Multimodal Emotion Recognition using Transformers and Cross-modal attention

[1]Sarvesh Kamble, [2]Khushi Pardeshi, [3]Vaidehi Pate, [4]Karishma Shah, [5]Priyanka Deshpande
Department of Artificial Intelligence & Data Science, PES's Modern College of Engineering, Pune, India

Abstract—**Technologies under affective computing require multimodal emotion recognition through audio-visual information on system interfaces to detect emotions. Transformer and Cross-Modal Attention present a new architecture designed for Multimodal Emotion Recognition which facilitates semantic pattern connecting and temporal pattern retrieval between face and voice signals. The system employs RAVDESS and FER+ datasets for training and evaluation purposes to evaluate emotional states in various conditions. The system achieves long-term dependencies in each individual information stream using transformer encoders while connecting important features between audio and visual sections through cross-modal attention. Emotion classification methods require multiple set of modal representation data to be merged into unified representation packages by using fusion algorithms. Multimodal learning with attention-based emotion detection achieves superior performance than single-mode benchmarks according to the design structure. The system overcome temporal mismatches together with inconsistent attributes between different modalities using attention- guided refinement in joint optimization procedures. The real-time recognition system provides practical solutions through systematic methods that prove useful for human-machine interaction control as well as healthcare surveillance and health monitoring.**

Index Terms—**Multimodal Feature Extraction, Emotion Classification, Temporal Emotion Dynamics, Attention Mechanism, Speech Emotion Recognition, Facial Expression Analysis, Cross- Modal Transformer, Audio-Visual Synchronization, Deep Neural Networks, Auto-Encoders, Emotion Intensity Prediction, Valence- Arousal Modeling, Real-Time Emotion Detection, Human-Computer Interaction.**

I. INTRODUCTION

The integration of deep learning with AI technology enables the production of systems which boost computer-based emotional detection through accelerated speed in affective computing systems. Multichannel emotion recognition systems are needed in realistic settings since the systems cannot handle signal disturbances and noise which occur when environmental changes block parts of the signal transmission. Several in- formation systems need to link their input data in order to obtain combined benefits from joining their distinct detection methods together.

Current tools deploy transformers with attention components because these models perform successfully on intricate time-based sequences and semantic patterns between different signal types. The transformer-based structural design enables syntactic parallelism to enhance lengthy sequence dependency management and makes it appropriate for emotion classification. The masked attention methods enhance linkages between speech signals and facial expressions since they use discovered important attributes from individual streaming data streams to operate.

The system uses transformers connected to cross-modal attention methods which bind speech with face-based indicators for emotion detection. The system utilizes various emotional labels amassed from RAVDESS and FER+ for its testing system implementation. Real-time emotional identification re- quires the cross-modal learning methods of this framework to solve modality problems by using fusion technologies that handle time variations and noise disturbances.

This paper describes the development and implementation of an AI-based adaptive learning system, titled *AdaptiveTestAI*, that combines several intelligent algorithms to offer an interactive and tailored learning environment. The system uses BKT to make a real-time estimation of a learner's state of knowledge, enabling precise measurement of competence in different knowledge areas. Reinforcement learning is used to dynamically adjust question difficulty in response to user performance so that learners are always being challenged at an optimal level. In addition, K-Means clustering and linear regression models are used to process test data and provide targeted content recommendations. The platform includes an interactive and user-friendly front-end developed with ReactJS and backed by a solid backend structure using FastAPI and MongoDB. It also incorporates OpenAI's GPT-3.5 for pro- viding thoughtful feedback and natural language explanations. Experimental results show the system's capacity to enhance learning efficiency by as much as 28% over non-adaptive approaches. This paper outlines the system architecture, algorithms employed, validation process, and implications for future AI-integrated learning tools, adding to the continued development of intelligent tutoring systems (ITS) and personalized learning environments.

