Cognitive Computation
https://doi.org/10.1007/s12559-024-10287-z

A Review of Key Technologies for Emotion Analysis Using Multimodal Information

Xianxun Zhu[1] · Chaopeng Guo[1] · Heyang Feng[1] · Yao Huang[1] · Yichen Feng[1] · Xiangyang Wang[1] · Rui Wang[1]

Received: 18 September 2023 / Accepted: 10 April 2024

Abstract

Emotion analysis, an integral aspect of human–machine interactions, has witnessed significant advancements in recent years. With the rise of multimodal data sources such as speech, text, and images, there is a profound need for a comprehensive review of pivotal elements within this domain. Our paper delves deep into the realm of emotion analysis, examining multimodal data sources encompassing speech, text, images, and physiological signals. We provide a curated overview of relevant literature, academic forums, and competitions. Emphasis is laid on dissecting unimodal processing methods, including preprocessing, feature extraction, and tools across speech, text, images, and physiological signals. We further discuss the nuances of multimodal data fusion techniques, spotlighting early, late, model, and hybrid fusion strategies. Key findings indicate the essentiality of analyzing emotions across multiple modalities. Detailed discussions on emotion elicitation, expression, and representation models are presented. Moreover, we uncover challenges such as dataset creation, modality synchronization, model efficiency, limited data scenarios, cross-domain applicability, and the handling of missing modalities. Practical solutions and suggestions are provided to address these challenges. The realm of multimodal emotion analysis is vast, with numerous applications ranging from driver sentiment detection to medical evaluations. Our comprehensive review serves as a valuable resource for both scholars and industry professionals. It not only sheds light on the current state of research but also highlights potential directions for future innovations. The insights garnered from this paper are expected to pave the way for subsequent advancements in deep multimodal emotion analysis tailored for real-world deployments.

Keywords Multimodal information · Emotional analysis · Multimodal fusion

Introduction

Background and Significance

Emotion is a complex interplay of physiological and psychological states, incorporating an array of thoughts, feelings, and behavioral manifestations [1]. Simply put, emotion is the response individuals have to external stimuli within a specific context. This response incorporates both physiological indicators such as muscle tension, elevated heart rate, and perspiration, as well as psychological elements like subjective experiences and emotional reactions [2]. Expressions of emotion can often be discerned through observable behaviors, known as expressions. Besides facial cues, emotional states can be communicated via language, body posture, vocal tonal changes, and other expressive mechanisms [3]. Emotion is not the result of a singular event; rather, it is a multifaceted process tightly interwoven with an individual's cognition, values, and personality. Essentially, it signifies a sequence of subjective cognitive experiences, attitudes toward objective phenomena, and corresponding behavioral reactions. Emotions are generally viewed as psychological activities influenced by an individual's desires and needs [4]. In the age of artificial intelligence, the expression of emotion has transcended human interactions to also include human–machine interactions. Transmitting voice, text, and images online to enable machines to comprehend human emotions more rapidly and accurately has become a focal point of research [5].

Emotion recognition, often termed sentiment analysis, involves discerning the emotional states expressed by others through observing various elements like facial expressions,

✉ Rui Wang
rwang@shu.edu.cn

[1] School of Communication and Information Engineering, Shanghai University, No. 99, Shangda Road, Shanghai 200444, China

Published online: 01 June 2024

language, and behaviors [6]. Accurate emotion recognition typically necessitates the assimilation of information from multiple modalities. These modalities can include facial expressions, body language, and vocal characteristics to analyze, process, and infer underlying emotions, attitudes, or perspectives embedded within emotionally charged data. Early research in this domain primarily relied on single-modal data, such as text, speech, and images, as the principal sources for analysis [7–9]. Yet, the expression of complex emotions via a single modality has its limitations. Text-based emotion analysis, reliant solely on the lexical and syntactical structure, can often result in ambiguity. Speech-based analysis poses its own set of challenges, from loss of human-like qualities to possible errors caused by health conditions or environmental noise. Likewise, image-based recognition methods are influenced by factors like lighting conditions, occlusion, and the tenuous link between facial expressions and actual emotional states. For instance, Fig. 1 illustrates an example of a "sad" emotional expression represented across three modalities: a positive facial expression, neutral text, and a sad tone, highlighting the limitations of relying on a single modality for accurate emotion recognition. Furthermore, humans naturally convey their emotional states using a blend of text, voice, and facial cues. For example, when individuals are happy, their mouth corners lift, their voice takes on a buoyant tone, and their choice of words carries a distinct emotional hue.

Integrating multiple types of data addresses the limitations of single-modality emotion recognition by capturing emotional cues from different angles. For example, the phrase "I don't know" can be interpreted differently depending on the vocal tone accompanying it. Multimodal emotion recognition offers a consolidated approach for identifying emotional states, pulling from diverse data streams like facial expressions, vocalizations, body language, and physiological markers. This approach is more accurate, robust, and flexible compared to single-modality methods, making it applicable across fields like human–machine interaction and virtual reality.

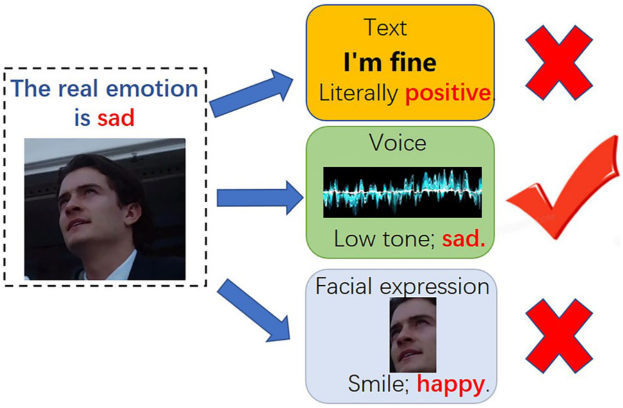


Fig. 1 Comparison of single mode identification results with real ones

The multimodal emotion recognition pipeline consists of six critical stages. First, data is collected from various modalities such as audio, visuals, and text. This is followed by preprocessing, which includes tasks like noise filtration and feature extraction. Next, the data streams are fused using techniques such as early fusion, late fusion, model fusion, and hybrid fusion. This fused data is then used to train and fine-tune an emotion recognition model. The model then classifies emotions, and the final step involves analyzing the detected emotions for trends and potential areas for improvement, as depicted in Fig. 2.

Related Surveys

Since multimodal emotions have been extensively studied by many scholars, there are now classic reviews available.

In 2023, Das et al. conducted an all-encompassing review focusing on the trends transitioning from unimodal to multimodal methods in emotion analysis [10]. Following a similar vein, Zhu et al. surveyed fusion methods in emotion recognition technologies, touching upon their advantages, limitations, and application scenarios, as well as suggesting avenues for future dataset and metric development [11]. Naveed et al. specialized in the exploration of multimodal emotion analysis on social media platforms. Their work involved a comparative review of popular emotion recognition datasets and an in-depth discussion of various machine learning classifiers [12]. Similarly, Jabeen et al.'s 2023 review delved into an array of modalities like images, text, audio, and physiological signals. They offered insights into recent advances in multimodal deep learning applications [13]. Gandhi et al. explored the analysis of users' emotions toward various products and services in their 2022 review. They paid special attention to cutting-edge multimodal fusion architectures and indicated gaps in the existing research [14]. Giovanna, also publishing in 2022, summarized the two-dimensional perspective of multimodal integration, describing its specific applications in fields such as speech recognition and image processing [15]. Zhao et al. assessed recent developments in deep learning-based multimodal emotion recognition, covering principles, progress, and future directions for this field in their 2022 review [16]. Luna et al. looked at developments in multimodal speech emotion recognition methods and discussed challenges related to validation processes and representation learning [17]. Focusing on human-computer interactions, Ganesh et al. addressed the latest advancements in analyzing emotions across text, audio, and visual data. They discussed feature extraction algorithms and classification techniques in detail [18]. In 2021, Zhao et al. turned their attention to critical

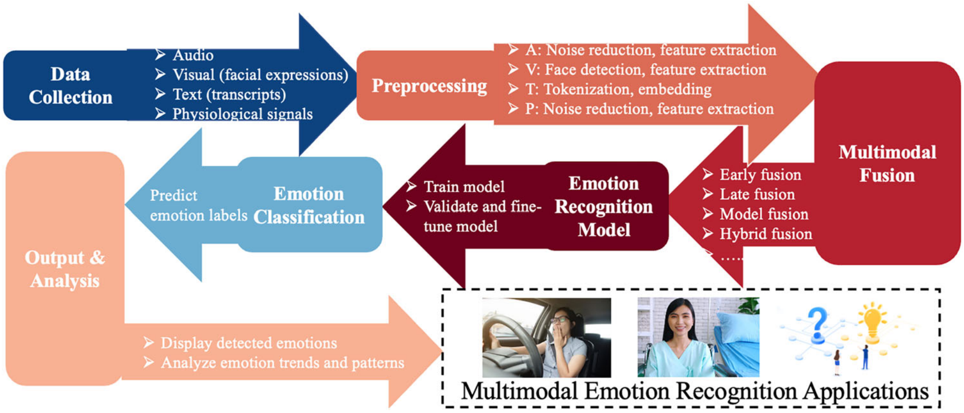

Data Collection
➤ Audio
➤ Visual (facial expressions)
➤ Text (transcripts)
➤ Physiological signals
Preprocessing
➤ A: Noise reduction, feature extraction
➤ V: Face detection, feature extraction
➤ T: Tokenization, embedding
➤ P: Noise reduction, feature extraction
Multimodal Fusion
➤ Early fusion
➤ Late fusion
➤ Model fusion
➤ Hybrid fusion
➤ .....
Emotion Recognition Model
➤ Train model
➤ Validate and fine-tune model
Emotion Classification
Predict emotion labels
Output & Analysis
➤ Display detected emotions
➤ Analyze emotion trends and patterns
Multimodal Emotion Recognition Applications

Fig. 2 Process diagram of multimodal emotion recognition

aspects of multimodal emotion recognition, from annotation strategies to computational tasks, and proposed several optimization methods [19]. Sarah et al. categorized and evaluated 35 of the most advanced models for video emotion analysis, scrutinizing their effectiveness on widely used datasets [20]. Taking a broader view, Garima et al.'s 2021 review covered various emotion recognition techniques and discussed current issues like privacy and fairness [21]. Published in 2020, Nandi reviewed advanced emotion recognition methods in e-learning environments, contrasting them with earlier works [22]. Zhang et al. studied emotion recognition techniques based on brain and physiological signals, discussing machine learning algorithms and outlining open problems [23]. Seng's 2019 review provided an extensive look into emotion and emotion modeling. A novel multimodal emotion and big data modeling architecture was proposed, and its performance was verified through experiments [24]. Baltruaitis et al. in 2018 surveyed multimodal machine learning, offering a new classification scheme that illuminated future research directions [25]. Poria et al. in 2017 focused on the potential performance improvements of multimodal analyses compared to unimodal analyses after conducting a comprehensive literature review [26]. Latha et al. concluded the historical narrative with their 2016 work, which reviewed deep learning techniques for improving classification accuracy in FEMG and speech signals [27]. While numerous reviews exist on multimodal emotion recognition, the majority of them primarily concentrate on fusion method research, with insufficient attention given to single-modal preprocessing and other crucial aspects of multimodal technology.

Our review's primary innovations and differentiations include

- Addressing the prevalent lack of resources in academic journals, conferences, and competitions in this field, our review offers an in-depth collation and comprehensive summary. This provides invaluable reference resources for scholars aiming to delve deeper into this domain.
- Unlike most reviews that primarily focus on technical introductions, our work emphasizes detailed discussions of readily available tools, offering a thorough compilation and elucidation of key tools in this field.
- While most existing literature primarily concentrates on multimodal fusion strategies and their scoring analyses, there is a significant research gap in comprehensive analysis of modal absence, data alignment, and contextual semantics. Our review not only encompasses existing methods in these areas but also delves into other critical technologies and their solutions.
- Existing literature largely focuses on multimodal sentiment analysis based on speech, text, and images, with relatively limited research on the fusion of physiological signals. Our review conducts an extensive analysis of this aspect, laying a solid foundation for further research and expansion in this field.

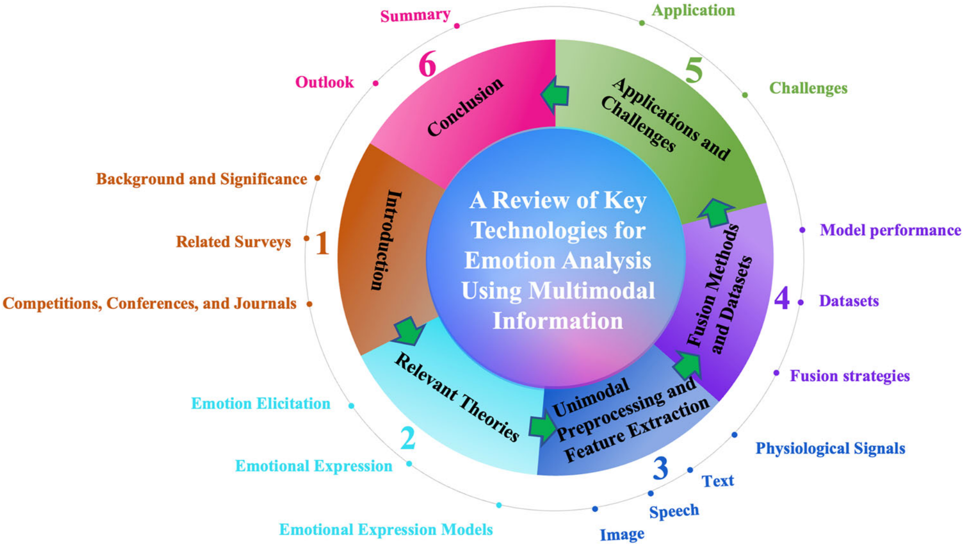
Chapter Arrangement

In order to provide a more comprehensive discussion and summary of deep learning-based multimodal emotion recognition, this paper is divided into six parts, as shown in Fig. 3. Part 1 introduces current conferences, competitions, and journals related to this field. Part 2 evaluates the causes and expressions of emotions and existing multimodal emotion assessment models. Part 3 introduces unimodal feature extraction methods. Part 4 discusses multimodal fusion models and multimodal datasets. Part 5 provides an analysis of challenges and relevant applications. Finally, Part 6 concludes the paper.

Competitions, Conferences, and Journals

In the rapidly evolving landscape of multimodal emotion recognition technology, a plethora of conferences, competitions, and journals are propelling advancements in the field. To help researchers readily navigate this diverse ecosystem, we have curated an overview of some of the most influen-

Fig. 3 Paper Structure Diagram


A Review of Key Technologies for Emotion Analysis Using Multimodal Information
1
Introduction
Background and Significance
Related Surveys
Competitions, Conferences, and Journals
2
Relevant Theories
Emotion Elicitation
Emotional Expression
Emotional Expression Models
3
Unimodal Preprocessing and Feature Extraction
Image
Speech
Text
Physiological Signals
4
Fusion Methods and Datasets
Fusion strategies
Datasets
Model performance
5
Applications and Challenges
Challenges
Application
6
Conclusion
Summary
Outlook

tial platforms, encapsulating key journals, groundbreaking competitions, and seminal conferences.

