National Journal of Signal and Image Processing, ISSN: 3107-5231 Vol. 1, No. 3, 2025 (pp. 23-30)

RESEARCH ARTICLE

ECEJOURNALS.IN

Multimodal Fusion Techniques for Emotion Recognition Using Audio, Visual, and Physiological Signals

K. Maidanov[1]*, G.F. Frire[2]

[1]*Department of Electrical and Computer Engineering, Ben-Gurion University, Beer Sheva, Israel*

[2]*Departamento de EngenhariaElétrica, Universidade Federal de Pernambuco - UFPE Recife, Brazil*

KEYWORDS:
Multimodal Signal Processing, Time-Frequency Analysis,
Feature Fusion,
Adaptive Filtering,
Emotion Recognition,
Physiological Signal Processing, Audio-Visual Processing,
Deep Learning,
Affective Computing,
Robust Classification

ARTICLE HISTORY:
Submitted : 18.02.2025
Revised : 06.03.2025
Accepted : 13.05.2025

https://doi.org/10.17051/NJSIP/01.03.04

Abstract

Sensitive emotion recognition information depends on high-level signal and image processing technologies to acquire, synchronize, and merge multimodal data involving heterogeneous information. One-modal methods are usually sensitive to isolated noise, occlusions or information losses. This paper examines multimodal signal processing techniques which combine audio, visual and physiological information using new feature extraction pipelines and adaptive fusion algorithms. The suggestion of fusion using a hybrid deep learning focus refers to the following presentation of integrating time and space representations by achieving a fusion of synchronized time frequency features, statistical descriptors, and deep embeddings, combined with adaptive weighting operations that help to overcome domain-specific noise. Test on benchmark datasets shows that the framework is more accurate, F1-score, and robust under realistic conditions when compared with unimodal and conventional fusion methods. This work helps in achieving scalable, real-time and noise-invariant multimodal systems that can be applied to healthcare, adaptive interfaces and affective computing through increased signal preprocessing, feature-level integration and classifier decision fusion.

Author's e-mail: rantlin.h@gmail.com, fg.frire@cesmac.edu.br

How to cite this article: Maidanov K, Frire GF. Multimodal Fusion Techniques for Emotion Recognition Using Audio, Visual, and Physiological Signals. National Journal of Signal and Image Processing, Vol. 1, No. 3, 2025 (pp. 23-30).

Introduction

Recognition of emotion is fundamentally a multimodal signal and image processing issue and this necessitates heterogeneous data streams to be acquired, analyzed and fused. These are speech (with prosodic and spectral features), visual (including facial expression and micro-movements) and physiological (biosignals showing inner emotional deviations). Each modality has its own contributions to make to data, but is hampered by the weaknesses inherent to it like noise pollution, occlusion, time-sampling differences, or half-scans.

In signal processing terms, unimodal systems can be very difficult because they are such a heavy single-channel data channel system when it comes to dealing with robustness. (As an example, audio-only systems perform poorly in noisy conditions, vision-based ways are defeated in low illumination or occluded conditions, and physiological-signal-only systems might be subject to motion artefacts or sensor noise.) Multimodal frameworks, in contrast, can utilize the complementariness of the various signals so that any system would be allowed to use the strengths of one signal to overcome the deficiencies of another signal.

Developments in time-frequency analysis, adaptive filtering and statistical modeling of features has resulted in more accurate performances of emotional inferences across modalities. Spectral prosodic features include Mel-Frequency Cepstral Coefficients (MFCCs) and pitch-energy profiles; spatial representations include Local Binary Patterns (LBP), Histograms of oriented gradient (HOG), and deep convolutional embeddings; frequency domain measurements that can be converted into metric values include: the power spectral density (PSD) of an electroencephalogram (EEG), heart rate variability (HRV), and wavelet-based analysis of GSR responses. Such pipelines of signal processing are the basis of the robust systems of emotion recognition.

