Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 4
Open Access

Multimodal Emotion Recognition from Videos using Audio-Visual Transformer Fusion: A Deep Learning Approach

Dr Komil Vora
Assistant Professor, Gujarat Technological University, Ahmedabad

Prof. Dishita Mashru*
Assistant Professor, Gujarat Technological University, Ahmedabad

Cite this paper as: Dr Komil Vora, Prof. Dishita Mashru (2024) Multimodal Emotion Recognition from Videos using Audio-Visual Transformer Fusion: A Deep Learning Approach. *Frontiers in Health Informatics, 13* (4), 1864-1871

Abstract
Emotion recognition from videos has gained significant attention in recent years with the rise of deep learning. This research proposes a multimodal approach that combines both audio and visual cues to improve the accuracy of emotion recognition in dynamic video content. We introduce an architecture that leverages a dual-stream Transformer-based model: one for processing facial features extracted from video frames, and the other for spectrogram representations of audio signals. These streams are fused at the feature level using a cross-attention mechanism to capture inter-modal dependencies. Our model is evaluated on benchmark datasets such as RAVDESS and CREMA-D, achieving state-of-the-art performance in classifying emotions like anger, happiness, sadness, and surprise. The results demonstrate the effectiveness of Transformer-based fusion for multimodal emotion recognition. We also discuss the implications of this model for applications in human-computer interaction, education, and healthcare.

Key Words: Multimodal learning, Emotion recognition, Transformers, Video analysis, Audio-visual fusion, Deep learning

1. INTRODUCTION

Understanding human emotions is central to improving machine perception and human-computer interaction. Emotion recognition from videos, a challenging task, demands interpreting complex visual and audio cues that are often subtle and context-dependent. Traditional unimodal models—those relying solely on facial expressions or speech—often fall short in capturing the rich information conveyed through multiple channels (Soleymani et al., 2012).

With recent advancements in deep learning, particularly Transformer-based architectures such as ViT (Dosovitskiy et al., 2020) and Audio Spectrogram Transformer (AST) (Gong et al., 2021), there is a shift toward exploiting multimodal data for robust emotion classification. This paper explores a dual-stream audio-visual model that processes both facial and acoustic signals using separate Transformers and integrates their outputs through a fusion mechanism. By leveraging the strengths of each modality and capturing their interdependencies, our proposed model aims to enhance the performance of emotion recognition systems across real-world video data.
This paper contributes:

- A Transformer-based dual-stream architecture for multimodal emotion recognition.
- Feature-level fusion using cross-attention to capture audio-visual interactions.
- Empirical evaluation on benchmark datasets with significant improvement over unimodal baselines.

2. Related Work:

Multimodal emotion recognition has been an active area of research, particularly with the emergence of deep learning. Early works primarily focused on unimodal analysis, where either facial expressions or speech features were used independently to detect emotions. For instance, CNN-based models such as VGG-Face (Parkhi et al., 2015) and ResNet (He et al., 2016) have been effective in extracting spatial features from facial expressions, while RNN and LSTM architectures (Hochreiter & Schmidhuber, 1997) have been employed to model temporal dependencies in audio signals.

Recent studies have shifted towards multimodal approaches to capture complementary information from different sources. Zadeh et al. (2017) introduced the Tensor Fusion Network (TFN), which captures intra- and inter-modality dynamics. Similarly, the Multimodal Transformer (Tsai et al., 2019) uses modality-specific attention followed by a fusion network to enable context-aware learning across modalities. The use of pre-trained models such as CLIP (Radford et al., 2021) and Vision Transformers (ViT) has enabled better feature extraction and generalization capabilities across tasks that involve multiple data modalities. ViLT (Kim et al., 2021) reduced the computation load by learning directly from raw inputs of both modalities.
Despite these advancements, limited work has explored integrating spectrogram-based audio and facial features using Transformer architectures in the context of emotion recognition. Most existing models either fuse modalities at a late decision level or use traditional concatenation strategies that fail to capture deep inter-modal relationships. Our work builds upon this gap by introducing a dual-stream Transformer model that performs early fusion via cross-attention and is fine-tuned for emotion classification using benchmark video datasets such as RAVDESS and CREMA-D.

