Multimedia Systems (2024) 30:115
https://doi.org/10.1007/s00530-024-01302-2

SURVEY

Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects

Umair Ali Khan[1] · Qianru Xu[2] · Yang Liu[2] · Altti Lagstedt[1] · Ari Alamäki[1] · Janne Kauttonen[1]

Received: 15 July 2023 / Accepted: 21 February 2024 / Published online: 6 April 2024

Abstract
In recent years, emotion recognition has received significant attention, presenting a plethora of opportunities for application in diverse fields such as human–computer interaction, psychology, and neuroscience, to name a few. Although unimodal emotion recognition methods offer certain benefits, they have limited ability to encompass the full spectrum of human emotional expression. In contrast, Multimodal Emotion Recognition (MER) delivers a more holistic and detailed insight into an individual's emotional state. However, existing multimodal data collection approaches utilizing contact-based devices hinder the effective deployment of this technology. We address this issue by examining the potential of contactless data collection techniques for MER. In our tertiary review study, we highlight the unaddressed gaps in the existing body of literature on MER. Through our rigorous analysis of MER studies, we identify the modalities, specific cues, open datasets with contactless cues, and unique modality combinations. This further leads us to the formulation of a comparative schema for mapping the MER requirements of a given scenario to a specific modality combination. Subsequently, we discuss the implementation of Contactless Multimodal Emotion Recognition (CMER) systems in diverse use cases with the help of the comparative schema which serves as an evaluation blueprint. Furthermore, this paper also explores ethical and privacy considerations concerning the employment of contactless MER and proposes the key principles for addressing ethical and privacy concerns. The paper further investigates the current challenges and future prospects in the field, offering recommendations for future research and development in CMER. Our study serves as a resource for researchers and practitioners in the field of emotion recognition, as well as those intrigued by the broader outcomes of this rapidly progressing technology.

Keywords Emotion recognition · Contactless multimodal emotion recognition · Human-computer interaction · Contactless data collection · Ethical and privacy considerations · Real-world implementation

1 Introduction

Emotions are complex and multifaceted mental and physical states that reflect an individual's situation and state of mind. There is no universally accepted definition of emotions. They are commonly understood as a range of mental or physical states such as anger, happiness, sadness, or surprise [1]. Various sources, including dictionaries and scholarly works, offer different perspectives on emotions. The Oxford Dictionary[1] describes emotion as "*a strong feeling such as love, fear, or anger; the part of a person's character that consists of feelings*." In contrast, the Encyclopedia Britannica[2] defines it as "*a complex experience of consciousness, bodily sensation, and behavior that reflects the personal significance of a thing, an event, or a state of affairs*." In addition, emotion has also been defined as "*A response to a particular stimulus (person, situation or event), which is generalized and occupies the person as a whole. It is usually an intense experience of short duration—seconds to minutes—and the person is typically well aware of it*" [2].

From a philosophical perspective, emotions can be viewed as states or processes. As a state, like being angry

Communicated by B. Bao.

✉ Umair Ali Khan
umairali.khan@haaga-helia.fi

[1] RDI & Competences, Haaga-Helia University of Applied Sciences, 00520 Helsinki, Finland

[2] Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland

[1] Oxford Learner Dictionary | Emotions.

[2] Encyclopedia Britannica | Emotions.

or afraid, an emotion is a mental state that interacts with other mental states, leading to specific behaviors (Internet Encyclopedia of Philosophy[3]). Neurologically, emotions can be defined as a series of responses originating from parts of the brain that affect both the body and other brain regions, utilizing neural and humoral pathways [3]. Psychologically, emotions are characterized as "*conscious mental reactions (such as anger or fear) subjectively experienced as strong feelings, usually directed toward a specific object, and typically accompanied by physiological and behavioral changes in the body*" (American Psychological Association[4]).

Interpreting emotional states involves assessing various components such as behavioral tendencies, physiological reactions, motor expressions, cognitive appraisals, and subjective feelings [4, 5]. However, capturing these signals often requires specialized equipment, posing substantial challenges to emotion recognition. In the field of human–computer interaction (HCI), emotion recognition plays a pivotal role in determining an individual's current situation and interaction context. It has garnered huge attention due to its extensive applications across sectors including video games [6], medical diagnosis [7], education [8, 9], patient care [10], law enforcement [11, 12], digital marketing and sales [13], entertainment [14], road traffic safety [15], autonomous vehicles [16, 17], smart home assistants [18], surveillance [19], robotics [20], and cognitive edge computing [21], to name a few. Consequently, the global market for emotion detection and recognition has witnessed significant growth recently with an estimated value of USD 32.95 billion in 2021 and a projected compound annual growth rate of 16.7% from 2022 to 2030 [22]. This growth can be attributed to various factors, including the increasing demand for advanced technologies in various industries, the need for enhanced customer experiences, and the rising significance of mental health and well-being. As a result, organizations across sectors are investing in emotion detection and recognition solutions to better understand and cater to their users' needs, driving market growth and innovation.

Human communication is inherently multimodal, involving textual, audio, and visual channels that work together to effectively convey emotions and sentiments during interactions. This underlines the significance of integrating various modalities for more accurate emotion recognition [23]. Multimodality refers to the presence of multiple channels or modalities [23], such as visual, audio, text, and physiology, encompassing a wide range of formats, including text, image, audio, video, numeric, graphical, temporal, relational, and categorical data [24].

Sensing techniques for Multimodal Emotion Recognition (MER) can be categorized into three main types: invasive, contact-based, and contactless [25]. Invasive methods, which are relatively less common, involve the neurosurgical placement of electrodes inside the user's brain to measure physiological signals such as stereotactic EEG or electrocorticographic [26]. On the other hand, contact-based methods are non-invasive but necessitate the use of sensors in direct contact with the skin. Examples of contact-based methods include scalp EEG, disposable adhesive ECG electrodes for heart rate measurement, finger electrodes for electrodermal activity assessment [27], armbands for EMG signal detection, respiration sensors worn around the chest or abdomen, and wearable cameras for capturing facial expressions or body language. Contact-based methods offer the advantage of providing authentic emotional data that is difficult to enact; however, these methods can cause discomfort or psychological distress for users due to the need to wear the equipment [28].

Contactless Multimodal Emotion Recognition (CMER) integrates various sensing techniques that eliminate the need for physical contact with the user, providing a non-invasive and unobtrusive approach to emotion detection. These contactless methods employ an array of sensors, including RGB cameras, infrared/near-infrared cameras, frequency-modulated continuous-wave radars, continuous-wave Doppler radars, and Wi-Fi [29]. Although there has been considerable progress in emotion recognition technology, current methods of data collection through contact-based methods restrict the effective implementation and widespread adoption of this technology. Moreover, the lack of a comprehensive and user-friendly CMER system, along with unaddressed ethical challenges, impedes the technology's potential to transform various industries.