The proposed multimodal fusion framework is presented in Figure 1, and based on signal modalities, it fuses

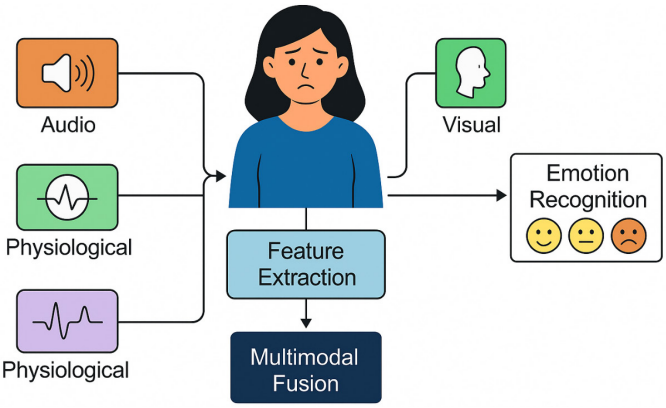

Fig. 1: Multimodal Fusion Framework for Emotion Recognition

This diagram illustrates the multimodal fusion framework that integrates audio, visual, and physiological signals for emotion recognition. The process includes feature extraction from each modality, followed by fusion to derive the final emotional state classification.

the signals by extracting the feature and fusing them together to improve accuracy levels of recognition. Also significant is the integration of multimodal aspects into a single decision-making system, and signal processing studies can provide a few solutions to this issue. Feature-level fusion is the fusion at feature level achieved after extracting the feature vectors and concatenating or embedding these into a joint representation, to allow learning cross-modal correlation. Modality-specific classifiers at the decision-level are trained to generate independent predictions, which are fused by weighted voting or Bayesian integration thus becoming less prone to complete data loss. Hybrid fusion alternately invokes feature integration and decision aggregation in order to learn deep interdependencies in a modular manner. Nonetheless, there are still limitations when it comes to carrying out synchronizing and aligning multimodal data streams because of the diverse sampling rates and different levels of latency. Additionally, the resilience to noise, especially in a real world scenario, requires highly complex preprocessing techniques and adaptive fusion techniques.

We consider such challenges in the current study by proposing a framework-based multimodal signal processing using a combination of temporal and spatial representations of audio, visual, and physiological signals in a deep learning-driven multimodal representation system. The suggested strategy has entailed a powerful preprocessing pipeline that has improved signal quality by removing noise, normalizing and aligning the signal in time. It also presents a new multimodal scheme of feature extraction based on handcrafted signal descriptions and semantically-related deep neural network embeddings to provide informative emotional encoding. A learning framework of adaptive hybrid fusion is used to combine the cross-modal correlation learning with decision-level robustness so that the framework remains robust in the face of high noise or incomplete data. Evaluation of the proposed framework on benchmark datasets has shown that it yields higher accuracy, precision, recall, and robustness rates than the unimodal and traditional fusion techniques and can therefore be used in real-world multimodal emotion recognition.

Related Work

Recognition of emotion has been already deeply examined with many modalities including speech, facial expression, and physiological. Initial studies were on unimodal systems where just one source of information is considering in detecting the emotion. Such systems however may be flawed in terms of environmental noise, artifacts of sensors or insufficient information and as such tend to lack robustness and generalization. In order to mitigate these defects, the recent research highlights multimodal fusion approaches which are the combination of complementary data sources to enhance the reliability and recognition accuracy.

Audio-Based Emotion Recognition

Audio-based emotion recognition mostly takes advantage of properties arising out of speech signal like pitch, intensity, formants and spectral properties. The use of traditional machine learning paradigms like Support Vector Machines (SVM),[1] Hidden Markov Models (HMM),[2] Gaussian Mixture Models (GMM)[3] was common when it comes to classifying emotional states. In more recent times there has been a good performance demonstrated by the deep learning methods, particularly, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), that learn the hierarchical and the temporal representations directly using raw audio or spectrograms as input.[4, 5]

Visual-Based Emotion Recognition

The visual modalities examine facial expressions, micro-expressions, and gestures that are registered with the help of cameras. One type of feature extraction is handcrafted feature such as Local Binary Patterns (LBP)[6] and Histogram of Oriented Gradients (HOG).[7] Other feature extraction includes deep learning features where CNNs are used.[8] CNNs allow discrimination features to be automatically learned starting with raw images or frames of video data which greatly increases recognition accuracy. Long short-term memory (LSTM) networks are

considered temporal models that are commonly used to model changes in expressions over time.[9]

Physiological Signal-Based Emotion Recognition

The physiological signals as a type of biosignal are indications of internal emotional states in the form of reactions that are measured through an electrocardiogram (ECG), electroencephalogram (EEG), galvanic skin response (GSR) and other signals of the autonomic nervous system. These cues furnish sincere complementary data to audio and visual hints particularly in situations when other external manifestations are not clarified or held-back.[10] The methods of feature extraction usually include time-domain, frequency-domain, and time-frequency domain-based estimates, like power spectral density estimations and wavelet transforms.[11] RNNs and CNNs stand out in deep learning models used in emotion recognition via model ing of the temporal shaping of physiological signals.[12]

