International Journal of Applied Operational Research
Vol. 12, No. 1, pp. 29-41, Winter 2024

Journal homepage: ijorlu.liau.ac.ir Review Paper

A review of the methods of recognition multimodal emotions in sound, image and text

S. S. Hosseini, M. R. Yamaghani*, S. Poorzaker Arabani

Received: 18 September 2023 **;** **Accepted:** 25 December 2023

Abstract The study of recognizing multifaceted emotions through auditory, visual, and textual cues is a rapidly growing interdisciplinary field, encompassing the domains of psychology, computer science, and artificial intelligence. This paper investigates the spectrum of methodologies utilized to isolate and identify complex emotional states across these modalities, with the objective of delineating advancements and identifying areas for future investigation. Within the realm of sound, we explore progress in signal processing and machine learning techniques that facilitate the extraction of nuanced emotional indicators from vocal inflections and musical arrangements. Visual emotion recognition is evaluated through the effectiveness of facial recognition algorithms, analysis of body language, and integration of contextual environmental information. Text-based emotion recognition is examined using natural language processing techniques to perceive sentiment and emotional connotations from written language. Moreover, the paper considers the amalgamation of these distinct sources of emotional data, contemplating the challenges in constructing coherent models capable of interpreting multimodal inputs. Our methodology encompasses a meta-analysis of recent studies, evaluating the effectiveness and precision of diverse approaches and identifying commonly employed metrics for their assessment. The results suggest a preference towards deep learning and hybrid models that harness the strengths of multiple analytical techniques to enhance recognition rates. However, challenges such as the subjective nature of emotion, cultural disparities in expression, and the necessity for extensive, annotated datasets persist as significant hurdles. In conclusion, this review advocates for more nuanced datasets, enhanced interdisciplinary cooperation, and an ethical framework to govern the implementation of emotion recognition technologies. The potential applications for these technologies are expansive, ranging from healthcare to entertainment, and necessitate a concerted endeavor to refine and ethically integrate emotion recognition into our digital interactions.

Keyword: Multimodal Emotions, Fusion, Machine Learning, Deep Learning, Regression, CNN, RNN.

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

* **Corresponding Author. (✉)**
E-mail: O_yamaghani@liau.ac.ir (M. R. Yamaghani)

S. S. Hosseini
Department of Computer Engineering and Information Technology, Lahijan branch, Islamic Azad University, Lahijan, Iran

M. R. Yamaghani
Department of Computer Engineering and Information Technology Lahijan branch, Islamic Azad University, Lahijan, Iran

S. Poorzaker Arabani
Department of Computer Engineering and Information Technology, Lahijan branch, Islamic Azad University, Lahijan, Iran

1 Introduction

Emotion recognition spans multiple disciplines, including psychology, computer science, and artificial intelligence. Humans express emotions through various channels auditory, visual, and textual. Accurately interpreting these expressions is pivotal for advancing human-computer interaction, mental health diagnosis, and security systems. The growing field of affective computing aims to equip machines with the ability to detect and respond to human emotions, facilitating a more natural interaction between humans and technology [1]. The complexity of emotional expression underscores the significance of recognizing multifaceted emotions. Emotions are multidimensional and are often expressed simultaneously across several channels, such as vocal tone, facial expressions, and language choice. This inter-modal expression of emotions presents a challenge for computational recognition systems, requiring integrated approaches to analyze and interpret complex emotional states from diverse data sources [2]. This review has three primary objectives: firstly, to assess the current methodologies for recognizing emotions in sound, image, and text; secondly, to synthesize the findings from these approaches to identify trends, challenges, and best practices; and thirdly, to offer insights into the future direction of multimodal emotion recognition research. This review will navigate through the intricate landscape of current research, exploring the strengths and limitations of existing technologies and methodologies, and proposing the integration of these modalities as a gateway to more empathetic and emotionally aware computing [3]. The general classification of multimodal emotions is depicted in Figure 1.

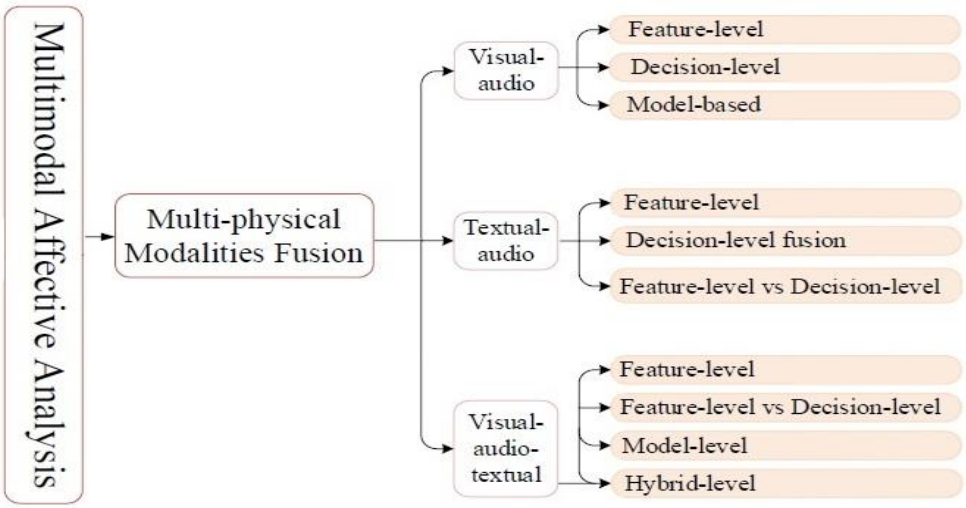

Fig. 1 This paper provides a classification of non-contact emotional calculations based on existing ethods and involving combinations of audio, text, and image.