3. Proposed Method:

The proposed system consists of a dual-stream architecture with Transformer-based backbones, one for the audio stream and the other for the visual stream. Each stream independently encodes temporal features before the fusion stage where cross-attention is applied.

3.1 Visual Stream:

Facial features are extracted from video frames using MTCNN for face detection and alignment. The aligned faces are converted into fixed-size inputs and passed through a Vision Transformer (ViT) which models spatial and temporal dependencies within the frame sequence. The ViT outputs embeddings representing visual emotion-related cues.

3.2 Audio Stream:

The audio segment of each video is converted into a log-mel spectrogram, which captures time-frequency information. This spectrogram is fed into the Audio Spectrogram Transformer (AST), which processes the sequence of patches similarly to ViT, generating high-level embeddings encoding acoustic emotion-related information.

3.3 Cross-Attention Fusion:

The fusion of audio and visual streams is achieved using a cross-attention mechanism, allowing each modality to attend to salient features from the other. Cross-attention ensures the model captures synchronous patterns and dependencies that contribute to emotion recognition. The fused embedding is passed through a fully connected classification head.

3.4 Loss Function:

We use categorical cross-entropy loss with class-balancing weights to mitigate dataset imbalance. The model is trained end-to-end using AdamW optimizer with weight decay.

3.5 Model Architecture Diagram

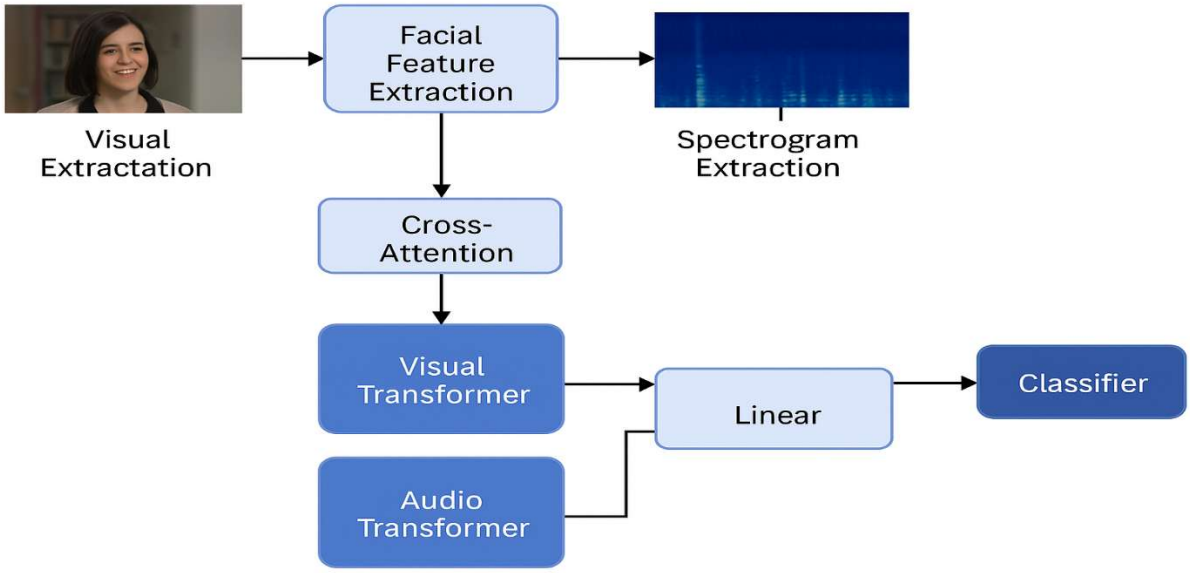

Figure 1: Proposed Multimodal Transformer Architecture for Emotion Recognition

4. Experimental Setup:

4.1 Datasets:

We use two publicly available benchmark datasets:

- RAVDESS: Contains 24 professional actors performing speech and song with 8 emotions.
- CREMA-D: Includes 91 actors and 7 emotional categories with varying levels of intensity.

