www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1461-1469

e-ISSN : 2583-1062
Impact Factor : 7.001

ADVANCES IN AUDIO-VISUAL EMOTION RECOGNITION: A COMPREHENSIVE REVIEW OF DEEP LEARNING APPROACHES

Jalluri Karthikeya Sai Sri Adithya[1], Y. Nagamani[2]

[1,2]Computer Science and Engineering GMR Institute of Technology Rajam, India
adithyajalluri2005@gmail.com
nagamani.y@gmrit.edu.in

DOI: https://www.doi.org/10.58257/IJPREMS37004

ABSTRACT

Audio-visual emotional analysis is an important part of affective computing, helping in areas like human-machine interaction, mental health, and autonomous systems. This review looks at the latest ways to use both audio and visual information to understand emotions and shows why combining these two types of information is helpful. However, many methods today don't fully use the shared information between audio and visual information, which makes them less effective. The review talks about different approaches, including advanced machine learning models that use attention mechanisms and methods to blend audio and visual information. It also examines new loss functions that help improve how features are learned from both types of information. The review includes methods like correlation analysis and joint loss strategies to combine audio and visual information better. It is based on studies using informationsets like RAVDESS, CREMA-D, eNTERFACE'05, and BAUM-1s. The review highlights both the strengths and weaknesses of current methods and suggests where more research is needed. Keywords: Audio-Visual Emotion Recognition, Attention Mechanisms, Feature Fusion, Deep Learning, Loss Functions.

Keywords- Audio-Visual Emotion Recognition, Attention Mechanisms, Feature Fusion, Deep Learning, Loss Functions)

1. INTRODUCTION

Emotion recognition using audio and visual information is a rapidly growing field that understands human emotions by analyzing signals like facial expressions, voice tone, and other non-verbal cues. Driven by advanced machine learning models, this technology is transforming areas such as healthcare, self-driving systems, and customer service by offering insights that enhance safety, improve user experiences, and allow for more personalized interactions. For example, in healthcare, it helps track mental health by spotting signs of distress, while in cars, it can detect driver fatigue, reducing accident risks. Previously limited to either video or audio information, today's emotional analysis systems combine both, allowing for a fuller understanding of complex emotions. Advanced models like 3D Convolutional Neural Networks (3D CNNs) and Operational Neural Networks (ONNs) have been essential for this progress, capturing emotions over time and identifying subtle patterns. However, challenges remain, including the high computing power these models require and how background noise or mixed emotions can reduce accuracy. Future improvements aim to make models more efficient, better at managing noise, and more culturally aware, preparing emotional analysis to transform areas like personalized learning, virtual reality, and human resources with richer, more intuitive interactions.

2. RELATED WORK

- This paper introduced MSER, a new system for multimodal speech emotional analysis that uses cross-attention to focus on important features from both audio and visual information. Cross-attention improves the interaction between audio and visual inputs by highlighting the most relevant emotional cues from each source. A deep fusion layer then combines these separately processed features into one emotional representation. The system's real-time capability is crucial for practical uses like virtual assistants, customer service, and emotion-aware human-computer interactions. [1]
- This paper developed a fuzzy logic-based system for recognizing children's emotions during computer games by combining both audio and visual information. The system used fuzzy rules to manage the uncertainty and variability in children's emotional expressions, which can differ from those of adults. The study showed that this fuzzy approach outperformed traditional methods in terms of accuracy and reliability, especially in noisy or uncertain environments. The authors concluded that this flexible method could be further expanded for use in a variety of applications, particularly in gaming contexts for children. [2]
- This paper proposed a deep CNN model with late fusion to improve real-time multimodal emotional analysis by combining audio and visual information at the decision level. They used a late fusion approach to merge emotional information from both modalities, which helped improve classification accuracy. The study compared this method

with early fusion and unimodal models, showing that processing audio and visual information separately before combining them leads to better performance. [3]

- This paper introduced a new multimodal architecture that combines audio and visual information for emotional analysis. The architecture included preprocessing steps to manage noisy information and improve the accuracy of emotional analysis. The paper also suggested future improvements, such as better synchronization methods and the ability to process information in real-time. [4]
- This paper introduced a deep operational neural network (ONN) model for emotional analysis, combining both audio and visual inputs. The model used audio features like pitch, tone, and rhythm, along with visual features such as facial expressions and movements, to capture a broad range of emotional signals. By using strong feature extraction and combining both types of information, the system was able to work effectively in dynamic settings, such as real-time emotion detection during live interactions or video analysis. [5]