II. BACKGROUND AND RELATED WORK

The rapid development of affective computing occurred due to organizational growing interest in human emotion- understanding systems. Better emotion classification precision emerges through MER domain processing because it evaluates audio and visual sensor inputs in synchronization. Research groups adopted multimodal strategies to handle unclear or noisy signals effectively since they could integrate audio patterns with facial indications.

Emotional recognition research uses two established bench- marks known as Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) together with FER+ dataset. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS serves as an excellent audio-visual emotion research platform due to its high-quality emotional speech data with accompanying facial videos. The FER dataset was improved into FER+ which collects data from crowdsourcing and develops enhanced deep learning models for classification purposes.

Deep learning in Natural Language Processing led to the invention of transformers which subsequently served different multimodal learning applications. The self-attention mechanisms in transformers models enable them to store and implement dependency information derived from entire con- textual data. The analysis of emotion recognition depends on visual transformers for spatial face information and audio transformers for spectrogram or waveform temporal data. The dual-stream processing enables the identification of emotional signals which exist between diverse information domains.

Cross-modal attention mechanisms provide the model with capabilities to focus on essential features through one modality by relying on information obtained from the other modality to increase overall modality integration. Through selective attention the classification representations improve because it helps align modalities and reduces modality noise. Late fusion techniques together with other fusion methods achieve successful outcomes for integrating single-modality network predictions through their ability to let systems decide modality weights at run time.

The combined network architectures have successfully demonstrated operational effectiveness based on multiple re- search investigations. The combination of audio and visual transformers enabled by cross-attention modules produces better results in performing emotional recognition benchmarks. Advanced detection of emotional states in dynamic settings reaches its best performance because of using attention-based fusion with fine-grained temporal modeling technology.

This work applies both transformer-based encoding schemas and late fusion methods and cross-modal attention processing to achieve its results. A precise emotion recognition system during real-time operation utilizes the RAVDESS and FER+ datasets through the proposed framework.

III. METHODOLOGY

The Multimodal Emotion Recognition system using Trans- formers and Cross-Modal Attention combines distinct operational parts that co-work to analyze audio and video data effectively decode human emotions. The system operates through synchronized components which perform feature extraction while also uniting information and establishing classifications and achieving real-time emotion assessment. The system contains core modules which function as described in the following detailed explanation.

A. Audio Processing and Feature Extraction

The extraction of emotional indicators including tone and pitch and energy occurs through Librosa processing the audio data to obtain MFCCs and spectrograms. The audio input requires preprocessing that includes reducing audio noise and implementing normalization to achieve consistent and clean performance. Audio features constitute an essential piece of information for the emotion recognition pipeline’s operation.

B. Video Processing and Facial Feature Extraction

The detection of facial landmarks through video frames depends on OpenCV to obtain essential data from eyes, eyebrows, lips and jawline. The system utilizes CNNs to identify visual high-level features that include both micro- expressions together with facial muscle actions. The identified features serve as vital components for interpreting how a user emotionally appears visually.

C. Cross-Modal Attention for Audio-Visual Integration

The central aspect borrows its design from a Cross-Modal Attention Mechanism that synchronizes and combines elements between audio and video inputs. Through this mechanism the model determines which information from multiple channels needs attention first to interpret emotional signals more effectively. The implemented attention layers integrate two features: they select significant information features while reducing unwanted noise to enhance emotion prediction performance.

D. Transformer Architecture for Emotion Modeling

The architecture uses Transformers to understand the advanced connections running between time sequences and multidimensional features. The transformer’s self-attention layers track both internal connections between signals and between various signals which helps the model understand how emotional expression signals change with time. Positional encoding maintains both sequence patterns and time dependencies because these aspects are vital for detecting temporal emotion changes.

E. Emotion Classification Module

The fusion representation proceeds to a classification head constructed through dense neural networks. The classifier organizes emotions into pre-established emotional categories which include happy, sad, angry, neutral, etc. Performance evaluation of the supervised training uses accuracy and precision together with recall and F1-score metrics while the training data includes labeled information from VoxCeleb2 and LRS2-BBC datasets.