Given these challenges, there is a pressing need for a systematic literature review that explores the potential of contactless data collection methods and identifies gaps in the existing research landscape. The extensive review presented in this paper aims to address the problem by critically analyzing and benchmarking the existing review studies on MER, identifying the individual modalities, cues, and modality combinations used in MER, highlighting the importance and need for a CMER system, benchmarking the existing review studies, discussing a comparative schema for selecting modality combinations, addressing ethical considerations and challenges, and providing future research directions. By delving into these concerns, the literature review helps advance the field of CMER systems and ensure their responsible and ethical application across different domains.

The research questions outlined below guide the systematic literature review process. By addressing these research questions, the review aims to provide a comprehensive understanding of the current state and future directions of

[3] Internet Encyclopedia of Philosophy | Emotions.

[4] American Physiological Association | Emotions.

CMER, ultimately contributing to its advancement and ethical implementation across various applications and domains.

RQ1:	How do different emotion recognition modalities function, what limitations do they face, and how can these be overcome through multimodal and contactless methods?
RQ2:	What are the existing gaps in multimodal emotion recognition research and how can they be bridged to improve the effectiveness of CMER systems?
RQ3:	How can CMER systems be adapted for various real-world scenarios, what criteria should guide the selection of modality combinations and cues, and what are the potential challenges and their respective solutions?
RQ4:	What ethical issues are associated with the deployment of CMER technology, and how can these challenges be effectively mitigated?
RQ5:	What are the current limitations and upcoming trends in CMER that could shape the direction of future research and development?

This comprehensive literature review stands apart from numerous existing studies on MER. Instead of adhering to traditional reviews that primarily focus on data collection methods, fusion techniques, datasets, and machine learning approaches, our study takes a unique approach to examine this field. We believe that MER has already been extensively covered from traditional perspectives, which is why we have opted for a two-tier approach. In the first tier, we conduct a thorough review of all recent reviews published on MER in general. Our motivation stems from the investigation of contemporary trends, challenges, and issues in this domain. By exploring only the review studies, we provide a more comprehensive understanding of the field than if we were to review other study types separately. In the second tier, we identify and review studies specifically addressing CMER, a more feasible and practical approach to emotion recognition. This tier considers all relevant study types to uncover current methods, developments, challenges, and limitations. We emphasize and justify the importance of contactless emotion recognition over contact-based methods.

The insights gained from this two-tier study enable us to identify the specific set of modalities, cues, and unique modality combinations used in CMER. A thorough review of the existing open datasets of contactless modalities further leads us to discuss the implementation of a CMER system in diverse use cases with the help of a comparative schema, an aspect that has been overlooked in the literature. Our comparative schema, which serves as an assessment model, is grounded in strong theoretical foundations and is supported by justifications derived from our comprehensive review of the relevant literature. Moreover, our study considers the ethical implications of contactless emotion recognition. We outline these challenges and provide potential solutions. Finally, we provide a detailed description of the challenges and limitations in implementing a CMER system while also shedding light on its future prospects, research, and development.

The contributions made by this systematic literature review can be summarized as follows:

1. A comprehensive examination of individual modalities, cues, and models for emotion recognition, underlining the significance of MER, and presenting the benefits of CMER as a solution to address existing limitations (Sect. 2, ***answer to RQ1***).
2. Benchmarking of the survey studies by the selected metrics and identification of unaddressed research gaps, accompanied by a critical review of these gaps, potential solutions, an exploration of available contactless open datasets, and an analysis of unique modality combinations, shedding light on new avenues for research (Sects. 4–6, ***answer to RQ2***).
3. The formulation of a comparative schema for a CMER system, demonstrating its applicability across a diverse range of use-cases, promoting its practical relevance and addressing the potential challenges (Sect. 7, ***answer to RQ3***).
4. A thorough investigation into the ethical implications of CMER systems, providing insights into data protection, privacy issues, and methods to address biases. This includes the proposal of key principles concerning ethics and privacy, framing guidelines for ethical CMER applications (Sect. 8, ***answer to RQ4***).
5. A detailed analysis of the current limitations, potential remedial measures, and emerging trends in the CMER landscape, providing guidance for future research and development in the field (Sect. 9, ***answer to RQ5***).

This study elucidates both the general principles of MER and the specifics of its contactless variant. It explores the roles of various modalities and their associated cues in emotion recognition, highlighting their advantages and disadvantages. Moreover, the study explores how traditionally contact-based cues can be procured through contactless methods. A thorough analysis of the dominance and frequency of usage of each modality, respective cues, modality combinations, and the relevant open datasets within the existing literature offers valuable insights into prevalent trends.

This comprehensive examination aids in identifying unique combinations of modalities and facilitating the development of a comparative schema for determining a particular modality combination suitable for a specific use case and its requirements. The discussions concerning the implementation of CMER on diverse use cases and the modality selection based on their unique requirements, challenges, ethical considerations, potential solutions, and future research directions highlighted in this study provide a

stimulating reference guide for AI researchers, AI developers seeking to incorporate these systems into applications, and industry professionals and policymakers who are responsible for shaping guidelines and practices in this emergent field.

The remainder of this article is structured to methodically address the proposed research questions. Section 2 discusses the theoretical background and related work in MER and serve as a comprehensive response to RQ1. Section 3 describes our review methodology, detailing the review protocol, research strategy, selection criteria, and the process of study selection. Thereafter, our study delves into RQ2, which is thoroughly explored and addressed through the discourse presented in Sects. 4–6. Section 4 offers an exhaustive review of existing literature on MER, whereas Sect. 5 is dedicated to CMER specifically. In Sect. 6, we critically examine existing datasets pertinent to CMER. Section 7 responds effectively to RQ3 by exploring the application of CMER systems across various use cases using a modality-selection schema. Section 8 is devoted to unpacking and addressing RQ4 by exploring ethical and privacy concerns associated with CMER, proposing key principles to ensure ethical practice and privacy protection. Finally, Sect. 9, which focuses on answering RQ5 in a detailed manner, consolidates the insights gathered from the study, highlights several challenges and potential solutions, and outlines future avenues for research.

2 Theoretical background and related work

Before exploring the specific challenges and limitations in multimodal emotion recognition, as well as identifying the existing research gaps through an in-depth systematic literature review, it is essential to establish a foundational understanding through a discussion of the theoretical background and a review of related work. This introductory section provides a thorough overview of the various types of emotions, followed by an in-depth analysis of the individual modalities, cues, and models pertinent to emotion recognition. Furthermore, this section also discusses the application of machine learning techniques and the role of pre-trained models in enhancing the efficacy and accuracy of CMER methodologies.