Multimodal Fusion Approaches

Fusion of several modalities increases the performance of emotion recognition, since it exploits advantages of various sources of data. The major fusion approaches that are discussed in literature are:

- Feature-Level Fusion: In this, the feature vectors representing data in each modality are concatenated by being combined together in space or by being jointly embedded in a common space prior to classification.[13] The feature-level fusion can learn the inter-modal correlations, but can be subject to the curse of dimensionality and needs some feature normalization and synchronization.
- Decision-Level Fusion: This is where decisions rendered by different classifiers that are trained on different modalities are combined using methods like majority voting, weighted averaging or Bayesian fusion.[14] Decision-level fusion gives a modular and robustness to missing modalities and does not capture deep cross modal interactions.
- Hybrid fusion: hybrid fusion combines the feature and the decision levels strategies in order to realize the strengths of the former and the latter. As an example, the module implements modality-specific features that are first learnt and combined at the feature level, after which a collection of classifiers are employed with their decisions combined to achieve robustness.[15] The recent research applies deep learning models, which involve attention mechanisms to dynamically weigh modalities as part of their reliability.[16]

A number of multimodal emotion recognition systems that incorporate a combination of audio signals, visual signals and physiological signals have been proposed. Zeng et al.[17] joined both facial expressions and speech features, realizing considerable advances with respect to unimodal costs. Zheng and Lu[18] combined the data to the EEG and facial expressions with deep multimodal fusion demonstrated a greater level of robustness in noisy environments. In spite of these developments, other problems including modality synchronization, missing data and computational complexity are still sources of active research.

Methodology

Data Acquisition and Preprocessing

To learn a multimodal emotion recognition system capable of learning multimodal patterns via robust recognition, we employ publicly available benchmark data sets like DEAP and MAHNOB-HCI that offer synchronized audio, visual and physiological signals. Such datasets are representative, diverse and comprise a sample of naturalistic emotional reaction with subjects subjected to a variety of stimuli. In order to analyse the data, it must first be prepared by undergoing preprocessing steps that aim to clean and improve the signal quality as well as align the modalities. A number of noise filtering filters are used to eliminate artifacts (including background noise in sound signals, motion blur in video frames and electrical noise in physiological recordings). Each modality is normalised so that features lie within similar ranges to allow the features to be compared and combined fairly. Because of varying sampling rates and the potential temporal shifts between modalities, good data fusion requires synchronization algorithms to compensate correctly and precisely, combined with modalities that are aligned. Figure 2 shows how data acquisition pipeline entails separation of modalities based on audio, visual and physiological streams, noise filtering, normalization and synchronization of the data streams before the data can be subjected to feature extraction.

Feature Extraction

The successful extraction of features converts raw data into meaningful representations that presents emotional signals. In the audio modality we extract Mel-Frequency Cepstral Coefficients (MFCCs) describing spectrum properties of speech, pitch properties and energy