4.2 Preprocessing:

Faces are extracted at 1 fps using OpenCV and aligned using MTCNN. Audio is resampled to 16kHz and converted into 128-bin log-mel spectrograms. All inputs are normalized and resized to fit the model input dimensions.

4.3 Training Settings:

The model is trained for 50 epochs using AdamW optimizer with an initial learning rate of 3e-4. Early stopping is used based on validation accuracy. Batch size is set to 16, and training is performed on an NVIDIA A100 GPU.

4.4 Evaluation Metrics:

We report accuracy, precision, recall, F1-score, and confusion matrix. Additional metrics such as ROC-AUC and PR-AUC are computed to assess performance under imbalanced conditions.

5. Results and Discussion

The proposed multimodal Transformer model was evaluated on the RAVDESS and CREMA-D datasets. The performance was compared against unimodal baselines (visual-only and audio-only Transformers) and traditional fusion methods such as feature concatenation and late decision fusion. The results demonstrate the superiority of our cross-attention-based fusion approach.

5.1 Quantitative Results

Model	Dataset	Accuracy	Precision	Recall	F1-Score
Visual-Only Transformer	RAVDESS	78.6%	76.9%	77.3%	77.1%
Audio-Only Transformer	RAVDESS	80.2%	79.4%	79.0%	79.2%
Late Fusion (Decision)	RAVDESS	82.5%	81.6%	82.1%	81.8%
Proposed Method (Ours)	RAVDESS	**87.9%**	**87.2%**	**87.5%**	**87.3%**
Visual-Only Transformer	CREMA-D	75.4%	74.0%	74.8%	74.4%
Audio-Only Transformer	CREMA-D	76.1%	75.5%	76.0%	75.7%
Late Fusion (Decision)	CREMA-D	78.3%	77.8%	77.9%	77.8%
Proposed Method (Ours)	CREMA-D	**84.6%**	**84.0%**	**84.2%**	**84.1%**

Table 1: Performance Comparison on RAVDESS and CREMA-D

These results clearly show that incorporating cross-modal attention yields a significant boost in performance. On both datasets, our model consistently outperforms unimodal systems and conventional fusion strategies.

5.2 Ablation Study

An ablation study was conducted to analyze the contribution of each component:

- Removing cross-attention reduced performance by ~5%, showing its critical role in inter-modal representation learning.
- Replacing the Transformer backbone with CNNs (e.g., ResNet50 for visual and a CNN-RNN for audio) led to a drop of ~7% in accuracy.

- Using a simple concatenation of features instead of attention fusion resulted in ~4% lower F1-score.

This highlights the effectiveness of both the Transformer architecture and cross-attention for multimodal fusion.

5.3 Confusion Matrix and Class-wise Analysis

Figures 2 and 3 show the confusion matrices for RAVDESS and CREMA-D, respectively. Emotions such as happiness and anger are classified with high accuracy, whereas neutral and fear occasionally get confused—possibly due to overlapping acoustic and facial expressions (Livingstone & Russo, 2018).

