International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 10, Issue 3, pp 169-174, March-2026
https://doi.org/10.47001/IRJIET/2026.103024

Multimodal Identification of Emotions Using Facial Expressions and Physiological Signals

Nayana S Patel

Sri Siddhartha Institute of Business Management, Maraluru, Tumkur, Karnataka, India

Abstract **- Through the simultaneous analysis of voice signals and facial expressions, multimodal identification seeks to comprehend individual behaviors. In order to represent information more richly across several modalities, feature fusion is essential to this process. However, temporal misalignment between modalities and overfitting brought on by high-dimensional feature spaces are frequent problems for multimodal systems. An attention technique is developed to solve these problems by enabling the network to automatically concentrate on the most instructive local features. This approach is used by the network for both audio-visual feature integration and temporal modeling. The two primary contributions of this work are: first, it uses a multi-head self-attention mechanism to fuse audio and video features, reducing the influence of prior assumptions on the fusion process; and second, it uses a bidirectional gated recurrent unit to model the temporal dynamics of the fused features, incorporating autocorrelation coefficients along the time dimension as attention weights. Experimental results show that the proposed attention-based approach significantly improves multimodal emotion recognition accuracy.**

Keywords: Multimodal identification of Emotion, Audio–Visual Fusion, Attention Mechanism, Self-Attention, Temporal Dynamics, emotion recognition.

I. Introduction

Multimodal identification of Emotion

Multimodal emotional awareness is the process of recognizing and understanding human emotions by integrating information from multiple sources. Instead than relying on a single indicator, such as speech or facial expressions, this approach combines multiple modalities to produce a more accurate and thorough knowledge of an individual's emotional state.

The key modalities commonly used in identification of Emotion include:

Facial Expressions: Facial analysis focuses on changes in facial features and muscle movements to recognize emotions such as happiness, sadness, anger, and surprise.

Voice or Speech: Speech-based emotion recognition examines vocal attributes such as pitch, tone, rhythm, and intensity to detect emotional patterns in spoken communication.

Physiological Signals: Physiological data such as skin conductance, heart rate, and EEG signals are monitored to observe biological responses that reflect changes in mood. Multimodal identification research attempts to improve the precision and reliability of emotion detection systems by combining data from multiple senses. Combining them allows one source to compensate for the weaknesses of another because each medium has its own limitations. This fusion-based approach is highly effective in real-world situations when emotions are expressed through a range of facial expressions, verbal variations, and physiological reactions. Affective computing, virtual reality, human-computer interaction, and healthcare are just a few of the fields that use multimodal identification extensively. For instance, it can enhance the emotional awareness of virtual assistants, create more engaging gaming experiences, and assist in the diagnosis and treatment of emotional and mental health conditions in clinical settings.

II. Literature survey

Numerous scholars have investigated multimodal identification using physiological signs and facial expressions. Castellano and Caridakis (2008, et al.): used a Bayesian classifier and two-level fusion (feature and decision) to analyze speech, body language, and facial expressions. Kessous et al. (2010) used a Bayesian model for multimodal identification based on facial expressions, speech, and body motions, emphasizing better performance. According to Hussein et al. (2020), Dino and Abdulrazzaq (2020), and Adil (2021), deep learning models: convolutional neural networks (CNN) are commonly employed for feature extraction from video/image. Fusion techniques: enhancing robustness by merging features at the data or decision level. Data types: as demonstrated by studies by Zhang et al. (2020), investigations frequently mix facial expressions with voice (audio), textual data (text), and occasionally body motions or physiological information like EEG. and Fisioterapia Cuestiones. A two-stage fuzzy fusion-based convolution neural network for

ISSN (online): 2581-3048
Volume 10, Issue 3, pp 169-174, March-2026
https://doi.org/10.47001/IRJIET/2026.103024

dynamic emotion recognition was presented by Wuetal (2022). Chen et al.(2021): focused on MER with temporal and semantic consistency, utilizing audio, speech, and language, as seen in Pubmed and Science Direct.

III. How Facial Expressions Can Be Recognized

The technique of identifying and deciphering human emotions using facial expressions and visual cues is known as facial expression recognition. The key steps in successfully identifying facial expressions are covered in this section.

1. Data Collection

Facial Landmarks: The first step in facial expression recognition usually involves identifying and tracking facial landmarks. These landmarks are key points on the face, including the eyes, eyebrows, nose, and mouth. Computer vision methods, such as facial landmark detection algorithms, are applied to locate and follow these points with precision.