2 Background and Significance

The origins of emotion recognition can be traced back to the pioneering work of psychologists such as Mehrabian (1972), who quantified the impact of non-verbal communication in conveying emotions. This foundational research laid the groundwork for what later became a cornerstone of affective computing, a term introduced by Picard (1995) to describe machines'

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

capability to detect and respond to human emotions [4]. Over the years, this interdisciplinary field has harnessed advancements in machine learning, signal processing, and data analytics to interpret human emotions from diverse sources, including the spoken word, facial expressions, and written text [1]. Multifaceted emotions, as described by Ekman (1993), are intricate and often involve a blend of multiple feelings that are not easily discerned. These emotions are conveyed through a spectrum of channels, such as the tone, pitch, and rhythm of a voice; the subtleties of facial movements; the posture and gestures of a body; and the choice of words and syntax in written communication [5]. Recognizing these emotions necessitates sophisticated models capable of interpreting not only overt expressions, but also nuance and context, which can vary widely across cultures and individuals. The significance of emotion recognition is underscored by its diverse applications across fields. In healthcare, emotion recognition systems have the potential to serve as tools for mental health assessment, offering supplementary data to inform diagnoses and treatment plans [6]. The automotive industry benefits from emotion-sensing technologies to enhance driver safety, utilizing physiological and behavioral cues to detect distraction or stress. Meanwhile, in education, affective computing has the potential to revolutionize learning environments by adapting in real-time to the emotional states of students to optimize engagement and retention. Furthermore, for businesses, sentiment analysis of customer feedback can yield valuable insights, informing product development and marketing strategies [7]. The field of emotion recognition is on the cusp of a new era with the emergence of deep learning and neural networks, providing models that excel in pattern recognition and learning from vast datasets. However, this journey is laden with challenges, including data privacy concerns, ethical considerations, and the potential for cultural bias in emotion recognition systems, necessitating a multidisciplinary approach that encompasses technical excellence, as well as ethical and cultural sensitivity [8]. As the technology matures, the integration of emotion recognition systems into daily life is likely to become more prevalent. This integration can lead to emotionally aware virtual assistants, responsive and adaptive media content, and even smart environments that react to the emotional states of their inhabitants. The implications for human-computer interaction, media, marketing, and even smart city initiatives are vast and largely untapped. In conclusion, the field of emotion recognition is poised to make significant contributions to our understanding and interaction with technology. By embracing the multifaceted nature of human emotions, developing robust and sensitive recognition systems, and considering the ethical implications of such technology, we stand to benefit from machines that understand us better and, in turn, serve us better [9].

3 Methodology of the Review

The methodology of this extensive review has been intricately fashioned to guarantee the inclusion of a wide range of research, covering diverse approaches to recognizing complex emotions in sound, image, and text. The selection of studies for this review has been guided by specific criteria that prioritize relevance, recency, and research quality.

3.1 Criteria for Study Selection

Relevance: Studies were selected based on their direct focus on emotion recognition methodologies within the realms of auditory, visual, and textual data. Priority was given to works that presented novel approaches or significant improvements on existing methods.

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

Recency: Emphasis was placed on studies published within the last five years to capture the most current state of the field. However, seminal works that provide essential background or have historically contributed significant insights were also included irrespective of their publication date. Research Quality: Peer-reviewed articles from reputable journals and conferences were prioritized. Each study was evaluated for methodological rigor, including the validity of the experimental design, the reliability of results, and the robustness of the conclusions drawn. Impact: Studies that have been widely cited and have influenced subsequent research were included to trace the development of the field and understand the evolution of emotion recognition methodologies.

3.2 Process of Data Analysis

The analysis process began with a systematic literature search using academic databases such as IEEE Xplore, PubMed, Scopus, and Google Scholar. Various combinations of keywords related to emotion recognition, including "affective computing," "emotion detection," "sentiment analysis," and "multimodal emotion recognition," were utilized to ensure the comprehensive capture of relevant studies. Following the initial collection of articles, an abstract review was conducted to assess their alignment with the selection criteria. Subsequently, full-text articles underwent thorough scrutiny by multiple reviewers to minimize bias and ensure a comprehensive evaluation based on the aforementioned criteria. Data extraction from each study focused on the methodologies employed for emotion recognition, the types of data analyzed (sound, image, text), the algorithms and models utilized, and the effectiveness of each approach as indicated by reported metrics such as accuracy, precision, recall, and F1 score. The extracted data were then synthesized to identify patterns, trends, and gaps in the research. When applicable, meta-analytic techniques were employed to quantitatively compare the effectiveness of different emotion recognition methods. Qualitative analysis was also conducted to comprehend the context of the research findings and the implications for future studies. This dual approach ensures that the review not only provides a statistical comparison of methods, but also offers a narrative that contextualizes the state of emotion recognition research. By adhering to a rigorous methodological framework, this review aims to provide a clear and comprehensive understanding of how multifaceted emotions are currently recognized across sound, image, and text, and to pave the way for future advancements in the field.

4 Emotion Recognition in Sound

Emotion recognition in sound, also known as acoustic emotion recognition (AER), is a complex field that intersects signal processing, machine learning, and psychology. Its primary objective is to interpret emotional states from vocal attributes. The intricacy of this task is heightened by the variability of speech and the subtlety of emotional expressions. The human voice carries a wealth of emotional information. During speech production, our vocal cords generate sounds that are shaped by the articulators into speech. This process is influenced by our emotional state, impacting various acoustic features such as pitch, volume, and rhythm. Emotion recognition in sound entails analyzing these features to recognize patterns that correspond to specific emotions [10]. Pitch, or fundamental frequency, stands out as one of the most studied acoustic features in AER. It reflects the rate of vocal cord vibration and is often associated with emotional intensity. Higher pitches are typically linked to heightened

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

emotions such as joy or anger, while lower pitches may indicate sadness or calmness. Volume, or intensity, is another pivotal feature, with louder speech commonly conveying stronger emotions. Speech rate can also serve as an indicator of emotional states; rapid speech may indicate excitement or anxiety, while slower rates can signal depression or calmness [11]. In addition to these fundamental features, researchers have explored more nuanced characteristics such as formant frequencies, which are influenced by the shape of the vocal tract and can impact the perceived quality of the voice. Another advanced feature is spectral entropy, which measures the randomness in the voice signal and can be associated with the clarity or confusion of emotional expression [12]. Modern AER systems utilize various models to interpret these acoustic features. Early efforts relied on simple machine learning algorithms such as decision trees and support vector machines. However, as the field has advanced, more sophisticated models like deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) have become prevalent. These models are capable of handling high-dimensional data and capturing the temporal dynamics of speech, which are crucial for understanding emotions [13]. Recent advancements have led to the emergence of end-to-end learning systems capable of autonomously learning relevant features from raw audio signals without manual feature engineering. This represents a significant advancement, as it enables models to identify and leverage complex patterns that human engineers might not discern. The performance of these methods is typically assessed using metrics such as accuracy, precision, recall, and the F1 score. Current research indicates varying levels of success, with some systems achieving high levels of accuracy under controlled conditions. However, real-world performance can be significantly lower due to factors such as background noise, the speaker's idiosyncrasies, and the context of the speech [14]. One of the most significant challenges in AER is the variability of emotional expression across different individuals and cultures. An emotion expressed in one culture may have different acoustic signatures than in another. Additionally, intra-speaker variability, which refers to differences in how the same person expresses the same emotion at different times, presents additional challenges. In addressing these issues, researchers have explored the use of personalized models that adapt to the specific characteristics of an individual's voice. There is also a growing interest in cross-cultural studies aimed at developing more universally applicable models. Transfer learning, in which a model trained on one dataset is adapted to work with another, is one approach to enhancing the robustness of models across different populations [15]. Another promising development is the integration of context-aware models that can combine acoustic signals with linguistic and situational context to improve emotion recognition. For example, integrating speech recognition with AER enables the model to consider not only how something is said, but also what is being said, providing a more holistic understanding of the speaker's emotional state. In conclusion, emotion recognition in sound is a dynamic field with applications ranging from mental health assessment to human-computer interaction. Despite the significant progress made, there is still much work to be done. Future research directions may include enhancing the robustness of AER systems in diverse and noisy environments, improving cross-cultural applicability, and integrating multimodal data to achieve a more comprehensive understanding of human emotions [16].

5 Emotion Recognition in Image

Visual emotion recognition is an intriguing aspect of affective computing, focusing on discerning emotional states from static images or video sequences. This field has experienced

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

substantial progress, driven by the advancement of machine learning and computer vision technologies. Facial expression analysis serves as a fundamental component of visual emotion recognition. The seminal work of Ekman and Friesen (1978) established the Facial Action Coding System (FACS), a comprehensive framework for classifying facial movements by their appearance on the face. Emotions are deduced from the activation of specific facial muscles, known as Action Units (AUs). Contemporary approaches have expanded upon this groundwork, utilizing sophisticated algorithms to automatically detect and interpret AUs from images. Deep learning models, particularly Convolutional Neural Networks (CNNs), have proven effective in learning the intricate patterns associated with facial expressions from extensive datasets [17]. In addition to facial expressions, body language provides crucial emotional cues. Posture, the positioning of limbs, and the orientation of the body can all convey emotional information. Early endeavors in this domain relied on manually crafted features and geometric models to represent body language. However, the emergence of pose estimation algorithms has transformed this domain. Leveraging datasets such as MPII Human Pose and COCO, pose estimation models can now identify and track key body joints in real-time, facilitating dynamic analysis of body language in various emotional contexts [18]. The incorporation of other visual cues, such as eye gaze, interpersonal distance, and contextual elements within an image, further enhances the interpretation of emotions. For example, eye tracking can unveil attentional focus and engagement levels, while proxemics - the study of personal space - can indicate comfort or anxiety levels in social settings [19]. Recent advancements in visual emotion recognition have concentrated on multimodal approaches that integrate facial expressions, body language, and contextual cues. These methods recognize the complementary nature of these signals and strive to establish a more comprehensive model of emotion detection. For instance, multimodal emotion recognition frameworks have been employed in the analysis of social media images, where the context provided by the environment and the interaction between individuals in an image can significantly influence the perceived emotion. The assessment of performance in visual emotion recognition systems is multifaceted, encompassing accuracy, speed, and robustness across diverse scenarios and populations. Benchmarks such as the Facial Expression Recognition (FER) Challenge and the Emotion Recognition in the Wild (EmotiW) have provided platforms for comparing the effectiveness of various methods. While laboratory-controlled datasets have exhibited high recognition accuracy, real-world data present challenges due to variability in lighting, occlusions, and the inherent subtlety of natural emotional expressions [20]. Despite remarkable technological advancements, several challenges persist. Variability in emotional expression across cultures and individual idiosyncrasies complicates the task of creating universally applicable models. Additionally, there are ethical considerations pertaining to privacy and consent, as emotion recognition systems often operate in personal and sensitive domains. Looking ahead, the field of visual emotion recognition is anticipated to continue evolving, with increased focus on ethical AI practices, the enhancement of multimodal models, and the exploration of unsupervised learning techniques to mitigate the need for extensive labeled datasets. As these technologies mature, they hold the potential to enhance applications in mental health, entertainment, user experience design, and beyond [21].

6 Emotion Recognition in Text

Text-based emotion recognition is a rapidly expanding field within natural language processing (NLP) that focuses on identifying and categorizing emotional states conveyed

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

through written language. It transcends traditional sentiment analysis, which categorizes text as positive, negative, or neutral, to encompass a spectrum of emotions such as joy, anger, sadness, and surprise [22]. The roots of sentiment analysis can be traced back to the early work of Pang and Lee (2004), who applied machine learning techniques to analyze movie reviews. Since then, the field has progressed to include more detailed emotion analysis, necessitating the development of more sophisticated NLP models. These models have evolved from bag-of-words approaches to more nuanced techniques that consider the syntactic and semantic context of language. Context-aware emotion detection represents a further advancement, acknowledging that the meaning of words and the emotions they convey can change dramatically with context. For example, newer models strive to accommodate challenges such as sarcasm and irony, which may reverse the apparent sentiment of a phrase [23]. The latest methods in text-based emotion recognition leverage deep learning models, such as Long Short-Term Memory (LSTM) networks and Transformers, to capture the complexities of language. LSTMs excel at understanding the sequential nature of text, which is crucial for comprehending the emotional progression within a narrative [24]. Transformers, introduced with the influential architecture known as the Attention Is All You Need, offer an even deeper understanding of context through self-attention mechanisms. Models like BERT (Bidirectional Encoder Representations from Transformers) and its variants such as RoBERTa and GPT (Generative Pre-trained Transformer) have set new benchmarks in various NLP tasks, including emotion recognition. Performance evaluation of text-based emotion recognition systems often involves metrics such as accuracy, recall, precision, and F1 score. Additionally, the area under the receiver operating characteristic curve (AUROC) and confusion matrices are utilized to gain insights into the models' performance across different emotional categories. Despite significant progress, challenges persist. Language is inherently ambiguous and culturally dependent. The same word or phrase may carry different emotional connotations across cultures, and even within the same culture, individuals may interpret language differently based on their experiences and contexts [25]. Furthermore, there is increasing concern regarding bias in the training datasets. If a model is primarily trained on text data from one demographic, it may struggle to perform well on data from another. Efforts to mitigate these issues involve curating more diverse and balanced datasets and developing models that can adapt to the idiosyncrasies of individual language use. Looking ahead, the integration of multimodal data sources, including audio and visual cues alongside text, holds promise for a more comprehensive approach to emotion recognition. Additionally, the development of unsupervised and semi-supervised learning models could alleviate the burden of manual data annotation and allow systems to adapt more fluidly to the nuances of human emotion expression in text [26].

7 Integrated Approaches

Multimodal emotion recognition represents an emerging frontier in affective computing, striving to develop a comprehensive understanding of human emotions by integrating data from sound, image, and text. This holistic approach recognizes that individuals' true sentiments and emotional states are often a combination of various signals and cues that can be captured across different sensory modalities.

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

7.1 Integration of Sound, Image, and Text

The integration of auditory, visual, and textual data into a unified model presents both opportunities and challenges. Multimodal Machine Learning frameworks demonstrate how data from each modality can complement the others, leading to more robust emotion recognition systems. For example, audio-visual speech recognition systems enhance the understanding of spoken language by incorporating lip movement analysis with auditory information. Similarly, text can provide context to vocal intonations, and facial expressions can add depth to spoken words [27].