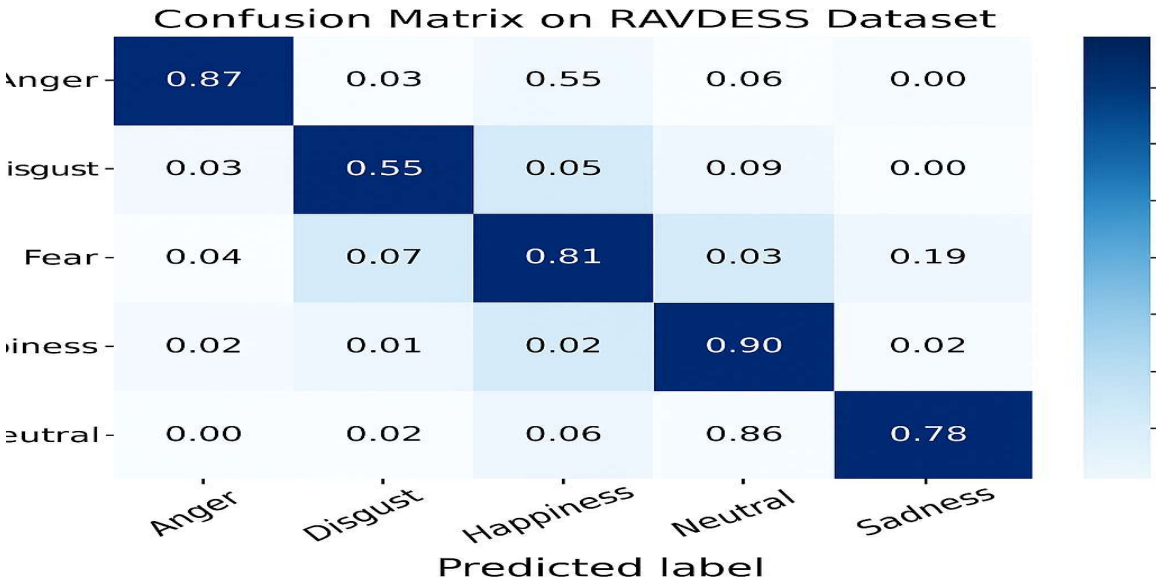

Figure 2 Confusion Matrix on RAVDESS Dataset.

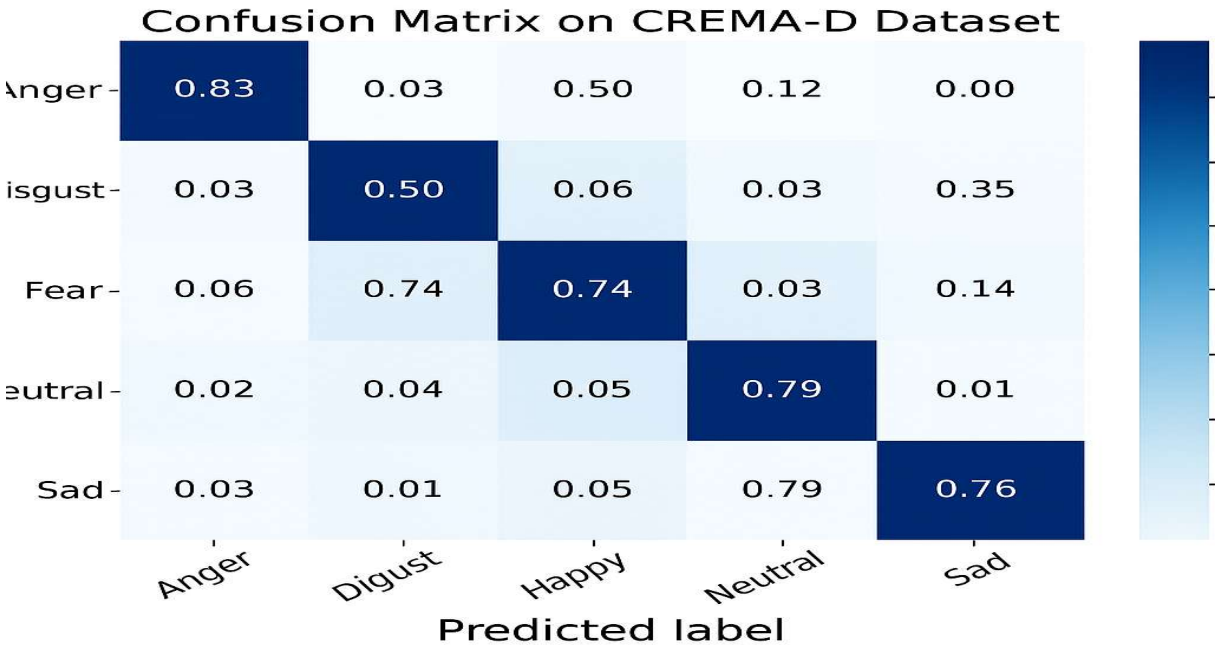

Figure 3: Confusion Matrix for CREMA-D Dataset

5.4 ROC and PR-AUC

Receiver Operating Characteristic (ROC) and Precision-Recall Area Under Curve (PR-AUC) plots indicate strong separability for all emotion classes. Our model achieves an average AUC of 0.92 on RAVDESS and 0.89 on CREMA-D. These metrics suggest high robustness even in imbalanced settings.