2.1 Emotion types and their impact

Comprehending the subtleties of emotion types and their impact is crucial for developing an effective emotion recognition system that accurately identifies and interprets various emotional states across diverse contexts. Delving into emotion types allows the system to capture the full spectrum of emotions experienced by users, encompassing subtle distinctions, variations, and complex emotions that may be challenging to categorize.

Emotions are intricate mental and physiological states that reflect an individual's circumstances and mindset, steering their actions and decisions. As a result, precise emotion identification is essential for analyzing behavior and anticipating a person's intentions, needs, and preferences. A fundamental aspect of human interaction that facilitates its natural flow is our ability to discern others' emotional states based on subtle and overt cues. This capacity enables us to tailor our responses and behaviors, ensuring effective communication and promoting mutual understanding [30].

Emotions generally fall into two primary categories: positive and negative [31]. Negative emotions, such as anger, sadness, fear, disgust, and loneliness, are linked to harmful consequences on an individual's well-being [32]. On the other hand, positive emotions, like love, happiness, joy, passion, and hope, are perceived as favorable and contribute to overall well-being [33]. While positive emotions reinforce resilience, enhance coping strategies, and foster healthy relationships [34], negative emotions can result in stress, anxiety, and various physical health problems. Consequently, understanding and managing emotions are integral aspects of personal development and maintaining a balanced, healthy lifestyle.

Recognizing the importance of emotion types and their impact is important for creating emotion recognition systems that are more accurate, effective, and better equipped to address the complexities of human emotions in various contexts. This knowledge ultimately leads to improved system performance, user satisfaction, and an enhanced understanding of the emotional landscape that supports human experiences.

2.2 Emotion models

Selecting an appropriate emotion model is fundamental for any emotion recognition system, as it dictates how emotions are represented, analyzed, and interpreted across different contexts. Two primary models, categorical and dimensional, have gained popularity [35]. The former, also known as the discrete model, identifies basic emotions like happiness, sadness, anger, fear, surprise, and disgust [36]. Even though its simplicity makes it extensively used in emotion recognition, it may fail to accurately capture an emotion's valence or arousal [37].

Contrarily, dimensional models view emotions as points or regions within a continuous space [38]. The circumplex affect model is a popular example of this methodology that classifies emotions into two independent neurophysiological systems: one related to valence and the other to arousal [15]. This perspective allows to encompass a wider range of emotions and represent subtle transitions between them.

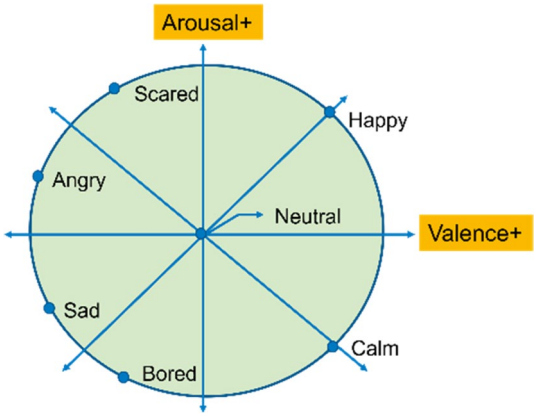

Fig. 1 Discrete emotion model mapped to continuous emotion model

Studies indicate that continuous dimensional models may provide a more accurate description of emotional states than categorical models [39–41]. Figure 1 illustrates the mapping of the discrete emotions model to the continuous dimensional model.

2.3 Emotion modalities

In HCI's realm, modality can be defined as a communication channel that enables interaction between users and a computer through a specific interface [42]. It can also be seen as a sensory input/output channel between a computer and a human [43]. Modalities represent a diverse range of information sources capable of offering various types of data and perspectives [31]. They offer diverse types of data that contribute to our understanding of emotional states. Modalities for emotion recognition encompass text, visuals, auditory signals, and physiological signals. Each offers unique insights and varies in terms of effectiveness and accessibility. While some data can be easily extracted, others might require specialized setups or equipment. An understanding of different modalities' characteristics can aid researchers in designing comprehensive and effective emotion recognition systems that respond accurately to subtle changes in human emotions.

2.3.1 Text modality

Text modality has been predominantly utilized in sentiment analysis [44, 45]; however, researchers have also discovered that it contains elements of emotions, particularly in data published on social media platforms [46]. Text data have been employed for emotion recognition as a standalone modality [47] or in conjunction with other modalities such as audio [48]. Convolutional neural networks have shown success in emotion recognition from text [49], and recent advances in large language transformers have further improved the accuracy of emotion recognition in text-based data [50]. Despite its potential, relying exclusively on text data for emotion recognition presents several challenges. One such challenge is the contrasting perceptions of emotions between writers and readers [51], which can lead to misinterpretations. Additionally, the absence of contextual meaning within text data can further complicate the process of correctly identifying emotions [52].

2.3.2 Visual modality

Videos provide a multifaceted source of data, combining visual, auditory, and textual elements to enable comprehensive identification of human emotions through multimodal analysis [53]. Video cues include facial expressions [54], body language [55], gestures [56], eye gaze [57], contextual elements within the scene [58], and socio-dynamic interactions between individuals [59]. Researchers have utilized these visual cues, either individually or in combination, to recognize and classify emotions more accurately. Visual modality plays a crucial role in continuous emotion recognition and serves as a methodology that aims to identify human emotions in a temporally continuous manner. By capturing the natural progression and fluctuations of emotions over time, continuous emotion recognition offers a more realistic and distinct understanding of human emotional states [60]. However, relying solely on visual modalities for emotion recognition, particularly a single cue, has its limitations, as certain individuals may not express their emotions as openly or clearly. Some people may display subtle or masked emotional cues, making it challenging for visual-based systems to accurately detect and interpret their emotions [61].

2.3.3 Audio modality

Acoustic and prosodic cues, such as pitch, intensity, voice quality features, spectrum, and cepstrum, have been widely studied and employed for emotion recognition [62]. Researchers have also utilized cues like formants, Mel frequency cepstral coefficients, pause, teager energy operated-based features, log frequency power coefficients, and linear prediction cepstral coefficients for feature extraction. With the advent of deep learning, new opportunities have arisen for extracting prosodic acoustic cues [63]. Convolutional neural networks have been effectively applied to extract both prosodic and acoustic features [64, 65]. In addition to these verbal prosodic-acoustic cues, non-verbal cues such as laughter, cries, or other emotion interjections have been investigated for emotion recognition using convolutional neural networks [66]. Since convolutional neural networks cannot capture temporal information, researchers have also employed alternative deep neural networks like dilated residual networks [67] and long short-term memory networks

[68] to preserve temporal structure in learning prosodic-acoustic cues.