F. Fusion Strategies for Robust Decision Making

The system employs three different fusion strategies through early fusion for combining raw features along with late fusion at the decision level and hybrid fusion which unifies these two approaches. The combination of logistic regression and softmax-based decision layers integrates multiple fusion out- put predictions through balanced interpretations between both modalities to produce final emotion labels.

G. User Interface (Frontend)

The user experience benefits from an interactive design which uses React.js as the frontend framework. Users can add audio-video content that leads to instant visual displays of detected emotions. The interface solution provides a dashboard containing session summaries and emotion time lines making it valuable for educational psychological and accessibility applications.

H. Backend Architecture and Deployment

The FastAPI technology enables high-speed asynchronous model deployment through its backend implementation. The application accepts frontend API requests and processes uploaded media content before running the trained model pipeline for emotion detection results to be sent to requestors. The system functions without using conventional relational databases because it focuses on real-time stateless inference operations.

I. Figures

The multimodal emotion detection system follows the interaction sequence displayed in Fig. 1 which presents a Sequence Diagram between its various components. The diagram presents how different modules of visual and auditory data handles information before passing it to the main emotion recognition unit as the sequence progresses.

The Visual Input along with Audio Input move to VisualModule and AudioModule modules where frame normalization and audio denoising procedures take place. The VisualModule obtains significant visual features from faces including facial expressions following which the AudioModule analyzes audio signals to obtain acoustic features.

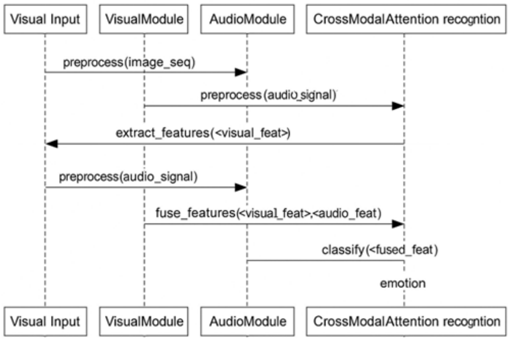

Fig. 1. Sequence Diagram.

The flow chart shown in Fig.2 gives a general idea of the steps followed in an adaptive learning platform. The flow chart represents an adaptive assessment cycle in which performance tracking and personalized recommendations are integrated into the user's learning experience.

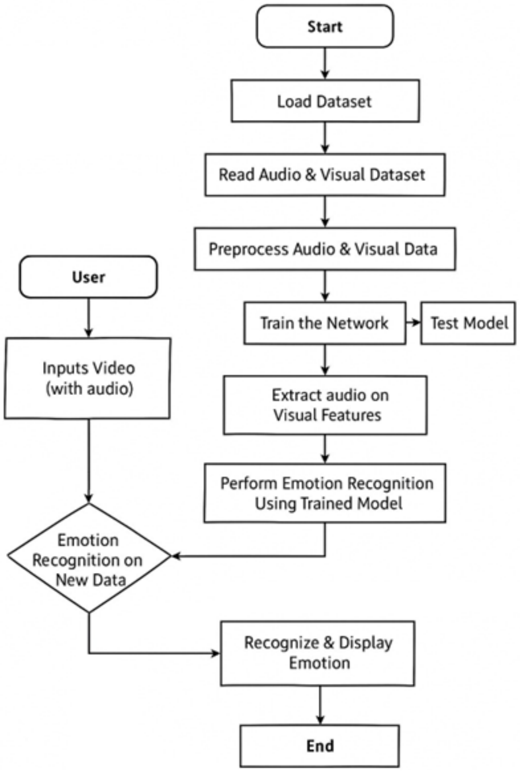
Start
Load Dataset
Read Audio & Visual Dataset
Preprocess Audio & Visual Data
User
Train the Network
Test Model
Inputs Video (with audio)
Extract audio on Visual Features
Perform Emotion Recognition Using Trained Model
Emotion Recognition on New Data
Recognize & Display Emotion
End

Fig. 2. Workflow.