7.2 Challenges in Multimodal Emotion Analysis

One of the primary challenges in multimodal emotion recognition is data fusion. Several strategies for integrating data exist, including early fusion, where raw data from all modalities are combined at the input stage; late fusion, where the final decision is made by combining the outputs of separate models for each modality; and hybrid approaches that combine both strategies. Temporal alignment poses another significant challenge. Emotional expressions in speech, text, and facial expressions occur over different time scales, and aligning these temporally can be challenging. Techniques such as Dynamic Time Warping (DTW) and sequence-to-sequence models have been developed to address this issue. Additionally, the availability of datasets containing synchronized multimodal emotional data is limited. The creation and annotation of such datasets are resource-intensive but crucial for the advancement of the field [28].

7.3 Solutions and Advances in Multimodal Analysis

To tackle these challenges, researchers have proposed several solutions. Advanced machine learning techniques such as deep learning have been particularly effective in learning joint representations of multimodal data [29]. A summary of the selected techniques for emotional analysis is presented in Table 1. Another solution involves the use of attention mechanisms, which enable models to prioritize the most relevant features from each modality when making predictions. This is especially valuable in scenarios where one modality may be more informative than the others, depending on the context.

7.4 Performance of Multimodal Emotion Recognition Systems

The performance of multimodal emotion recognition systems is typically assessed using a combination of accuracy, precision, recall, and F1 score. However, given the complexity of these systems, researchers also employ more sophisticated measures such as the concordance correlation coefficient to evaluate the agreement between the predicted and actual emotion labels. Despite the advancements, the performance of these systems in real-world scenarios is still a topic of active research. The inter-subject and intra-subject variability of emotional expressions, the subtlety of certain emotions, and the influence of cultural factors remain significant obstacles [29]. In conclusion, although multimodal emotion recognition presents a promising path toward establishing more empathetic and context-aware computing systems, substantial challenges persist. Future research endeavors may involve the advancement of

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

more sophisticated integration techniques, the establishment of larger and more diverse multimodal datasets, and the exploration of unsupervised learning methods to lessen reliance on annotated data.

Table 1 Summary of chosen techniques for conducting emotional analysis

Publication	Feature Representation	Classifier	Fusion Strategy	Database	(%) Performance
		audio Emotion Recognition- Visual			
[30]	CNN, Resnet	LSTM	level-Feature	RECOLA	A/V: 78.80/73.20
[31]	C3D + DBN	level Fusion-Score	based-Model	eNTERFACE	classes: 89.39 6
[32]	-Acoustic, Geometric and HOG TOP	Multiple kernel SVM	level-Feature	CK; AFEW	classes: 95.7 7 classes: 45.20 7
[33]	D CNN, 3D CNN2	based fusion-ELM SVM	based-Model	Big Data eNTERFACE	classes: 91.30 3 classes: 78.42 6
[34]	Multitask CNN	Classifier-Meta	level-Decision	eNTERFACCE	classes: 81.36 6
[35]	D ResNet+Attention2 D esNet+Attention3	FC	level-Feature	-VideoEmotion 6-Ekman 8	classes: 54.50 8 classes: 55.30 6
		audio Emotion Recognition- Text			
[36]	prosodic Semantic - Acoustic labels	,Base classifiers MDT, MaxEnt	level-Decision	utterances 2033	classes: 83.55 4 classes: 85.79 4
[37]	attention-DNN Self-DCNN, T-A	FC	level-Feature	IEMOCAP	classes: 80.51 4 classes: 79.22 4
[38]	Acoustic features Word embeddings	Pooling Scalar weight fusion	level -Decision level-Feature	-IEMOCAP; MSP PODCAST	58.20/65.10 58.00/63.90
		text Emotion Recognition Emotion Recognition- audio- Visual			
[39]	Facial movement, MFCC	SVM, BiLSMT	level-Hybrid	SEMAINE	class:80.50,65.202
[40]	CNN, handcrafted, CFS, PCA	MKL	level-Feature	HOW	classes: 88.60 3
[41]	GRU-Three Bi	attention-CIM	level -Decision based-Model	MOSEI-CMU	classes: 86.27 3 label: 62.80-Multi
[42]	Proxy and Attention Multiplicative fusion	FC	level-Feature	-IEMOCAP CMU MOSEI	classes:82.70 4 classes:89.00 6
[43]	CNN,RNN,BI-LSTM	Lstm ,Softmax Regration	level-Feature level -Decision	IEMOCAP	lasses:82.90

8 Discussion

The field of emotion recognition has undergone substantial transformation in recent decades, as evidenced by the extensive body of literature. The evolution from rule-based systems to sophisticated machine learning algorithms signifies a notable shift toward more accurate and nuanced emotion detection. Examining the findings from sound, image, and text modalities, it becomes clear that while each modality offers unique insights, their integration holds the promise of a more comprehensive understanding of human emotions. In sound, the transition from simple spectral features to complex models like Deep Neural Networks (DNNs) reflects an emphasis on capturing the subtleties of vocal expressions. Similarly, image-based recognition has benefited from advancements in Convolutional Neural Networks (CNNs) for facial and body language analysis, moving away from geometric feature extraction. Text-based recognition has experienced a significant leap forward with the advent of contextual models like BERT and GPT, which offer a deeper understanding of semantic nuances. Comparing these methods reveals a trade-off between the granularity of emotion recognition and the computational complexity of the models. While deep learning approaches provide state-of-the-art performance, they require significant computational resources and large annotated datasets, which may not be feasible in all application contexts. Furthermore, the