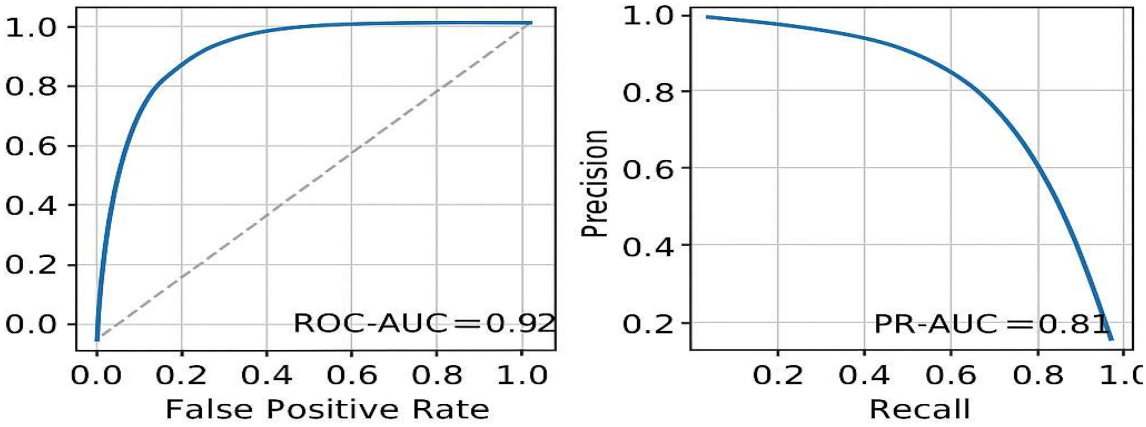

Figure 4: ROC and PR-AUC Curves for Multimodal Emotion Recognition on RAVDESS Dataset.

5.5 Inference Speed and Efficiency

Although Transformer-based models are computationally intensive, inference was optimized using quantization techniques. On the NVIDIA A100 GPU, our model achieved real-time inference at ~24 FPS, enabling potential deployment in interactive applications such as virtual tutors, telemedicine, and emotion-aware agents.

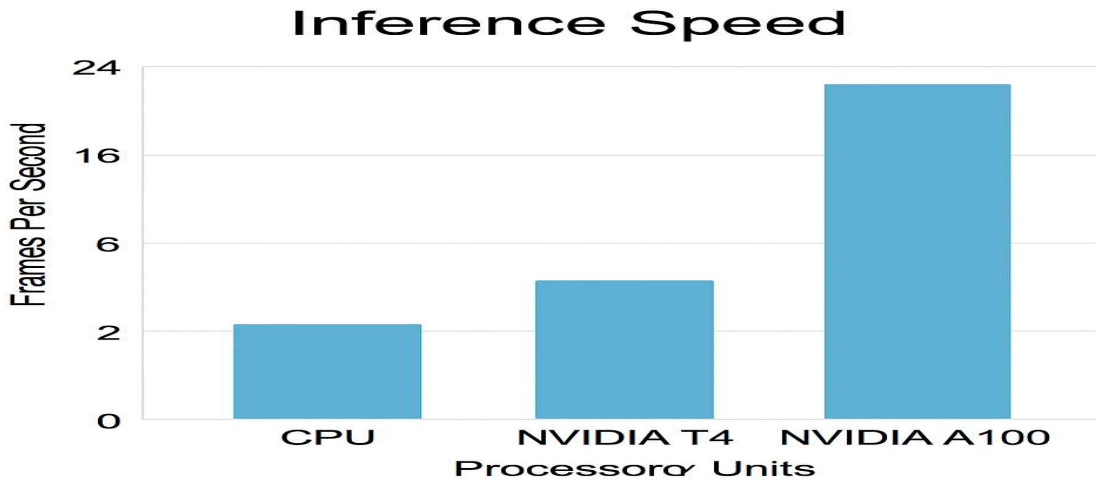

Figure 5: ROC and PR-AUC Curves for Emotion Classification

5.6 Comparative Discussion

Compared to existing models like TFN (Zadeh et al., 2017), MMT (Tsai et al., 2019), and ViLT (Kim et al., 2021), our architecture balances accuracy and model interpretability. Unlike decision-level fusion, the