This paper used multimodal fusion techniques to combine audio and visual information, enhancing the model's ability to process emotional information from both sources at the same time. The CNN architecture was improved to handle complex inputs, allowing the system to efficiently extract emotional features from both audio and visual information. The study showed that analysing both speech and facial expressions helped the system recognize more complex emotions than using just one type of information alone. [6]

This paper aimed to create a system that recognizes emotions by combining audio and video information using cross-modal fusion techniques. To improve performance, the system used attention mechanisms to focus on the most important features from both audio and video. Audio information was processed using mel -spectrograms, while video frames were analysed to detect facial expressions. When tested on standard information sets, this system outperformed other methods that only used one type of information or simpler fusion techniques. [7]

This paper created a advanced machine learning model using a convolutional neural network (CNN) to recognize emotions from both audio and visual information. The study focused on combining audio and video because emotions are shown through both facial expressions and voice tone. A fusion method was used to blend the features from both types of information, helping the model better understand emotions. The model was tested on standard emotional analysis information sets and performed better than models that used only audio or video. [8]

This paper focused on combining advanced machine learning features with a mixture of brain emotional learning (BEL) to improve emotional analysis from both audio and visual information. The system was designed to handle the multi-sensory nature of emotions, using both audio cues (such as speech tone and rhythm) and visual cues (like facial expressions and movements). The researchers emphasized that using biologically-inspired models like BEL helps make emotional analysis more reliable by mimicking how humans process emotional information from multiple sources. [9]

This paper developed a system called DARE, which was designed to trick multimedia-based speech recognition (AVSR) models by adding small, deceptive changes. The goal was to test how vulnerable AVSR models are to attacks in both the audio and visual parts of the system. DARE made subtle changes to lip movements in the video and speech signals in the audio to confuse the models. The paper highlighted the need to address these weaknesses, especially for AVSR systems used in security and authentication[10]

This paper worked on improving multimedia-based emotional analysis by finding and using the shared information between audio and video information. The model used audio features like voice tone and speech rhythm, along with visual features like facial expressions, to identify emotions. When tested on different emotional analysis information sets, the system performed better than models that didn't use the shared information between audio and video.[11]

This paper aimed to improve audio emotional analysis (AER) with limited labelled audio information by using large amounts of labelled facial expression information. They explored the connection between visual and audio information in an unpaired way, using a semi-supervised adversarial network. The model worked by estimating feature density and adding low-density generated samples to better define decision boundaries for more accurate emotional analysis. [12]

This paper developed a multi-task and ensemble learning framework for emotional analysis, using both visual and audio information. The system combined features from speech signals and facial expressions to improve emotion classification accuracy. It also explored how ensemble learning, which combines the results from multiple classifiers, can make the model more reliable and robust. By merging the outputs of different classifiers, the system produced stronger and more accurate predictions. [13]

This paper developed a hybrid advanced machine learning model for multimedia-based emotional analysis by combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs were used to extract spatial features from visual information, like facial expressions, while LSTMs (a type of RNN) modeled the changes in emotions over

time in audio information. The system combined both audio and visual information using fusion techniques, improving the accuracy of emotional analysis by using the strengths of both types of information. [14]

This paper developed a system for multimedia-based emotional analysis by combining speech features with facial expressions from video clips. The goal was to create a strong system that could recognize emotions in dynamic video clips, similar to real-world situations where emotions are shown through both speech and facial movements. The system was trained on standard emotion informationsets and tested on video clips, where it performed well in recognizing basic emotions such as happiness, anger, and sadness. [15]

3. METHODOLOGIES

1. Attention-Enhanced Multimodal Emotion Recognition Using Cross-Modal Fusion and Emotion-Constrained Loss [7].

The paper's methodology describes an advanced machine learning-based innovative approach. Approach to multimodal emotional analysis with complex attention mechanisms and a special loss function to raise classifying accuracy over the various emotional categories.

Here's a closer look at the primary components:

1. Feature Extraction from Visual and Audio Modalities

Visual Stream (3D-CNN): The method uses a 3D convolutional neural network (CNN) to analyze video information. This model, called ResNet-3D, breaks down each video into smaller parts, or snippets, to capture both the movement and details of what is happening over time. By performing calculations in three dimensions—height, width, and time—the model creates detailed feature maps that represent the visual content effectively.