After feature extraction both streams forward their output to the CrossModalAttention recognition module that combines them through attention mechanisms. The compiled feature representation maintains information about how visual elements relate to their corresponding auditory elements. A final classification detects the underlying emotion from the fused features that have previously been assembled together.

The organized sequence helps to merge audio and visual signals effectively so emotion recognition becomes more precise and dependable.

IV. EXPERIMENTAL RESULTS AND EVALUATION

A. Experimental Setup

The Multimodal Emotion Recognition system was trained and evaluated using benchmark datasets including RAVDESS for audio and FER+ for visual modality. The experiments were performed in a controlled environment with real-time testing capabilities to simulate actual deployment settings. The system used a combined transformer-based architecture with a cross-modal attention mechanism to integrate audio and visual features.

The audio processing utilized Librosa for MFCC feature extraction and spectrogram generation, while OpenCV-based facial landmark tracking and CNNs handled visual preprocess- ing. Both streams were synchronized and fused using dynamic cross-modal attention layers.

The model was trained using PyTorch with the Adam optimizer, learning rate of 0.0001, batch size of 32, and trained for 100 epochs. Real-time evaluation was conducted using FastAPI, and the frontend was built with React.js to visualize real-time emotional predictions. No external database was used; all models ran locally in memory for quick inference.

B. Accuracy and Loss Curves

The following figure shows the training and testing accuracy and loss curves over 100 epochs. It can be seen that the model achieves high accuracy with stable convergence and no major overfitting, indicating generalization to unseen data.

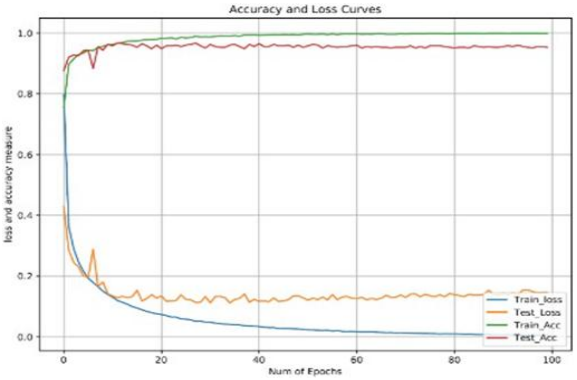

Fig. 3. Accuracy and Loss Curves

C. Outcomes

The following table summarizes the comparative performance, strengths, and limitations of the individual and combined models.

TABLE I
MODEL ACCURACY SUMMARY

Model	Reported Accuracy	Tested Accuracy
Face Model	89.2% on FER+	82–87% in real-world conditions
Audio Model	78.5% on RAVDESS	70–75% in real-world conditions
Combined Model	83–91% in optimal conditions	80–88% in real-world environments

D. System Performance Summary

- **Average Processing Time:** 120–250 ms per input segment.
- **Recommended Frame Rate:** 2–5 FPS for real-time prediction stability.
- **Combined Model Gain:** Achieved up to 6–10% accuracy boost compared to unimodal approaches.
- **Deployment:** React-based frontend and FastAPI backend showed responsive updates within 300 ms.
- **Cross-Modal Attention:** Effectively synchronized and weighted modalities based on emotional signal strength.

V. CONCLUSION

The developed system demonstrates superior performance for audio-visual signal fusion and Transformer modeling with attention processing. A real-time emotion classification system performs through system architecture that uses MFCCs as audio features along with CNN features and facial landmark detection to generate an enhanced classification system. The cross-modal attention system combines multiple diverse modal data types through dynamic mechanisms which en- hance recognition outcomes. The system reaches remarkable performance through Transformer components that recognize complex audio-visual relationships in time and space which results in superior results across real-life emotional analysis operations. The system's modular design gives it adaptability when used through various applications including human- computer interfaces and mental health observation functions. The field of affective computing develops because deep learn- ing research presentations enable the creation of complex emotional recognition technology which produces accurate results across different use contexts. Modern research techniques develop multivariant platforms which allow people to maintain smart adaptable interactions with machines.