2. Feature Extraction

Action Units (AUs): Action Units, which correlate to particular facial muscle movements, are frequently used to depict face emotions. To recognize expressions, these units are necessary. For instance, the muscles surrounding the lips and eyes are activated when you smile. Geometric and Texture Features: In addition to Action Units, other features are retrieved, such as patterns of facial texture and geometric correlations between facial landmarks. These characteristics offer useful data for examining emotional responses.

3. Model Training

Machine learning and deep learning models are trained to categorize facial expressions following feature extraction. Traditional machine learning methods, such as Support Vector Machines (SVMs) and Random Forests, depend on manually generated features. On the other hand, deep learning techniques like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) automatically extract important features from raw image or video data.

4. Labeling and Annotation

Dataset Preparation: To train face expression recognition algorithms, a labeled dataset is needed. The relevant emotion labels are labeled on the images or video frames in this collection. For this, popular datasets like FER2013 and CK+ (Cohn–Kanade) are frequently utilized.

5. Model Evaluation

Performance Assessment: An alternate test dataset is used to assess the model's generalizability. Accuracy, precision, recall, and F1-score are common evaluation metrics.

6. Real-Time Detection

Application: The model feasible to use in real-time applications after it has been trained and assessed. This allows the system to analyze live video streams or static images to determine a person’s emotional state.

7. Post-Processing and Interpretation

Contextual Information: Incorporating contextual information, such as the user's background or the environment, can occasionally increase recognition accuracy.

Temporal Analysis: Examining facial expressions over time enables a better understanding of emotional changes because emotions often manifest gradually rather than rapidly. Deep learning techniques have greatly advanced facial expression recognition, resulting in increased robustness and accuracy. System performance is still impacted by issues including variable lighting conditions, different head positions, and the requirement for a variety of training datasets, despite these developments.

IV. How Voice or Speech Can Be Recognized

Voice or speech recognition is the process of translating spoken words into text or other useful representations. This section describes the main techniques for correctly recognizing and deciphering human speech.

1. Acoustic Feature Extraction

Spectral Analysis: Analyzing the acoustic properties of the voice signal is the initial stage in speech recognition. Fourier analysis and other methods are used to separate the signal into its frequency components. Mel-Frequency Cepstral Coefficients, or MFCCs, are one of the most widely used features in sound processing. They portray the short-term power spectrum of sound and faithfully reflect the essential components of human speech.

2. Signal Preprocessing

Noise Reduction: To enhance recognition performance in noisy environments, noise reduction techniques are employed to raise the signal-to-noise ratio..

Normalization: Normalization is used to maintain consistency in extracted features and to manage variations in signal amplitude.

ISSN (online): 2581-3048
Volume 10, Issue 3, pp 169-174, March-2026
https://doi.org/10.47001/IRJIET/2026.103024

3. Language Modeling

Phonetic and Language Models: Phonetic and linguistic models are used by speech recognition systems to forecast appropriate word and sound sequences. To increase recognition accuracy, language models include grammatical and structural knowledge.

4. Acoustic Modeling

Markov models that are hidden (HMMs): The acoustic behavior of speech is frequently represented by hidden Markov models. These models explain changes between various states, each of which is associated with a distinct phoneme or sound.

5. Machine Learning and Deep Learning

A labeled dataset is required in order to train face expression recognition algorithms. The pictures or video frames in this collection are labeled with the appropriate emotion descriptors. Popular datasets like CK+ (Cohn–Kanade) and FER2013 are often used for this.

6. Model Training

Supervised Learning: Labeled datasets with coupled audio recordings and text transcriptions are used to train speech recognition models. The model learns to link auditory features to textual outputs during training.

Adaptation and Transfer Learning: Models can adjust to various speakers, accents, and acoustic surroundings through the use of transfer learning and fine-tuning approaches.

7. Model Evaluation

Testing with Unseen Data: Unseen data is used to test the trained model's generalization performance. Accuracy, word error rate (WER), and character error rate (CER) are common evaluation criteria.

8. Real-Time Recognition

Application: After training and evaluation, speech recognition systems can be deployed in real-time applications to convert spoken language into text or other required outputs.

9. Post-Processing and Natural Language Interpretation (NLP)

NLP Techniques: Additional natural language processing methods, such as syntactic and semantic analysis, can be applied to improve the interpretation and understanding of the recognized speech. Speech recognition technologies are widely used in applications such as voice-activated devices, virtual assistants, transcription services, and automated customer support systems. Recent advances in deep learning have greatly increased the accuracy, robustness, and adaptability of modern voice recognition systems.