Audio Stream (2D-CNN on Spectrograms): For audio analysis, the raw sound waves are transformed into spectrogram images that show how sound frequencies change over time. These images are processed using a 2D CNN, specifically ResNet-18. Similar to the visual stream, the audio is also divided into snippets, allowing the network to capture important audio features that help recognize emotions.

2. Attention Mechanisms for Detailed Focus

Spatial Attention (Visual Data): This mechanism helps the model focus on important areas in the visual information, such as facial expressions. It assigns more importance to parts of the video that likely convey emotions while downplaying less relevant background areas.

Channel Attention (Visual Data): This component evaluates which feature maps generated by the CNN are most significant for emotional analysis. It highlights channels that contain useful emotional information and diminishes those that do not contribute meaningfully.

Temporal Attention (Visual and Audio Data): Temporal attention adjusts the focus on different frames in both video and audio, recognizing that not all moments are equally important for conveying emotions. This is particularly useful for aligning key audio signals with visual cues, such as matching vocal tones with facial expressions during emotional peaks.

3. Cross-Modal Attention Fusion Cross-Attention Mechanism: The framework uses a cross-modal attention mechanism to effectively combine audio and visual information. This allows the audio and visual features to interact with each other, taking advantage of their complementary strengths. The model learns how these two types of information depend on one another, which improves the overall feature representation. The cross-attention process evaluates how relevant paired multimedia-based snippets are to each other and assigns weights based on their connection, ensuring that important features from one type enhance the features of the other.

Fusion Layer: After calculating the weights, the features from both audio and visual modalities are combined into a single representation. This unified representation is then sent to the classification layers. This fusion enhances the model's ability to classify emotions accurately by merging emotion-related information from audio and visual cues.

4. Emotion-Constrained Loss Function

Triplet Loss Foundation: The paper presents an emotion-constrained loss function that builds on the traditional triplet loss approach. This method ensures that for a specific anchor point (like a video snippet labeled as "happy"), similar samples (other "happy" snippets) are closer together than dissimilar samples (like those labeled "sad" or "angry"), maintaining a certain margin between them.

Emotional Metric Constraint: To improve this loss function further, an additional constraint is introduced that organizes emotions based on their similarities. Similar emotions (like happiness and surprise) are grouped closer together while opposing emotions (like happiness and sadness) are kept farther apart. This helps the model better distinguish between different emotions, leading to more accurate emotion representations.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1461-1469
e-ISSN : 2583-1062
Impact Factor : 7.001
www.ijprems.com
editor@ijprems.com

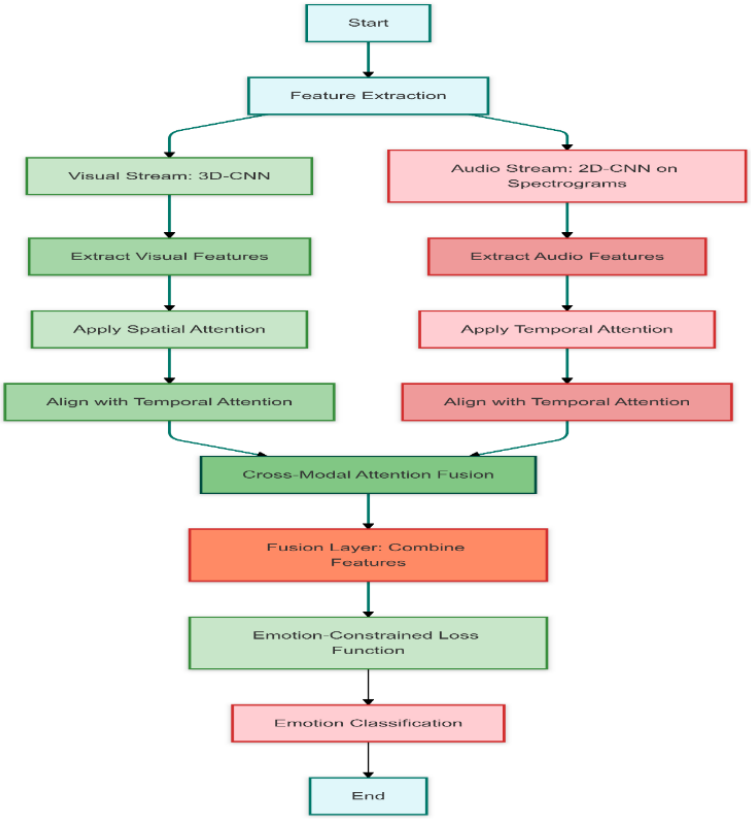

fig-(I)-Attention-Enhanced Multimodal Emotion Recognition Using Cross-Modal Fusion and Emotion-Constrained Loss