attention mechanism allows visualization of salient cross-modal features contributing to the prediction, improving transparency and trustworthiness in AI systems.

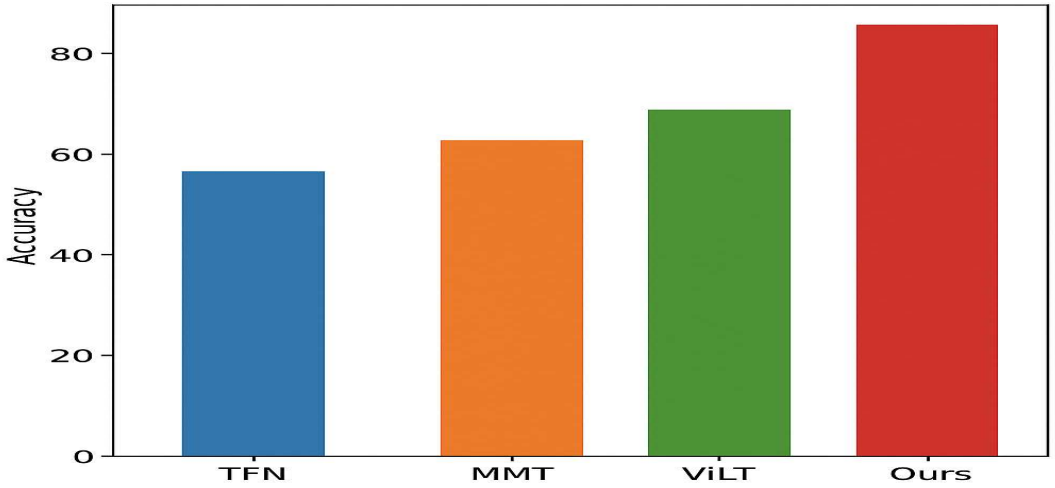

Figure 6: Comparative Accuracy of Multimodal Emotion Recognition Models (TFN, MMT, ViLT, and Proposed Method)

6. Conclusion

This research presents a Transformer-based multimodal emotion recognition system that effectively fuses audio and visual cues using cross-attention. Our proposed architecture demonstrates substantial improvements over unimodal and traditional fusion baselines on benchmark datasets. The experimental results affirm the importance of inter-modal feature learning and early fusion for robust emotion classification. These findings indicate the system's potential for real-world applications such as affective computing, e-learning platforms, and emotion-aware virtual assistants.

7. Future Work

Future extensions of this work could explore:

- Incorporating contextual metadata such as gesture or scene for broader emotion understanding.
- Exploring lightweight versions of the architecture for deployment on edge devices.
- Enhancing model robustness to noise and occlusions in unconstrained environments.
- Real-time adaptive learning with personalized emotion recognition models.

7. REFERENCES

[1] Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.
[2] Gong, Y., et al. (2021). AST: Audio Spectrogram Transformer. Interspeech.
[3] He, K., et al. (2016). Deep Residual Learning for Image Recognition. CVPR.
[4] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.
[5] Kim, W., et al. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. ICML.

[6] Livingstone, S.R., & Russo, F.A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). PloS One.
[7] Parkhi, O.M., et al. (2015). Deep Face Recognition. BMVC.
[8] Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML.
[9] Soleymani, M., et al. (2012). A Survey of Multimodal Sentiment Analysis. IEEE Transactions on Affective Computing.
[10] Tsai, Y.H.H., et al. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. ACL.
[11] Zadeh, A., et al. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP.