected layers with 256 and 128 units. Both architectures were trained using the Adam optimizer with learning rates of 0.001 (1DCNN) and 0.0005 (3DCNN), cross-entropy loss, and L2 weight regularization ($\lambda = 0.0001$ for 1DCNN and $\lambda = 0.0005$ for 3DCNN) over 50 epochs. Regularization techniques, including batch normalization, dropout, and early stopping, were employed to prevent overfitting, while data augmentation (time shifting for audio, flipping for video) enhanced generalization. Together, these architectures effectively captured multimodal features, allowing robust emotion recognition when integrated with the CBGP mechanism.

The results clearly show that the focused weighting of key features from the correlation weight matrix through dynamic gating is critical to improve the performance in audio-video-based emotion recognition tasks. The primary advantage of

CBGP lies in its gating mechanism, which enhances the model's ability to focus on the most relevant features for emotion recognition. On applying the softmax function, this mechanism provides the ability to measure the weight of features that are more important and helps to predict the correct emotion. This is particularly important in datasets such as IEMOCAP and CMU-MOSEI, where emotions are often expressed subtly and can overlap with sentiments. While FBP offers computational efficiency and CBP provides a good balance between complexity and performance, CBGP's ability to adaptively gate feature interactions makes it a superior choice for complex emotion recognition tasks.

Tables IV and V illustrate the trade-offs between computational efficiency and accuracy for all fusion mechanisms. This suggests that future work in multimodal emotion recognition should consider the integration of gating mechanisms to further enhance model performance.

TABLE IV. TRADE-OFFS BETWEEN COMPUTATIONAL EFFICIENCY AND ACCURACY FOR ALL FUSION MECHANISMS

Fusion technique	Memory usage	Processing time
Early Fusion	High	High
Late Fusion	Low	Low
Hybrid Fusion	Moderate	Moderate
CBGP	Moderate	Moderate

TABLE V. TRADE-OFFS BETWEEN COMPUTATIONAL EFFICIENCY AND ACCURACY FOR ALL FUSION MECHANISMS

Fusion technique	Model complexity	Computational cost	Accuracy
Early Fusion	High	High	Moderate to High
Late Fusion	Low	Low to Moderate	Moderate
Hybrid Fusion	Moderate	Moderate	High
CBGP	Moderate	Moderate to High	High

E. Future Directions

The insights gained from this study open several avenues for future research and development in the domain of multimodal emotion recognition.

1) Extended Dataset Exploration

Although this study focused on three datasets, future research could explore the application of these fusion mechanisms to other datasets that contain different cultural contexts, languages, or more complex multimodal interactions, such as those involving text or physiological signals (e.g., EEG, heart rate).

2) Real-Time Emotion Recognition

Implementing these fusion mechanisms in real-time emotion recognition systems is a promising direction. This would involve optimizing the computational efficiency of CBGP and other fusion techniques to meet the requirements of real-time applications, such as human-computer interaction, social robots, or virtual reality environments.

3) Multi-Task Learning

Future studies could explore the integration of these fusion mechanisms within a multitask learning framework, where the model is simultaneously trained to perform related tasks, such as emotion recognition, sentiment analysis, and action recognition. This could potentially improve the model's robustness and generalizability across tasks.

4) Fusion Mechanism Enhancements

Further refinement of the CBGP mechanism could involve experimenting with different gating functions or learning strategies, such as attention mechanisms, to dynamically adjust the importance of different features based on context. Additionally, exploring hybrid approaches that combine the strengths of different fusion techniques could lead to even more powerful emotion recognition systems.

5) Cross-Domain Transfer Learning

Given the variability in emotional expression across different domains (e.g., movies, social media, conversational agents), future work could investigate the use of transfer learning techniques to adapt the CBGP model to new domains with limited annotated data, enhancing its applicability and effectiveness in diverse real-world scenarios.