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

performance of these methods varies significantly across domains and is highly dependent on the quality and diversity of the training data. Current methodologies are not without limitations and challenges. One of the most pressing issues is the subjectivity of emotions themselves, which can be interpreted differently across cultures and individuals. This subjectivity presents a challenge for the creation of universal models that perform consistently across diverse user groups. Moreover, the reliance on labeled data for supervised learning poses a significant bottleneck. The labor-intensive process of annotating data with emotional labels is prone to human error and bias, which can propagate through the models. Additionally, privacy concerns arise when dealing with sensitive personal data required for training these systems. The challenges extend to the integration of multimodal data, where issues such as data alignment, fusion, and the temporal synchronicity of emotional expressions across modalities come to the fore. The current state-of-the-art multimodal systems, while promising, still struggle with achieving real-time performance and maintaining high accuracy in uncontrolled environments. In light of these limitations, the field is moving towards more innovative solutions such as unsupervised and semi-supervised learning, which can reduce the dependency on labeled datasets. Transfer learning and domain adaptation techniques are also gaining traction as means to enhance the generalizability and robustness of emotion recognition models. The discussion of these findings underscores the dynamic nature of emotion recognition research, highlighting the need for continued innovation in computational models, data collection practices, and the ethical use of emotion recognition technologies. As the field progresses, it will be crucial to address these challenges to realize the full potential of affective computing in real-world applications.

9 Conclusions

This comprehensive review systematically examined a variety of methodologies utilized in the recognition of complex emotions across sound, image, and text modalities. The evolution from rudimentary emotion recognition systems to advanced deep learning frameworks signifies the remarkable progress within the field. In sound, the refinement of acoustic feature analysis and the integration of sophisticated models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have significantly improved the subtlety and accuracy of emotion detection. Visual emotion recognition has likewise benefited from advancements in facial and body language analysis, capitalizing on high-dimensional data and temporal dynamics. Moreover, text-based emotion recognition has advanced considerably with the implementation of context-aware models like BERT and GPT, which take into account the semantic and syntactic nuances of language. The key insights gleaned from the literature underscore the complexity of human emotions and the challenges associated with automating their recognition. Despite substantial progress, current methodologies encounter limitations related to data dependency, cultural and contextual variability, and the inherently subjective nature of emotions. Multimodal emotion recognition, which integrates cues from sound, image, and text, emerges as a promising approach to address these challenges, seeking a holistic understanding of emotional states. However, persistent issues such as data fusion, temporal alignment, and model interpretability need to be addressed. The significance of advancing emotion recognition methods cannot be overstated. The applications are extensive, encompassing enhancements to user experience in technology interfaces, as well as offering critical insights in mental health assessment. Emotion recognition systems hold the potential to revolutionize various sectors, including healthcare, automotive safety, customer service,

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

and education. Moving forward, the field must navigate the ethical landscape of emotion recognition, ensuring privacy, consent, and the unbiased use of technology. The development of more sophisticated, culturally aware, and ethical emotion recognition systems will necessitate concerted efforts from interdisciplinary teams, combining expertise from computer science, psychology, linguistics, and ethics. Anticipating the future of emotion recognition, attention turns towards unsupervised learning techniques, the exploration of transfer learning for improved model generalization, and the creation of more robust, real-world datasets. These advancements will pave the way for more accurate, reliable, and empathetic affective computing systems.