V. How Physiological Signals Can Be Used in Facial Recognition

The accuracy and resilience of emotion recognition and general human condition analysis can be significantly increased by combining physiological information with facial recognition methods. Systems are able to obtain a more thorough grasp of both underlying physiological responses and emotional reactions by integrating physiological indications. This types describes how facial recognition techniques can be successfully integrated with physiological inputs.

1. Types of Physiological Signals

Heart Rate (HR): Since heart rate variations reflect the action of the autonomic nervous system, they are important indicators of emotional arousal.

Electrodermal Activity (EDA): Skin conductance measurements show changes in sweating, which are strongly associated with stress and emotional intensity.

Electroencephalography (EEG): EEG recordings provide information about emotional and cognitive states by analyzing brainwave patterns.

Respiration Rate: Breathing patterns offer additional information about emotional and physiological responses, complementing other signals.

2. Data Collection

Simultaneous Recording: Facial expressions and physiological signals are captured concurrently to maintain consistency. This is typically achieved using wearable devices such as heart rate monitors, skin conductance sensors, and EEG equipment.

3. Signal Synchronization

Time Alignment: Precise synchronization between facial data and physiological signals is essential. Proper time alignment enables accurate mapping of physiological changes to specific facial expressions.

4. Feature Extraction

Physiological Features: Physiological signals are the source of key features such skin conductance responses, heart rate variability, and frequency-based EEG components.

ISSN (online): 2581-3048
Volume 10, Issue 3, pp 169-174, March-2026
https://doi.org/10.47001/IRJIET/2026.103024

Correlation Analysis: To find patterns linked to various emotional states, the relationship between face traits and physiological markers is examined.

5. Multimodal Fusion

Feature-Level Fusion: Features from physiological data and facial expressions are combined to form a single representation.

Model-Level Fusion: Emotion identification algorithms evaluate physiological and facial information simultaneously. Neural networks and other multimodal machine learning techniques are widely employed for this.

6. Training and Cross-Validation

Dataset Preparation: Labeled datasets containing synchronized facial expressions and physiological signals are prepared for training and evaluation.

Model Training: Machine learning and deep learning models are trained to associate multimodal inputs with corresponding emotional states.

7. Real-Time Recognition

Online Processing: Real-time processing systems are built to concurrently evaluate physiological inputs and facial emotions for practical applications such as affective computing.

8. Applications

Human–Computer Interaction: By incorporating both facial and physiological cues, systems can respond more naturally and adaptively, improving human–computer interaction.

Healthcare: In medical and healthcare applications, this multimodal approach can support stress detection, emotional monitoring, and mental health assessment.

9. Interpretation and Analysis

Contextual Understanding: Examining physiological and facial data takes situational and environmental context into consideration. The significance of observed emotional patterns is made clear by contextual information. Creating comprehensive models that can faithfully capture the complex relationship between physiological responses and emotions is the aim of fusing physiological information with facial recognition. This multimodal approach holds great potential for affective computing, mental health monitoring, and human-computer interaction.

VI. Integrating Physiological, Voice, and Facial Signals to Identify Emotions

The accuracy and dependability of emotion recognition can be significantly increased by combining voice, facial expressions, and physiological cues. Systems can produce a more comprehensive and nuanced picture of a person's emotional state by integrating data from many modalities. The following steps outline a standard multimodal multimodal identification of emotions workflow.

1. Data Collection

Simultaneous Recording: All modalities should be captured at the same time to ensure temporal alignment. This typically means using sensors such as cameras for facial expressions, microphones for speech, and heart rate monitors or EEG devices for physiological signals.

2. Preprocessing

Signal Alignment: Data from different modalities must be synchronized to correspond to the same time intervals. Proper alignment is essential for accurate feature fusion and analysis.

3. Feature Extraction

Voice Features: Among the acoustic features that are obtained from speech data are pitch, intensity, and formant frequencies.

Facial Features: Face landmarks, expressions, and movement patterns are examples of visual elements that are extracted from images or video frames.

Physiological Features: Physiological signals, such as variations in skin conductance, EEG patterns, and heart rate variability, are the source of features.

4. Individual Modality Processing

Modality-Specific Models: Different deep learning or machine learning models are trained for each modality, such as specific algorithms for physiological inputs, facial expression models for visual data, and speech recognition models for voice.

5. Fusion Strategies

Early Fusion: Create a single joint feature representation early on by combining features from all modalities. This representation is then fed into a single unified model.

Late Fusion: To determine the ultimate feeling, process each modality separately and then combine the outputs of each model.