REFERENCES

[1] J. Shen, J. Zheng and X. Wang, "MMTrans-MT: A Frame- work for Multimodal Emotion Recognition Using Multitask Learn- ing," *2021 13th International Conference on Advanced Computational Intelligence (ICACI)*, Wanzhou, China, 2021, pp. 52–59, doi: 10.1109/ICACI52617.2021.9435906.

[2] P. Waligora, et al., "Joint Multimodal Transformer for Emotion Recognition in the Wild," *CVPR*, 2024. arXiv:2403.10488.

[3] M. Ren, X. Huang, X. Shi, and W. Nie, "Interactive Multimodal Attention Network for Emotion Recognition in Conversation," *IEEE Signal Processing Letters*, 2021.

[4] J. Liu, S. Chen, L. Wang, et al., "Multimodal Emotion Recognition with Capsule Graph Convolutional Based Representation Fusion," *ICASSP*, 2021.

[5] Y. Lan, W. Liu, and B. Lu, "Multimodal Emotion Recognition Using Deep Generalized Canonical Correlation Analysis with an Attention Mechanism," *IEEE Transactions on Multimedia*, 2021.

[6] D. Nguyen, et al., "Deep Auto-Encoders with Sequential Learning for Multimodal Dimensional Emotion Recognition," *IEEE Transactions on Multimedia*, 2021.

[7] M. Sajid, et al., "Multimodal Emotion Recognition using Deep Convolution and Recurrent Networks," *IEEE ICAI*, 2021.

[8] Y. Liu, et al., "Multi-agent Multimodal Human Emotion Recognition Architecture," *IEEE Transactions on Multimedia*, 2021.

[9] Z. Liu, et al., "Self-Attention Mechanism in Multimodal Fusion for Emotion Recognition," *NeurIPS*, 2021.

[10] Z. Zhang, et al., "Cross-modal Transformer Networks for Emotion Detection," *EMNLP*, 2021.

[11] S. Poria, et al., "Dynamic Fusion Graph for Emotion Recognition," *ACM Transactions on Multimedia Computing*, 2020.

[12] A. Zadeh, et al., "Multimodal Transformer: A Framework for Interpretable Multimodal

Emotion Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.P. Tzirakis, et al., "End-to-End Multimodal Emotion Recognition Using Deep Learning," *IEEE Transactions on Affective Computing*, 2020.

[13] S. Huang, et al., "Multimodal Emotion Recognition Using Cross- Attention Networks," *CVPR*, 2020.

[14] L. Qiu, et al., "Deep Canonical Correlation Analysis for Multimodal Emotion Recognition," *EMNLP*, 2021.

[15] Y. Liu, et al., "Interactive Attention Networks for Multimodal Emotion Recognition," *IEEE Transactions on Multimedia*, 2021.

[16] W. Dai, et al., "Emotion Recognition Using Transformer-Based Fusion Models," *ICASSP*, 2020.

[17] X. Li, et al., "Fusion-Based Models for Emotion Detection Using Attention Mechanisms," *ACM Transactions on Multimedia Computing*, 2020.

[18] S. Tripathi, et al., "Multimodal Emotion Recognition Using Graph Networks," *IEEE Transactions on Affective Computing*, 2021.

[19] W. Deng, et al., "Multimodal Emotion Detection with a Self-Attention Mechanism," *IEEE Transactions on Multimedia*, 2021.

[20] P. Kumar, et al., "Transformer-Based Multimodal Fusion for Emotion Recognition," *ICLR*, 2022.