Several key initiatives have shaped the field of emotion recognition through challenges and competitions. For instance, the Audio-Visual Emotion Challenge (AVEC), initiated in 2011 by the European Science Research Center, has seen multiple iterations focusing on diverse emotional data types like audio, video, and text [28]. Similarly, the annual Emotion Recognition in the Wild (EmotiW) Competition, part of the ACM International Conference on Multimodal Interaction, emphasizes naturalistic emotion recognition using data like facial images and videos [29]. In 2021, the IEEE Signal Processing Society launched the MISP Challenge, focusing on multimodal speech processing and intelligent multimedia interactions [30]. The Affect in the Wild Challenge, which started at the CVPR Conference in 2017, aims to predict emotional states in complex video scenes, leveraging the Aff-Wild dataset [31]. The LIRIS-ACCEDE Emotional Video Classification Challenge is centered around the LIRIS-ACCEDE video database and involves tasks like binary emotion classification and continuous emotion regression [32]. Another notable event is the Multimodal Sentiment Analysis Challenge (MuSe), held annually since 2016. The latest iteration, organized by ACM Multimedia in 2022, featured a range of tasks covering facial expressions, speech, and body gestures [33]. These challenges collectively propel advancements in multimodal sentiment analysis, humor detection, and emotion recognition. The Multimodal Emotion Recognition Challenge (MEC) was initiated in 2018 as an integral segment of the inaugural Asia-Pacific Conference on Affective Computing and Intelligent Interaction (ACII Asia). Aimed at improving emotion recognition efficacy in real-life scenarios, MEC 2017 utilized the China Natural Affective Video Database (CHEAVD) 2.0, incorporating three distinct sub-challenges and attracting wide-ranging global participation [34]. In the ensuing years, the prominence of the 5th Automated Facial Behaviour Analysis Workshop and Competition (ABAW) [35] and the Multimodal Emotion Recognition Challenge (MER 2023) [36], organized under the auspices of ACMMM, has substantially surged in the domain of multimodal emotion recognition. These platforms have not merely catalyzed advancements in this field but have also nurtured a plethora of distinguished research outcomes. Notably, the Muse2022's victorious entry by Li J et al. adopted the TEMMA framework and a self-attention equipped GRU for an innovative amalgamation of multimodal features, integrating groundbreaking audio, facial expression attributes, and paragraph-level text embeddings to boost predictive precision [37]. This methodology markedly enhanced both the accuracy and dependability of emotion predictions by amalgamating multimodal features and effectively leveraged data augmentation techniques to mitigate sample imbalance issues. Further, at the MER 2023 convention, Daoming Zong et al. introduced a multimodal training strategy that amalgamated data augmentation and weighted supervision signal fusion, adeptly countering potential underfitting challenges inherent in single-modality training [38]. The accomplishments underscored by these competitions have been instrumental in propelling forward the research in multimodal emotion recognition (Table 1).

The Neural Information Processing Systems Conference (NeurIPS) is a premier forum in machine learning and computational neuroscience, established in 1987 [39]. Organized by the Association for Computing Machinery, the annual ACM Multimedia Conference focuses on multimedia computing, AI, computer vision, and related technologies [40]. Hosted by the CCF Computer Vision Professional Committee, the International Conference on CVPR is a key event in

Table 1 Overview of competitions related to multimodal emotion recognition

Abbreviation	Website
AVEC	https://avec2021.github.io/
EmotiW	https://cs.anu.edu.au/few/emotiw.html
MISP	https://mispchallenge.github.io/
Aff-Wild	https://ibug.doc.ic.ac.uk/resources/first-affect-wild-challenge
LIRIS-ACCEDE	http://liris-accede.ec-lyon.fr/
MuSe	https://www.muse-challenge.org/
MEC	https://www.acii-conf.org/
ABAW	https://ibug.doc.ic.ac.uk/resources/cvpr-2023-5th-abaw/
MER	http://merchallenge.cn/mer2023

computer vision and image processing, founded in 1983 [41]. The IEEE Computer Society's International Conference on Data Engineering (ICDE) explores advances in data management and AI. Topics range from data modeling to big data computing [42]. The annual conference organized by the Association for Computational Linguistics (ACL) attracts global experts in computational linguistics and increasingly focuses on multimodal research [43]. The International Joint Conference on Artificial Intelligence (IJCAI) has, since 1989, covered a broad array of AI topics, including multimodal research [44]. The IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), hosted by the IEEE Signal Processing Society, is a cornerstone event in signal and speech processing [45]. The ACM International Conference on Multimodal Interaction (ICMI), established in 1908, is a leading forum in multimodal human-computer interaction, affiliated with the International Mathematical Union since 1952 [46]. Initiated in 1998, the IEEE Conference on Multimedia and Expo (ICME) focuses on multimedia technologies, systems, and applications [47]. The Pacific Rim International Conference on Artificial Intelligence (PRICAI) is a significant event in AI, particularly in Asia, with a focus on multimodal data processing [48]. The International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS) is a platform for multimodal data discussions, including intelligent video and smart speaker designs [49].

Multimodal emotion recognition technology benefits significantly from specialized gatherings that serve as hubs for showcasing research, insights, and collaboration. Several journals are pivotal in this field, such as Information Fusion, IEEE Transactions on Affective Computing, and Elsevier's Image and Vision Computing. Other publications like Pattern Recognition, Journal of Ambient Intelligence and Humanized Computing, and International Journal of Human-Computer Interaction offer valuable platforms for related research. Springer and Elsevier offer other notable journals, such as Multimedia Tools and Applications, Universal Access in the Information Society, and Artificial Intelligence Review. These cover emotion recognition from various angles, including medical and transportation contexts. These journals, along with conferences and competitions, play a crucial role in advancing the field. They facilitate the exchange of research findings, encourage innovation, and drive progress by fostering competition. The collective aim is to enhance multimodal emotion recognition technology to meet growing real-world demands.

Related Theories

Emotional elicitation is one of the sources for recognizing and perceiving emotions. By studying emotional elicitation, people can better understand the causes and influencing factors of emotions. Appraisal theory focuses on individuals' cognition and expression of emotions. Through the study of appraisal theory, people can better identify their own and others' emotions. Appraisal theory emphasizes that individuals' representations of emotions vary, indicating that individuals may have different cognitions and expressions of the same emotion [50]. In summary, both emotional elicitation and appraisal theory make significant contributions to emotion recognition.

Emotional Elicitation

When conducting multimodal emotion recognition research, the primary consideration is how to effectively elicit emotions in participants. At present, there are two main methods of emotion elicitation, namely situational elicitation and stimulus-based elicitation [51]. As shown in Fig. 4.

Situational Elicitation

Situational emotions arise from specific contexts, influenced by environmental, personal, and social factors. These emotions are often studied in modern scenarios like work or

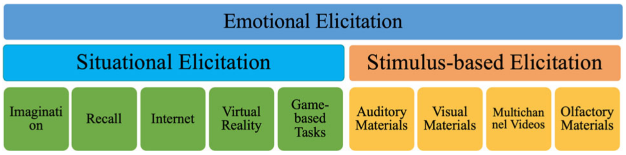


Fig. 4 Emotion induction diagram

leisure, using various elicitation methods, particularly useful for multimodal dataset construction [52].

Figure 4 outlines five key elicitation techniques: Imagination tasks prompt participants to visualize scenarios based on guided directions to evoke specific emotions. While straightforward and ecologically valid, they rely heavily on clear guidance and participant engagement. In Recall tasks, participants revisit past experiences to reproduce associated emotions. This approach offers better control over emotion timing and intensity but necessitates a high level of involvement from the participant. The Internet approach combines conventional elicitation methods with online platforms, streamlining data collection. It is scalable and gives participants the discretion to withdraw. However, web-based games, which enhance emotional engagement, hinge on reliable technology and sizable participation. Virtual Reality (VR) combines lab environments with immersive tech, facilitating quick transitions between emotional stimuli [53]. Its primary challenges lie in achieving lifelike scenarios and maintaining the necessary tech infrastructure. Game Tasks use gameplay to induce varied behaviors and emotions. They double up as emotional stimuli and behavior markers, especially when rewards surpass participant expectations. While each technique has its merits, they commonly rely on the active engagement of participants for successful emotional elicitation.

Stimulus-Based Elicitation

Stimulus-based elicitation uses a range of sensory prompts, such as visuals and sounds, to provoke specific emotional reactions in participants [54]. The link between event-associated emotions and particular scenarios means varied stimuli can effectively evoke distinct emotions. Figure 4 outlines five principal methods: Auditory Stimuli: Natural sounds, including bird calls or rain, along with music, serve well in evoking emotions. Notably, the NIMH has a catalog of these sounds assessed for their emotional resonance. Visual Stimuli: Widely utilized, this method employs emotionally resonant text and images to prompt specific emotional reactions. Several repositories are available, especially for textual stimuli. Multichannel Videos: Merging visual and auditory cues, film snippets are potent emotional triggers, albeit with less precision. For instance, "Mr. Bean" segments are popular for eliciting joy, though standardization is a challenge. Olfactory Stimuli: Scents have the capability to provoke emotions, influencing both thought processes and actions. Many available datasets lean towards film clips and scripted dialogues to capture multimodal emotional data, mirroring natural emotional expression closely.

Emotion Expression

Facial expressions are a vital aspect of emotional communication, accounting for 55% of how emotions are conveyed according to psychological research [55]. These non-verbal cues are essential in human interactions and can significantly impact first impressions. Different emotions elicit unique facial responses, involving specific muscle contractions and relaxations. These are universally understood to signify basic emotions like anger, happiness, or fear [56]. Therefore, facial expressions serve as the primary medium for emotional expression.

Emotion Expression Models

Emotion recognition is essentially about discovering the mapping relationship between features related to emotions and the underlying emotional states. Emotion modeling refers to describing emotional states by establishing mathematical models and scientifically classifying or even quantifying emotions. The establishment of emotion models has significant implications for emotion measurement, as it allows for more accurate assessment of emotional states. In 2003, Picard discussed emotion modeling descriptions. Many researchers have proposed corresponding emotion representation methods, which can be divided into discrete emotion models and continuous emotion models based on different representation methods [57].

Discrete Emotion Representation

Discrete emotion representation has been central to psychology and affective computing since the mid-twentieth century. Initially, Tomkins categorized basic emotions into positive and negative groups [58]. Ekman later identified universal expressions for six basic emotions, and Plutchik expanded this to an eight-emotion model, capturing nuances like anticipation and joy [59]. In the twenty-first century, the field has benefited from AI and machine learning advancements, leading to more refined models for representing discrete emotions. While the prevailing view focuses on a combination of six basic emotions, including sadness, anger, and happiness, discrete models have limitations in capturing the full spectrum and intensity of human emotions.

Continuous Emotion Representation

Continuous emotion representation offers a nuanced alternative to discrete emotion models. Russell's circumplex model is the most recognized, representing emotions in a two-dimensional space with axes for arousal and valence [60]. This allows for a broad mapping of emotional states. The PAD model by Mehrabian extends this by adding a third dimension-dominance. It has been widely employed in fields like affective computing, virtual reality, and human-computer interaction [61]. Continuous models offer the advantage of capturing both the type and intensity of emotions. However, they present the challenge of requiring sophisticated algorithms to map continuous features from multimodal data onto their underlying emotional dimensions. Both discrete and continuous models have their own merits and limitations. The choice between them depends on the research goals, data type, and desired granularity. Researchers need to carefully evaluate these factors to select an appropriate emotion representation model [60, 61].

Unimodal Preprocessing and Feature Extraction

Image Modality

Pretreatment

Preprocessing in facial emotion recognition plays a crucial role in improving model accuracy and efficiency. The main purposes of preprocessing are to enhance image quality by removing noise, performing histogram equalization, and enhancing contrast; to expand the dataset through data augmentation techniques, increasing data diversity, and making the model more robust; and to address the issue of data imbalance by adopting class-balancing methods. For complex image datasets, preprocessing techniques can simplify and optimize data, compressing the dimensionality of the data, thereby reducing processing time and computational complexity. Preprocessing techniques can also standardize and ensure consistency in input data, minimizing the impact of noise and interference, thus improving model generalizability and accuracy [62, 63].

Common preprocessing methods include image normalization, which standardizes image size, brightness, and color [64]; image enhancement, which employs filters to improve image contrast and clarity [65]; histogram equalization, which adjusts image brightness distribution for a more even distribution [66]; color space transformation, which converts images from RGB or other color spaces to HSV, Lab, etc., for better processing and feature extraction [67]; denoising, which uses filters to remove noise from images [68]; edge detection, which employs edge detection algorithms to locate edges and contours in images [69]; feature point detection, which uses feature point detection algorithms to identify keypoints and corner points in images [70]; face detection, which uses face detection algorithms to locate faces in images and analyze them based on face position and pose information [71]; keypoint annotation, which manually or semi-automatically annotates and locates keypoints on facial features such as eyes and mouth [72]; and face alignment, which calibrates and aligns face images for better extraction and analysis of facial expression information [73].

Facial information, which constantly changes based on facial features, contains rich emotional information. Image/video feature extraction for facial emotion recognition is mainly based on geometric and texture features. Geometric features represent faces using a set of vectors based on the position, size, and proportion of facial features. Relevant facial expressions and body language information are extracted to further mine underlying emotional information. These features can serve as a supplement to enhance the accuracy and robustness of emotion recognition systems. Moreover, in certain scenarios, text and audio information may not provide sufficient emotional information, while image modality can provide a more direct and richer emotional presentation, making image modality feature extraction crucial in multimodal emotion recognition [74].

Feature Extraction

In this paper, we delve deeply into various dimensions of facial expression features. Among these are geometric positions like eye, mouth, and nose placements; curvatures of facial components such as eyebrows and lips; and temporal aspects such as the onset and duration of expressions. We also examine facial motion metrics, including velocity and angle changes, and textural elements like wrinkles and skin tone. Furthermore, we evaluate symmetry, internal motion relationships between facial parts, fine-grained details like eyelid tremors, area and volume changes, amplitude, and dynamic properties. We even consider the relationship between facial expressions and speech prosody, and explore multiple feature extraction techniques.

For feature extraction, we compare manual and deep learning-based approaches. In manual extraction, we cover an array of traditional methods including Gabor filters for texture, Local Binary Pattern (LBP) for local texture, and Histogram of Oriented Gradients (HOG) for shape and edge features [75–77]. Other techniques such as Speeded Up Robust Features (SURF), Local Phase Quantization (LPQ), and Haar-like features are discussed for their relevance in capturing local facial characteristics [78–80]. We also delve into color and spatial features, using methods like Color Histogram and Color Co-occurrence Matrix [81, 82]. Advanced statistical methods like Non-negative Matrix Factorization (NMF) and Principal Component Analysis (PCA) are discussed for their utility in feature extraction [83, 84].

The advantages of manual feature extraction include its strong interpretability and lower computational needs, making it particularly suitable for small datasets. However, it has limitations such as sensitivity to noise and occlusions, and its feature selection often hinges on domain-specific expertise, impacting its generalizability.

Our work aims to serve as a comprehensive guide, providing insights into the nuanced landscape of facial expression feature extraction for both academic and industry research [85].

We discuss various deep learning methods for feature extraction, namely Convolutional Neural Networks (CNNs), AlexNet, VGGNet, Inception (GoogLeNet), ResNet, DenseNet, MobileNet, EfficientNet, 3D CNNs, and C3D. CNNs are capable of automatic feature learning in images, including facial expressions, thanks to their architecture of convolution, pooling, and activation layers [86]. AlexNet, equipped with five convolutional layers and three fully connected layers, excels in facial recognition tasks through pre-training and fine-tuning [87]. VGGNet offers enhanced feature extraction capabilities, typically featuring between 13 and 19 layers, and is also adaptable to facial expression recognition through transfer learning. Inception employs parallel Inception modules to capture facial expression features at various scales [88]. ResNet uses residual connections to tackle the vanishing gradient issue, enabling deeper and more feature-rich models [89]. DenseNet's architecture ensures efficient feature propagation and gradient backpropagation [90]. MobileNet focuses on balancing performance and computational resources, making it suitable for mobile and real-time applications [91]. EfficientNet dynamically adjusts its architecture to maintain high performance with reduced complexity [92]. For video data, 3D CNNs and C3D provide effective methods to capture dynamic facial expressions over time [93].