2. Enhanced Emotion Recognition through Visual-Audio Feature Fusion and Non-Linear Operational Neural Networks [5]

1. Visual Input – Key Frame Selection

Purpose:

To choose a single frame from the video that best shows the individual's emotional expression. Selection Process: Color Space Transformation: Video frames are converted into the LUV color space, which separates lightness from color information. This helps in consistently identifying frames with significant visual features.

Frame Filtering:

Color Differences: The model calculates differences in LUV values between consecutive frames, keeping only those that show significant changes.

Brightness Analysis: Frames are ranked based on their brightness, focusing on well-lit frames that likely show clear expressions. Entropy Calculation: Entropy measures how much information is in each frame. Frames with high entropy are prioritized as they likely contain more expressive features.

Contrast and Clustering: High-contrast frames are filtered and grouped using histograms to identify similar frames.

Laplacian Variance Sorting: This technique finds the frame with the most detail (least blur) by measuring pixel intensity variance, ensuring the selected frame has clear visual details for accurate facial feature extraction.

2. Audio Input – Mel Spectrogram Generation

Purpose:

Mel spectrograms visually represent sound frequencies, capturing pitch and tone nuances that relate to emotional expression.

Process of Generating Spectrograms:

Segmentation and Fourier Transform: The audio signal is divided into segments and transformed into frequency components using the Fourier transform, allowing frequency details to be mapped over time.

Mel Scale Conversion: The frequency information is converted to the Mel scale, which reflects how humans perceive pitch, highlighting frequencies important for detecting emotions.

Spectrogram Representation: The resulting Mel spectrogram is visualized as an image, enabling the model to use image-based techniques (like CNNs) for audio analysis.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1461-1469
e-ISSN : 2583-1062
Impact Factor : 7.001

3. Operational Neural Network (ONN) for Non-Linear Weight Calculations

Non-Linear Operations:

Nodal and Pool Operators: ONNs use non-linear nodal operators (like exponential or sinusoidal functions) in place of traditional linear weights. This simulates biological neuron processes, allowing the model to capture diverse and complex feature interactions.

Mathematical Transformation: Instead of simply summing pixel intensities after basic multiplications like CNNs do, ONNs apply advanced mathematical transformations across different non-linear functions. This includes using Taylor series expansions within each receptive field, enhancing the network's ability to learn intricate patterns in information.

4. Training and Fusion of Audio-Visual Features:

Parallel Branches for Audio and Visual Inputs: The model has two branches one for each input that run simultaneously. The key frame image and Mel spectrogram are processed independently in each branch using operational layers.

Feature Extraction in Each Branch:

In the visual branch, features related to facial expressions' spatial and temporal aspects are extracted. In the audio branch, spectral features indicating tone, pitch, and emotional intonations are captured.

Fusion Layer: The outputs from both branches are combined into a single feature vector. This fusion allows the model to integrate multimedia-based cues before classification, enhancing its ability to classify subtle emotions.

5. Multi-Input and Single-Input Comparison

To evaluate how effective the multi-input model is, a single-input ONN model was also tested. Results indicate that the dual-input architecture significantly outperforms the single-input model, confirming that using both visual and audio information improves classification accuracy