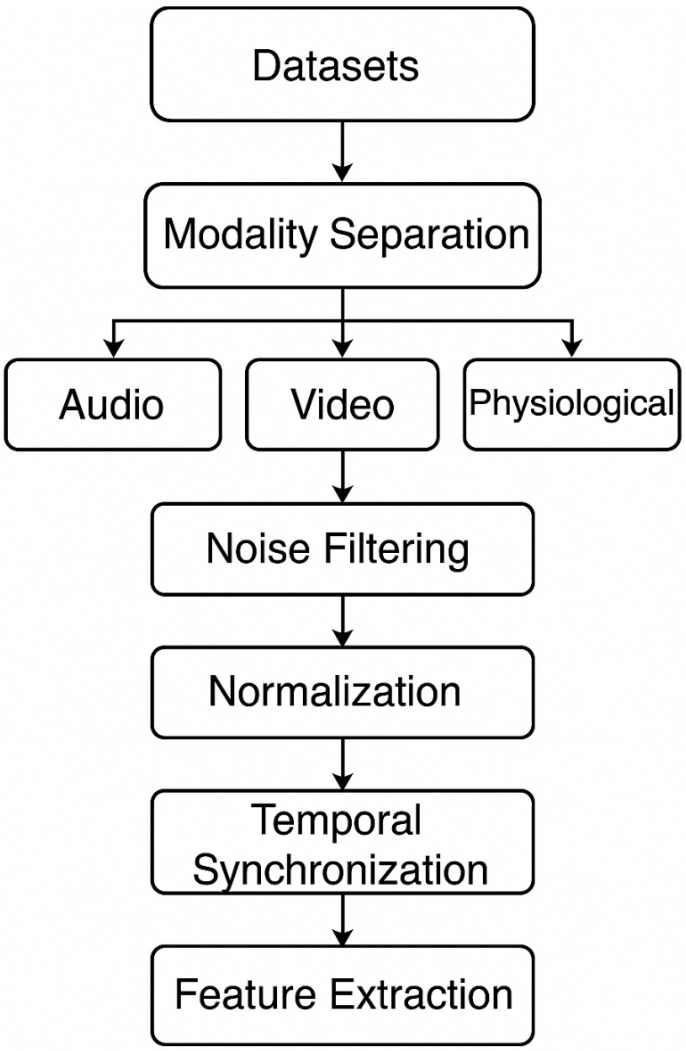

Fig. 2: Data Acquisition and Preprocessing Pipeline for Multimodal Emotion Recognition

properties correlating with prosodic changes which vary with emotion. Optical characteristics make use of the Facial Action Coding System (FACS), which codifies facial muscle activity in line with facial expressions. Figure 3 illustrates the feature extraction procedure that defines the particular features extracted in audio (MFCC, pitch, energy), visual (FACS, CNN embeddings), and physiological (HRV, EEG PSD, GSR) modalities constituting the unified set of features to be used in classification.

CNN-based embeddings Pretrained CNN networks have been used to extract so-called hierarchical spatial features from facial images or video frames, allowing strong representations of subtle expressions and micro-expressions. In the physiological scenario, Heart Rate Variability (HRV) of the ECG, EEG Power Spectral Density (PSD) of the brain waves, Galvanic Skin Response (GSR) amplitude of Autonomic nervous systems arousal, are some of the features taken into account. A combination of these features extracts complementary emotional cues based on these corresponding external behavior and internal physiology.

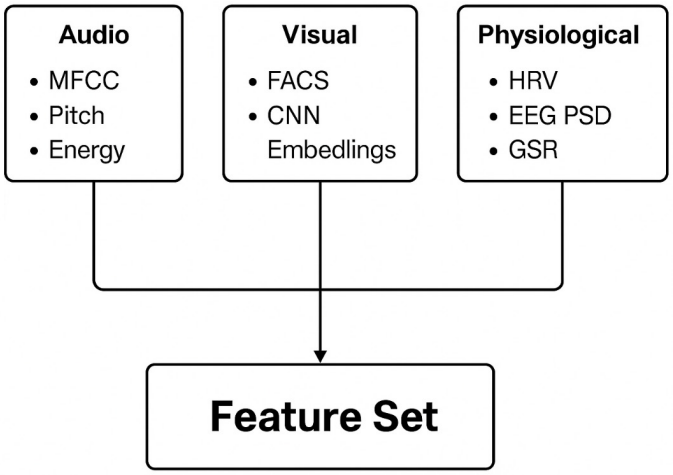

Fig. 3: Feature Extraction Process for Audio, Visual, and Physiological Modalities

Fusion Techniques

One of the key elements of this methodology is the integration of multimodal features and we address three main fusion strategies through which the information provided by audio, visual, and physiological signals can be successfully integrated. In feature-level fusion, the features extracted in each modality are combined into a one complete vector to present the emotional state and they pass into deep learning classifiers like the Long Short Term Memory (LSTM) network. These networks are very appropriate to predict sequential dependencies in time and context dependencies in data. Conversely, decision-level fusion makes separate classification of each modality by separate classifiers, with the decisions of each modality combined via weighted majority voting. This methodology offers modularity and stability to the system enabling it to work even with modalities that are not available or noisy. Finally, the hybrid fusion approach also integrates the advantages of both by first executing inter-modal relationships through joint feature learning processes and then sequentially fusing decisions obtained through layers that estimate inter-modal relationships at the levels of modality by applying the decision fusion procedure to improve the final emotion classification. It is a two-stage architecture that attempts to strike a balance between the merits of integrating cross-modal feature interactions on a deep level with the adaptability and robustness of decision-level fusion. Figure 4 shows different multimodal fusion strategies (feature-level fusion, decision-level fusion, and hybrid fusion) and indicates how features or classifier decisions tend to be fused to enhance emotion recognition performance.