Deep learning for feature extraction comes with its own set of pros and cons. Among the advantages are automatic feature learning, high performance, and robustness against image deformations and noise. However, the drawbacks include a need for large training datasets, significant computational resources, and limited interpretability.

Tools

Table 2 Facial processing-related tools

Tool	Website
OpenFace	www.github.com/TadasBaltrusaitis/OpenFace
Affectiva	www.affectiva.com
Rekognition	www.aws.amazon.com/rekognition
OpenCV	www.opencv.org
scikit-image	www.scikit-image.org
Pillow (PIL)	www.python-pillow.org
imageio	www.imageio.github.io
TensorFlow	www.tensorflow.org
Keras	www.keras.io
PyTorch	www.pytorch.org
torchvision	www.pytorch.org/vision
Caffe	www.caffe.berkeleyvision.org
Dlib	www.dlib.net
face_recognition	www.github.com/ageitgey/face_recognition
MTCNN	www.github.com/ipazc/mtcnn
FER	www.github.com/kdhht2334/awesome-SOTA-FER

Table 2 shows the common facial expression processing tools. Common facial expression feature extraction tools include OpenFace, an open-source software library for face recognition and facial expression analysis, capable of extracting features related to facial pose, expression, eyes, mouth, and head rotation [94]. Affectiva is a deep learning-based facial expression and emotion recognition software, capable of performing face detection, facial expression analysis, and emotion recognition [95]. Rekognition is Amazon AWS's facial analysis API, supporting face recognition, facial analysis, face search, and person tracking [96]. OpenCV is an open-source computer vision library for face detection, facial expression analysis, and other operations [97]. scikit-image is an image processing library based on SciPy, offering various image processing and analysis functions [98]. Pillow (PIL, Python Imaging Library) is a Python library for handling and manipulating images [99]. imageio is a Python library for reading and writing various image formats [100]. TensorFlow is an open-source library for machine learning and deep learning, suitable for handling image expression feature extraction tasks [101]. Keras is a high-level deep learning library based on TensorFlow, making it easy to build and train neural networks [102]. PyTorch is an open-source library for machine learning and deep learning, featuring dynamic computation graphs and a simple API [103]. torchvision is a computer vision library based on PyTorch, including pre-trained models, datasets, and image transformations [104]. Caffe is an open-source framework for deep learning, supporting various deep learning models, including CNN and RNN [105]. Dlib is a C++ library containing machine learning, image processing, and computer vision functionalities, offering a Python interface [106]. face-recognition is a Python library for face recognition and processing tasks, implemented based on dlib [107]. MTCNN is a Python library for real-time face detection and alignment, based on multi-task cascaded convolutional networks

[108]. FER is a Python library for expression recognition, implemented based on Keras, capable of recognizing facial expression features [109].

Speech Modality

Pretreatment

In speech emotion recognition, preprocessing is pivotal for enhancing model accuracy and robustness. Preprocessing chiefly aims at noise removal, signal enhancement, segmentation, rhythm analysis, feature extraction, data balancing, and refining recognition precision. Effective preprocessing can notably elevate model performance, expanding the practical applications of speech emotion recognition.

Prominent preprocessing techniques encompass preemphasis, framing, and applying window functions. Techniques like Discrete Fourier Transform (DFT) and Mel filter bank fine-tune the audio signal for processing. The extraction of Mel Frequency Cepstral Coefficients (MFCC) or Discrete Cosine Transform (DCT) is also common. Dynamic features, including first-order and second-order differences, are procured by high-order differencing of the speech signal. Feature vectors are scaled via normalization to negate variations across recording scenarios. Forward-inverse filtering attenuates short-term energy dominance, while high-pass and low-pass filters retain essential signal details. The adaptive predictive filter focuses on enhancing the signal-to-noise ratio, while spectral smoothing fortifies feature robustness. Techniques like spectral subtraction clarify speech signals, and channel error compensation corrects transmission distortions. Dynamic threshold setting, entropy calculations, and spectral centroid calculations offer nuanced signal analysis. Time-domain parameter calculation, endpoint detection, and sound pressure level normalization further refine the recognition process. Addressing environmental factors includes recording noise reduction, echo cancellation, and environment matching. Volume normalization and standard noise reduction methods, such as median and mean filtering, culminate in clearer speech signals for emotion recognition [75, 76].

Feature Extraction

In the realm of speech emotion recognition, feature extraction serves as a pivotal step to convert raw speech signals into computable formats, thereby efficiently capturing emotional cues. Feature extraction algorithms generally fall into mathematical and signal processing categories, aiming to capture variables such as pitch, rhythm, tone, and energy.

Traditionally, Mel Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), and Gammatone Filterbanks have been the go-to methods for extracting speech features [75]. Other techniques such as Wavelet Transform, Discrete Cosine Transform (DCT), and Teager Energy Operator (TEO) are also employed. Features like pitch and energy serve as auxiliary inputs. Zero Crossing Rate, Short-time Energy and Entropy, and Linear Spectral Frequency (LSF) are also notable [78]. Handcrafted features are interpretable, computationally efficient, and do not require large annotated datasets. However, they may lack representational power, require domain expertise, and are sensitive to noise and variability. On the flip side, deep learning methods like Deep Belief Network (DBN), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs) have been effective in capturing complex features [80]. These methods are robust but are considered "black box" models, requiring substantial computational resources and large annotated datasets [110].

Tools

The tools commonly used in speech processing are summarized in Table 3. Among them, Praat offers both basic and advanced speech feature extraction and is open-source [81]. OpenSMILE is a versatile C++ toolkit that can capture a range of acoustic, linguistic, and emotional features [82]. LIUM SpkDiarization specializes in identifying different speakers within a single audio stream [111]. Kaldi is renowned for its capabilities in speech feature extraction and model training [112]. Commercially available HTK provides various algorithms for feature extraction and model training [113]. YAAFE and Librosa are Python-based tools suitable for both speech and music feature extraction [83, 114]. DeepSpeech, developed by Mozilla, leverages deep learning for speech recognition [115]. VoiceSauce and COVAREP focus on extracting prosodic and resonance features, respectively [116, 117]. IS10 sets the standard for a variety of acoustic and linguistic speech features [83]. PyAudioAnalysis and SoX offer broad capabilities, while specialized tools like PRAISE-1 and MFA focus on advanced algorithms and neural network-based features. FAVEExtract is geared toward accent research, and SPHINX offers a probabilistic approach to speech recognition [118].

Text Modality

Pretreatment

This paper presents 20 text preprocessing techniques to enhance the accuracy of text emotion recognition models. First, remove HTML tags using Python's Beautiful Soup library as they don't contribute to sentiment analysis. Second, eliminate non-letter characters like numbers and symbols using regular expressions.

Table 3 Speech processing-related tools

Tool	Website
Praat	www.praat.org
OpenSMILE	www.audeering.com/opensmile
LIUM SpkDiarization	www.projets-lium.univ-lemans.fr
Kaldi	www.kaldi-asr.org
HTK	www.htk.eng.cam.ac.uk
YAAFE	www.github.com/Yaafe
DeepSpeech	www.github.com/mozilla/DeepSpeech
VoiceSauce	www.github.com/voicesauce
COVAREP	www.github.com/covarep
Librosa	www.librosa.org
PyAudioAnalysis	www.github.com/tyiannak/pyAudioAnalysis
SoX	www.sox.sourceforge.net
FAVEExtract	www.fave.ling.upenn.edu
WaveSurfer	www.wavesurfer-js.org
ELAN	www.tla.mpi.nl/tools/tla-tools/elan
SPHINX	www.cmusphinx.github.io
EMU Speech Database System	www.emu.r-forge.r-project.org

Numbers should be stripped from text because they hold no emotional value. Likewise, web links need to be removed to prevent model interference. Punctuation marks, generally insignificant in sentiment analysis, can be deleted using string functions. Converting text to lowercase using string functions reduces redundancy. Common non-contributory words, or stop words, can be filtered out with libraries like NLTK. Words can be reduced to their base forms, or lemmatized, using NLTK or spaCy. Assign part-of-speech tags to words using these libraries to assist in sentiment analysis. Tokenize the text into individual words, also achievable through NLTK or spaCy. Gather word frequency statistics using Python's Counter class for pattern insights. Limit whitespace between words to a single space and remove line breaks using string functions. Confusing abbreviations can be expanded to their full forms with the pyenchant library. Eliminate consecutive duplicate words and words with very low occurrence after setting a frequency threshold. Similarly, remove high-frequency words that offer little value after establishing a higher threshold. Interestingly, reversing sentence word order can make models more sensitive to textual nuances. Remove emotionally polarized words like "happy" or "sad" using a sentiment word library. Finally, replace less expressive emotion-indicative words with synonyms via WordNet [119].

Feature Extraction

In this study, we conduct a thorough exploration of text feature extraction, a cornerstone in multimodal emotion recognition. The process transforms raw text into numerical or vector formats that encapsulate the emotional undertones, making them suitable inputs for machine learning algorithms focused on sentiment classification or regression.

Key text features for sentiment analysis encompass sentiment word frequency, which quantifies the emotional tone; negation word frequency for assessing emotional polarity; and degree adverb frequency to evaluate emotional intensity. Other noteworthy features include ordering word frequency for directional emotions, sentiment word ratio, and more intricate metrics such as sentiment score modes and standard deviations, sentence length, text length, and various types of word frequencies. Punctuation marks such as exclamation and question marks are also considered, alongside advanced metrics like sentence sentiment score and distribution. For speech emotion recognition, two primary schools of feature extraction exist: manual and deep learning. Manual methods, while interpretable and computationally efficient, often fail to capture the complexity of emotional undertones. Techniques range from count vector and TF-IDF to more sophisticated methods like Word2Vec and FastText [120]. On the flip side, deep learning methods such as RNNs, LSTMs, and CNNs offer robust feature representation but come with high computational costs and data requirements. These methods include a wide array of architectures, including but not limited to GloVe, BiLSTM, TextCNN, RCNN, Transformer-based models like BERT, and its variants like ALBERT and T5 [84, 86, 121–123].

While deep learning techniques excel in automatic feature learning and robustness, they are computationally intensive and require extensive labeled datasets, limiting their applicability in data-scarce environments.

Tools

Text Emotion Recognition toolkits, as delineated in Table 4, comprise collections of models and algorithms specifically geared toward textual feature extraction. Unlike feature extraction toolkits, these toolkits focus primarily on sentiment and emotion recognition. TextBlob is a Python library offering services such as sentiment polarity detection, while NLTK, another Python library, provides a broad suite of natural language processing tools including sentiment analysis and syntactic parsing. Similarly, spaCy specializes in text parsing and attribute extraction, and Pattern offers an expansive range of natural language processing capabilities. In the Java ecosystem, Stanford CoreNLP provides functionalities such as sentiment analysis and entity recognition.

For cloud-based solutions, IBM Watson offers a comprehensive AI platform featuring sentiment analysis and entity recognition, and Google Cloud Natural Language API delivers a wide variety of text processing features, including sentiment analysis. Amazon Comprehend incorporates machine learning-based sentiment recognition and entity extraction. TextRazor serves as an API-focused service offering a wide array of features, while MonkeyLearn is an online tool proficient in both Chinese and English sentiment analysis and custom model training. Various SaaS platforms such as Tencent Cloud, Alibaba Cloud, and Huawei Cloud also offer robust sentiment analysis services. OpenAI's GPT-3 not only excels in text generation but can also analyze text sentiment based on its generated outputs [124].

Physiological Signals Modality

Pretreatment

Physiological signals are prominently significant in emotion analysis, with numerous researchers utilizing various biosensors and monitoring devices to capture the physiological responses of the human body in different emotional states. The primary physiological signals used for emotion recognition and their methods of acquisition are as follows:

Electrocardiogram (ECG) measures the electrical activity of the heart through electrodes attached to the skin. It reveals variations in heart rate and irregularities, features closely linked to emotional states. Electrodermal activity (EDA), also known as galvanic skin response (GSR), assesses increased sweat gland activity by measuring the electrical conductivity at the fingers or other skin areas, typically associated with emotional states of tension or excitement [125]. Electroencephalogram (EEG) records brain activity through electrodes placed on the scalp, capturing changes in brain activity patterns under different emotional states. Heart rate variability (HRV), derived from analyzing ECG data, refers to variations in intervals between heartbeats and is indicative of stress and emotional arousal. Respiratory pattern monitoring relies on breath sensors to track changes in breathing frequency and depth, reflecting changes in emotional states. Electromyography (EMG) measures electrical activity in specific muscle groups (like facial muscles) to reflect emotional states, with facial EMG being particularly useful in analyzing expressions such as smiling or frowning [126].

To ensure the quality and accuracy of data analysis for these physiological signals, their preprocessing involves several key steps. Signal denoising is crucial, including the use of filters to remove electrical noise and interference caused by muscle movements. Normalization helps reduce physiological differences between individuals, typically scaling data to a certain range (e.g., 0 to 1) [127]. Artifact identification and removal are essential to eliminate abnormal signals caused by movement or equipment failure, ensuring data quality. Data segmentation and windowing involve dividing continuous signals into short segments, each representing a time window, facilitating focused analysis on specific signal characteristics over time. Time-frequency analysis, especially for

Table 4 Text processing-related tools

Toolkit	Website
TextBlob	www.textblob.readthedocs.io
NLTK	www.nltk.org
spaCy	www.spacy.io
Pattern	www.github.com/clips/pattern
Stanford CoreNLP	www.stanfordnlp.github.io/CoreNLP
IBM Watson	www.cloud.ibm.com/docs/watson
Google Cloud Natural Language API	www.cloud.google.com/natural-language
Amazon Comprehend	www.aws.amazon.com/comprehend
TextRazor	www.textrazor.com
MonkeyLearn	www.monkeylearn.com
OpenAI GPT-3	www.openai.com/gpt-3

signals like EEG, can reveal frequency features associated with different emotional states [128].

Feature Extraction

In multimodal emotion analysis, traditional methods for extracting features from physiological signals are crucial. For electrocardiogram (ECG) signal, features such as heart rate (HR), which is the count of heartbeats per minute, heart rate variability (HRV), analyzing the time variation between heartbeats, and QRS waveform, examining the shape and duration of the QRS complex in ECG, are indicative of cardiac electrophysiological properties [129]. Electrodermal activity (EDA) features include peak count, measuring the number of EDA peaks within a specific timeframe, response amplitude, assessing the strength of EDA responses, and recovery time, the duration it takes for EDA peaks to return to baseline [130].

Electroencephalogram (EEG) feature extraction involves energy spectrum, analyzing the energy of different frequency bands like alpha and beta waves, coherence, measuring the correlation of electrical activity between different brain areas, and event-related potentials (ERP), analyzing brain responses post specific stimuli or events. Heart rate variability (HRV) features in time domain, such as average intervals between heartbeats and standard deviation, and in frequency domain, analyzing the contribution of different frequency components to HRV, are also examined [131].

Respiratory pattern features include respiratory rate, counting the number of breaths per minute, depth of breathing, analyzing the shallowness or depth of breaths, and regularity, assessing the consistency of the breathing pattern. Electromyography (EMG) features involve amplitude characteristics, analyzing the magnitude of the EMG signal, frequency characteristics, extracting the primary frequency components of the EMG signal, and duration of muscle activity, measuring the length of specific muscle activities.