6) Interpretability and Explainability

As multimodal emotion recognition systems are increasingly deployed in sensitive applications, ensuring their interpretability and explainability becomes crucial. Future research could focus on developing techniques that provide insight into how and why the model arrives at certain emotional classifications, making the systems more transparent and trustworthy.

IV. CONCLUSION

This study investigated the effectiveness of fusing features by applying the gated pooling mechanism and a robust deep learning model for audio-video-based artificial emotion recognition. The performance of the proposed gated fusion mechanism was evaluated and compared with state-of-the-art fusion techniques, such as early fusion, late fusion, and hybrid fusion mechanisms on the IEMOCAP, RAVDESS, and CMU-MOSEI datasets using accuracy and F1-score. The performance of each fusion technique was evaluated across various emotional categories, including happiness, sadness, fear, anger, neutrality, and disgust. The experimental results demonstrate that the CBGP mechanism outperforms the other fusion techniques across all datasets, consistently achieving higher accuracy and F1 scores. The gating mechanism integrated within the CBGP enables the model to selectively emphasize relevant feature interactions that are crucial to accurately recognizing complex and nuanced emotional expressions. Overall, the findings of this study suggest that incorporating a gating mechanism in multimodal fusion processes can significantly enhance the performance of emotion recognition systems, making CBGP a promising approach for future developments in various fields and applications, such as mental health and well-being, improving communication, identifying security threats and consumer behavior, and in cross-disciplinary research.

REFERENCES

[1] C. Busso *et al.*, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008, https://doi.org/10.1007/s10579-008-9076-6.

[2] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, 2018, Art. no. e0196391, https://doi.org/10.1371/journal.pone.0196391.

[3] S. Sahay, S. H. Kumar, R. Xia, J. Huang, and L. Nachman, "Multimodal Relational Tensor Network for Sentiment and Emotion Classification," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, Melbourne, Australia, 2018, pp. 20–27, https://doi.org/10.18653/v1/W18-3303.

[4] A. T. Eyam, W. M. Mohammed, and J. L. Martinez Lastra, "Emotion-Driven Analysis and Control of Human-Robot Interactions in Collaborative Applications," *Sensors*, vol. 21, no. 14, Jul. 2021, Art. no. 4626, https://doi.org/10.3390/s21144626.

[5] S. Saxena, S. Tripathi, and S. Tsb, "Deep Robot-Human Interaction with Facial Emotion Recognition Using Gated Recurrent Units & Robotic Process Automation," in *Frontiers in Artificial Intelligence and Applications*, A. J. Tallón-Ballesteros and C. H. Chen, Eds. IOS Press, 2020.

[6] M. Amorim, Y. Cohen, J. Reis, and M. Rodrigues, "Exploring Opportunities for Artificial Emotional Intelligence in Service Production Systems," *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 1145–1149, 2019, https://doi.org/10.1016/j.ifacol.2019.11.350.

[7] S. Zhou, X. Wu, F. Jiang, Q. Huang, and C. Huang, "Emotion Recognition from Large-Scale Video Clips with Cross-Attention and Hybrid Feature Weighting Neural Networks," *International Journal of Environmental Research and Public Health*, vol. 20, no. 2, Jan. 2023, Art. no. 1400, https://doi.org/10.3390/ijerph20021400.

[8] A. P. Shah, V. Vaibhav, V. Sharma, M. Al Ismail, J. Girard, and L.-P. Morency, "Multimodal Behavioral Markers Exploring Suicidal Intent in Social Media Videos," in *2019 International Conference on Multimodal Interaction*, Suzhou, China, Oct. 2019, pp. 409–413, https://doi.org/10.1145/3340555.3353718.

[9] S. C. Venkateswarlu, S. R. Jeevakala, N. U. Kumar, P. Munaswamy, and D. Pendyala, "Emotion Recognition From Speech and Text using Long Short-Term Memory," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11166–11169, Aug. 2023, https://doi.org/10.48084/etasr.6004.