Despite the potential of the audio modality for emotion recognition, several challenges remain. One major challenge is selecting the most suitable prosodic-acoustic cues for effective emotion recognition [69]. Another active area of research is determining the appropriate duration of speech to analyze [70]. Furthermore, emotion recognition using the audio modality faces issues such as background noise [71], variability in speech [72], and ambiguity in emotional expression [73], which can impact the accuracy and reliability of the emotion recognition system.

2.3.4 Physiological modality

The physiological modality captures data from the human body and its systems. These responses arise from both the central and autonomic nervous systems [12]. A range of physiological cues can be used to monitor these responses, such as electroencephalography (EEG), electrocardiography (ECG), galvanic skin response, heart rate variability, respiration rate analysis, skin temperature measurements, electromyogram, electrooculography [25], and photoplethysmography or blood volume pulse [74].

Physiological signals, which are involuntary and often unnoticed by individuals, offer a more reliable approach to emotion recognition since they reflect dynamic changes in the central neural system that are difficult to conceal [61]. Recent studies have extensively explored physiological signals for emotion recognition [75–77]. Among these signals, EEG has gained significant attention due to its fine temporal resolution, which captures the rapid changes in emotions [78], and its proximity to human emotions as a neural signal [79].

While physiological signals offer valuable insights into emotion recognition, they also present several challenges. Data acquisition typically require specialized setups or wearable sensors, which may be invasive, uncomfortable, or inconvenient for users, especially when worn for extended periods. A standard lab setting often involves subjects wearing earphones and sitting relatively motionless in front of a screen displaying emotional stimuli [80]. Despite the advancements in ubiquitous computing that enable the collection of physiological data through electronic devices (e.g., skin conductivity, skin temperature, heart rate), some signals, such as EEG and ECG, still necessitate costly sensors or laboratory environments. Moreover, the reliance on contact-based sensing and voluntary user participation can pose practical difficulties. Another challenge lies in the variability of participants' responses to different stimuli, which can lead to lower emotion recognition accuracy. This issue necessitates a well-designed experimental setup and the careful selection of stimuli to ensure reliable results [81].

2.3.5 Multimodal systems

Multimodal systems do not necessarily require cues from different modalities. Cues may come from a single modality, forming a multimodal system. For example, the visual modality can combine facial expressions and gaits [59], while the integration of different acoustic cues such as 1D raw waveform, 2D time–frequency Mel-spectrogram, and 3D temporal dynamics can be used for speech emotion recognition [82]. Similarly, different cues from various modalities can form a multimodal system, such as sentence embeddings (text modality), spectrogram (audio modality), and face embeddings (visual modality) as used in [83], or facial expressions and EEG-based emotion recognition used in [84]. However, multiple inputs for the same cue do not constitute a multimodal system (e.g., facial expression recognition using multiple camera views). This leads us to define multimodality as '*the data obtained from more than one cue pertaining to single or multiple modalities, where modalities refer to textual, auditory, visual, and psychological channels*.'

2.3.6 Contactless multimodal emotion recognition (CMER)

The modalities, including visual cues (face, posture, and gestures), oculography (visible light and infrared), skin temperature, heart rate variability, and respiratory rate can all be obtained via contactless sensors [25]. The primary advantage of CMER is that it minimizes discomfort and inconvenience typically associated with wearable devices or skin-contact sensors, enhancing user experience and promoting natural interactions. By seamlessly integrating into various environments, contactless approaches are well-suited for applications where user comfort and unobtrusive emotion recognition are essential.

The study of CMER encompasses a range of methodologies, including traditional machine learning, deep learning, and the utilization of pre-trained models. These various approaches are elaborated upon in the subsequent discussion.

2.3.6.1 Employing traditional machine learning for CMER Machine learning has become a popular tool for CMER across a variety of fields, employing both traditional and deep learning techniques. Over the last one decade, several machine learning techniques have been employed for CMER. Here, we discuss a few of them proposed in the last few years. These methods first extract the modality features using a feature extractor and then train machine learning algorithms to learn the emotions. These algorithms include Support Vector Machines (SVM), K-nearest neighbors (KNN), random forests (RF), and decision trees (DT), among others. The extracted fea-

tures included different combinations of modalities studied for different applications. For instance, the authors in [85] train SVM models on acoustic and facial features for emotion recognition from surgical and fabric masks. It has been found that the performance of machine learning algorithms vary significantly for the same data and task. For instance, the authors in [86] evaluated both traditional machine learning algorithms (such as KNN, random forests RF, decision trees DT, among others) as well as extreme learning machines (ELMs) and Long-Short-Term Memory networks (LSTM) for board game interaction analysis. The results show that ELMs and LSTM perform best for expressive moment detection whereas random forests perform best for emotional expression classification.

Earlier techniques did not consider the temporal information and environmental conditions in emotion recognition. However, temporal and environmental factors have been deemed increasingly important for emotion recognition as shown by some recent techniques. For example, the authors in [87] showed the importance of temporal information in emotion detection by proposing a Coupled Hidden Markov Model (CHMM)-based multimodal fusion approach to modeling the contextual information based on the temporal change of facial expressions and speech responses for mood disorder detection. The speech and facial features were used to construct the SVM-based detector for emotion profile generation. The CHMM was applied for mood disorder detection. Similarly, environmental factors were studied for emotion recognition in combination with physiological signals by a hybrid approach in [88] by developing RF, KNN, and SVM models. These techniques showed that temporal and environmental context has significant impact on emotions.

Besides visual and acoustic features, text features have also been studied in combination with these two modalities and have been identified as an important modality. For instance, the authors in [88] used a Gabor filter and SVM for facial expression analysis, regression analysis of audio signals, and sentiment analysis of audio transcriptions using several existing emotion lexicons for tension level estimation in news videos. This work showed that the integration of sentiments extracted from text into visual and acoustic features improve the performance of CMER. This also motivated authors to study physiological signals in combination with visual and acoustic modalities. For example, a study [89] exploited facial expressions and physiological responses such as heart rate and pupil diameter for emotion detection, using an extreme learning machine algorithm. However, the setup required for measuring the physiological signals makes it infeasible for practical applications.

2.3.6.2 Employing deep learning for CMER Due to the growing performance of deep learning algorithms, CMER's focus has greatly shifted to employing deep learning architectures. The notable deep learning architectures used in emotion recognition include CNNs, Recurrent Neural Networks (RNNs), LSTMs, and transformer-based architectures. These architectures have been used individually or in combination. Most deep learning approaches to CMER emphasize on speech signals and audio-visual features [90, 91]. Attention-based methods in deep learning have further enhanced the deep learning performance. A notable instance is a study [92] that utilized an attention-based fusion of facial and speech features for CMER. Frustration detection in gaming scenarios was another application where deep neural networks were used on audio-visual features [93]. Text features, in combination with visual and audio features were also studied using deep learning. For instance, depression detection using audio, video, and text features was achieved through deep learning-based CMER [94]. It was found that the integration of text features into audio and visual features using deep learning further improved the CMER performance.