Deep learning methods have emerged as advanced and effective for feature extraction from physiological signals. Convolutional Neural Networks are particularly suited for data with spatial structure, like EEG and ECG signals, where hierarchical features of the signals are extracted through multiple convolutional and pooling layers. For EEG signals, 2D-CNNs can be used to process the signals represented as time-frequency maps [132]. Recurrent Neural Networks and Long Short-Term Memory networks are appropriate for time-series data such as ECG and HRV signals, capturing temporal dependencies and long-term relationships within the signals. For respiratory patterns and EMG signals, these networks can effectively identify and analyze temporal pattern variations [133].

Autoencoders, useful for data dimensionality reduction and feature learning, are particularly applicable for EDA signals, learning higher-level and abstract features of the signals. Attention mechanisms and Transformers allow the model to focus on key parts of the signal, like specific waveforms in ECG or EEG, and are effective for analyzing long sequence data in HRV or EMG signals. Hybrid models, combining CNNs with RNNs/LSTMs, can extract both spatial and temporal features, making them particularly effective for analyzing multi-channel EEG or EMG data.

Tools

The field of physiological signal processing is well-established, boasting a plethora of powerful toolboxes, each with its unique functionalities and specialties. As shown in Table 5, typical toolboxes include BioSPPy, an open-source Python library designed for processing physiological signals such as ECG, EDA, EMG, and PPG. It offers functionalities for signal denoising, feature extraction, and graphical representation. EEGLAB, a MATLAB-based open-source toolbox, focuses on EEG data analysis and visualization, incorporating features like ICA, time-frequency analysis, and event-related potentials.

FieldTrip, another MATLAB-based open-source toolbox, is used for complex analyses of EEG and MEG signals, providing time-frequency analysis, source localization, and statistical testing. HeartPy, a Python library, is dedicated to heart rate signal analysis and is suitable for both ECG and PPG signals, offering baseline detection and heart rate variability analysis. PyEEG, a Python module, specializes in EEG data feature extraction including frequency analysis, entropy features, and synchronization measures.

LabChart is a commercial software suitable for various physiological signals, offering data recording, analysis, and visualization, ideal for educational and research settings. Kubios HRV focuses on heart rate variability analysis, providing detailed HRV analysis in time domain, frequency domain, and non-linear analysis, suitable for clinical and research applications. BrainVision Analyzer, a commercial software for EEG and ERP data analysis, offers data preprocessing, analysis, and visualization tools, widely used in neuroscience research.

MNE-Python, an open-source Python package, is dedicated to MEG and EEG data analysis, including data preprocessing, source estimation, and statistical analysis. OpenSignals, software for Biosignalsplux physiological signal acquisition devices, offers signal acquisition, real-time analysis, and visualization, applicable to ECG, EDA, EMG, and ACC signals.

These toolboxes, with their varied focus areas, provide robust support for physiological signal processing and emotion analysis research, ranging from basic signal processing to complex statistical analysis and pattern recognition.

Table 5 Physiological signal processing toolkits

Toolkit	Website
BioSPPy	www.biosppy.readthedocs.io
EEGLAB	www.sccn.ucsd.edu/eeglab
FieldTrip	www.fieldtriptoolbox.org
HeartPy	www.python-heart-rate-analysis-toolkit.readthedocs.io
PyEEG	www.pyeeg.readthedocs.io
LabChart	www.adinstruments.com/products/labchart
Kubios HRV	www.kubios.com/hrv
BrainVision Analyzer	www.brainproducts.com/productdetails.php?id=17
MNE-Python	www.mne.tools/stable/index.html
OpenSignals	www.biosignalsplux.com/en/software/opensignals

Fusion Strategies and Datasets

Fusion Strategies

In multimodal emotion recognition, feature fusion is an essential step. Multimodal emotion recognition requires extracting the most representative and distinctive features from different modalities. Since the features between various modalities are not entirely the same, these features need to be fused. Feature fusion can combine different features from multiple modalities, thereby improving the model's judgment accuracy for emotional states.

Early Fusion

In multimodal emotion recognition, early fusion refers to the process of feature extraction and processing from different modalities and then combining the information for emotion classification. Early Fusion of Feature Vectors is one of the earliest multimodal emotion recognition fusion algorithms, proposed by Bazzanella et al. in 1997. The idea is to concatenate each modality's feature vector directly as input for the classifier. As shown in Fig. 5, all modalities' features are combined into a new comprehensive feature vector, which is then fed into the classifier for emotion classification. Feature-level fusion is the shallow fusion of different modalities after feature extraction, directly connecting multiple modalities through shallow concatenation, addition, or weighted summation. Before deep learning, feature engineering is often used to extract modality features. Feature fusion requires integrating various features from different modalities into a common space. Due to the differences between modalities, this often includes a lot of redundant information. Dimensionality reduction techniques, such as Principal Component Analysis, are used to eliminate redundant information.

The paper [134] introduces a Contrastive Learning and Multilayer Fusion (CLMLF) for multimodal emotion detection, encoding text and images to fuse their features. Yoon et al. address the limitations of feature-based models in learning emotional cues from speech signals by proposing a multimodal dual recursive encoding model [135]. Reference [100] suggests a Multi-modal Fusion Model with multi-level attention mechanisms for analyzing depression-related audio-visual and textual data. Hazarika et al. employ convolutional neural networks to create a 300-dimensional multimodal mapping [136]. Mai et al. design a Hierarchical Feature Fusion Network (HFFN) that employs an ABS-LSTM network and attention mechanisms [137]. You et al. combine CNN and LSTM to capture relationships between text and images [138]. Chen et al.'s GME-LSTM model innovatively fuses multimodal information at the word level [139]. Zadeh et al. present a tensor fusion network for online video emotion recognition, considering both intra- and inter-modality aspects [140]. Rosas et al. use feature-level fusion for emotion polarity determination using SVM [141]. Poria and colleagues propose an LSTM-based model and a convolutional recurrent multiple kernel learning (CRMKL) model [142, 143]. Deng et al. fuse time-varying visual information with audio and text for emotion analysis [144].

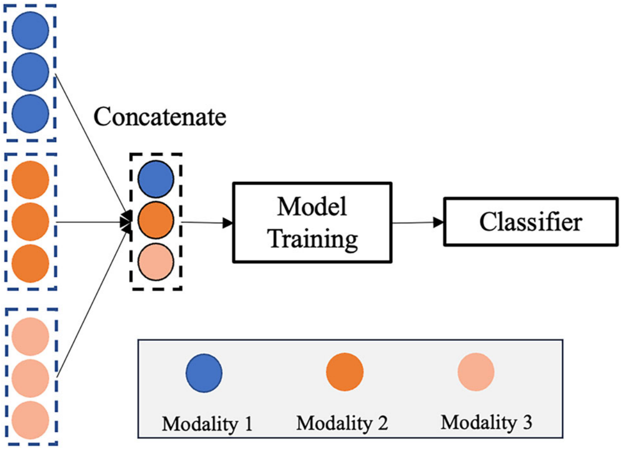


Fig. 5 Early-Stage Fusion Process Flowchart

Other notable models include the Hierarchical Stacked Graph Convolutional Framework (HSGCF) [91], multi-feature and multi-modality fusion models [145], and self-attention mechanisms [146]. Techniques also involve deep semantic feature extraction [147], entity-level multimodal emotion classification [148], and text-dominated frameworks [149]. Xu et al. leverage aspect-guided attention mechanisms [150], Liu et al. employ a deep belief network [151], and Siriwardhana et al. use a Transformer-based self-supervised feature fusion approach [152].

The advantages of early fusion include simplicity, full utilization of modality information, and efficiency. Its drawbacks are high dimensionality, risk of overfitting, and inability to utilize intermodal relationships effectively.

Late Fusion

Definition of late fusion: Late fusion methods refer to the process of training classifiers separately for each modality and then fusing the output results of multiple classifiers to obtain the final classification result. As shown in Fig. 6, classic late fusion decision-level fusion strategies:

Huang et al. examined emotion recognition performance through two decision-level fusion methods: enumeration weights and adaboost. They used facial expression and electroencephalogram classifiers and found satisfactory results on DEAP and MAHNOB-HCI datasets [153]. A neural network-based framework leveraging transfer learning for emotion recognition from speech and text is presented in [154]. Another study introduces the Hierarchical Fusion Graph Convolutional Network (HFGCN), which incorporates modality dependencies for more accurate arousal and valence recognition [95]. Lu et al. improved emotion recognition accuracy to 87.59% using a fuzzy integral fusion strategy that combines eye movement and electroencephalograms [154]. In contrast, some research, such as by Yu, J et al., focuses on single-modality feature extraction for sentiment analysis on Weibo using deep CNNs [155]. For handling heterogeneous data, Yu et al. proposed a deep CNN-based method that employs separate classifiers for text, video, and audio [156]. Poria, S addressed dataset limitations through a select-additive learning (SAL) procedure, achieving satisfactory generalization across different datasets [157]. Wang et al. used a new loss function and attention mechanisms to focus on significant aspects of video sequences for emotion recognition [158]. Williams et al. demonstrated that both feature and decision fusion can outperform single modality approaches, especially when dimensionality reduction is considered [159]. Gkoumas et al. introduced a quantum cognition-driven fusion strategy, modeling single modality classifiers in a complex-valued Hilbert space [160]. Sun et al. combined voice and micro-expression recognition and found improved accuracy through tensor fusion networks and soft fusion techniques [161].

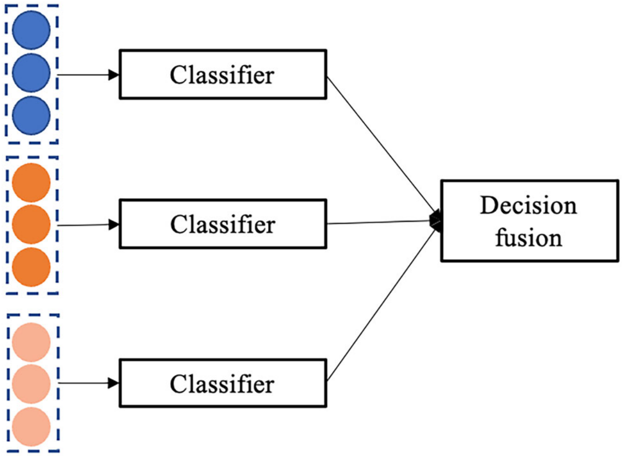


Fig. 6 Late fusion process flowchart

Advantages of late fusion methods include scalability, time-efficiency, and modular adaptability for diverse tasks. However, they suffer from correlation ignorance between modalities, making them potentially prone to misclassification. They also necessitate complex weight adjustments and extensive data support for method selection.

Model Fusion

In multimodal emotion recognition, model fusion combines different models together, extracting the essence of each model to achieve the best results. Classic model-level fusion methods do not rely on the three levels of fusion architecture mentioned above. Decision-level fusion focuses on the credibility of different modalities at the decision stage, but model-level fusion does not need to focus on the importance of each modality. Instead, it is necessary to establish appropriate models based on modality characteristics and jointly learn related information. Feature-level fusion primarily constructs feature sets or mixed feature spaces before sending them to classification models for decision making. As shown in Fig. 7, model-level fusion can input different modality features into different model structures for further feature extraction.

Huang et al. examined emotion recognition using enumeration weights and adaboost in decision-level fusion strategies. Both methods showed strong performance on public datasets such as DEAP and MAHNOB-HCI [153].

Reference [88] introduced M2Seq2Seq, a multi-modal and multi-task learning model with specialized attention mechanisms. Another study, citing the arousal model in cognitive science, proposed a Deep Emotional Arousal Network (DEAN) that simulates emotional consistency in transformers [162]. GA2MIF, a two-stage multi-modal fusion

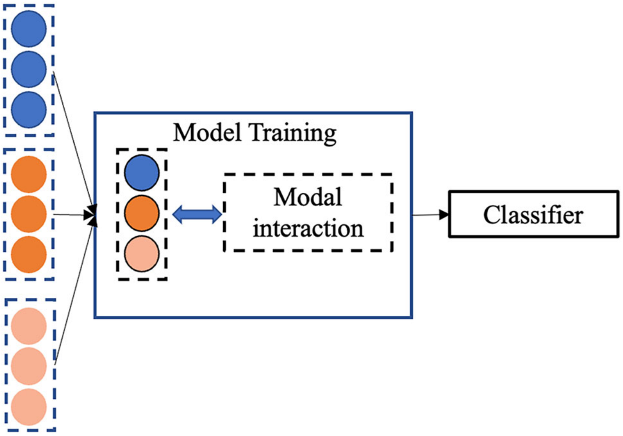


Fig. 7 Model layer fusion flowchart

approach, was presented for emotion detection in conversations, utilizing advanced graph attention networks [90]. Comparative learning was introduced as a regularization technique to reduce noise in multimodal emotion recognition using audio and text [87]. TETFN, a Text-Enhanced Transformer Fusion Network, leverages text-based multi-head attention to integrate emotion-related nonverbal cues effectively [163]. Zheng et al. used deep belief networks for model-level fusion, incorporating EEG and eye movement features to improve emotion recognition [164]. Zhang et al. took a hybrid deep learning approach, integrating visual and auditory data through CNN and 3DCNN, respectively, and then fusing them in a Deep Belief Network for joint feature learning [165].

Model fusion enhances classification accuracy and robustness but requires greater computational resources and domain expertise. It also runs the risk of error propagation due to model interactions. Late fusion methods are scalable and computationally efficient but may ignore intermodal correlations, leading to potential misclassifications.

Hybrid Fusion

Hybrid fusion strategies harness the strengths of early, mid-level, and late fusion methods throughout various data processing stages. Chen's multimodal LSTM model, equipped with temporal attention gating for word-level fusion, emphasizes both communication context and timing. This groundbreaking approach adeptly reduces noise using attention and reinforcement learning [166].

In a quest to identify frustration in gaming, research introduced two fusion modes—decision-level and feature-level—using Deep Neural Networks. The study deployed diverse architectures fine-tuned to specific data types and tested them against authentic audio-visual datasets [89]. Shenoy's end-to-end RNN model for sentiment evaluation deploys sGRU and cGRU units to discern intermodal relationships and emotional conditions. Dual attention mechanisms further amplify the model's efficacy, setting new performance benchmarks on multimodal datasets [167]. VISTA-Net blends early and late fusion, autonomously fine-tuning intermediate output weights. The KAAP method is utilized to analyze feature contributions to distinct emotion classes [96]. Cimtay's multi-modal system integrates CNNF with CNNV and CNNA models, training on facial, GSR, and EEG data. These outputs merge to deduce emotional states [168]. Gunes fashioned a visual-centric framework amalgamating facial expressions and body gestures at the feature level before decision-level fusion. Due to dataset scarcity, the researchers curated their own [169]. Hybrid fusion's adaptability aligns with diverse task necessities, enhancing the model's resilience and broad applicability. But, it demands intricate parameter adjustments, uses substantial computational power, and necessitates rigorous validation [170].

While modern deep learning techniques lean towards bottom-up fusion, human cognitive models underscore the value of top-down fusion where thought guides perception. A trailblazing neural design was suggested to capture these top-down interplays using forward-feedback schemes, marking advancements in multimodal sentiment analysis on the CMU-MOSEI dataset [170].