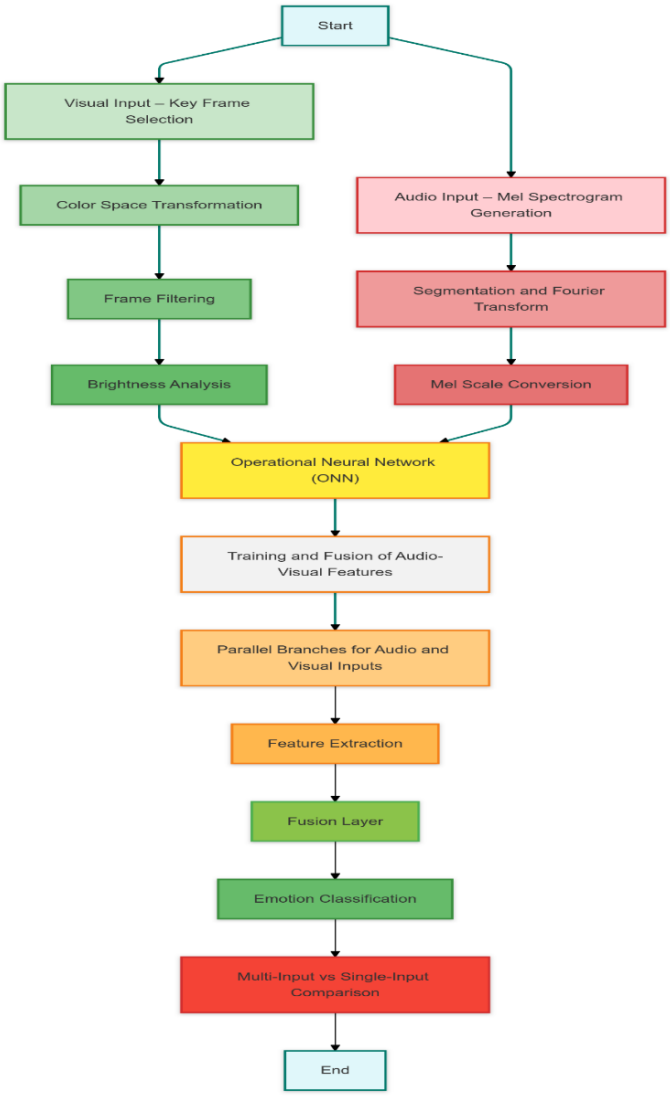

fig-2 - Enhanced Emotion Recognition through Visual-Audio Feature Fusion and Non-Linear Operational Neural Networks

3. Multimodal Emotion Recognition through Video, Audio, and Text Fusion Using Deep Convolutional Neural Networks [6]

This paper was designed to introduce and discuss emotion recognition using advanced machine learning. Specifically, deep convolutional neural networks (DCNNs), focusing on three types of Information:

Video: facial expression; audio: speech; text: spoken or written words. Here is a simplified breakdown of how it works:

1. Emotion Detection for Each Type of Data

Each type of information has its own process for identifying emotions. Here's how each one works:

A. Video Module: Emotion Classification from Facial Expressions

Dataset: FER-2013 info set of gray images numbered by tens of thousands labeled by emotion, on which the system is trained.

Steps:

Face Detection: The major facial parts including the eyes and mouth are detected because those parts play the most critical role in a facial emotion recognition system.

Image Simplification: An image is reduced to grayscale to reduce complexity and speed up its processing.

Image Processing: Techniques like batch normalization and information augmentation help the system learn better and avoid overfitting (learning too much from one set of information).

Model Structure: The video model makes use of three layers to process the images, incrementally identifying features before categorizing the emotion.

Performance: This system's performance side correctly identifies emotions 69% of the time. It struggles a little bit more to use these emotions that are poorly represented by the information, like fear and disgust.

B. Audio Module: Recognizing Emotions from Speech

Dataset: The RAVDESS informationset, containing audio files labeled by emotion, is used to train this part. Steps: **Noise Reduction**: Unwanted sounds are removed, and only the meaningful parts of the audio are kept.

Audio Segmentation: The audio files are divided into 3-second segments, capturing important speech features like pitch and rhythm.

Feature Extraction: The system focuses on specific audio characteristics (like pitch and energy) that change with different emotions.

Model Structure: This model uses Conv2D layers to process the audio features, capturing patterns that correlate with emotions.

Performance: This audio module achieves 100% accuracy, meaning it classifies emotions correctly every time, thanks to clear and effective audio features.

C. Text Module: Recognizing Emotions from Words

Dataset: The Emotion Text Dataset (ETD) with sentences labeled by emotion is used to train this module. Steps: **Text Cleaning:** Unnecessary elements (hashtags, emojis, special characters) are removed so the system can focus on the actual words.

Feature Engineering: Libraries like NLTK and neat text are used to process and analyze the words.

Model Testing: Several models (Logistic Regression, Support Vector Machine, etc.) are tested, and Logistic Regression performs best with 64% accuracy.

Performance: Recognizing emotions in text alone is challenging, so accuracy is slightly lower at 64%, due to the complexity and variety in expressing emotions through words.