CMER employed deep learning in a number of interesting applications including education. For example, CMER was applied to younger students during the COVID-19 lockdown by using two convolutional neural networks (CNNs) on acoustic and facial features combined [95]. Another CMER technique used CNN on eye movement and audiovisual features in Massive Open Online Courses (MOOC) environments to classify learners' emotions during video learning scenarios [96]. Further research included student interest exploration, where deep learning was used on head pose as well as facial expression and class interaction analysis [97]. Some methods have also probed into contactless physiological signals with deep learning for CMER; a study employed a method that combined heart rate and spatiotemporal features to recognize micro-expressions [98].

2.3.6.3 Employing pre-trained models for CMER The progressive strides in deep learning have paved the way for sophisticated, large-scale, pre-trained models that are now finding application across diverse fields. Pre-trained models are the deep learning models trained on large datasets. These models can be fine-tuned or re-trained for a similar task without training from scratch. These models have emerged as a particularly effective tool for CMER, achieving superior results when fine-tuned for CMER. Several studies have leveraged the power of these pre-trained models for emotion recognition using different modalities such as speech and facial features. Research has also demonstrated the effectiveness of finetuning multiple pre-trained models for CMER tasks. Fine-tuning pre-trained models eliminates the need for vast datasets, extensive computational resources, and prolonged training times. Several pre-trained models have been finetuned for CMER.

Among such models, one model is Wav2Vec2.0 which is a transformer based speech model trained on large-scale unlabeled automatic speech recognition data that has been extensively employed for speech recognition-based CMER [99]. Other notable models include the CNN architectures such as AlexNet, ResNet, VGG and Xception which have been employed for both audio and visual features [82, 100]. Similarly, pre-trained LLMs such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pretrained Transformer (GPT) and Robustly Optimized BERT Pretraining Approach (RoBERTa) have also been used for CMER using text features [101].

The use of multiple pre-trained models to handle each modality by its respective model has also been studied in several studies for improving CMER's performance. For instance, AlexNet was used to extract audio features which, along with visual features extracted by CNN + LSTM architectures, formed an effective combination for CMER tasks [102]. Similarly, the authors in [103] utilized both Wav-2Vec2.0 and BERT pre-trained models for CMER while extracting speech and text features respectively. Similar multi-model approaches have been adopted by other researchers as well [104]. A notable example is [105] in which a pre-trained model trained on the ImageNet dataset was used to generate segment-level speech features before applying them to a Bidirectional Long Short-Term Memory with Attention (BLSTMwA) model.

Other techniques involved integrating audio and visual features with the BERT model [106], or fine-tuning separate pre-trained models such as GPT, WaveRNN, and FaceNet + RNN to extract domain-specific text, audio, and visual features, respectively [83]. More recent studies involved finetuning an audiovisual transformer for CMER [107], and finetuning ResNet and BERT pre-trained models for audio-text-based CMER [108]. Pre-trained models were also extended to study continuous emotions to detect arousal and valence, where multiple pre-trained models were tested for feature extraction across different modalities [109].

While finetuning single or multiple pre-trained models has been extensively used for CMER, some studies have also proposed specialized pre-trained models, such as MEmoBERT [110], that make them ideal candidates for CMER methods. This is further supported by recent initiatives such as an open-source, multimodal LLM based on a video-LLAMA architecture designed explicitly to predict emotions using multimodal inputs [111].

Given the wide range of CMER techniques, their relative performance comparison becomes a complex task due to the use of varied datasets and evaluation metrics. This complexity is enhanced by different research approaches, such as deep learning versus machine learning, and performance metrics such as accuracy, concordance correlation coefficient, and F1 score among others, and the scenarios considered. Furthermore, the application of these research techniques varies across studies, with some using publicly available datasets while others utilize self-generated or privately held data. Adding to this complexity is the diverse focus and emphasis found within each article which makes a balanced comparison nearly impossible without additional comprehensive investigation. Future studies should aim for a thorough exploration presenting a benchmark comparison among various methods. Despite these complexities in comparative analysis, there is an observable trend toward improved efficiency in existing studies that adopt more advanced models and methods over time. In general, several recent studies (e.g. [112, 113],) show that the performance of deep learning-based multimodal emotion recognition for speech and image data is better than traditional machine learning algorithms due to the fact that deep learning can find intricate patterns in complex data more efficiently than traditional machine learning.

As many contactless methods rely on audiovisual modalities, these are prone to manipulations [28] and can be affected by factors like occlusion, lighting conditions, noise, and cultural differences. To enhance CMER accuracy, it is essential to combine audiovisual modalities with other cues from different modalities. The contact-based nature of physiological modality-based methods has traditionally limited their use in CMER; however, recent research has demonstrated promising advancements.

Several studies have successfully acquired physiological cues, traditionally dependent on contact-based methods, using contactless techniques. Examples include measuring heart rate and respiratory rate with remote photoplethysmography [114] and continuous-wave Doppler radars [115]. These breakthroughs have been made possible by recent advancements in multispectral imaging, computing, and machine learning, which facilitate the transformation of contact-based cues into non-contact ones.

3 Review method

This section outlines the methodological approach employed to conduct a systematic literature review on CMER. We detail the steps taken to ensure a comprehensive, organized, and rigorous examination of the existing literature. The following subsections cover the review protocol, research strategy, inclusion/exclusion criteria, and the study selection process. These elements collectively provide a clear and transparent framework for the systematic review, ensuring that the findings are robust and relevant to the research questions and objectives.