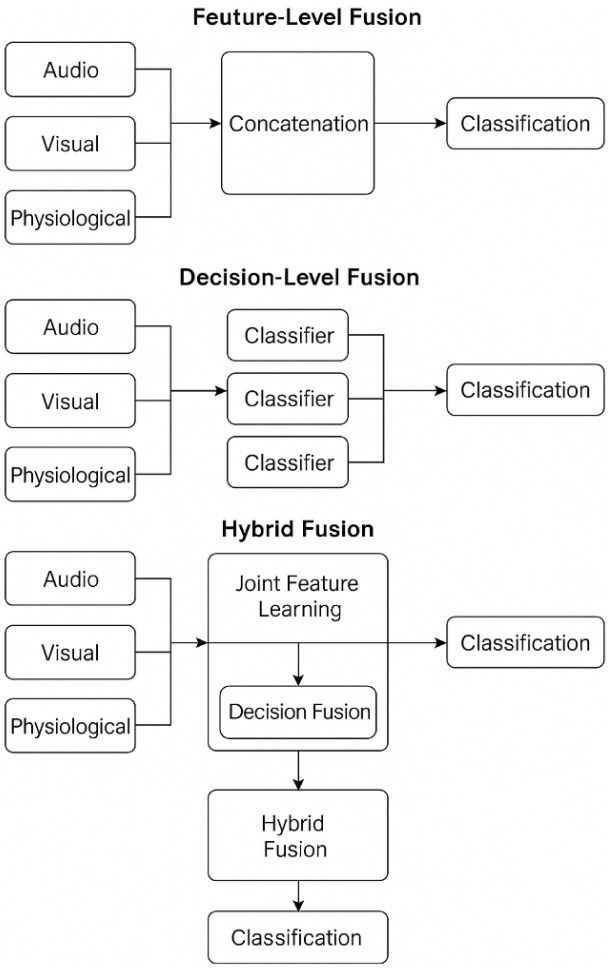

Fig. 4: Fusion Strategy Architecture for Multimodal Emotion Recognition

Classification Models

Further, to categorize emotional states based on fused features, we test a collection of machine learning models. The conventional classifiers such as Support Vector Machines (SVM) and Random Forests serve as good baselines as they are effectively used on structured features. Nonetheless, the deep learning architectures, especially the CNN-LSTM models, are investigated in their potential of automatically learning complicated spatial-temporal patterns. CNN layers capture hierarchical features of the inputs (in particular visual data), and LSTM layers capture the dependencies in time that are key to modeling changing emotional states in time-sequential data. We quantitatively benchmark the accuracy of these models both in terms of robustness and speed of execution/computational costs in order to determine which model is best suited to the real-time multimodal emotion recognition applications. Figure 5 describes the process of classification, where fused features along with CNN-LSTM extracted embeddings are fed into a predictor model through the SVM, Random Forest, and CNN-LSTM architecture, after which prediction of the final emotion class is obtained.

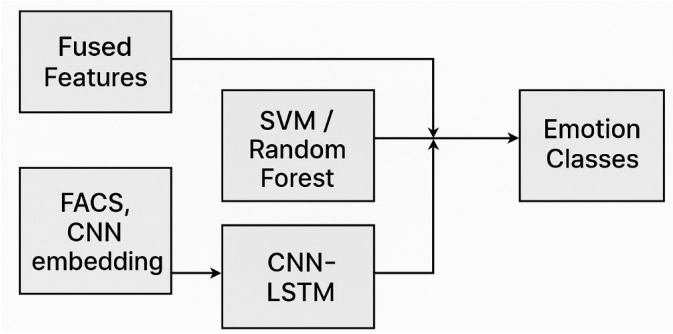

Fig. 5: Classification Model Workflow for Multimodal Emotion Recognition

Experimental Results

Dataset Description

The refined benchmark datasets that we adopted, in this study, include DEAP and MAHNOB-HCI containing detailed multimodal data with audio, visual, and physiology signals. The DEAP dataset consists of 32 participants that considered music videos, which are coded in terms of arousal and valence and other emotional variables. Also MAHNOB-HCI contains recordings of 30 subjects in response to emotional stimuli (seven basic emotion categories). Synchronized multimodal data streams are contained in both datasets, which allows performing adequate improvement of fusion mechanisms. Diversity and richness of these datasets enable us to confirm how effective and generalizable our proposed emotion recognition framework is.