2. Combining the Results from Video, Audio, and Text (Fusion)

Fusion Method: The system combines the predictions from each modality using a weighted approach, where each modality's importance is adjusted based on how people generally communicate emotions. Weighting: Video (Visual): 55%

Audio: 38%

Text: 7%

Calculation: The system calculates a final score by weighing each modality's accuracy based on these percentages. Performance: The combined model, with all three information types, reaches an 80% accuracy rate, showing that merging multiple information types makes the system more reliable.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1461-1469
e-ISSN : 2583-1062
Impact Factor : 7.001

3. Training and Evaluating the System

Setup: The models are trained using TensorFlow on a computer with a standard processor and graphics card. Training Process: Each model is trained for 100 rounds, using techniques to prevent overfitting. Metrics Used: The main evaluation measure is accuracy, with each modality tested separately before combining the results.

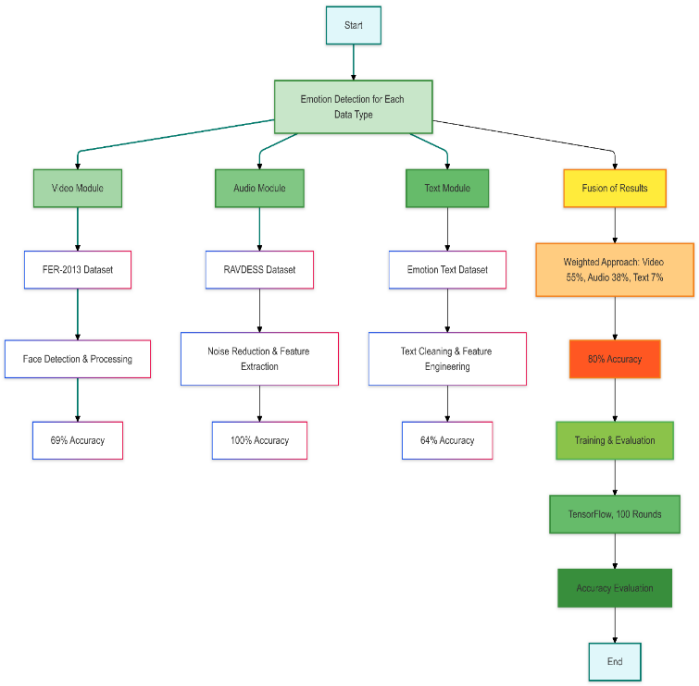

fig-3 - Multimodal Emotion Recognition through Video, Audio, and Text Fusion Using Deep Convolutional Neural Networks

4. RESULTS AND DISCUSSIONS

1. Main Models and Their Performance

Model 1: 3D CNNs with Cross-Modal Attention

- **Accuracy**: Achieved up to 89.25% on the RAVDESS dataset and 84.57% on CREMA-D.
- **F1 Score:** Averaged 0.88 on RAVDESS, showing it performs consistently well across different emotions.
- **Strengths:** Excellent at capturing dynamic facial expressions and focusing on key areas, like the eyes and mouth, that show emotions.
- **Drawbacks:** High computational needs, so it's difficult to use in real-time without a powerful GPU.

Model 2: Deep CNN with Model-Level Fusion (e.g., Dixit & Satapathy, 2024)

- **Accuracy:** Reached 86% on RAVDESS and 99% on SAVEE, proving its effectiveness across different emotional expressions.
- **F1 Score:** Averaged 0.92 on SAVEE, indicating high accuracy in identifying emotions in controlled settings.
- **Real-Time Use:** Efficient enough to work in real-time applications.
- **Drawbacks:** Faces challenges in noisy environments or with emotions that are hard to tell apart, like surprise and fear.

Model 3: Operational Neural Networks (ONNs)

- **Accuracy**: Achieved 96% on MEAD, 93% on RAVDESS, and 75% on MELD, demonstrating strong results across different datasets.
- **Single vs. Multi-Input:** Models using both audio and visual inputs were 25–45% more accurate than those using only one input.
- **Efficiency:** Training and testing times are comparable to CNNs but deliver better accuracy, making ONNs well-suited for real-time use.
- Strengths: ONNs' non-linear processing allows them to pick up on complex patterns in audio-visual data, outperforming CNNs by over 40% on MEAD and 52% on RAVDESS.
- **Drawbacks:** Their complexity and larger model size can be challenging for devices with limited resources.