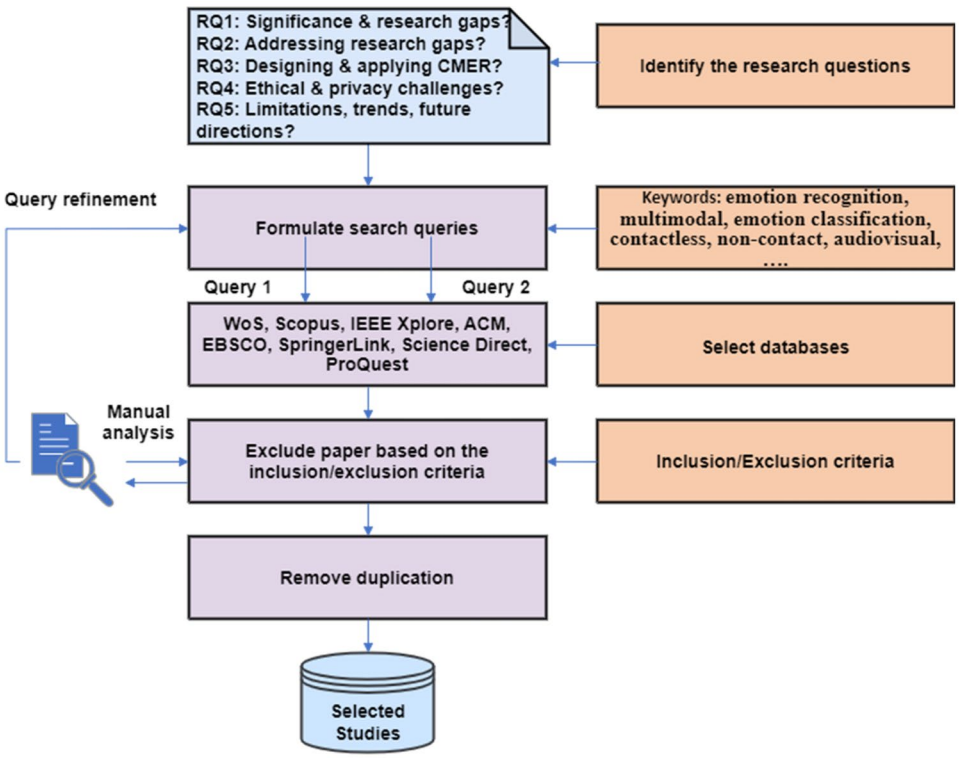

Fig. 2 A depiction of our review protocol

3.1 Review protocol

Our review protocol, developed in accordance with the established guidelines for conducting systematic literature reviews as outlined in [116], is illustrated in Fig. 2. The process begins with the identification of research questions that target the primary theme of our review. These questions guide the scope and objectives of the systematic review. After establishing the research questions, we create search queries based on them and select relevant databases to search for relevant literature. These search queries are individually applied to each chosen database, ensuring comprehensive coverage of the relevant studies. The retrieved papers then undergo a manual analysis to assess their relevance based on our inclusion and exclusion criteria. In an iterative process, we refine the search queries and repeat the search until no significant changes in the results are observed. This approach ensures that we collect the most relevant studies and minimize the risk of overlooking crucial papers. After the iterative refinement, we remove any duplicate papers from the search results and consolidate the remaining papers into a final list of relevant studies. This step reduces redundancy and allows for a more focused analysis of the literature. Finally, we conduct a thorough review of the selected papers, extracting essential insights, identifying trends, and drawing conclusions.

4 Research strategy

Considering our research questions and the objectives of our review, we first conduct a thorough examination of review articles that address MER in general. Following this initial analysis, we shift our focus specifically to CMER. To accommodate these two distinct aspects of our review, we formulate two separate search queries, the details of which can be found in Table 1. It is important to note that our selected keywords may not always appear in the title fields of the articles; therefore, we extend our search to include titles, abstracts, and keywords to identify relevant articles. However, we refrain from searching the entire article text, as this approach could yield numerous irrelevant results. That said, we meticulously analyze the retrieved articles to ensure their relevance to our study. By using this two-tiered search strategy, we are able to comprehensively review the literature on MER while also delving deeper into the specific area of CMER. This approach allows us to effectively address our research questions and gain a deeper understanding of the current state and future directions of this rapidly evolving field.

It is crucial to emphasize that the selection of keywords is meticulously carried out, considering contemporary trends in MER. This process begins with a basic search query, which is refined iteratively as we analyze the retrieved articles. The goal is to ensure that the search query becomes

Table 1 Details of the search queries

Query	Description	Query structure	Search fields
Query 1	Articles related to MER	("*multimodal*" **OR** "*multiple modalities*" **OR** "*multiple channels*" **OR** "*multiple sensors*") **AND** ("*emotion recognition*" **OR** "*emotion classification*" **OR** "*emotion detection*" **OR** "*emotion analysis*" **OR** "*emotion sensing*")	Title, abstract, author keywords, and keywords plus
Query 2	Article related to contactless emotion recognition	("*multimodal*" **OR** "*multiple modalities*" **OR** "*multiple channels*" **OR** "*multiple sensors*") **AND** ("*emotion recognition*" **OR** "*emotion classification*" **OR** "*emotion detection*" **OR** "*emotion analysis*" **OR** "*emotion sensing*") **AND** ("*contactless*" **OR** "*non-contact*" **OR** "*non-invasive*" **OR** "*audiovisual*") **NOT** "*EEG*"	

more focused and aligned with the most relevant and current trends in the field. In our second query, we specifically include the keyword "audiovisual" to fetch more articles related to contactless emotion recognition. The addition of this keyword helps to capture a broader range of studies that focus on contactless approaches, reflecting the increasing popularity and effectiveness of audiovisual-based emotion recognition methods in the field. In order to maintain consistency and align with the objective of investigating contactless-based methods, we deliberately excluded the term "EEG" from our search in the second query. Although "EEG" often appeared in conjunction with "non-invasive" in the search results, it predominantly refers to a contact-based or even invasive method, which deviates from our research focus. By refining our search queries in this manner, we can create a comprehensive and targeted collection of articles that reflect the current state of research in CMER. This approach ensures that our review is both accurate and up-to-date, allowing us to draw meaningful conclusions and identify areas for future exploration.

4.1 Inclusion/exclusion criteria

In a systematic literature review, well-defined inclusion and exclusion criteria play a crucial role in identifying papers relevant to the research questions for inclusion in the review. Our inclusion/exclusion criterion is based on relevance, specific databases, date of publication, study type, review method, completeness of study, and publication language. We selected studies from prominent and pertinent databases which were chosen due to their comprehensive coverage of literature in the relevant fields, ensuring access to a wide range of high-quality research articles. The selected studies included journal articles and papers published in conference proceedings and workshops. To focus on the most recent developments and advancements in the area, we set the publication range from 1st January 2017 to 1st March 2023. This timeframe was chosen to provide a comprehensive overview of the recent trends, methods, and findings while maintaining the review's relevance and currency.

For the first query, our focus is exclusively on review articles. We filter the results by using the available option within the database to specifically select review articles or, when such an option is not provided, by manually analyzing and selecting the review articles. This approach aligns with the scope of our literature review, as our primary objective is to focus on the review articles related to MER, enabling us to identify research gaps unaddressed in existing reviews. In recent years, review articles on MER have already assessed various study types, including experimental studies, case studies, qualitative, and quantitative research. Consequently, we chose not to include these study types in the first tier of our literature review, as they would not offer any new insights. By focusing on review articles, we can obtain a more comprehensive understanding of the existing research landscape, allowing us to recognize trends, gaps, and potential areas for future exploration in the field of MER.