Other Typical Fusion

In addition to the typical fusion strategies previously discussed, there are other influential mainstream methods in multimodal fusion worth noting. These include multimodal fusion methods based on the Transformer and attention mechanisms [171–173], as well as multimodal fusion strategies tailored for multitask learning [174]. Furthermore, strategies that incorporate contrastive learning [175], exchange mechanisms [176], and large-model multimodal fusion [177] also play a significant role in this field. These methods are not only theoretically innovative but have also demonstrated exceptional performance and broad applicability in practical applications.

Multimodal fusion methods based on the Transformer and attention mechanisms focus on dynamically allocating the importance of different modal features using attention mechanisms. This approach thoroughly considers the characteristics and contributions of each modality, achieving efficient information fusion through refined weight distribution [178]. For instance, modality-specific attention mechanisms focus on analyzing the contribution of each modality to the final fusion outcome, while modality fusion attention mechanisms consider the features of all modalities for more comprehensive information fusion [179]. Additionally, spatial and physiological attention mechanisms respectively concentrate on the spatial areas of visual data and related physiological signals, enhancing the model's recognition capabilities in these dimensions [180]. Temporal attention mechanisms focus on the temporal dynamics of modalities, while cross-modal temporal attention captures the interactions between different modalities over time. Context-aware and spatiotemporal context attention further enrich the model's emotion recognition and understanding.

Multitask learning-based multimodal fusion strategies are another efficient method in the field of emotion recognition. By handling multiple related tasks simultaneously (such as emotion recognition, sentiment analysis, or gender identification), they improve the model's performance and generalization ability [181]. This learning setup enables the model to understand and process emotions from multiple perspectives, thus enhancing the accuracy of emotion recognition and the understanding of complex emotional expressions. For example, single-task learning focuses on training models for specific emotion recognition tasks, while multitask learning extends the learning tasks to handle multiple emotion-related tasks simultaneously [182]. The key to this approach is considering the shared features between tasks to achieve joint optimization. In practical applications, models like the M2Seq2Seq proposed by Zhang et al., combining the encoder-decoder structure with attention mechanisms, are typical examples. This structure enables models to more effectively handle context dependencies and interactions between modalities, significantly improving performance.

Contrastive learning-based multimodal fusion strategies learn robust feature representations by comparing the similarities and differences between different modalities [183]. This self-supervised learning method trains models by maximizing the similarity between positive samples and minimizing the similarity between negative samples. In the multimodal domain, this often involves taking data from different modalities as input to learn a joint feature space where similar multimodal samples are brought closer and dissimilar samples are pushed apart. This strategy shows significant advantages in handling the heterogeneity of different modal data. Recent studies have made significant progress in tasks like visual question answering and image captioning by using contrastive learning to synchronize learning visual and textual representations [184].

Exchange-based multimodal fusion strategies, an advanced method, focus on exchanging and fusing information between different modalities to enhance the accuracy and efficiency of emotion recognition. The key to these strategies is designing effective exchange mechanisms that allow emotional features from different modalities, such as text, voice, image, or video, to be interchanged. Learning mechanisms determine which features are shared across modalities and which are modality-specific, thus achieving more efficient information fusion [185]. Additionally, the system needs to process and understand information from different modalities, often involving deep learning, machine learning, or other intelligent algorithms to learn how to transfer information from one modality to another and integrate this information for more accurate emotion recognition [186]. In multimodal emotion recognition, it's also necessary to consider the interactions of modality information in different contexts. For example, the tone and speed of voice might closely relate to facial expressions in visual information, jointly contributing to a more comprehensive understanding of a certain emotional state [187]. As emotions are dynamically changing, such technology needs to process and respond to information changes in different modalities in real-time, requiring the system to be highly flexible and adaptable, able to adjust and optimize information fusion strategies immediately.

In the current trend of technological development, large-model multimodal fusion is becoming a major trend in the field of artificial intelligence. As of 2023, several tech giants and research institutions have launched milestone large-scale models in the field of multimodal fusion [188]. These models have not only achieved significant breakthroughs in theory but also demonstrated outstanding performance in practical applications. Especially in the field of multimodal sentiment analysis, these large models have shown unique advantages and potential. Here are some of the notable large models and their characteristics:

OpenAI's GPT-3: As a pioneer in the field of natural language processing, GPT-3 demonstrates the powerful capabilities of deep learning in language understanding and generation with its massive model size and number of parameters. Google's BERT: This Transformer-based model achieved revolutionary results in natural language processing, especially excelling in text classification and information extraction tasks. Huawei's MindSpore: An all-scenario AI computing framework that supports multimodal data processing, suitable for various deployment scenarios from edge to cloud. Microsoft's Turing-NLG: A large-scale natural language generation model, Turing-NLG demonstrates exceptional performance in various natural language processing tasks. Baidu's ERNIE: By integrating knowledge graphs, ERNIE enhances the language representation model's understanding and expression capabilities in specific domains. Google's Gemini: As a native multimodal model, Gemini can handle various data types including language, auditory, and visual. Its flexibility and powerful processing capabilities make it highly promising in the field of multimodal fusion [189–193]. In the application of multimodal sentiment analysis, these large models significantly enhance the accuracy and efficiency of emotion recognition. For example, in complex sentiment analysis tasks, these models can comprehensively consider the semantics of language, the emotional tone of voice, and subtle changes in facial expressions, thus achieving a deeper level of emotional understanding.

Dataset and Model Performance

Datasets

Multimodal emotion datasets are fundamental resources for emotion recognition and analysis. They provide a way to study and explore how various types of sensory inputs, such as speech, facial expressions, body postures, text, etc., affect human emotions and affect. By using these datasets, more accurate and universal emotion recognition and analysis systems can be developed, which can be widely used not only at the individual level, but also in industries such as social media, online customer service, healthcare, and more. Furthermore, multimodal emotion datasets also provide important resources for emotion computing research, helping to understand the neural mechanisms underlying human emotional information processing and promoting further development of deep learning and AI technologies in emotion recognition and related fields.

The existing classic multimodal emotional dataset is shown in Table 6: IEMOCAP is a leading dataset for emotion recognition, featuring 10 actors and up to 12 emotional states across 12 h of English dialogue. It includes high-quality facial, head, and gesture markers and is commonly used for evaluating multimodal emotion models [194]. MOSI comprises 1200 video clips with features from speech, text, and visual information. The dataset contains movie review videos labeled for sentiment intensity on a scale from -3 to +3 [195]. MELD includes 10,000 dialogues from movies and TV shows, particularly the series “Friends.” Each utterance has an emotion and mood label, and the data is divided into training, validation, and test sets [196]. CMU-MOSEI is the largest dataset for speech-level sentiment and emotion analysis, containing over 65 h of annotated videos from YouTube. It includes annotations from three different annotators for reduced bias [197]. CH-SIMS is a Chinese dataset featuring 2,281 annotated video clips. It allows both multimodal and single-modal emotion recognition studies [198]. Aff-Wild is a large, cross-cultural dataset that includes speech, text, and facial expression data, along with emotional labels [199]. RAVDESS provides 1440 audio clips from 73 actors, covering 8 different emotions. It also includes some visual information [200]. SEMAINE contains conversational data between robots and humans, including facial expressions and speech [201]. HEU Emotion is the most extensive multimodal emotion database to date, featuring 19,004 video clips and 9,951 subjects. It improves emotion recognition accuracies by using traditional machine learning and deep learning methods [202]. MEmoR is a novel dataset designed for multimodal affective reasoning. It includes fine-grained emotion annotations for both speakers and non-speakers and requires machines to reason about human emotions from context. The dataset has shown to outperform relevant baselines in experimental results [203].

Model performance

The accuracy of the current classic multimodal emotion recognition model is shown in Table 7.

Challenges and Applications

Challenges

Dataset Construction

Privacy concerns in multimodal emotion recognition are significant, as the collection of emotion-related data can reveal sensitive personal information. Nandi A and team have tackled this by introducing a federated learning framework,

Table 6 Multimodal emotion datasets

Dataset	Year	Samples	Modality	Website
IEMOCAP	2008	10000	Text, visual, audio	https://sail.usc.edu/iemocap
MOSI	2016	2199	Text, visual, audio	http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset
MELD	2018	13708	Text, visual, audio	https://github.com/SenticNet/MELD
CMU-MOSEI	2018	23453	Text, visual, audio	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset
RAVDESS	2018	7356	Audio	https://zenodo.org/record/1188976
SEMAINE	2012	959	Audio, visual	http://semaine-db.eu
AFEW	2016	330	Audio, visual	https://cs.anu.edu.au/few/AFEW.html
RECOLA	2016	46	Audio, visual, ECG, EDA	https://diuf.unifr.ch/main/diva/recola/
BAUM-1	2017	31	Audio, visual	https://www.payititi.com/opendatasets/show-25963.html
EMOEEG	2017	8	EEG, EOG, EMG, ECG, EDA	https://ieeexplore.ieee.org/abstract/document/8081305
WESAD	2018	15	ECG, EDA, EMG, RESP, TEMP, ACC	https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/

Table 7 Multimodal emotion datasets

Ref.	Year	Fusion type	Dataset	F1 Score
[136]	2018	Early	IEMOCAP	77.62%
[137]	2019	Early	CMU-MOSI	80.19%
[139]	2018	Early	CMU-MOSI	76.50%
[140]	2021	Early	CMU-MOSI	77.10%
[142]	2017	Early	CMU-MOSI	80.30%
[143]	2016	Early	IEMOCAP	79.20%
[91]	2023	Early	IEMOCAP	73.24%
[204]	2022	Early	DEAP	85.34%
[204]	2023	Early	DEAP	96.63%
[132]	2018	Early	IEMOCAP	72%
[146]	2020	Early	CMU-MOSI	83.90%
[205]	2018	Early	CMU-MOSI	77.10%
[206]	2018	Early	CMU-MOSI	77.40%
[149]	2021	Early	CMU-MOSI	81.80%
[152]	2020	Early	IEMOCAP	86.50%
[153]	2019	Late	CMU-MOSI	69.75%
[95]	2022	Late	IEMOCAP	74.68%
[157]	2017	Late	CMU-MOSI	73%
[97]	2022	Late	IEMOCAP	82.92%
[158]	2017	Late	CMU-MOSI	75.10%
[159]	2018	Late	CMU-MOSI	76%
[160]	2021	Late	CMU-MOSI	84.90%
[161]	2021	Late	CMU-MOSI	92.28%
[135]	2018	Early	IEMOCAP	71.80 %
[88]	2023	Model	IEMOCAP	60.59%
[162]	2022	Model	CMU-MOSI	83.20%
[90]	2023	Model	IMEOCAP	70.00%
[87]	2022	Model	IEMOCAP	76.06%
[163]	2023	Model	CMU-MOSI	83.83%
[164]	2018	Model	EEG	72.39%
[166]	2017	Hybrid	CMU-MOSI	76.5%
[167]	2020	Hybrid	CMU-MOSI	80.01%
[96]	2022	Hybrid	IEMOCAP	76.06%
[170]	2022	Hybrid	CMU-MOSI	80.65%
[207]	2023	Other	IMEOCAP	67.5%
[208]	2022	Other	IMEOCAP	65.72%
[209]	2022	Other	CMU-MOSEI	75.9%
[177]	2022	Other	IMEOCAP	72.27%
[178]	2023	Other	CREMA-D	67.2%
[210]	2023	Late	EMO-DB	94.36%
[211]	2023	Late	IEMOCAP	69.36%
[212]	2023	Late	IEMOCAP	81.20%
[213]	2023	Late	IEMOCAP	73.95%
[214]	2023	Late	IEMOCAP	72.84%
[215]	2023	Late	DEAP	91.59%
[216]	2023	Early	IEMOCAP	69.13%
[217]	2023	Model	EMOTIC	81.2%
[218]	2023	Early	CMU-MOSI	83.5%
[219]	2023	Late	CMU-MOSI	67.8%

Fed-ReMECS, for real-time emotion classification using wearables. This framework provides scalability and privacy by ensuring that no sensitive user data leaves the IoT environment [220]. The issue of data scarcity is being addressed by leveraging data augmentation techniques, both traditional and advanced, like Generative Adversarial Networks. These methods have been found effective in improving the accuracy of the A VSR framework when tested on noisy, real-world datasets [221]. Manual annotation remains the gold standard for labeling emotions in datasets. It requires clear guidelines and consistent quality checks[222].

Modality Alignment

Multimodal alignment aims to sync heterogeneous data from text, audio, and video, but faces several challenges including data heterogeneity, modality differences, alignment criteria, non-linear transformations, and scalability [223]. In emotion recognition, an attention-based approach has shown state-of-the-art performance on the IEMOCAP dataset by effectively aligning speech frames and text words. Another model employs hierarchical architecture with attention to provide visual interpretability for utterance-level emotions [220]. Wang et al. introduced Deep Canonical Correlation Autoencoder (DCCAE), a model that balances canonical correlation learning with autoencoder reconstruction to generate highly correlated features across modalities [224]. Similarly, Yu et al. presented a variant of DCCA tailored for scene recognition by correlating field photos and text [225]. Liu et al. extended DCCA for emotion recognition and demonstrated its superiority over traditional methods like Bimodal Deep Autoencoder (BDAE) [226]. Meanwhile, Deshmukh et al. leveraged DCCA and its multiset extension, DMCCA, for biometric systems, highlighting its unsupervised learning capabilities as an advantage over cross-modal similarity-based methods [227]. This version preserves the references and encapsulates the main ideas and contributions from each study.

Model Lightweighting

The challenge of designing efficient multimodal visual tracking methods lies in balancing complexity and performance. While existing multimodal fusion models offer high precision, their computational and storage demands make them impractical for real-time and embedded applications. Several lightweight approaches for multimodal emotion recognition have emerged, such as network pruning, model distillation, and quantized networks, which aim to reduce model size and computational overhead without significantly compromising accuracy. In a notable study [228], a lightweight multimodal deep learning approach is proposed that fuses audiovisual data streams for real-time emotion classification. The system employs four compact neural networks to concurrently analyze visual and auditory features [229]. These features are then integrated into a single data stream, utilizing an exponentially weighted moving average to accumulate evidence over time for the final emotion prediction. Tested on the RAVDESS database, this method outperforms single-modality analyses, achieving an accuracy rate of 90.74%. This presents a promising avenue for real-time emotion analysis systems with hardware or software constraints [230].

Lack of Contextual Information

The challenge of lacking contextual information is significant in multimodal dialogue emotion analysis, where emotion recognition often relies on the context and background information from the communication history. Without adequately mining and leveraging this contextual information, it becomes difficult to make accurate emotion judgments. To address this issue, typical approaches include: Sequential Modeling: Employing models such as Recurrent Neural Networks, Long Short-Term Memory networks, or Gated Recurrent Units to capture the temporal information in dialogues. These models are capable of processing data at different time points and learning the temporal dependencies therein [231]. Attention Mechanisms: Incorporating attention mechanisms into models allows them to focus on parts of the dialogue most relevant to the emotional state. This aids in highlighting critical information during analysis and ignoring irrelevant or redundant content. Contextual Embeddings: Utilizing pre-trained language models like BERT to generate embeddings that represent the dialogue history. These embeddings can capture the deeper semantics and contextual relationships of language [232]. Graph Neural Networks: Using GNNs to model the entity relationships and interaction patterns within dialogues. This helps understand the complex structures and contextual relationships in the conversation. Emotion Lexicons and Phrase Banks: Combining dictionaries or banks of emotional phrases to identify and weight the expression of emotions in dialogue. This enhances the model's understanding of emotional tendencies within the context [233]. Situational Modeling: Building models to capture and simulate the scenarios of dialogue scenes. This helps models to understand the nuances of emotional expression in different situations.

