International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 10 | Oct - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Cross-Modal Harmony: Low-Rank Fusion for Enhanced Artificial Emotion Recognition

Himanshu Kumar[1*] and A. Martin[2]

[1]Department of Computer Science, Central University of Tamil Nadu, India, himanshukphd20@students.cutn.ac.in
[2]Department of Computer Science, Central University of Tamil Nadu, India,
martin@cutn.ac.in
***Corresponding author:** Himanshu Kumar, Department of Computer Science, Central University of Tamil Nadu, India. Email, himanshukphd20@students.cutn.ac.in

Abstract: Emotion recognition has become a subject of considerable interest in recent times, owing to its diverse and far-reaching applications in various fields. These applications span from enhancing human-computer interactions to assessing mental health and improving entertainment systems. The proposed study presents a novel approach for emotion recognition by fusing audio and video modalities using low-rank fusion techniques. The proposed methodology leverages the complementary nature of audio and video data in capturing emotional cues. Audio data often encapsulates tone, speech patterns, and vocal nuances, while video data captures facial expressions, body language, and gestures. However, the challenge lies in effectively integrating these two modalities to enhance recognition accuracy. To address the challenge, it employs low-rank fusion, a dimensionality reduction technique that extracts the most informative features from both modalities while minimizing redundancy. Furthermore, it presents the implementation of the chosen low-rank fusion algorithm in a real-world emotion recognition system. The results can contribute to advancing the field of emotion recognition by providing a practical and efficient solution for combining audio and video data to achieve more robust and accurate emotion classification.

***Keywords**: Deep Learning; Emotion Recognition; Human-Computer Interaction; Low-Rank Fusion; Multimodal Fusion.*

I. INTRODUCTION

Emotions, as a fundamental aspect of human expression and interaction, have captured the attention of researchers and technologists across various domains. Recognizing and understanding emotions plays a pivotal role in human- computer interaction, mental health assessment, entertainment systems, and beyond. Among the myriads of approaches to emotion recognition, the fusion of audio and video data has emerged as a promising avenue, leveraging the synergy between auditory and visual cues to enhance the accuracy and robustness of emotion classification. The proposed work embarks on an exploration into the dynamic realm of emotion recognition, delving deep into the multifaceted world of audio and video data fusion. In the era of ever-evolving technology, the proposed research endeavors to shed light on the novel paradigm of low-rank fusion—an innovative approach poised to revolutionize the landscape of emotion recognition. The integration of low-rank fusion techniques represents a sophisticated endeavor aimed at synergizing the strengths of both audio and video modalities, discerning the subtlest nuances that define emotional states. With audio data encapsulating the tonal subtleties of speech and video data capturing the rich tapestry of facial expressions and body language, the research seeks to harness the complementary nature of these modalities.

Low-rank fusion techniques, known for their prowess in dimensionality reduction and informative feature extraction, provide a powerful framework for amalgamating these rich sources of emotional cues. The proposed study is not only a theoretical exploration but a practical implementation of low-rank fusion in the real-world context of emotion recognition. Through

meticulous analysis and experimentation, the proposed work aims to showcase the efficacy of low-rank fusion algorithms, elucidating their impact on the accuracy and reliability of emotion recognition systems. By addressing the challenges of data integration, feature selection, and algorithm optimization, the proposed work endeavor to provide a comprehensive roadmap for individuals looking to make strides in this field of emotion recognition. In a world where human-computer interaction is increasingly intertwined with emotional experiences, the outcomes of the research hold promise for applications ranging from healthcare, where emotional well-being assessment is vital, to education and entertainment, where tailored user experiences are the hallmark of success.

The exploration and implementation of the proposed work will provide insight into the deep learning methodologies and contribute significantly to the evolving landscape of audio-video-based emotion recognition. The Contribution of the paper includes section 2 as a literature review of the low-rank fusion approach, section 3 as a Proposed methodology that leverages audio and video modality using a deep learning model, section 4 as the implementation of low-rank fusion approach, section 5 discusses evaluation metrics and key sensitive parameters in audio-video modality, and section 6 concludes the paper.