Since our study specifically emphasizes CMER, it is essential to consider other study types in addition to review papers in the second query. By doing so, we aim to conduct a comprehensive review of the existing modalities, data collection methods, fusion techniques, and machine learning approaches relevant to the topic. This thorough examination enables us to find plausible answers to our research questions and better understand the current state of the field. Furthermore, incorporating other study types, such as experimental studies, case studies, qualitative, or quantitative research, allows us to explore the nuances and complexities of CMER. This inclusive approach also helps us identify any potential gaps or shortcomings in the existing literature, paving the way for further advancements in the field. Moreover, considering various study types provides valuable insights into the practical implementation of CMER, ensuring that our research is both relevant and applicable to real-world scenarios.

Our selection criteria focus on comprehensive, peer-reviewed studies available in full text. This encompasses journal articles, conference proceedings, and workshop papers, all published in English. We source these materials from a mix of open-access platforms, subscription-based databases, and preprint repositories, including Web of Science (WoS), IEEE Xplore, ACM, EBSCO, SpringerLink, Science Direct, and ProQuest. The subject areas under consideration span a diverse range, including computer science, psychology, neuroscience, signal processing, human–computer interaction, and artificial intelligence.

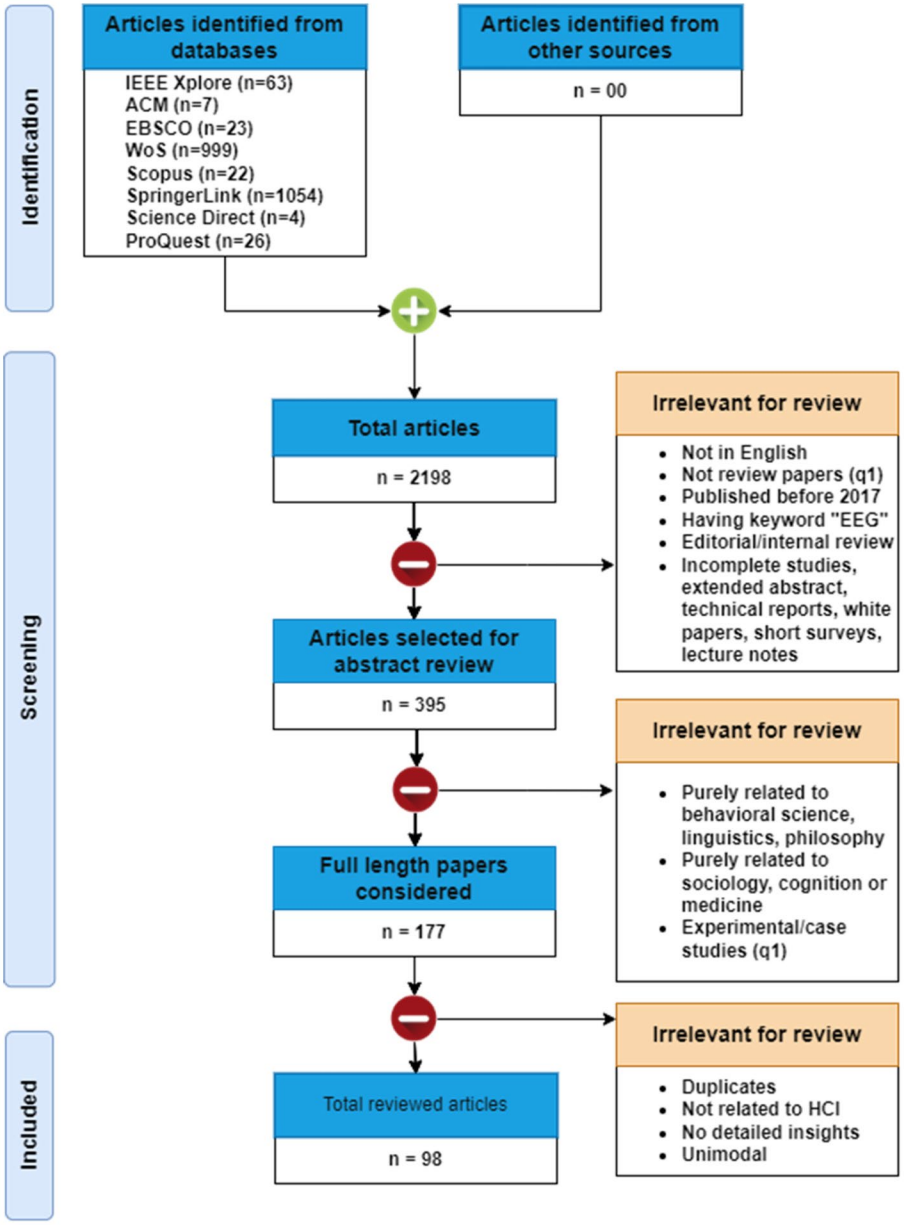

Fig. 3 PRISMA study flow diagram

We exclude certain types of content to maintain the focus of our study. Editorial and internal reviews, incomplete studies, extended abstracts, technical reports, white papers, short surveys, and lecture notes are left out of the selection. In addition, we do not include articles that were primarily concerned with sociology, cognition, medicine, behavioral science, linguistics, and philosophy to maintain a specific thematic focus.

4.2 Study selection process

Figure 3 presents the PRISMA flow diagram demonstrating the study selection process used in our systematic review and meta-analysis. Both Query 1 and Query 2 initially yielded 2196 articles across all databases. Some databases, such as WoS, allowed direct filtering for review articles. In contrast, databases such as SpringerLink required manual examining of search results for articles featuring the term "review" in their titles. In this phase, we excluded non-English articles, studies published before 2017, and incomplete studies, reducing the article count to 395. Additionally, for query 2, we manually excluded studies that utilized electroencephalograph/EEG or other brain imaging techniques as the primary recording channel. This exclusion was essential due to the current limitations in recording such data in a contactless manner. Following this, all references were collated within the Mendeley Reference Manager. Subsequently, in the second phase, abstracts underwent additional scrutiny to