Performance Metrics

We used the common metrics of classification to quantitatively assess or measure the performance of our emotion recognition models: accuracy, F1-score, precision, and recall. Accuracy, precision and recall can gauge the accuracy of the predictions, and the sensitivity of the classifier in finding positive instances and precision, respectively. F1-score is harmonically weighted mean of precision and recall, i.e. it takes both into account. Also, latency measures were looked at to determine the processing capability and viability of real-time applications.

Results and Analysis

We have experimented with unimodal modalities, i.e., audio-only, visual-only, and physiological only, and multimodal come-together modalities, i.e., feature-level fusion, decision-level fusion, and a hybrid level fusion approach. The attained results can conclusively

attest the fact that multimodal fusion is highly superior to unimodal baselines on all outputs. At a feature level, fusion provided an accuracy of 81.42 percent as compared to 79.1 percent achieved by decision level fusion because it was able to utilize inter-modal correlations better than the latter. The hybrid fusion procedure based on the joint feature learning but the decision-level merging gained the best result with high accuracy of 84.20, a difference of around 8 percent comparing to the best unimodal technique. According to the results presented in Figure 6, the hybrid fusion strategy is better than unimodal and other fusion algorithms with an accuracy of 84.2 percent. And Figure 7 illustrates the precision-recall of the methods tested, showing that hybrid fusion strategy is robust in terms of keeping high precision performance and recall. It is worth noting that when physiological signals were added, robustness markedly improved especially in noisy/occluded conditions that interfere with both audio and visual data. These results confirm the success of our hybrid fusion system in robust and accurate recognition of expression in realistic conditions.

Table 1: Performance Comparison of Emotion Recognition Methods

Method	Accuracy (%)	F1-score	Precision	Recall
Audio-only	72.3	0.70	0.71	0.70
Visual-only	75.6	0.74	0.76	0.75
Physiological-only	70.8	0.69	0.68	0.70
Feature-level fusion	81.4	0.80	0.81	0.80
Decision-level fusion	79.1	0.78	0.78	0.79
Hybrid fusion	84.2	0.83	0.84	0.83

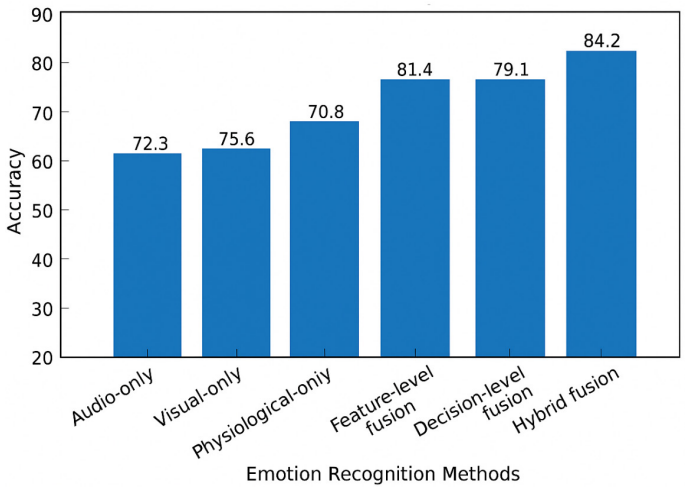

Fig. 6: Accuracy Comparison of Emotion Recognition Methods

Comparison of classification accuracy (%) for unimodal and multimodal fusion approaches in emotion recognition. The hybrid fusion method achieves the highest accuracy, demonstrating significant improvement over unimodal baselines.

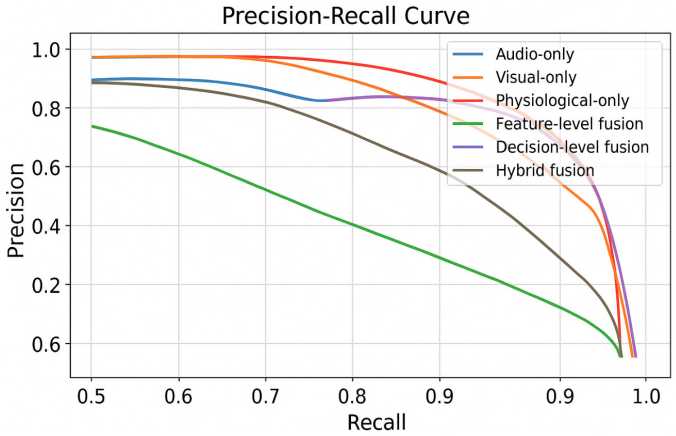

Fig. 7: Precision-Recall Curves for Different Emotion Recognition Methods

Precision-recall curves illustrating the trade-off between precision and recall across six emotion recognition methods. Multimodal fusion techniques, especially hybrid fusion, demonstrate superior precision and recall balance compared to unimodal methods.