References

1. Spezialetti M, Placidi G, Rossi S. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. Front Robot AI. 2020 Dec 21;7:532279. doi: 10.3389/frobt.2020.532279. PMID: 33501307; PMCID: PMC7806093.
2. Trinh Van L, Dao Thi Le T, Le Xuan T, Castelli E. Emotional Speech Recognition Using Deep Neural Networks. Sensors (Basel). 2022 Feb 12;22(4):1414. doi: 10.3390/s22041414. PMID: 35214316; PMCID: PMC8877219.
3. Hasnul MA, Aziz NAA, Alelyani S, Mohana M, Aziz AA. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare-A Review. Sensors (Basel). 2021 Jul 23;21(15):5015. doi: 10.3390/s21155015. PMID: 34372252; PMCID: PMC8348698.
4. Elhamdadi H, Canavan S, Rosen P. AffectiveTDA: Using Topological Data Analysis to Improve Analysis and Explainability in Affective Computing. IEEE Trans Vis Comput Graph. 2022 Jan;28(1):769-779. doi: 10.1109/TVCG.2021.3114784. Epub 2021 Dec 29. PMID: 34587031.
5. Hwang S, Hwang J, Jeong H. Study on Associating Emotions in Verbal Reactions to Facial Expressions in Dementia. Healthcare (Basel). 2022 Jun 1;10(6):1022. doi: 10.3390/healthcare10061022. PMID: 35742073; PMCID: PMC9222752.
6. Smith E, Storch EA, Vahia I, Wong STC, Lavretsky H, Cummings JL, Eyre HA. Affective Computing for Late-Life Mood and Cognitive Disorders. Front Psychiatry. 2021 Dec 23;12:782183. doi: 10.3389/fpsyt.2021.782183. PMID: 35002802; PMCID: PMC8732874.
7. Fusar-Poli P, Manchia M, Koutsouleris N, Leslie D, Woopen C, Calkins ME, Dunn M, Tourneau CL, Mannikko M, Mollema T, Oliver D, Rietschel M, Reininghaus EZ, Squassina A, Valmaggia L, Kessing LV, Vieta E, Correll CU, Arango C, Andreassen OA; PSMD EBRA cluster (annex 1). Ethical considerations for precision psychiatry: A roadmap for research and clinical practice. Eur Neuropsychopharmacol. 2022 Oct;63:17-34. doi: 10.1016/j.euroneuro.2022.08.001. Epub 2022 Aug 27. PMID: 36041245.
8. Park CY, Cha N, Kang S, Kim A, Khandoker AH, Hadjileontiadis L, Oh A, Jeong Y, Lee U. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci Data. 2020 Sep 8;7(1):293. doi: 10.1038/s41597-020-00630-y. PMID: 32901038; PMCID: PMC7479607.
9. Chen M, Liang X, Xu Y. Construction and Analysis of Emotion Recognition and Psychotherapy System of College Students under Convolutional Neural Network and Interactive Technology. Comput Intell Neurosci. 2022 Sep 17;2022:5993839. doi: 10.1155/2022/5993839. PMID: 36164423; PMCID: PMC9509236.
10. Kamiloğlu RG, Boateng G, Balabanova A, Cao C, Sauter DA. Superior Communication of Positive Emotions Through Nonverbal Vocalisations Compared to Speech Prosody. J Nonverbal Behav. 2021;45(4):419-454. doi: 10.1007/s10919-021-00375-1. Epub 2021 Jul 24. PMID: 34744232; PMCID: PMC8553689.
11. Selosse G, Grandjean D, Ceravolo L. Influence of bodily resonances on emotional prosody perception. Front Psychol. 2022 Dec 8;13:1061930. doi: 10.3389/fpsyg.2022.1061930. Erratum in: Front Psychol. 2023 Mar 06;14:1170276. PMID: 36571062; PMCID: PMC9773097.
12. Graf S, Schwiebacher J, Richter L, Buchberger M, Adachi S, Mastnak W, Hoyer P. Adjustment of Vocal Tract Shape via Biofeedback: Influence on Vowels. J Voice. 2020 May;34(3):335-345. doi: 10.1016/j.jvoice.2018.10.007. Epub 2018 Nov 15. PMID: 30448316.
13. Wang D, Lin M, Zhang X, Huang Y, Zhu Y. Automatic Modulation Classification Based on CNN-Transformer Graph Neural Network. Sensors (Basel). 2023 Aug 20;23(16):7281. doi: 10.3390/s23167281. PMID: 37631817; PMCID: PMC10459892.