[10] R. Madhok, S. Goel, and S. Garg, "SentiMozart: Music Generation based on Emotions:," in *Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, Funchal, Madeira, Portugal, 2018, pp. 501–506, https://doi.org/10.5220/0006597705010506.

[11] J. Cho and H. Hwang, "Spatio-Temporal Representation of an Electoencephalogram for Emotion Recognition Using a Three-Dimensional Convolutional Neural Network," *Sensors*, vol. 20, no. 12, Jun. 2020, Art. no. 3491, https://doi.org/10.3390/s20123491.

[12] O. El Hammoumi, F. Benmarrakchi, N. Ouherrou, J. El Kafi, and A. El Hore, "Emotion Recognition in E-learning Systems," in *2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, Rabat, Morocco, May 2018, pp. 1–6, https://doi.org/10.1109/ICMCS.2018.8525872.

[13] D. Lakshmi and R. Ponnusamy, "Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders," *Microprocessors and Microsystems*, vol. 82, Apr. 2021, Art. no. 103834, https://doi.org/10.1016/j.micpro.2021.103834.

[14] D. Hu, C. Chen, P. Zhang, J. Li, Y. Yan, and Q. Zhao, "A Two-Stage Attention Based Modality Fusion Framework for Multi-Modal Speech Emotion Recognition," *IEICE Transactions on Information and Systems*, vol. E104.D, no. 8, pp. 1391–1394, Aug. 2021, https://doi.org/10.1587/transinf.2021EDL8002.

[15] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational Transformer Network for Emotion Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021, https://doi.org/10.1109/TASLP.2021.3049898.

[16] I. M. Revina and W. R. S. Emmanuel, "A Survey on Human Face Expression Recognition Techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, Jul. 2021, https://doi.org/10.1016/j.jksuci.2018.09.002.

[17] W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung, "Multimodal End-to-End Sparse Model for Emotion Recognition," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5305–5316, https://doi.org/10.18653/v1/2021.naacl-main.417.

[18] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition." arXiv, Jul. 06, 2019, https://doi.org/10.48550/arXiv.1907.03196.

[19] A. Petrova, D. Vaufreydaz, and P. Dessus, "Group-Level Emotion Recognition Using a Unimodal Privacy-Safe Non-Individual Approach," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, Oct. 2020, pp. 813–820, https://doi.org/10.1145/3382507.3417969.

[20] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017, https://doi.org/10.1016/j.inffus.2017.02.003.

[21] W. Ismaiel, A. Alhalangy, A. O. Y. Mohamed, and A. I. A. Musa, "Deep Learning, Ensemble and Supervised Machine Learning for Arabic Speech Emotion Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13757–13764, Apr. 2024, https://doi.org/10.48084/etasr.7134.

[22] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Machine Vision and Applications*, vol. 30, no. 5, pp. 975–985, Jul. 2019, https://doi.org/10.1007/s00138-018-0960-9.

[23] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial Emotion Recognition Using Transfer Learning in the Deep CNN," *Electronics*, vol. 10, no. 9, Apr. 2021, Art. no. 1036, https://doi.org/10.3390/electronics10091036.

[24] N. Hajarolasvadi and H. Demirel, "Deep facial emotion recognition in video using eigenframes," *IET Image Processing*, vol. 14, no. 14, pp. 3536–3546, Dec. 2020, https://doi.org/10.1049/iet-ipr.2019.1566.

[25] K. H. Cheah, H. Nisar, V. V. Yap, C.-Y. Lee, and G. R. Sinha, "Optimizing Residual Networks and VGG for Classification of EEG Signals: Identifying Ideal Channels for Emotion Recognition," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–14, Mar. 2021, https://doi.org/10.1155/2021/5599615.

[26] M. Quiroz, R. Patiño, J. Diaz-Amado, and Y. Cardinale, "Group Emotion Detection Based on Social Robot Perception," *Sensors*, vol. 22, no. 10, May 2022, Art. no. 3749, https://doi.org/10.3390/s22103749.