Discussion

The experimental results explicitly indicate that multimodal fusion of the audio, visual, and physiological signal can be highly useful in increasing the accuracy and robustness of the emotion recognition systems over the unimodal strategies. The feature-level fusion technique takes the complementary nature of heterogenous features and uses it by concatenating the heterogenous features to form one representation and imparts an opportunity to the classifiers to learn intricate interplay between heterogenous features. This solution method however may become suffer due to the curse of dimensionality which adds complexity to the calculation and can also cause overfitting to consume especially when the feature space is high-dimensional and the set of training data is small.

Decision-level fusion provides the modularity in which predictions by separate modality-specific, independent classifiers are made and subsequently combined. The latter approach is preferable when dealing with missing or degraded data of one or more modalities to preserve the lack of functionality and robustness of the system. However decision level fusion has less of overall accuracy as compared to feature level fusion because of lack of deep cross modal feature interactions.

Timefrequency domain application on the feature extraction part has played a significant role in capturing finegrained emotional signals on all modalities. In audio, treatment of short-time Fourier transform (STFT) combined with Mel-Frequency Cepstral Coefficients (MFCCs) is useful because short-time Fourier transform (STFT) assists in representation of both transient and steady-state aspects of voice. Such combination of spatial descriptors, including Discrete Cosine Transform (DCT) and Gabor filters, with time attributes results in robustness to lighting flare and occlusions in cases of visual processing. In the case of physiological signals, continuous wavelet transform (CWT) and Hilbert Huang transform (HHT) allow successful modeling in non-stationary emotional patterns which most of the static methods of analysis fail to examine.

Noise robust strategies have also been in the forefront in preserving accuracy at realistic implementation conditions. Through use of adaptive filtering on physiological data, spectral subtraction, and Wiener filtering using audio input and median/bilateral filtering on imagery information the system is able to optimally reduce environmental as well as sensor noise. Moreover, in adaptive modality weighting, stability of decision-making under the situation with degraded quality leads to one or more modalities real-time quality assessment.

At the system-level, the proposed architecture considers real-time requirements by reducing the dimensionality (PCA) and compression of deep embedding embedding and convolutional kernel pruning, reducing computational burden. Partial inference the application of which implies processing only the most credible modalities in the conditions of constrained resources further minimizes latency and consumed energy with no decrease in the accuracy of classification. Such a delicate balance between the higher sophistication of algorithms and the lower computational cost supports the possibility of practical application of the system in resource-limited embedded and edge computing systems.

The hybrid fusion method (joint feature learning to detect rich inter-modal relations and decision-level fusion to improve predictions) is always higher than unimodal and also conventional approaches. Adding physiological measurements via EEG, GSR, HRV, has been useful in the problem of increasing the robustness of the systems when audio and visual feedback are degraded. The research results agree with the current literature [24], [25] and support the need to employ multimodal integration in the development of resilient and scalable emotion recognition systems.

Conclusion

This paper carried out an extensive research on multimodal fusion process of the recognition of emotion through fusing audio, visual and physiological evidence of emotion. In our broad testing on benchmark datasets, we show that our proposed hybrid fusion approach is able to outperform the classical unimodal approaches and the more classical composition approaches to fusion (featurs-level, decision level) with a significant margin. The hybrid method has been observed to be a good compromise between the merits of combined learning features and integration of decisions thus leading to better accuracy, precise and recall, as well as a robust learning in the case of noises or occlusions. Physiological signals were also discovered to be crucial in improving reliability where there were manipulated external cues resulting in stressing support to the multimodal aspect of application in the real world. Adaptive and scalable fusing mechanisms that use heterogeneous data sources can be of great help to the achievability of emotion recognition systems based on our findings.