II. LITERATURE REVIEW

The performance of emotion recognition in multimodal paradigm have been greatly improved with deep learning models. Recently, machine learning algorithms and deep neural network architectures have been revitalized, such as CNN, RNN, LSTM, Bi-LSTM, models [1]. These models have been used with various fusion mechanism in multimodal recognition with the extracted features of audio modality, text modality, and video modality. Previous studies have shown the more effective fusion mechanism with optimum results in emotion recognition using deep learning models. There has been a wide range of fusion mechanisms. [2] proposed an Early fusion or Feature fusion based deep learning model, feature fusion is most widely used fusion model in emotion recognition The proposed technique combines features or attributes from multiple input modalities to create a single feature vector and it obtained the 89.53% of accuracy.

Feature level fusion fuse the features at an early stage to enhance the information content and improve the performance of subsequent analysis and deep learning model. [3] proposed a late fusion or Decision level fusion based deep learning model. Decision level fusion combines the decisions or outputs of multiple sources or models to make a final decision or prediction. It is oftens used when different sources provide complementary information, and their decisions need to be integrated to reach a consensus, achieved 76.27% of accuracy by using attention based BLSTM.

[4] proposed a hybrid fusion based deep learning model. Hybrid fusion refers to a combination of different fusion techniques. For example, a system might use both feature level fusion and decision level fusion in a coordinated manner to optimize information integration. [5] Proposed a work on hybrid fusion and achieved a mean accuracy of 74.2% on three emotion classes from a cross subject multimodal dataset. [6] proposed a model level fusion, here different models or algorithms are trained independently on the same or different data sources, and their outputs are combined to make a final decision or prediction. Model level fusion can be useful when there are multiple models specialized in different aspects of a problem. [7] Proposed end to end multimodal learning and obtained 73% of accuracy using model level fusion [8] proposed a Rule-based fusion, it involves combining information from multiple sources with predefined rules or logical conditions.

These rules can be designed to reflect domain knowledge or expert opinions on how the fusion should be performed. [9] proposed a classification-based fusion, the outputs of multiple classifiers or models are combined to perform a classification task. The combination can be done using various techniques, such as majority voting, weighted voting, or stacking. [10] proposed an estimation-based fusion, the goal is to estimate a particular parameter or state. Different sources or sensors provide estimates, and these estimates are combined to produce a more accurate or robust estimation. [11] proposed audio-video-based emotion recognition using rule-based fusion and achieved 70% of an average accuracy.

The emotion recognition application has impactfully benefited with the advent of deep learning models along

with fusion mechanism[12]. The basic architecture of deep neural network is to learn the features from the audio and video modality Some good deep learning models trained on audio and video modality has performed with outstanding accuracy. Earlier extracted features were hand crafted, but after involvement of deep learning learned features are also performing better with fusion strategies. Basically, it can be affirmed, Fusion mechanism plays a vital role in accuracy performance of emotion classification. Some studied shows how the deep neural networks can extract the features and fuse with the different modalities.

III. PROPOSED METHODOLOGY

The proposed methodology introduces the low rank fusion model that extracts complimentary features from the audio and video modalities, and providing an approach to improve the accuracy in multimodal emotion recognition.

A. Audio Modality

The use of the audio modality in conjunction with feature extraction and deep learning models has proven to be highly beneficial in various applications. For instance, in emotion recognition tasks, the incorporation of audio information allows for more precise and robust identification of spoken words. Additionally, in areas such as audio enhancement and sound classification, the utilization of audio data enables the extraction of meaningful features that can be effectively utilized by deep learning models to improve the accuracy and performance of the task at hand. In the context of indoor scene classification, the combination of visual and audio information extracted from data collected from social media can provide valuable insights.

The non-verbal information embedded in the speech signal, known as para-lingual cues, has been identified as a valuable indicator of an individual's emotional state. While traditional approaches to emotion recognition from voice have relied on manually crafted features like MFCC and global features [13][14]. Recent years have witnessed substantial advancements with the advent of deep learning (DL) models. Spectrograms generate a visual representation of the spectrum of frequencies present in a signal as it varies with time. It captures important para-lingual information related to a person's emotional state have gained prominence and are now integrated into DL frameworks for speech-based emotion recognition. Various 2D CNNs have been employed in previous studies to harness the potential of spectrograms for emotion recognition [15].

B. Video Modality

The Appearance of the video modality in multimodal using deep learning models has proven to be a powerful approach. By analyzing the visual information in videos, deep learning models can accurately detect and classify facial expressions, allowing for the recognition of various emotions. The multimodal deep learning approach enables more comprehensive and accurate emotion recognition compared to using any single modality alone. Efficiently modeling the spatial and temporal dynamics within video sequences is of paramount importance for extracting robust features, leading to enhancements in the overall system performance.