[27] K. L. Lakshmi *et al.*, "Recognition of emotions in speech using deep CNN and RESNET," *Soft Computing*, Mar. 2023, https://doi.org/10.1007/s00500-023-07969-5.

[28] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention Augmented Convolutional Networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 3285–3294, https://doi.org/10.1109/ICCV.2019.00338.

[29] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets," *Electronics*, vol. 11, no. 22, Nov. 2022, Art. no. 3831, https://doi.org/10.3390/electronics11223831.

[30] Q. Liu, L. Dong, Z. Zeng, W. Zhu, Y. Zhu, and C. Meng, "SSD with multi-scale feature fusion and attention mechanism," *Scientific Reports*, vol. 13, no. 1, Dec. 2023, Art. no. 21387, https://doi.org/10.1038/s41598-023-41373-1.

[31] H. J. Li, Z. Wang, J. Pei, J. Cao, and Y. Shi, "Optimal estimation of low-rank factors via feature level data fusion of multiplex signal systems," *IEEE Transactions on Knowledge and Data Engineering*, 2020, https://doi.org/10.1109/TKDE.2020.3015914.

[32] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-Attentive Feature-Level Fusion for Multimodal Emotion Detection," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Miami, FL, USA, Apr. 2018, pp. 196–201, https://doi.org/10.1109/MIPR.2018.00043.

[33] C. Salazar, E. Montoya-Múnera, and J. Aguilar, "Analysis of different affective state multimodal recognition approaches with missing data-oriented to virtual learning environments," *Heliyon*, vol. 7, no. 6, Jun. 2021, Art. no. e07253, https://doi.org/10.1016/j.heliyon.2021.e07253.

[34] N. H. Ho, H. J. Yang, S. H. Kim, and G. Lee, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020, https://doi.org/10.1109/ACCESS.2020.2984368.

[35] A. Ghorbanali and M. K. Sohrabi, "A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis," *Artificial Intelligence Review*, vol. 56, no. S1, pp. 1479–1512, Oct. 2023, https://doi.org/10.1007/s10462-023-10555-8.

[36] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020, https://doi.org/10.1109/ACCESS.2020.3023871.

[37] H. Zhu, Z. Wang, Y. Shi, Y. Hua, G. Xu, and L. Deng, "Multimodal Fusion Method Based on Self-Attention Mechanism," *Wireless Communications and Mobile Computing*, vol. 2020, pp. 1–8, Sep. 2020, https://doi.org/10.1155/2020/8843186.

[38] D. Sharma, M. Jayabalan, N. Sultanova, J. Mustafina, and D. N. L. Yao, "Multimodal Emotion Recognition Using Attention-Based Model with Language, Audio, and Video Modalities," in *Data Science and Emerging Technologies*, vol. 191, Y. Bee Wah, D. Al-Jumeily Obe, and M. W. Berry, Eds. Springer Nature Singapore, 2024, pp. 193–210.

[39] O. Buza, G. Toderean, and J. Domokos, "A rule-based approach to build a text-to-speech system for Romanian," in *2010 8th International Conference on Communications*, Bucharest, Romania, Jun. 2010, pp. 83–86, https://doi.org/10.1109/ICCOMM.2010.5509108.

[40] Y. Huang, J. Yang, P. Liao, and J. Pan, "Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 1–8, 2017, https://doi.org/10.1155/2017/2107451.

[41] A. M. Aubaid and A. Mishra, "A Rule-Based Approach to Embedding Techniques for Text Document Classification," *Applied Sciences*, vol. 10, no. 11, Jun. 2020, Art. no. 4009, https://doi.org/10.3390/app10114009.

[42] Z. Fu *et al.*, "A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition." arXiv, Nov. 03, 2021, https://doi.org/10.48550/arXiv.2111.02172.

[43] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 167–176, Jul. 2021, https://doi.org/10.1016/j.eij.2020.07.005.