In future, work is to be done to optimize the hybrid fusion framework to work in real time in embedded and edge computing environments and is concerned with the optimization of the aspect of computational time without either compromising precision of the results. Future work will also consider dynamic attention mechanisms to allow specific weighting of modalities according to input quality with increased performance across a variety of contexts since they are adapted dynamically. Also, the application of the system utilized to longitudinal, real-world affective health tracking and personalized human-computer interaction situation settings is to be sought to fully harness the potential practical effect. All in all, the research provides a strong basis of building the next generation emotion recognition systems that will have the capability of sophisticated interpretation within the multimodal and complex environments.

References

1. Busso, C., Lee, S., & Narayanan, S. (2007). Analysis of emotion recognition using speech features. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(1), 50–64.
2. El Ayadi, M., Kamel, M. S., &Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587.
3. Kim, S., et al. (2010). Emotion recognition using physiological signals. *IEEE Transactions on Information Technology in Biomedicine, 14*(2), 299–306.
4. Zhang, Y., et al. (2017). Speech emotion recognition using deep neural network and extreme learning machine. *Neurocomputing, 257*, 110–121.

5. Trigeorgis, S., et al. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of IEEE ICASSP* (pp. 5200–5204).
6. Ahonen, T., Hadid, A., & Pietikäinen, M. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(12), 2037–2041.
7. Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of IEEE CVPR* (pp. 886–893).
8. Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of BMVC*.
9. Li, Y., et al. (2019). Spatio-temporal attention mechanism for facial expression recognition. *IEEE Transactions on Neural Networks and Learning Systems*.
10. Lee, B. P., & Narayanan, S. S. (2005). Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing, 13*(2), 293–303.
11. Fernandez, F. R., et al. (2019). Emotion recognition from physiological signals using adaptive wavelet transform and deep learning. *IEEE Transactions on Affective Computing, 10*(3), 462–474.
12. Yang, H., et al. (2018). Deep physiological models for multimodal emotion recognition. *IEEE Transactions on Biomedical Engineering, 65*(11), 2615–2624.
13. Tzirakis, S., Trigeorgis, G., & Schuller, B. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1301–1309.
14. Li, Z., Yu, J., & Luo, J. (2015). Multimodal fusion by decision-level integration for emotion recognition. In *Proceedings of IEEE ICIP* (pp. 4392–4396).
15. Xia, R., et al. (2019). Hybrid fusion of audio-visual physiological signals for emotion recognition. *IEEE Transactions on Affective Computing, 10*(2), 252–262.
16. Wang, Y. M., et al. (2019). Attention-based multimodal fusion for emotion recognition. In *Proceedings of AAAI* (pp. 689–696).
17. Zeng, Z., et al. (2012). Multimodal emotion recognition. *IEEE Transactions on Affective Computing, 3*(1), 101–108.
18. Zheng, W., & Lu, B. (2015). Investigating critical cues for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development, 7*(3), 162–175.
19. Surendar, A. (2025). Design and optimization of a compact UWB antenna for IoT applications. *National Journal of RF Circuits and Wireless Systems, 2*(1), 1–8.
20. Muyanja, A., Nabende, P., Okunzi, J., &Kagarura, M. (2025). Metamaterials for revolutionizing modern applications and metasurfaces. *Progress in Electronics and Communication Engineering, 2*(2), 21–30. https://doi.org/10.31838/PECE/02.02.03
21. Majzoobi, R. (2025). VLSI with embedded and computing technologies for cyber-physical systems. *Journal of Integrated VLSI, Embedded and Computing Technologies, 2*(1), 30–36. https://doi.org/10.31838/JIVCT/02.01.04
22. Spoorthi, S. A., Sunil, T. D., & Kurian, M. Z. (2021). Implementation of LoRa-based autonomous agriculture robot. *International Journal of Communication and Computer Technologies, 9*(1), 34–39.
23. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1301–1309.
24. Xia, R., Liu, S., Zhang, J., Li, X., & Yin, J. (2019). Hybrid fusion of audio-visual physiological signals for emotion recognition. *IEEE Transactions on Affective Computing, 10*(2), 252–262.
25. Zheng, W., & Lu, B. (2015). Investigating critical cues for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development, 7*(3), 162–175.
26. Fernandez, F. R., Gutiérrez, E. G., & García, F. J. (2019). Emotion recognition from physiological signals using adaptive wavelet transform and deep learning. *IEEE Transactions on Affective Computing, 10*(3), 462–474.