Typically, state-of-the-art results are attained by combining Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs), which allows for capturing latent appearance representations and temporal dynamics effectively [16]. While several approaches have been investigated for dimensional emotion recognition using Long Short-Term Memory networks (LSTMs) [17], 3D-CNNs have demonstrated efficiency in capturing the spatiotemporal dynamics inherent in videos. Specifically, the deep learning model adopts the Inflated 3D- CNN (I3D) [18] [19] to extract spatiotemporal features from facial clips within video sequences. Compared to traditional 3D CNNs, I3D achieves efficient modeling of the spatiotemporal dynamics in the visual modality while utilizing fewer parameters during training.

C. Low Rank Fusion

In the context of data fusion methodologies, low-rank fusion pertains to a transformative process wherein the dimensionality of individual modalities data, which encompass domains such as imagery, textual content, or acoustic signals, is systematically diminished. It can be accomplished using well-established methods such as Principal Component Analysis (PCA) or Singular Value Decomposition (SVD), among others [20]. Subsequently, these resultant lower-dimensional representations are concatenated in the building of a unified singular feature vector that encapsulates the essence of the fused data. The

proposed work introduced an approach that integrates audio and video modality with the fusion of low-rank tensors. The model framework associated with the proposed method is illustrated in figure 1.

The objective of multimodal fusion is to amalgamate individual unimodal representations into a concise multimodal representation for subsequent predictive tasks. It involves generating a tensor through the outer product of input modalities. Furthermore, tensors are employed to capture the interactions among subsets of modalities.

IV. IMPLEMENTATION OF LOW RANK FUSION

The approach involves utilizing unimodal characteristics and their associated weights directly to approximate the complete multi-tensor outer product operation. The operation is based on low-rank matrix factorization, can be effortlessly applied to scenarios with an extensive interaction space, whether it pertains to the feature space or the number of modalities involved and the proposed technique implemented. Similar to the prior research, it employs Long Short-Term Memory (LSTM) networks [21][22] to compress the time-series data from individual modalities. Subsequently, it extracts the hidden state context vector for modality- specific fusion.

The objective of multimodal fusion is to preserve the model's initial approach while incorporating multiple impactful modalities into a cohesive high-dimensional representation. Consequently, the primary challenge in dealing with multimodal data lies in handling data heterogeneity. Normalized high-dimensional data plays a crucial role in capturing the fundamental aspects of various recognition tasks, such as speech recognition, facial emotion recognition and self-supervised learning [23]. The data normalization enables the extraction of multiple features for emotion recognition.

The proposed study illustrates the Low rank fusion to reduce the tensor weight 'W' into modality factors (M), so from the equation of tensor fusion-

$$Z = M_{n=1} \otimes Z_m \qquad (1)$$

Here, Z: Input tensors, Z_m : Input representation with extra constant 1, M: outer product of tensor

Tensor Z actually decomposes into Z_m and $M_{n=1}$ of low-rank factors, which directly allows to compute the output 'h'. So, Decomposition of tensor can be formulated as following equation-

$$W'_X = (\sum_{i=1} R \otimes Z_{n=1}^{m} W_j^i) \qquad (2)$$

Here, $W_j^i \in R_n^d$; R is the Rank of the tensor, and $Z_{n=1}^{m} W_j^i$: decomposition factor of tensor. From equation (1) and (2) it can be obtained by the low rank tensor-

$$h = \left(\sum R \otimes Z_{n=1}^{m} W_N^i\right).Z \qquad (3)$$

here, W_N^i : corresponding low rank factors.

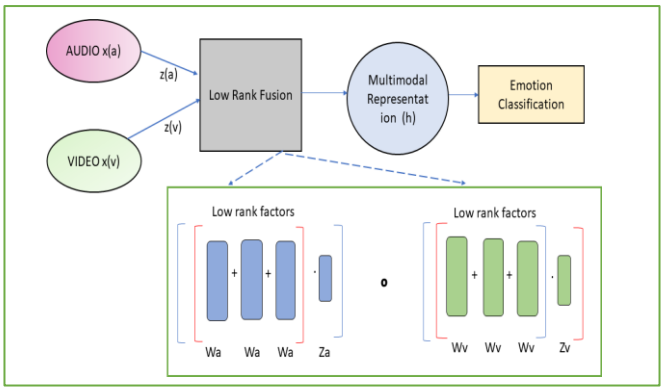

Fig. 1. Low rank fusion architecture

Figure 1 shows the Low rank fusion architecture where x(a) and x(v) are the audio and video unimodal input vectors which are obtained by z(a) and z(b) modalities. Low rank fusion generates the multimodal output with modality- specific factors, which is required to predict and classify the emotions. It can be implemented with the following fusion function-

$$f: v_1 * v_2 * v_3 * \ldots\ldots * v_m \rightarrow h \qquad (4)$$

where 'v' is input vector spaces, and 'h' is output vector space.