Table.1 – results and discussion comparison table

Study	Method	Dataset	Accuracy	F1Score	Comments
[7]	Cross-modal fusion+ Attention	RAVDESS	89.25%	0.88	Strong fusion model, but computationally demanding
[5]	ONN with Non-linear Weights	MEAD, RAVDESS, MELD	96%,93%, 75%	0.95 (MEAD)	Achieved higher accuracy than CNNs in emotion classification across all datasets
[6]	Deep CNN (DCNN) + Text	FER-2013, RAVDESS, ETD	80% (Fusion)	0.79	Text integration improved accuracy, but computationally intensive for three-modality fusion
[14]	Hybrid CNN-RNN	RAVDESS	85%	0.84	CNN + RNN effective for temporal sequences, but less suitable for real-time due to RNN complexity
[3]	Late Fusion CNN	SAVEE	99%	0.92	Late fusion demonstrated superiority in real-time applications
[9]	Brain Emotional Learning	Custom	92%	0.90	Biologically inspired, robust in uncertainty, yet computationally intense

5. CONCLUSIONS

This paper reviews the latest progress in audio-visual emotion recognition, focusing on deep learning models that combine audio and visual data using fusion and attention techniques. By integrating both types of data, these models capture a wide range of emotional signals and reach high accuracy, with some models achieving up to 99% in controlled environments. Operational Neural Networks (ONNs) further enhance emotion detection by handling complex patterns, making them suitable for real-time applications. However, high computational requirements and difficulties with noisy

6. REFERENCES

[1] Khan, M., Gueaieb, W., El Saddik, A., & Kwon, S. (2024). MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. Expert Systems with Applications, 245, 122946.

[2] Kozlov, P., Akram, A., & Shamoi, P. (2024). Fuzzy approach for audio-video emotion recognition in computer games for children. Procedia Computer Science, 231, 771-778.

[3] Dixit, C., & Satapathy, S. M. (2024). Deep CNN with late fusion for real time multimodal emotion recognition. Expert Systems with Applications, 240, 122579.

[4] Goncalves, L., Leem, S. G., Lin, W. C., Sisman, B., & Busso, C. (2024). Versatile Audio-Visual Learning for Emotion Recognition. IEEE Transactions on Affective Computing.

[5] Aktürk, K., & Keçeli, A. S. (2024). Deep operational audio-visual emotion recognition. Neurocomputing, 588, 127713.

[6] Almulla, M. A. (2024). A multimodal emotion recognition system using deep convolution neural networks. Journal of Engineering Research.

[7] Mocanu, B., Tapu, R., & Zaharia, T. (2023). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image and Vision Computing, 133, 104676.

[8] Aghajani, K. (2022). Audio-visual emotion recognition based on a deep convolutional neural network. Journal of AI and Data Mining, 10(4), 529-537.

[9] Farhoudi, Z., & Setayeshi, S. (2021). Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. Speech Communication, 127, 92-103.

[10] Mishra, S., Gupta, A. K., & Gupta, P. (2021). DARE: Deceiving audio–visual speech recognition model. Knowledge-Based Systems, 232, 107503.

[11] Ma, F., Zhang, W., Li, Y., Huang, S. L., & Zhang, L. (2020). Learning better representations for audio-visual emotion recognition with common information. Applied Sciences, 10(20), 7239.

[12] He, G., Liu, X., Fan, F., & You, J. (2020). Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 912-913).

[13] Hao, M., Cao, W. H., Liu, Z. T., Wu, M., & Xiao, P. (2020). Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. Neurocomputing, 391, 42-51.

[14] Zhang, S., Zhang, S., Huang, T., & Gao, W. (2016, June). Multimodal deep convolutional neural network for audio-visual emotion recognition. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (pp. 281-284).

[15] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing, 10(1), 60-75.

[16] Shixin, P., Kai, C., Tian, T., & Jingying, C. (2022). An autoencoder-based feature level fusion for speech emotion recognition. Digital Communications and Networks.

[17] Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowledge-Based Systems, 244, 108580.

[18] Wang, Z., Wang, L., & Huang, H. (2020). Joint low rank embedded multiple features learning for audio–visual emotion recognition. Neurocomputing, 388, 324-333.

[19] Ghaleb, E., Popa, M., & Asteriadis, S. (2019). Metric learning-based multimodal audio-visual emotion recognition. Ieee Multimedia, 27(1), 37-48.

[20] Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q. F., & Lee, C. H. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2617-2629.