Modality Missing

In multimodal emotion recognition, the issue of modality absence arises when data from certain modalities are unavailable due to various reasons such as device malfunction or signal distortion. Addressing this challenge requires innovative solutions for compensating for missing information. Strategies include cascade models that predict missing data, conventional and label-based interpolation techniques, multitask learning, and the use of adversarial networks to generate missing data for training. Other approaches involve ensemble learning, weight learning to prioritize modalities, feature selection to mitigate the impact of missing data, credibility prediction for better fusion, and data augmentation to generate new training data. A study [234] provides a formal framework for understanding modality absence in terms of flexibility and efficiency. The authors propose a novel approach called SMIL, which leverages Bayesian meta-learning to address both aspects [235]. This method was tested on popular benchmarks like MM-IMDb, CMU-MOSI, and avMNIST, outperforming existing techniques and generative baselines. The findings indicate that SMIL offers a state-of-the-art solution for handling modality absence in multimodal emotion recognition [236].

Applications

Driver Emotion Recognition

As cars become more prevalent in households, driving has evolved into an essential daily activity. However, the complexity of this task raises concerns about traffic safety, which continues to grow annually. Driving involves intricate cognitive processes, and even minor distractions can have severe consequences [101]. Emotions significantly influence cognitive functions, elevating the cognitive load on drivers. Research indicates that negative emotional states while driving increase the likelihood of road accidents, leading to economic, physical, and sometimes fatal losses [98]. In this context, multimodal emotion recognition offers promise for enhancing both driving safety and experience. Applications range from real-time emotional state monitoring through facial expressions, voice, and physiological signals to adjusting the vehicle's settings based on driver emotions [102], [103]. Such technologies can also facilitate more nuanced human–machine interactions, adapting navigation suggestions based on drivers' reactions. Furthermore, they can assist in recognizing the emotional states of other drivers, thereby improving overall road safety. This pool of emotional data serves as a rich resource for optimizing autonomous driving systems and traffic management [237].

Medical Assessment

Improving Emotional Communication: Emotion is pivotal in advancing machine intelligence and human–machine interaction [100]. A benchmark for robot intelligence includes the integration of emotional cognition capabilities. Medical emotional chatbots exemplify this application, leveraging algorithms to monitor health while offering artificial emotional support to vulnerable demographics like the elderly and children with depression [99]. In medical contexts, emotion computing has been instrumental in diagnosing and treating various conditions [107, 108]. It aids not only in mental health care, such as in treating autism and depression, but also supports professionals in delivering psychological counseling during public health crises. Multimodal emotion recognition methods provide more precise emotional analysis [109]. For instance, a deep learning model for real-time emotional state detection using data from the Emotional Internet of Things (E-IoT) has been proposed [238]. Named MEmoR, the model uses visual and psychophysiological data. Video frames are sampled, and a pre-trained ResNet50 model is adapted for emotion classification. Concurrently, a CNN is trained on psychophysiological signals, with decision-level weighted fusion employed to integrate the two data streams. The model's performance, evaluated using the Bio Vid Emo DB dataset, shows promising applications.

Question and Answer System

Multimodal emotion recognition technology enhances a plethora of insurtech products and services, from humanizing interactions in marketing and customer service to emotion-driven quality checks for staff performance. It even aids in product design by gauging public sentiment on specific insurance offerings [104]. In Q and A systems, this technology analyzes facial expressions and vocal tones to gauge user emotions, facilitating more accurate and personalized responses [105]. Key applications include sentiment analysis for understanding user intentions and question-answer matching for optimized response retrieval [106]. For instance, EmoChat is an online chat system that employs real-time emotion recognition, combining facial expressions with text messages [239]. It employs an information entropy-based method for fusing these multimodal cues and a Hidden Markov Model for context-sensitive emotion recognition. Experimental results indicate EmoChat's effectiveness, with an accuracy of 76.25% in emotion polarity recognition and 51.64% in categorizing emotions, all with a latency below 50 milliseconds.

Future Application Scenarios

The future of multimodal emotion recognition promises remarkable breakthroughs in fields like criminal investigation, medical diagnosis, and mental health. Here are some typical future application scenarios, some of which have already begun to be explored:

Interrogation in Criminal Investigations: In the field of criminal investigation, multimodal emotion analysis technology can accurately assess a suspect's genuine emotions and psychological state by analyzing their language, facial expressions, voice, and physiological signals. This not only aids in enhancing the efficiency and accuracy of interrogations but also helps prevent miscarriages of justice due to misinterpretation of a suspect's emotional state. Diagnosis of Pediatric Tic Disorders: The application of multimodal emotion analysis in the medical field, especially in diagnosing pediatric tic disorders, holds significant potential. By comprehensively analyzing various modal data like facial expressions, speech, and body movements of children, a more thorough and precise diagnosis can be made, providing vital information for subsequent treatment. Mental Health Early Warning and Tracking in Schools: In the educational domain, multimodal emotion analysis technology can be utilized to monitor and analyze students' emotional and psychological states, identifying potential mental health issues early on. For instance, by analyzing students' speech, facial expressions, and writing habits, stress, anxiety, or symptoms of depression can be effectively identified for early intervention, thereby ensuring the mental well-being of students. Customer Emotion Analysis in the Business Sector: In the commercial arena, multimodal emotion analysis technology can be used to analyze customer emotional responses, helping businesses better understand customer needs and preferences and optimize product design and marketing strategies. For example, in customer service, analyzing customers' speech and facial expressions can lead to a more accurate understanding of their emotional state and needs, enabling more personalized and efficient service provision. Each of these application scenarios demonstrates the profound impact and potential of multimodal emotion recognition technology in various sectors. As the technology continues to advance and evolve, it is poised to unlock even more extensive and profound applications in the future.

Conclusion

This paper critically examines the pivotal technologies in emotion analysis, with a focus on multi-modal information sources. Initially, we delve into the mechanisms of emotional elicitation, expression, and representation models, establishing a foundational comprehension for the field. Subsequent sections detail the preprocessing and feature extraction techniques for each modality, along with associated tools. We then explore various fusion strategies and datasets, assessing the advantages and limitations of early, late, model, and hybrid fusion methods. This is complemented by comparative analyses of different datasets and model performances. The paper concludes by highlighting key challenges in emotion analysis, including dataset creation, modality alignment, and model optimization, and discusses their implications in practical scenarios like driver emotion recognition and medical assessments.

Key Findings and Future Directions: The study reaffirms that integrating multimodal information significantly enhances the accuracy and robustness of emotion analysis. However, several challenges persist, necessitating future research focus on: Diversifying multimodal emotion datasets to bolster model generalization and robustness. Developing more efficient alignment and fusion techniques to fully harness the potential of multimodal data. Crafting lightweight models suitable for mobile and embedded systems. Applying transfer learning and reinforcement learning strategies to overcome the limitations posed by small sample sizes. Investigating novel approaches to manage missing modalities, especially in real-world application contexts.

Author Contributions Xianxun Zhu drafted and wrote the manuscript. Rui Wang served as the corresponding author, overseeing and coordinating the entire study. Chaopeng Guo created figures and charts for the article. Heyang Feng retrieved and organized relevant literature. Yao Huang typeset and formatted the manuscript. Yichen Feng and Xiangyang Wang assisted with language editing. All authors have reviewed and approved the final version of the manuscript.

Funding This work was supported by the National Natural Science Foundation of China(61771299).

Data Availability The data are available from the corresponding author on reasonable request.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest The authors declare no conflict of interest.

References

1. Foa EB, Kozak MJ. Emotional processing of fear: exposure to corrective information[J]. Psychol Bull. 1986;99(1):20.
2. Ernst H, Scherpf M, Pannasch S, et al. Assessment of the human response to acute mental stress-An overview and a multimodal study[J]. PLoS ONE. 2023;18(11): e0294069.
3. Liu EH, Chambers CR, Moore C. Fifty years of research on leader communication: What we know and where we are going[J]. The Leadership Quarterly. 2023:101734.
4. Russell JA. Core affect and the psychological construction of emotion[J]. Psychol Rev. 2003;110(1):145.
5. Abdullah SMSA, Ameen SYA, Sadeeq MAM, et al. Multimodal emotion recognition using deep learning[J]. J Appl Sci Technol Trends. 2021;2(02):52–8.
6. Marechal C, Mikolajewski D, Tyburek K, et al. Survey on AI-Based Multimodal Methods for Emotion Detection[J]. High-performance modelling and simulation for big data applications. 2019;11400:307–24.
7. Shoumy NJ, Ang LM, Seng KP, et al. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals[J]. J Netw Comput Appl. 2020;149:102447.
8. Zhao S, Yao X, Yang J, et al. Affective image content analysis: Two decades review and new perspectives[J]. IEEE Trans Pattern Anal Mach Intell. 2021;44(10):6729–51.
9. Christian H, Suhartono D, Chowanda A, et al. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging[J]. J Big Data. 2021;8(1):1–20.
10. Das R, Singh T D. Multimodal Sentiment Analysis: A Survey of Methods, Trends and Challenges[J]. ACM Comput Surv. 2023.
11. Zhu L, Zhu Z, Zhang C, et al. Multimodal sentiment analysis based on fusion methods: A survey[J]. Inform Fusion. 2023.
12. Ahmed N, Al Aghbari Z, Girija S. A systematic survey on multimodal emotion recognition using learning algorithms[J]. Intell Syst Appl. 2023;17: 200171.
13. Jabeen S, Li X, Amin MS, et al. A Review on Methods and Applications in Multimodal Deep Learning[J]. ACM Trans Multimed Comput Commun Appl. 2023;19(2s):1–41.
14. Gandhi A, Adhvaryu K, Poria S, et al. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Inform Fusion. 2022.
15. Dimitri GM. A Short Survey on Deep Learning for Multimodal Integration: Applications, Future Perspectives and Challenges[J]. Computers. 2022;11(11):163.
16. Xiaoming Z, Yijiao Y, Shiqing Z. Survey of Deep Learning Based Multimodal Emotion Recognition[J]. J Front Comput Sci Technol. 2022;16(7):1479.
17. Luna-Jimenez C, Kleinlein R, Griol D, et al. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset[J]. Appl Sci. 2021;12(1):327.
18. Chandrasekaran G, Nguyen TN, Hemanth DJ. Multimodal sentimental analysis for social media applications: A comprehensive review[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2021;11(5): e1415.
19. Zhao S, Jia G, Yang J, et al. Emotion recognition from multiple modalities: Fundamentals and methodologies[J]. IEEE Signal Process Mag. 2021;38(6):59–73.
20. Abdu SA, Yousef AH, Salem A. Multimodal video sentiment analysis using deep learning approaches, a survey[J]. Inform Fusion. 2021;76:204–26.
21. Sharma G, Dhall A. A survey on automatic multimodal emotion recognition in the wild[J]. Advances in data science: Methodol Appl. 2021:35-64.
22. Nandi A, Xhafa F, Subirats L, et al. A survey on multimodal data stream mining for e-learner's emotion recognition[C]. In: 2020 International Conference on Omni-layer Intelligent Systems (COINS). IEEE; 2020. p. 1–6.
23. Zhang J, Yin Z, Chen P, et al. Emotion recognition using multimodal data and machine learning techniques: A tutorial and review[J]. Inform Fusion. 2020;59:103–26.

24. Seng JKP, Ang KLM. Multimodal emotion and sentiment modeling from unstructured Big data: Challenges, architecture, and techniques[J]. IEEE Access. 2019;7:90982–98.
25. Baltru?aitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy[J]. IEEE Trans Pattern Anal Mach Intell. 2018;41(2):423–43.
26. Poria S, Cambria E, Bajpai R, et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Inform Fusion. 2017;37:98–125.
27. Latha CP, Priya M. A review on deep learning algorithms for speech and facial emotion recognition[J]. APTIKOM J Comput Sci Inf Technol. 2016;1(3):92–108.
28. Schuller B, Valstar M, Eyben F, et al. Avec 2011-the first international audio/visual emotion challenge[C]. Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part II. Springer Berlin Heidelberg, 2011:415-424.
29. Schuller B, Valstar M, Eyben F, McKeown G, Cowie R, Pantic M. Avec 2011-the first international audio/visual emotion challenge. In Affective Computing and Intelligent Interaction, 2011, p. 415-424. Springer Berlin Heidelberg.
30. Chen H, Zhou H, Du J, et al. The first multimodal information based speech processing challenge:Data, tasks, baselines and results. In Processing ICASSP. 2022, p. 9266-9270. IEEE.
31. Zafeiriou S, Kollias D, Nicolaou M A, et al. Aff-wild: valence and arousal'In-the-Wild'challenge[C]. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017:34-41.
32. Baveye Y, Dellandrea E, Chamaret C, et al. LIRIS-ACCEDE: A video database for affective content analysis[J]. IEEE Trans Affect Comput. 2015;6(1):43–55.
33. Stappen L, Baird A, Rizos G, et al. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild[C]. Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop. 2020:35-44.
34. Li Y, Tao J, Schuller B, et al. Mec 2017: Multimodal emotion recognition challenge[C]. 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). IEEE, 2018:1-5.
35. Kollias D. Abaw: valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:2328-2336.
36. Lian Z, Sun H, Sun L, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning[C]. In: Proceedings of the 31st ACM International Conference on Multimedia. 2023:9610-9614.
37. Li J, Zhang Z, Lang J, et al. Hybrid multimodal feature extraction, mining and fusion for sentiment analysis[C]. In: Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge. 2022:81-88.
38. Zong D, Ding C, Li B, et al. Building robust multimodal sentiment recognition via a simple yet effective multimodal transformer[C]. In: Proceedings of the 31st ACM International Conference on Multimedia. 2023:9596-9600.
39. Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference[M]. Mit Press, 1998.
40. Amsaleg L, Huet B, Larson M, et al. Proceedings of the 27th ACM International Conference on Multimedia[C]. 27th ACM International Conference on Multimedia. ACM Press, 2019.
41. Lomonaco V, Pellegrini L, Rodriguez P, et al. Cvpr 2020 continual learning in computer vision competition: Approaches, results, current challenges and future directions[J]. Artif Intell. 2022;303: 103635.
42. Gatterbauer W, Kumar A. Guest Editors' Introduction to the Special Section on the 33rd International Conference on Data Engineering (ICDE 2017)[J]. IEEE Trans Knowl Data Eng. 2019;31(7):1222-1223.
43. Liu Y, Paek T, Patwardhan M. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations[C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2018.
44. Lang J. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)[J]. 2018.
45. Reddy C K A, Dubey H, Gopal V, et al. ICASSP 2021 deep noise suppression challenge[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021:6623-6627.
46. Morency L P, Bohus D, Aghajan H, et al. ICMI'12: Proceedings of the ACM SIGCHI 14th International Conference on Multimodal Interaction[C]. 14th International Conference on Multimodal Interaction, ICMI 2012. Association for Computing Machinery (ACM), 2012.
47. Nitta N, Hu A, Tobitani K. MMArt-ACM 2022: 5th Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia[C]. Proceedings of the International Conference on Multimedia Retrieval. 2022;2022:692–3.
48. PRICAI 2022: Trends in Artificial Intelligence: 19th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2022, Shanghai, China, November 10-13, 2022, Proceedings, Part III[M]. Springer Nature, 2022.
49. Gabbouj M. Proceedings of WIAMIS 2001: Workshop on Image Analysis for Multimedia Services[J]. 2001.
50. Strike PC, Steptoe A. Behavioral and emotional triggers of acute coronary syndromes: a systematic review and critique[J]. Psychosom Med. 2005;67(2):179–86.
51. Hubert W, de Jong-Meyer R. Autonomic, neuroendocrine, and subjective responses to emotion-inducing film stimuli[J]. Int J Psychophysiol. 1991;11(2):131–40.
52. Bhattacharyya MR, Steptoe A. Emotional triggers of acute coronary syndromes: strength of evidence, biological processes, and clinical implications[J]. Prog Cardiovasc Dis. 2007;49(5): 353–65.
53. Scopa C, Contalbrigo L, Greco A, et al. Emotional transfer in human-horse interaction: New perspectives on equine assisted interventions[J]. Animals. 2019;9(12):1030.
54. Hong JK, Gao L, Singh J, et al. Evaluating medical device and material thrombosis under flow: current and emerging technologies[J]. Biomater Sci. 2020;8(21):5824–45.
55. Werheid K, Alpay G, Jentzsch I, et al. Priming emotional facial expressions as evidenced by event-related brain potentials[J]. Int J Psychophysiol. 2005;55(2):209–19.
56. Matsumoto D, Ekman P. The relationship among expressions, labels, and descriptions of contempt[J]. J Pers Soc Psychol. 2004;87(4):529.
57. Picard R W. Affective computing[M]. MIT press, 2000.
58. Tomkins S S. Affect imagery consciousness: the complete edition: two volumes[M]. Springer publishing company, 2008.
59. Mehrabian A. Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression[J]. J Psychopathol Behav Assess. 1997;19:331–57.
60. Russell JA. Core affect and the psychological construction of emotion[J]. Psychol Rev. 2003;110(1):145.
61. Posner J, Russell JA, Peterson BS. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology[J]. Dev Psychopathol. 2005;17(3):715–34.
62. Bleicher RJ, Ciocca RM, Egleston BL, et al. Association of routine pretreatment magnetic resonance imaging with time to

surgery, mastectomy rate, and margin status[J]. J Am Coll Surg. 2009;209(2):180–7.