While prior research primarily emphasizes the latent adaptation of one modality to another, the focus centers on the adaptation of the latent multimodal signal itself. The application of single-head cross-modal attention to individual modalities enables the achievement. The proposed approach offers the advantage of reducing excessive model parameterization, as it avoids using all possible combinations of modality-to-modality cross-modal attention for each modality. Instead, it opts for a more efficient strategy, utilizing a linear number of cross-modal attention operations for each modality and the

resulting fused signal representation. To ensure that the input sequences maintain awareness of neighboring elements, low rank fusion introduce Temporal Convolutions following the Latent Multimodal Fusion (LMF) operation.

V. EVALUATION METRICS

Root Mean Square Error (RMSE): RMSE quantifies the average error between the observed data and the predictions made by the fused low-rank model. It is commonly used in recommendation systems and collaborative filtering [24]

RMSE = $\sqrt{((\sum(y_i) - y_i')^2 / n)}$

Here, n: number of observations, y: actual value, y': expected value.

Pearson Correlation: The Pearson correlation coefficient involves the ratio of the covariance between the two variables to the product of their individual standard deviations [25].

The formula is as follows:

$r = (\Sigma(u_i - u_i')(v_i - v_i') / \sqrt{(u_i - u_i')(v_i - v_i')^2})$

Here, r: Pearson correlation coefficient, ui: value of any variable u, u': mean of the variable u, vi: value of any variable v, and v': mean of the variable v.

F1 Score: The F1 score is a measure of a model's accuracy in binary classification problems, where the goal is to predict whether an input belongs to one of two possible categories. It is the harmonic mean of precision and recall [26].

Accuracy-k: The Accuracy-k measure the accuracy of binary class and multiclass emotion classification. Here, K is the class type for binary (Accuracy-2) and multiclass (Accuracy-7) [27].

Specificity: Specificity assesses the model's capacity to accurately recognize instances of the negative class. It is calculated by dividing the count of true negatives by the total number of actual negative instances.

$$\text{Specificity} = (\text{True Negatives}) / (\text{Total Actual Negatives})$$

Cohen's Kappa: Cohen's Kappa is a statistic that measures the agreement between predicted and actual class labels, considering the possibility of agreement occurring by chance.

Recall = (True Positives) / (True Positives + False Negatives)

Sensitive Parameters in Audio and Video based emotion recognition.

Table 1. illustrates the key sensitive parameters of audio and video modality. The choice of a parameter depends on the specific requirements of your model and fusion approach in emotion recognition application. It also must be preferable for the available dataset, the desired performance metrics, and computational and learning model complexity. Experimentation and validation on relevant datasets are challenging but advantageous to the application, which requires the most effective fusion approach for the application

Changes in pitch and F0 can be indicative of emotional states. High or low pitch variations can signify excitement or sadness, respectively. Mel-frequency Cepstral Coefficients (MFCCs) coefficients capture spectral characteristics of the audio signal and are commonly used for emotion recognition. Spectral Centroid represents the "center of mass" of the audio spectrum and can reflect changes in emotional intensity. Audio energy or loudness can indicate emotional arousal. High energy may correspond to strong emotions like anger or happiness. The speed of speech and the presence of pauses can be informative. Rapid speech may signify excitement, while long pauses might indicate hesitation or uncertainty.

Table 1. The Key Sensitive Parameters in Audio and Video based emotion recognition

Modality	**Sensitive Parameters**
Audio	Mel-frequency Cepstral Coefficients (MFCCs):
	Speech Rate and Pauses:
	Pitch and Fundamental Frequency (F0)
	Spectral Centroid
	Energy
Video	Head Movements
	Gestures
	Speech-Related Visual Cues
	Facial Expressions
	Electrodermal Activity (EDA)

The movement of the head, including nodding and shaking, can convey agreement or disagreement, which may be linked to emotions. Hand and body gestures can provide context to emotional expressions. For example, a clenched fist may signal anger. Lip movement and synchronization with audio speech can help identify emotions such as happiness or sadness. Key facial features, such as eyebrow movement, lip corners, and eye gaze, can be analyzed to detect emotions like happiness, anger, or surprise. In some cases, sensors measuring EDA, which is related to sweat gland activity, can provide additional emotional data.