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

14. Kodish-Wachs J, Agassi E, Kenny P 3rd, Overhage JM. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. AMIA Annu Symp Proc. 2018 Dec 5;2018:683-689. PMID: 30815110; PMCID: PMC6371385.
15. Nordström H, Laukka P, Thingujam NS, Schubert E, Elfenbein HA. Emotion appraisal dimensions inferred from vocal expressions are consistent across cultures: a comparison between Australia and India. R Soc Open Sci. 2017 Nov 15;4(11):170912. doi: 10.1098/rsos.170912. PMID: 29291085; PMCID: PMC5717659.
16. Atmaja BT, Sasou A. Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations. Sensors (Basel). 2022 Aug 24;22(17):6369. doi: 10.3390/s22176369. PMID: 36080828; PMCID: PMC9460459.
17. Dong Z, Wang G, Lu S, Li J, Yan W, Wang SJ. Spontaneous Facial Expressions and Micro-expressions Coding: From Brain to Face. Front Psychol. 2022 Jan 4;12:784834. doi: 10.3389/fpsyg.2021.784834. PMID: 35058850; PMCID: PMC8763852.
18. Glonek G, Wojciechowski A. Hybrid Orientation Based Human Limbs Motion Tracking Method. Sensors (Basel). 2017 Dec 9;17(12):2857. doi: 10.3390/s17122857. PMID: 29232832; PMCID: PMC5751617.
19. Suslow T, Hoepfel D, Günther V, Kersting A, Bodenschatz CM. Positive attentional bias mediates the relationship between trait emotional intelligence and trait affect. Sci Rep. 2022 Dec 1;12(1):20733. doi: 10.1038/s41598-022-25317-9. PMID: 36456618; PMCID: PMC9715682.
20. Dinkler L, Rydberg Dobrescu S, Råstam M, Gillberg IC, Gillberg C, Wentz E, Hadjikhani N. Visual scanning during emotion recognition in long-term recovered anorexia nervosa: An eye-tracking study. Int J Eat Disord. 2019 Jun;52(6):691-700. doi: 10.1002/eat.23066. Epub 2019 Mar 4. PMID: 30828832.
21. Wang X, Kou L, Sugumaran V, Luo X, Zhang H. Emotion Correlation Mining Through Deep Learning Models on Natural Language Text. IEEE Trans Cybern. 2021 Sep;51(9):4400-4413. doi: 10.1109/TCYB.2020.2987064. Epub 2021 Sep 15. PMID: 32413938.
22. Wang X, Kou L, Sugumaran V, Luo X, Zhang H. Emotion Correlation Mining Through Deep Learning Models on Natural Language Text. IEEE Trans Cybern. 2021 Sep;51(9):4400-4413. doi: 10.1109/TCYB.2020.2987064. Epub 2021 Sep 15. PMID: 32413938.
23. Kazeminejad G, Palmer M, Brown SW, Pustejovsky J. Componential Analysis of English Verbs. Front Artif Intell. 2022 May 30;5:780385. doi: 10.3389/frai.2022.780385. PMID: 35707764; PMCID: PMC9189303.
24. Yu Y, Si X, Hu C, Zhang J. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. Neural Comput. 2019 Jul;31(7):1235-1270. doi: 10.1162/neco_a_01199. Epub 2019 May 21. PMID: 31113301.
25. Souter NE, Lindquist KA, Jefferies E. Impaired emotion perception and categorization in semantic aphasia. Neuropsychologia. 2021 Nov 12;162:108052. doi: 10.1016/j.neuropsychologia.2021.108052. Epub 2021 Oct 12. PMID: 34624259.
26. Nguyen TL, Kavuri S, Lee M. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. Neural Netw. 2019 Oct;118:208-219. doi: 10.1016/j.neunet.2019.06.010. Epub 2019 Jul 2. PMID: 31299625.
27. Treille A, Vilain C, Hueber T, Lamalle L, Sato M. Inside Speech: Multisensory and Modality-specific Processing of Tongue and Lip Speech Actions. J Cogn Neurosci. 2017 Mar;29(3):448-466. doi: 10.1162/jocn_a_01057. Epub 2016 Oct 19. PMID: 28139959.
28. Mamieva D, Abdusalomov AB, Kutlimuratov A, Muminov B, Whangbo TK. Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. Sensors (Basel). 2023 Jun 9;23(12):5475. doi: 10.3390/s23125475. PMID: 37420642; PMCID: PMC10304130.
29. Lian H, Lu C, Li S, Zhao Y, Tang C, Zong Y. A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. Entropy (Basel). 2023 Oct 12;25(10):1440. doi: 10.3390/e25101440. PMID: 37895561; PMCID: PMC10606253.
30. P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Signal Process. 11 (2017) 1301–1309, https://doi.org/10.1109/ JSTSP.2017.2764438.
31. D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, C. Fookes, Deep spatio- temporal features for multimodal emotion recognition, in: 2017 IEEE Winter Conf. Appl. Comput. Vis. WACV, 2017, pp. 1215–1223, https://doi.org/10.1109/ WACV.2017.140.
32. J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, IEEE Trans. Affect. Comput. 9 (2018) 38–50, https://doi. org/10.1109/TAFFC.2016.2593719.
33. M.S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio–visual emotional big data, Inf. Fusion. 49 (2019) 69–78, https://doi. org/10.1016/j.inffus.2018.09.008.
34. M.S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio–visual emotional big data, Inf. Fusion. 49 (2019) 69–78, https://doi. org/10.1016/j.inffus.2018.09.008.

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]

35. M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, P. Xiao, Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features, Neurocomputing 391 (2020) 42–51, https://doi.org/10.1016/j. neucom.2020.01.048.
36. S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, K. Keutzer, An end- to-end visual-audio attention network for emotion recognition in user-generated videos, Proc. AAAI Conf. Artif. Intell. 34 (2020) 303–311, https://doi.org/ 10.1609/aaai.v34i01.5364.
37. C.-H. Wu, W.-B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, IEEE Trans. Affect. Comput. 2 (2011) 10–21, https://doi.org/10.1109/T-AFFC.2010.16.
38. D. Priyasad, T. Fernando, S. Denman, S. Sridharan, C. Fookes, Attention driven fusion for multi-modal emotion recognition, in: ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, 2020, pp. 3227–3231, https://doi. org/10.1109/ICASSP40776.2020.9054441.
39. L. Pepino, P. Riera, L. Ferrer, A. Gravano, Fusion approaches for emotion recognition from speech using acoustic and text-based features, in: ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP, IEEE, Barcelona, Spain, 2020, pp. 6484–6488, https://doi.org/10.1109/ ICASSP40776.2020.9054709.
40. M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, YouTube movie reviews: sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (2013) 46–53, https://doi.org/10.1109/MIS.2013.34.
41. S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: Proc. 2015 Conf. Empir. Methods Nat. Lang. Process., Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 2539–2544, https://doi. org/10.18653/v1/D15-1303.
42. M.S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, P. Bhattacharyya, Multi- task learning for multi-modal emotion recognition and sentiment analysis, in: Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Vol. 1 Long Short Pap., Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 370–379, https://doi.org/10.18653/v1/N19- 1034.
43. Hosseini, S.S., Yamaghani, M.R. & Poorzaker Arabani, S. Multimodal modelling of human emotion using sound, image and text fusion. SIViP (2023). https://doi.org/10.1007/s11760-023-02707-8

[Downloaded from mail.ijaor.ir on 2026-11-27]

[DOI: 10.71885/ijorlu-2024-1-657]