63. Swathi C, Anoop B K, Dhas D A S, et al. Comparison of different image preprocessing methods used for retinal fundus images[C]. 2017 Conference on Emerging Devices and Smart Systems (ICEDSS). IEEE, 2017:175-179.
64. Finlayson G D, Schiele B, Crowley J L. Comprehensive colour image normalization[C]. Computer Vision-ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June, 2-6, 1998 Proceedings, Volume I 5. Springer Berlin Heidelberg, 1998:475-490.
65. Vishwakarma AK, Mishra A. Color image enhancement techniques: a critical review[J]. Indian J Comput Sci Eng. 2012;3(1):39–45.
66. Celik T. Two-dimensional histogram equalization and contrast enhancement[J]. Pattern Recogn. 2012;45(10):3810–24.
67. Jayaram S, Schmugge S, Shin M C, et al. Effect of colorspace transformation, the illuminance component, and color modeling on skin detection[C]. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. IEEE, 2004, 2:II-II.
68. Pandey M, Bhatia M, Bansal A, An anatomization of noise removal techniques on medical images[C]. international conference on innovation and challenges in cyber security (iciccs-inbush). IEEE. 2016;2016:224–9.
69. Maini R, Aggarwal H. Study and comparison of various image edge detection techniques[J]. Int J Image Process (IJIP). 2009;3(1):1–11.
70. Eltanany AS, SAfy Elwan M, Amein AS. Key point detection techniques[C]. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019. Springer International Publishing. 2020:901-911.
71. Yang MH, Kriegman DJ, Ahuja N. Detecting faces in images: a survey[J]. IEEE Trans Pattern Anal Mach Intell. 2002;24(1): 34–58.
72. Qin J, He ZS. ASVM, face recognition method based on Gabor-featured key points[C]. international conference on machine learning and cybernetics. IEEE. 2005;2005(8):5144–9.
73. Xiong X, De la Torre F. Supervised descent method and its applications to face alignment[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2013:532-539.
74. Kalyuga S, Chandler P, Sweller J. Incorporating learner experience into the design of multimedia instruction[J]. J Educ Psychol. 2000;92(1):126.
75. Bezoui M, Elmoutaouakkil A, Beni-hssane A. Feature extraction of some Quranic recitation using mel-frequency cepstral coeficients (MFCC)[C]. 5th international conference on multimedia computing and systems (ICMCS). IEEE. 2016;2016:127–31.
76. Shrawankar U, Thakare V M. Adverse conditions and ASR techniques for robust speech user interface[J]. arXiv preprint arXiv:1303.5515, 2013.
77. Liu L, He J, Palm G. Signal modeling for speaker identification. In: Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (vol. 2). IEEE; 1996. pp. 665–8.
78. Bozkurt B, Couvreur L, Dutoit T. Chirp group delay analysis of speech signals[J]. Speech Commun. 2007;49(3):159–76.
79. Seman N, Bakar ZA, Bakar NA. An evaluation of endpoint detection measures for Malay speech recognition of an isolated words[C]. International Symposium on Information Technology, IEEE. 2010;2010(3):1628–35.
80. Hua Y, Guo J, Zhao H. Deep belief networks and deep learning[C]. Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things, IEEE. 2015:1-4.
81. Owren MJ. GSU Praat Tools: scripts for modifying and analyzing sounds using Praat acoustics software[J]. Behav Res Methods. 2008;40(3):822–9.
82. Eyben F, Wllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor[C]. Proceedings of the 18th ACM international conference on Multimedia. 2010:1459-1462.
83. Hossan M A, Memon S, Gregory M A. A novel approach for MFCC feature extraction[C]. In: 2010 4th International Conference on Signal Processing and Communication Systems. IEEE, 2010:1-5.
84. Acheampong F A, Nunoo-Mensah H, Chen W. Transformer models for text-based emotion detection: a review of BERT-based approaches[J]. Artificial Intelligence Review, 2021:1-41.
85. Mishra B, Fernandes SL, Abhishek K, et al. Facial expression recognition using feature based techniques and model based techniques: a survey[C]. In: 2nd international conference on electronics and communication systems (ICECS), IEEE. 2015;2015:589–94.
86. Mastropaolo A, Scalabrino S, Cooper N, et al. Studying the usage of text-to-text transfer transformer to support code-related tasks[C]. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). IEEE, 2021:336-347.
87. Qian F, Han J. Contrastive regularization for multimodal emotion recognition using audio and text[J]. arXiv preprint arXiv:2211.10885, 2022.
88. Zhang Y, Wang J, Liu Y, et al. A Multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations[J]. Inform Fusion. 2023.
89. Fuente C, Castellanos FJ, Valero-Mas JJ, et al. Multimodal recognition of frustration during game-play with deep neural networks[J]. Multimed Tools Appl. 2023;82(9):13617–36.
90. Li J, Wang X, Lv G, et al. GA2MIF: graph and attention based two-stage multi-source Information Fusion for Conversational Emotion Detection[J]. IEEE Trans Affect Comput. 2023.
91. Wang B, Dong G, Zhao Y, et al. Hierarchically stacked graph convolution for emotion recognition in conversation[J]. Knowledge-Based Systems, 2023:110285.
92. Padi S, Sadjadi S O, Manocha D, et al. Multimodal emotion recognition using transfer learning from speaker recognition and Bert-based models[J]. arXiv preprint arXiv:2202.08974, 2022.
93. Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]. In: Proceedings of the IEEE international conference on computer vision. 2015:4489-4497.
94. Bansal K, Agarwal H, Joshi A, et al. Shapes of emotions: multimodal emotion recognition in conversations via emotion shifts[C]. In: Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models. 2022:44-56.
95. Tang S, Luo Z, Nan G, et al. Fusion with hierarchical graphs for multimodal emotion recognition[C]. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE. 2022;2022:1288–96.
96. Qian F, Han J. Contrastive regularization for multimodal emotion recognition using audio and text[J]. arXiv preprint arXiv:2211.10885, 2022.
97. Wei Q, Huang X, Zhang Y. FV2ES: a fully end2end multimodal system for fast yet effective video emotion recognition inference[J]. IEEE Transactions on Broadcasting, 2022.
98. Wu Y, Li J. Multi-modal emotion identification fusing facial expression and EEG[J]. Multimed Tools Appl. 2023;82(7):10901–19.

99. Reid MJ, Omlin X, Espie CA, et al. The effect of sleep continuity disruption on multimodal emotion processing and regulation: a laboratory based, randomised, controlled experiment in good sleepers[J]. J Sleep Res. 2023;32(1): e13634.
100. Fang M, Peng S, Liang Y, et al. A multimodal fusion model with multi-level attention mechanism for depression detection[J]. Biomed Signal Process Control. 2023;82: 104561.
101. Stappen L, Baird A, Rizos G, et al. Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: emotional car reviews in-the-wild[C]. In: Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop. 2020:35-44.
102. Miranda J A, Canabal M F, Portela Garca M, et al. Embedded emotion recognition: autonomous multimodal affective internet of things[C]. In: Proceedings of the cyber-physical systems workshop. 2018, 2208:22-29.
103. Caesar H, Bankiti V, Lang A H, et al. nuscenes: a multimodal dataset for autonomous driving[C]. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020:11621-11631.
104. Mangano G, Ferrari A, Rafele C, et al. Willingness of sharing facial data for emotion recognition: a case study in the insurance market[J]. AI & SOCIETY. 2023:1-12..
105. Boyd KL, Andalibi N. Automated emotion recognition in the workplace: How proposed technologies reveal potential futures of work[J]. Proceedings of the ACM on Human-Computer Interaction. 2023;7(CSCW1):1–37.
106. Dubey A, Shingala B, Panara JR, et al. Digital content recommendation system through facial emotion recognition[J]. Int J Res Appl Sci Eng Technol. 2023;11:1272–6.
107. Holding B C, Laukka P, Fischer H, et al. Multimodal emotion recognition is resilient to insufficient sleep: results from cross-sectional and experimental studies[J]. Sleep. 2017;40(11):zsx145.
108. Egger M, Ley M, Hanke S. Emotion recognition from physiological signal analysis: a review[J]. Electron Notes Theor Comput Sci. 2019;343:35–55.
109. Andrews SC, Staios M, Howe J, et al. Multimodal emotion processing deficits are present in amyotrophic lateral sclerosis[J]. Neuropsychology. 2017;31(3):304.
110. O'Shea K, Nash R. An introduction to convolutional neural networks[J]. arXiv preprint arXiv:1511.08458, 2015.
111. Meignier S, Merlin T. LIUM SpkDiarization: an open source toolkit for diarization[C]. CMU SPUD Workshop. 2010.
112. Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]. IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011 (CONF).
113. Gaida C, Lange P, Petrick R, et al. Comparing open-source speech recognition toolkits[C]. 11th International Workshop on Natural Language Processing and Cognitive Science. 2014.
114. Moffat D, Ronan D, Reiss J D. An evaluation of audio feature extraction toolboxes[J]. 2015.
115. Karkada D, Saletore VA. Training speech recognition models on HPC infrastructure[C]. IEEE/ACM Machine Learning in HPC Environments (MLHPC), IEEE. 2018;2018:124–32.
116. Syed M S S, Stolar M, Pirogova E, et al. Speech acoustic features characterising individuals with high and low public trust[C]. 2019 13th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, 2019:1-9.
117. Degottex G, Kane J, Drugman T, et al. COVAREP-a collaborative voice analysis repository for speech technologies[C]. In: IEEE international conference on acoustics, speech and signal processing (icassp), IEEE. 2014;2014:960–4.
118. Yadav U, Sharma AK, Patil D. Review of automated depression detection: social posts, audio and video, open challenges and future direction[J]. Concurrency and Computation: Practice and Experience. 2023;35(1): e7407.
119. Vijayarani S, Ilamathi MJ, Nithya M. Preprocessing techniques for text mining-an overview[J]. International Journal of Computer Science and Communication Networks. 2015;5(1):7–16.
120. Thelwall M, Buckley K, Paltoglou G, et al. Sentiment strength detection in short informal text[J]. J Am Soc Inform Sci Technol. 2010;61(12):2544–58.
121. Wu Z, King S. Investigating gated recurrent networks for speech synthesis[C]. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016:5140-5144.
122. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery[J]. NPJ Digital Medicine. 2021;4(1):93.
123. Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
124. Zahidi Y, El Younoussi Y, Al-Amrani Y. Different valuable tools for Arabic sentiment analysis: a comparative evaluation[J]. International Journal of Electrical and Computer Engineering (2088-8708), 2021, 11(1).
125. Cai H, Lin Q, Liu H, et al. Recognition of human mood, alertness and comfort under the influence of indoor lighting using physiological features[J]. Biomed Signal Process Control. 2024;89: 105661.
126. Tan E, Hamlin JK. Toddlers' affective responses to sociomoral scenes: Insights from physiological measures[J]. J Exp Child Psychol. 2024;237: 105757.
127. Awada M, Becerik Gerber B, Lucas GM, et al. Stress appraisal in the workplace and its associations with productivity and mood: Insights from a multimodal machine learning analysis[J]. PLoS ONE. 2024;19(1): e0296468.
128. Guo W, Li Y, Liu M, et al. Functional connectivity-enhanced feature-grouped attention network for cross-subject EEG emotion recognition[J]. Knowl-Based Syst. 2024;283: 111199.
129. Naeini EK, Sarhaddi F, Azimi I, et al. A deep learning-based PPG quality assessment approach for heart rate and heart rate variability[J]. ACM Transactions on Computing for Healthcare. 2023;4(4):1–22.
130. Panjaitan F, Nurmaini S, Partan RU. Accurate prediction of sudden cardiac death based on heart rate variability analysis using convolutional neural network[J]. Medicina. 2023;59(8):1394.
131. Nashiro K, Yoo HJ, Cho C, et al. Effects of a randomised trial of 5-week heart rate variability biofeedback intervention on cognitive function: possible benefits for inhibitory control[J]. Appl Psychophysiol Biofeedback. 2023;48(1):35–48.
132. Qi N, Piao Y, Yu P, et al. Predicting epileptic seizures based on EEG signals using spatial depth features of a 3D-2D hybrid CNN[J]. Medical & Biological Engineering & Computing, 2023:1-12.
133. Cho D, Lee B. Automatic sleep-stage classification based on residual unit and attention networks using directed transfer function of electroencephalogram signals[J]. Biomed Signal Process Control. 2024;88: 105679.
134. Li Z, Xu B, Zhu C, et al. CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection[J]. arXiv preprint arXiv:2204.05515, 2022.
135. Yoon S, Byun S, Jung K, Multimodal speech emotion recognition using audio and text[C]. In,. IEEE Spoken Language Technology Workshop (SLT). IEEE. 2018;2018:112–8.
136. Hazarika D, Poria S, Zadeh A, et al. Conversational memory network for emotion recognition in dyadic dialogue videos[C]. In: Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting. NIH Public Access, 2018, 2018:2122.