Table 2: Results for fusion mechanisms for multimodal emotion recognition on CMU-MOSI dataset

	CMU-MOSI				
Metric	**Mean Absolute Error**	**Pearson Correlation**	**Accuracy-2**	**Accuracy -7**	**F1 Score**
Tensor Fusion	0.940	0.623	73.9	51.4	73.4
Low rank Fusion	0.910	0.678	77.4	55.7	75.7

Table 2 and 3 shows the performance and result of both the fusion mechanism on different datasets, CMU-MOSI and IEMOCAP, respectively. Table 3 presents the MAE, Pearson correlation, Accuracy-k, and F1 score. Table 3 shows the result on IEMOCAP dataset and shows F1 score metrics on four emotions Happy, sad, Angry, Neutral in both the mechanism.

Table 3: Results for fusion mechanisms for multimodal emotion recognition on IEMOCAP dataset

	IEMOCAP			
Metric	**F1 Score (Happy)**	**F1 Score (Sad)**	**F1 Score (Angry)**	**F1 Score (Neutral)**
Tensor Fusion	83.6	82.6	84.6	63.4
Low rank Fusion	85.9	85.6	89.3	71.4

The following Table 4 illustrates the speeds of Tensor Fusion and Low rank Fusion during training and testing**.** The speed measures in unit of inferences per second (IPS). These models are implemented in the PyTorch and TensorFlow frameworks.

Table 4: Evaluation of the training data and test data speeds of Tensor Fusion and Low rank Fusion

Fusion Mechanism	**Training speed (IPS)**	**Testing speed (IPS)**
Tensor Fusion	321.67	1138.33
Low rank Fusion	1153.54	2232.82

In future research, advanced techniques can be taken for temporal sequence compression that do not rely on LSTM-generated hidden state context vectors, as these vectors tend to discard temporal information. Instead, using convolutional layers, temporal information can be restored. Low rank fusion approach can be anticipated that these model variants can be useful for application leverage to resource-constrained environments, further, these applications will also be helpful in refinements and optimizations. Additionally, as per the keen interest, the feature sets for audio, text, and visual data pipelines can also be enhanced. The proposed approach expands the feature extraction utilization and could open up new possibilities for various use cases to access more abundant computational resources.

VI. CONCLUSION

Audio and video data, although different in their nature, are naturally complementary when it comes to capturing the richness of human emotions. While audio data captures the nuances of speech, intonation and intonation, video data reveals the complex world of facial expressions, gestures and body language. Their coordination is necessary for comprehensive emotion recognition. Low-rank fusion approaches have demonstrated their efficacy in bridging the gap between

audio and video modalities. By applying dimensionality reduction techniques and feature extraction methods, it has harnessed the strengths of both data sources while reducing redundancies. It makes the emotion classification process more robust and accurate. The proposed work intends to demonstrate the feasibility of adopting it into real applications. The findings underline the immense potential of multimodal approaches in emotion recognition. Integrating audio and video data is not limited to the scope of the research but extends to various fields where human emotions matter, including health care, education, and entertainment. Future research may explore novel fusion techniques, incorporate additional modalities such as physiological data, or delve deeper into the interpretability of emotion recognition models.

REFERENCES

[1] J. Ashok Kumar, T. E. Trueman, and E. Cambria, "A Convolutional Stacked Bidirectional LSTM with a Multiplicative Attention Mechanism for Aspect Category and Sentiment Detection," *Cognit. Comput.*, vol. 13, no. 6, pp. 1423–1432, 2021, doi: 10.1007/s12559-021-09948-0.

[2] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," *Inf. Fusion*, vol. 51, no. October 2018, pp. 10–18, 2019, doi: 10.1016/j.inffus.2018.10.009.

[3] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manag.*, vol. 57, no. 3, p. 102185, 2020, doi: 10.1016/j.ipm.2019.102185.

[4] A. Ghorbanali and M. K. Sohrabi, "A comprehensive survey on deep learning-based approaches for multimodal sentiment analysis," *Artif. Intell. Rev.*, 2023, doi: 10.1007/s10462-023-10555-8.

[5] 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, doi: 10.1109/ACCESS.2020.3023871.

[6] 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," in *IEEE Access*, IEEE, 2020, pp. 61672–61686. doi: 10.1109/ACCESS.2020.2984368.