ISSN (online): 2581-3048
Volume 10, Issue 3, pp 169-174, March-2026
https://doi.org/10.47001/IRJIET/2026.103024

Decision-Level Fusion: To reach a final judgment, combine model projections using methods like averaging or voting.

6. Model Training

Multimodal Model Training: To understand the relationship between multimodal inputs and emotional states, train a model using the integrated characteristics.

7. Cross-Validation

Performance Evaluation: To verify generalization, assess the trained model on an independent dataset. Usually, metrics like confusion matrices, accuracy, and F1-score are employed.

8. Real-Time Processing

Online Implementation: Create pipelines that process voice, face, and physiological data in real-time. This is crucial for apps that need to recognize emotions instantly.

9. Contextual Information

Context Integration: Environmental and situational context, such as surroundings or user history, can be incorporated to enhance recognition accuracy.

10. Applications

Human–Computer Interaction: Enable virtual assistants, gaming platforms, and educational tools to respond adaptively to user emotions.

Healthcare: In clinical or wellness applications, assist with stress detection, emotional tracking, and mental health evaluation.

VII. Conclusion

By merging voice, facial expressions, and physiological information, this multimodal approach offers a more thorough and accurate representation of human emotions. Such a paradigm is especially helpful in real-world scenarios where emotional states are frequently expressed simultaneously through multiple channels. By combining behavioral and physiological research, this method increases the accuracy of emotion identification. Additionally, it provides a helpful foundation for further research in human-computer interaction, emotional computing, and related areas.

VIII. Future Scope

Despite the impressive outcomes of the suggested multimodal emotion detection framework, there are still several areas that could be improved. In order to better comprehend human emotions, future research can improve the system by incorporating more information sources, such as context, body posture, eye movements, and environmental cues. Furthermore, the model may be able to function well across a variety of ethnicities and cultures while requiring less large labeled datasets because to recent advancements in deep learning and self-supervised learning. The real-time implementation of the framework in useful applications, such as social robots, driver assistance systems, personalized learning platforms, and mental health monitoring, is another crucial route. For dependable operation in real-world scenarios, the system's capacity to manage noise, sensor faults, and missing data must be improved. Future research should concentrate on ethical issues such safeguarding user privacy, minimizing model bias, and simplifying system evaluations. Multimodal emotion recognition systems can have a greater overall impact if long-term, cross-cultural research is conducted to better understand how emotions evolve over time.

REFERENCES

[1] B. Cheng and G. Y. Liu, "Emotion recognition from surface EMG signal using wavelet transform and neural network," *J. Comput. Appl.*, vol. 28, no. 2, pp. 1363–1366, 2008.

[2] Y. Xie, R. Liang, Z. Liang, X. Zhao, and W. Zeng, "Speech emotion recognition using multihead attention in both time and feature dimensions," *IEICE Trans. Inf. Syst.*, 2023.

[3] R. Harper and J. Southern, "A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat," 2019.

[4] S. Tomar, A. Gupta, and S. Rastogi, "Human behaviour recognition through AI," *GLIMPSE - J. Comput. Sci.*, vol. 2, no. 2, pp. 36–37, Jul.–Dec. 2023.

[5] R. D. Lane, P. M. Chua, and R. J. Dolan, "Common effects of emotional valence, arousal, and attention on neural activation during visual processing of pictures," *Neuropsychologia*, vol. 37, no. 9, pp. 989–997, 1999.

[6] J. Pan, Y. Li, and J. Wang, "An EEG-based brain–computer interface for emotion recognition," *in Proc. Int. Joint Conf. Neural Netw.*, pp. 2063–2067, 2016.

[7] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomed. Signal Process. Control*, vol. 70, 103029, 2021.

[8] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ECG-based emotion recognition in music listening," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 85–99, Jan.–Mar. 2017.

ISSN (online): 2581-3048
Volume 10, Issue 3, pp 169-174, March-2026
https://doi.org/10.47001/IRJIET/2026.103024

[9] R. Sharma, "Analysis of human sentiments using machine learning," *GLIMPSE - J. Comput. Sci.*, vol. 2, no. 2, pp. 46–51, Jul.–Dec. 2023.

[10] T. Tunce, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge-Based Syst.*, vol. 211, 106547, 2021.

Citation of this Article:

Nayana S Patel. (2026). Multimodal Identification of Emotions Using Facial Expressions and Physiological Signals. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(3), 169-174. Article DOI https://doi.org/10.47001/IRJIET/2026.103024

****