137. Mai S, Hu H, Xing S. Divide, conquer and combine: hierarchical feature fusion network with local and global perspectives for multimodal affective computing[C]. In: Proceedings of the 57th annual meeting of the association for computational linguistics. 2019:481-492.
138. You Q, Luo J, Jin H, et al. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia[C]. In: Proceedings of the Ninth ACM international conference on Web search and data mining. 2016:13-22.
139. Chen M, Wang S, Liang P P, et al. Multimodal sentiment analysis with word-level fusion and reinforcement learning[C]. In: Proceedings of the 19th ACM international conference on multimodal interaction. 2017:163-171.
140. Zadeh A, Chen M, Poria S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250, 2017.
141. Zhang Y, Yu Y, Wang M, et al. Self-adaptive representation learning model for multi-modal sentiment and sarcasm joint analysis[J]. Communications and Applications: ACM Transactions on Multimedia Computing; 2023.
142. Poria S, Cambria E, Hazarika D, et al. Context-dependent sentiment analysis in user-generated videos[C]. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers). 2017:873-883.
143. Poria S, Chaturvedi I, Cambria E, et al. Convolutional MKL, based multimodal emotion recognition and sentiment analysis[C]. In: IEEE 16th international conference on data mining (ICDM), IEEE. 2016;2016:439–48.
144. Deng D, Zhou Y, Pi J, et al. Multimodal utterance-level affect analysis using visual, audio and text features[J]. arXiv preprint arXiv:1805.00625, 2018.
145. Chen F, Luo Z, Xu Y, et al. Complementary fusion of multi-features and multi-modalities in sentiment analysis[J]. arXiv preprint arXiv:1904.08138, 2019.
146. Kumar A, Vepa J. Gated mechanism for attention based multi modal sentiment analysis[C]. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020:4477-4481.
147. Xu N, Mao W. Multisentinet: a deep semantic network for multimodal sentiment analysis[C]. In: Proceedings of the. ACM on Conference on Information and Knowledge Management. 2017;2017:2399–402.
148. Yu J, Jiang J, Xia R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;28:429–39.
149. Mai S, Xing S, Hu H. Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021;29:1424–37.
150. Xu N, Mao W, Chen G. Multi-interactive memory network for aspect based multimodal sentiment analysis[C]. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01):371-378.
151. Liu D, Chen L, Wang Z, et al. Speech expression multimodal emotion recognition based on deep belief network[J]. Journal of Grid Computing. 2021;19(2):22.
152. Wang F, Tian S, Yu L, et al. TEDT: transformer-based encoding-decoding translation network for multimodal sentiment analysis[J]. Cogn Comput. 2023;15(1):289–303.
153. Kumar A, Vepa J. Gated mechanism for attention based multi modal sentiment analysis[C]. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020:4477-4481.
154. Lu Y, Zheng W, Li B, et al. Combining eye movements and EEG to enhance emotion recognition. In: Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2015:1170-1176.
155. Yu Y, Lin H, Meng J, et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms. 2016;9(2):41.
156. Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015:2539-2544.
157. Wang HH, Meghawat A, Morency LP, et al. Select-additive learning: improving generalization in multimodal sentiment analysis. In: Proceedings of the 2017 IEEE International Conference on Multimedia and Expo, IEEE Computer Society, 2017:949-954.
158. Yu HL, Gui LK, Madaio M, et al. Temporally selective attention model for social and affective state recognition in multimedia content. In: Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017:1743-1751.
159. Williams J, Comanescu R, Radu O, et al. DNN multimodal fusion techniques for predicting video sentiment. In: Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), 2018:64-72.
160. Gkoumas, D., Li, Q., Dehdashti, S., et al. Quantum cognitively motivated decision fusion for video sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(1):827-835.
161. Sun, J., Yin, H., Tian, Y., et al. Two-level multimodal fusion for sentiment analysis in public security. Security and Communication Networks, 2021.
162. Zhang F, Li XC, Lim CP, et al. Deep emotional arousal network for multimodal sentiment analysis and emotion recognition[J]. Inform Fusion. 2022;88:296–304.
163. Wang D, Guo X, Tian Y, et al. TETFN: a text enhanced transformer fusion network for multimodal sentiment analysis[J]. Pattern Recogn. 2023;136: 109259.
164. Zheng W, Liu W, Lu Y, et al. Emotionmeter: a multimodal framework for recognizing human emotions. IEEE Transactions on Cybernetics. 2018;49(3):1110–22.
165. Zhang S, Zhang S, Huang T, et al. Learning affective features with a hybrid deep model for audio-visual emotion recognition. IEEE Trans Circuits Syst Video Technol. 2017;28(10):1–1.
166. Chen M, Wang S, Liang P P, et al. Multimodal sentiment analysis with word-level fusion and reinforcement learning[C]. In: Proceedings of the 19th ACM international conference on multimodal interaction. 2017:163-171.
167. Shenoy A, Sardana A. Multilogue-net: a context aware RNN for multi-modal emotion detection and sentiment analysis in conversation[J]. arXiv preprint arXiv:2002.08267, 2020.
168. Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S. Cross-subject multimodal emotion recognition based on hybrid fusion[J]. IEEE Access. 2020;8:168865–78.
169. Gunes H, Piccardi M. Bi-modal emotion recognition from expressive face and body gestures[J]. J Netw Comput Appl. 2007;30(4):1334–45.
170. Paraskevopoulos G, Georgiou E, Potamianos A. Mmlatch: bottom-up top-down fusion for multimodal sentiment analysis[C]. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022:4573-4577.
171. Qu L, Liu S, Wang M, et al. Trans2Fuse: empowering image fusion through self-supervised learning and multi-modal transformations via transformer networks[J]. Expert Syst Appl. 2024;236: 121363.
172. Fan H, Zhang X, Xu Y, et al. Transformer-based multimodal feature enhancement networks for multimodal depression detection

integrating video, audio and remote photoplethysmograph signals[J]. Inform Fusion. 2024;104: 102161.
173. Zhu X, Huang Y, Wang X, et al. Emotion recognition based on brain-like multimodal hierarchical perception[J]. Multimed Tools Appl. 2023:1-19.
174. Huang J, Pu Y, Zhou D, et al. Dynamic hypergraph convolutional network for multimodal sentiment analysis[J]. Neurocomputing. 2024;565: 126992.
175. Wang X, Guan Z, Qian W, et al. CS2Fusion: contrastive learning for self-supervised infrared and visible image fusion by estimating feature compensation map[J]. Inform Fusion. 2024;102: 102039.
176. Han Y, Nie R, Cao J, et al. IE-CFRN: information exchange-based collaborative feature representation network for multimodal medical image fusion[J]. Biomed Signal Process Control. 2023;86: 105301.
177. Ni J, Bai Y, Zhang W, et al. Deep equilibrium multimodal fusion[J]. arXiv preprint arXiv:2306.16645, 2023.
178. Li H, Zhao J, Li J, et al. Feature dynamic alignment and refinement for infrared-visible image fusion: translation robust fusion[J]. Inform Fusion. 2023;95:26–41.
179. Liu J, Capurro D, Nguyen A, et al. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities[J]. J Biomed Inform. 2023;145: 104466.
180. Zhang X, Wei X, Zhou Z, et al. Dynamic alignment and fusion of multimodal physiological patterns for stress recognition[J]. IEEE Trans Affect Comput. 2023
181. Zhang Y, Wang J, Liu Y, et al. A multitask learning model for multimodal sarcasm, sentiment and emotion recognition in conversations[J]. Inform Fusion. 2023;93:282–301.
182. Liu Y, Zhang X, Kauttonen J, et al. Uncertain facial expression recognition via multi-task assisted correction[J]. IEEE Trans Multimed. 2023.
183. Liu J, Lin R, Wu G, et al. Coconet: coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion[J]. Int J Comput Vis. 2023:1-28.
184. Liu K, Xue F, Guo D, et al. Multimodal graph contrastive learning for multimedia-based recommendation[J]. IEEE Trans Multimed. 2023.
185. Song J, Chen H, Li C, et al. MIFM: multimodal information fusion model for educational exercises[J]. Electronics. 2023;12(18):3909.
186. Zhang S, Yang Y, Chen C, et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: a systematic review of recent advancements and future prospects[J]. Expert Syst Appl. 2023:121692.
187. Dogan G, Akbulut FP. Multi-modal fusion learning through biosignal, audio, and visual content for detection of mental stress[J]. Neural Comput Appl. 2023;35(34):24435–54.
188. Liu W, Zuo Y. Stone needle: a general multimodal large-scale model framework towards healthcare[J]. arXiv preprint arXiv:2306.16034, 2023.
189. Zhao X, Li M, Weber C, et al. Chat with the environment: interactive multimodal perception using large language models[J]. arXiv preprint arXiv:2303.08268, 2023.
190. Kim K, Park S. AOBERT: all-modalities-in-one BERT for multimodal sentiment analysis[J]. Inform Fusion. 2023;92:37–45.
191. Tong Z, Du N, Song X, et al. Study on mindspore deep learning framework[C]. In: 2021 17th International Conference on Computational Intelligence and Security (CIS). IEEE, 2021:183-186.
192. Rasley J, Rajbhandari S, Ruwase O, et al. Deepspeed: system optimizations enable training deep learning models with over 100 billion parameters[C]. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020:3505-3506.
193. Huang J, Wang H, Sun Y, et al. ERNIE-GeoL: a geography-and-language pre-trained model and its applications in Baidu maps[C]. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022:3029-3039.
194. Busso C, Bulut M, Lee CC, et al. IEMOCAP: interactive emotional dyadic motion capture database[J]. Lang Resour Eval. 2008;42:335–59.
195. Zadeh A, Zellers R, Pincus E, et al. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[J]. arXiv preprint arXiv:1606.06259, 2016.
196. Poria S, Hazarika D, Majumder N, et al. Meld: a multimodal multi-party dataset for emotion recognition in conversations[J]. arXiv preprint arXiv:1810.02508, 2018.
197. Zadeh A A B, Liang P P, Poria S, et al. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph[C]. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018:2236-2246.
198. Yu W, Xu H, Meng F, et al. Ch-sims: a Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]. In: Proceedings of the 58th annual meeting of the association for computational linguistics. 2020:3718-3727.
199. Zafeiriou S, Kollias D, Nicolaou M A, et al. Aff-wild: valence and arousal'In-the-Wild'challenge[C]. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017:34-41.
200. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English[J]. PLoS ONE. 2018;13(5): e0196391.
201. McKeown G, Valstar M, Cowie R, et al. The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent[J]. IEEE Trans Affect Comput. 2011;3(1):5–17.
202. Chen J, Wang C, Wang K, et al. HEU Emotion: a large-scale database for multimodal emotion recognition in the wild[J]. Neural Comput Appl. 2021;33:8669–85.
203. Shen G, Wang X, Duan X, et al. Memor: a dataset for multimodal emotion reasoning in videos[C]. In: Proceedings of the 28th ACM International Conference on Multimedia. 2020:493-502.
204. Wu X, Zheng WL, Li Z, et al. Investigating EEG-based functional connectivity patterns for multimodal emotion recognition[J]. J Neural Eng. 2022;19(1): 016012.
205. Zadeh A, Liang P P, Poria S, et al. Multi-attention recurrent network for human communication comprehension[C]. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
206. Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning[C]. In: Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).
207. Liu S, Gao P, Li Y, et al. Multi-modal fusion network with complementarity and importance for emotion recognition[J]. Inf Sci. 2023;619:679–94.
208. Chen F, Shao J, Zhu S, et al. Multivariate, multi-frequency and multimodal: rethinking graph neural networks for emotion recognition in conversation[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:10761-10770.
209. Khan M, Gueaieb W, El Saddik A, et al. MSER: multimodal speech emotion recognition using cross-attention with deep fusion[J]. Expert Syst Appl. 2023:122946.
210. Pan J, Fang W, Zhang Z, et al. Multimodal emotion recognition based on facial expressions, speech, and EEG[J]. IEEE Open Journal of Engineering in Medicine and Biology, 2023.
211. Meng T, Shou Y, Ai W, et al. Deep imbalanced learning for multimodal emotion recognition in conversations[J]. arXiv preprint arXiv:2312.06337, 2023.

212. Fu Z, Liu F, Xu Q, et al. LMR-CBT: learning modality-fused representations with CB-transformer for multimodal emotion recognition from unaligned multimodal sequences[J]. Front Comp Sci. 2024;18(4): 184314.
213. Ma H, Wang J, Lin H, et al. A transformer-based model with self-distillation for multimodal emotion recognition in conversations[J]. IEEE Trans Multimed. 2023.
214. Shi T, Huang S L. MultiEMO: an attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations[C]. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023:14752-14766.
215. Li X. TACOformer: token-channel compounded cross attention for multimodal emotion recognition[J]. arXiv preprint arXiv:2306.13592, 2023.
216. Li J, Wang X, Lv G, et al. Graphcfc: a directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition[J]. IEEE Trans Multimed. 2023.
217. Palash M, Bhargava B. EMERSK–explainable multimodal emotion recognition with situational knowledge[J]. arXiv preprint arXiv:2306.08657, 2023.
218. Li Y, Wang Y, Cui Z. Decoupled multimodal distilling for emotion recognition[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:6631-6640.
219. Le HD, Lee GS, Kim SH, et al. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning[J]. IEEE Access. 2023;11:14742–51.
220. Tang J, Ma Z, Gan K, et al. Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment[J]. Inform Fusion. 2024;103: 102129.
221. He Y, Seng KP, Ang LM. multimodal sensor-input architecture with deep learning for audio-visual speech recognition in wild[J]. Sensors. 2023;23(4):1834.
222. Stappen L, Schumann L, Sertolli B, et al. Muse-toolbox: the multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox[M]. In: Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge. 2021:75-82.
223. Tang J, Ma Z, Gan K, et al. Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment[J]. Inform Fusion. 2024;103: 102129.
224. Wang W, Arora R, Livescu K, et al. On deep multi-view representation learning[C]. In: International conference on machine learning. PMLR, 2015:1083-1092.
225. Yu Y, Tang S, Aizawa K, et al. Category-based deep CCA for fine-grained venue discovery from multimodal data[J]. IEEE transactions on neural networks and learning systems. 2018;30(4):1250–8.
226. Liu W, Qiu JL, Zheng WL, et al. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition[J]. IEEE Transactions on Cognitive and Developmental Systems. 2021;14(2):715–29.
227. Deshmukh S, Abhyankar A, Kelkar S. DCCA and DMCCA framework for multimodal biometric system[J]. Multimed Tools Appl. 2022;81(17):24477–91.
228. Cevher D, Zepf S, Klinger R. Towards multimodal emotion recognition in German speech events in cars using transfer learning[J]. arXiv preprint arXiv:1909.02764, 2019.
229. Xi D, Zhou J, Xu W, et al. Discrete emotion synchronicity and video engagement on social media: a moment-to-moment analysis[J]. Int J Electron Commerce. 2024:1-37.
230. Lv Y, Liu Z, Li G. Context-aware interaction network for RGB-T semantic segmentation[J]. IEEE Trans Multimed. 2024.
231. Ai W, Zhang F C, Meng T, et al. A two-stage multimodal emotion recognition model based on graph contrastive learning[J]. arXiv preprint arXiv:2401.01495, 2024.
232. Wan Y, Chen Y, Lin J, et al. A knowledge-augmented heterogeneous graph convolutional network for aspect-level multimodal sentiment analysis[J]. Comput Speech Lang. 2024;85: 101587.
233. Tiwari P, Zhang L, Qu Z, et al. Quantum Fuzzy Neural Network for multimodal sentiment and sarcasm detection[J]. Inform Fusion. 2024;103: 102085.
234. Li J, Li L, Sun R, et al. MMAN-M2: multiple multi-head attentions network based on encoder with missing modalities[J]. Pattern Recogn Lett. 2024;177:110–20.
235. Zuo H, Liu R, Zhao J, et al. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities[C]. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023:1-5.
236. Li M, Yang D, Zhang L. Towards robust multimodal sentiment analysis under uncertain signal missing[J]. IEEE Signal Process Lett. 2023.
237. Mou L, Zhao Y, Zhou C, et al. Driver emotion recognition with a hybrid attentional multimodal fusion framework[J]. IEEE Trans Affect Comput. 2023.
238. Kumar A, Sharma K, Sharma A. MEmoR: a multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries[J]. Image Vis Comput. 2022;123: 104483.
239. Chong L, Jin M, He Y. EmoChat: bringing multimodal emotion detection to mobile conversation[C]. In: 2019 5th International Conference on Big Data Computing and Communications (BIGCOM). IEEE, 2019:213-221.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.