[7] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1301–1309, 2017, doi: 10.1109/JSTSP.2017.2764438.

[8] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," *2016 IEEE Students' Technol. Symp. TechSym 2016*, pp. 7–12, 2017, doi: 10.1109/TechSym.2016.7872646.

[9] J. Akram and A. Tahir, "Lexicon and heuristics based approach for identification of emotion in text," *Proc. - 2018 Int. Conf. Front. Inf. Technol. FIT 2018*, pp. 293–297, 2019, doi: 10.1109/FIT.2018.00058.

[10] H. Kumar and A. Martin, "Comparison and Performance Evaluation of Fusion Mechanism for Audio–Video Based Multimodal Emotion Recognition," *Lect. Notes Networks Syst.*, vol. 864, pp. 213–225, 2024, doi: 10.1007/978-981-99-8628-6_19.

[11] H. Kumar and A. Martin, "Audio-Video Based Fusion Mechanism Using Deep Learning for Categorical Emotion Recognition," *Int. Conf. Self Sustain. Artif. Intell. Syst. ICSSAS 2023 - Proc.*, pp. 595–602, 2023, doi: 10.1109/ICSSAS57918.2023.10331779.

[12] H. Madhuranath and S. V. S. T. Ravindra Babu, "Modified Adaboost method for efficient face detection," pp. 415–420, 2012.

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

[14] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," *Proc. 2011 Int. Conf. Electron. Mech. Eng. Inf. Technol. EMEIT 2011*, vol. 2, pp. 621–625, 2011, doi: 10.1109/EMEIT.2011.6023178.

[15] S. S. Hosseini, M. R. Yamaghani, and P. Arabani, Soodabeh, "Multimodal modeling of human emotions using sound , image and text fusion," *Res. Sq.*, vol. 1, 2023, doi: https://doi.org/10.21203/rs.3.rs-2579610/v1.

[16] J. Cho and H. Hwang, "Spatio-temporal representation of an electoencephalogram for emotion

recognition using a three-dimensional convolutional neural network," *Sensors (Switzerland)*, vol. 20, no. 12, pp. 1–18, 2020, doi: 10.3390/s20123491.

[17] Q. Wei, X. Huang, and Y. Zhang, "FV2ES: A Fully End2End Multimodal System for Fast Yet Effective Video Emotion Recognition Inference," *IEEE Trans. Broadcast.*, pp. 1–11, 2022, doi: 10.1109/tbc.2022.3215245.

[18] R. G. Praveen *et al.*, "A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2022-June, pp. 2485–2494, 2022, doi: 10.1109/CVPRW56347.2022.00278.

[19] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic Speech Emotion Recognition Using Machine Learning," in *Social Media and Machine Learning*, no. March, IntechOpen, 2020, p. 16. doi: 10.5772/intechopen.84856.

[20] S. Sahay, E. Okur, S. H. Kumar, and L. Nachman, "Low rank fusion based transformers for multimodal sequences," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 29–34, 2020, doi: 10.18653/v1/2020.challengehml-1.4.

[21] Z. Hao, Z. Li, X. Dang, Z. Ma, and G. Liu, "MM-LMF: A Low-Rank Multimodal Fusion Dangerous Driving Behavior Recognition Method Based on FMCW Signals," *Electron.*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223800.

[22] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognit. Lett.*, vol. 139, pp. 157–165, 2017, doi: 10.1016/j.patrec.2017.06.025.

[23] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020, doi: 10.1109/ACCESS.2020.3026823.

[24] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomed. Signal Process. Control*, vol. 78, no. June, p. 103970, 2022, doi: 10.1016/j.bspc.2022.103970.

[25] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition: Enhancing film sound design using audio features, regression models and artificial neural networks," *Pers. Ubiquitous Comput.*, vol. 25, no. 4, pp. 637–650, 2021, doi: 10.1007/s00779-020-01389-0.

[26] E. Mohammadi, H. Amini, and L. Kosseim, "Neural feature extraction for contextual emotion detection," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2019, pp. 785–794. doi: 10.26615/978-954-452-056-4_091.

[27] K. Candra Kirana, S. Wibawanto, and H. Wahyu Herwanto, "Facial Emotion Recognition Based on Viola-Jones Algorithm in the Learning Environment," in *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication*, IEEE, 2018, pp. 406–410. doi: 10.1109/ISEMANTIC.2018.8549735.