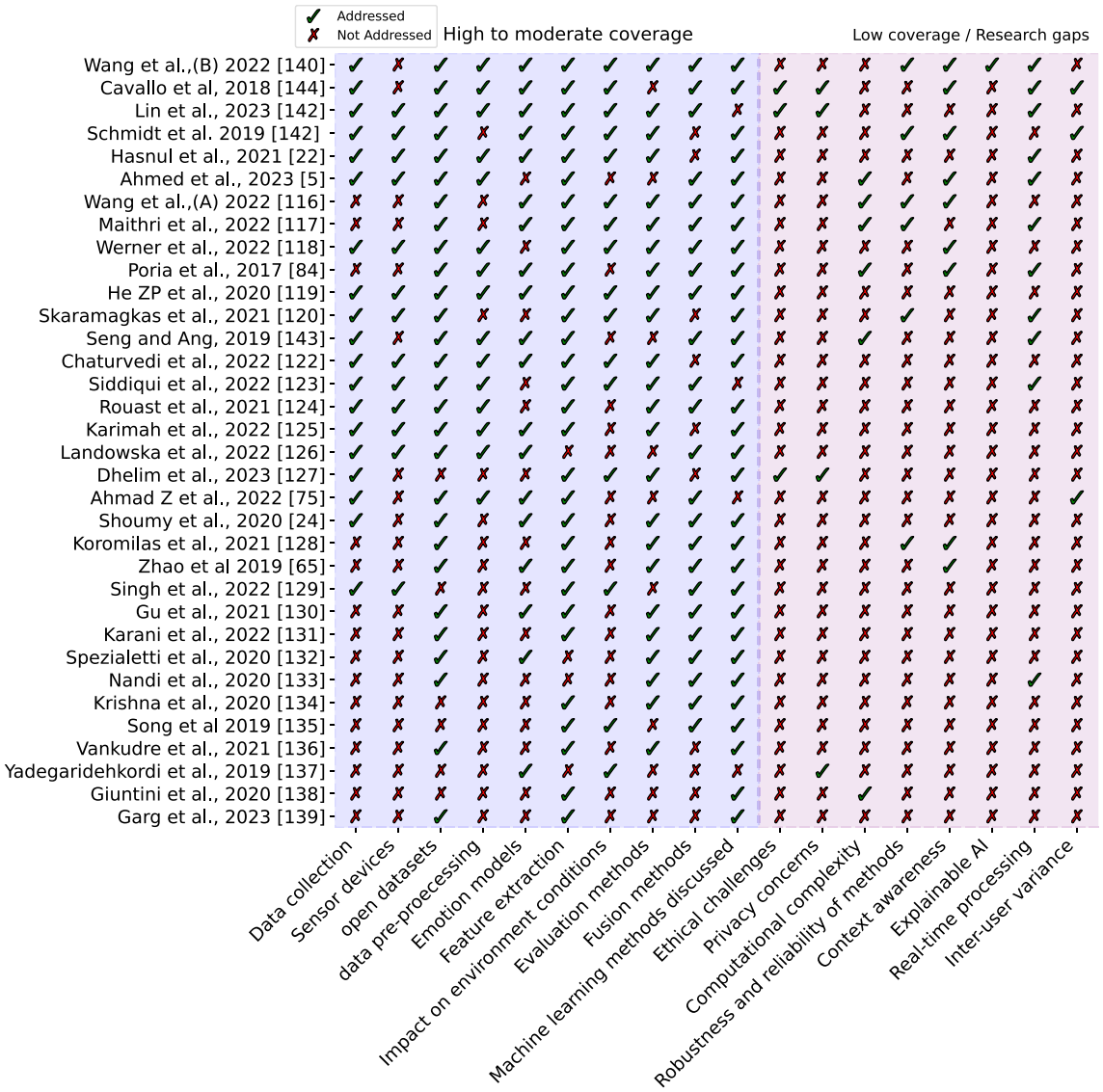

Fig. 4 Benchmarking of existing studies on MER with respect to the selected metrics

exclude experimental studies from Query 1 and papers relating to other disciplines such as behavioral sciences. This left us with 177 full papers for detailed examination.

Upon reviewing these papers thoroughly, we further excluded 79 articles. Exclusion criteria included duplication, lack of direct relevance to Human–Computer Interaction (HCI), absence of significant insights (often found in short conference papers), or exclusive focus on unimodal techniques. Ultimately, this meticulous process led to the inclusion of 98 papers in our final study.

Our meticulous selection process ensures that the articles included in our review are highly relevant to our research questions and objectives, providing a comprehensive understanding of MER and addressing any existing gaps in the literature.

5 Benchmarking multimodal emotion recognition: analysis of key metrics and research gaps

To explore the state-of-the-art in MER, we conduct a comprehensive review of existing literature on the subject. Our first objective is to identify relevant comparison and analysis metrics that could effectively evaluate each study. For this purpose, we perform a thorough examination of prior work in this field [10, 12, 23, 31, 72, 81, 117–144], identifying the key metrics shown in Fig. 4. These metrics have been explained in detail in Table 2. Following the identification of these metrics, we carefully assess each review article using these criteria as a guiding framework. It is worth acknowledging that while it may not be feasible to address all these

Table 2 Description of the metrics used for benchmarking MER studies

Metric	Description
Data collection	The process of gathering and measuring data from multiple sources
Sensing devices	Equipment used to capture, measure, or detect emotional cues and expressions
Open datasets	Publicly available datasets containing multimodal data for emotion recognition research and development
Data preprocessing	Techniques and methods applied to clean, normalize, annotate, and transform raw data into a suitable format for analysis
Emotion models/types	Different frameworks of emotions such as categorical, dimensional models, or appraisal theories
Modalities	Types of data or channels used in emotion recognition, such as audio, visual, physiological, or textual
Cues	Observable indicators or markers that provide information about emotions, such as facial expressions, speech, or body movements
Feature extraction	Techniques used to extract relevant and discriminative characteristics from raw data for emotion recognition
Impact of environmental conditions	The influence of external factors, such as lighting, background noise, or weather, on the performance of emotion recognition methods
Evaluation methods	Methods used to assess the performance and effectiveness of emotion recognition methods such as accuracy, precision, recall, F1-score, and ROC AUC
Fusion Methods	Techniques for combining information from multiple modalities
Machine learning algorithms/models	Algorithms and models employed to learn and recognize emotions
Ethical concerns	Issues related to the ethical implications of emotion recognition such as consent, biases, or potential misuse
Privacy and security	Considerations related to the protection of personal information, data confidentiality, and security in emotion recognition applications
Computational complexity	The number of computational resources required by the methods, including processing time and memory usage
Robustness and reliability	Ability to maintain accurate and consistent performance in diverse conditions
Context awareness	Assessing if the methods incorporate contextual information to enhance emotion recognition accuracy
Interpretability and explainability	Considering the interpretability and explainability of the machine learning models used in the methods, which can be important for understanding and trust-building in human-AI interactions
Real-Time Processing	Assessing whether the methods can process and analyze data in real-time or if they require offline processing
Inter-Subject Data Variance	Considering differences in emotional expressions, cues, and physiological responses among individuals when experiencing and expressing emotions
Generic/Specific	The degree to which the methods are general-purpose and applicable to various scenarios, or tailored for specific applications or contexts

metrics in any single study, our aim in identifying them and comparing them with the existing literature is to shed light on potential research gaps that could be addressed in future work. In addition to this, identifying and comparing these metrics across existing literature results in establishing a common framework for evaluation and benchmarking, as well as highlighting best practices and successful methodologies. Some acronyms, used in the analysis, are given in the table given in Appendix (supplementary information).

By analyzing the distribution of modalities and cues used in the study set, our comprehensive review provides a thorough understanding of the current landscape of MER research (see Figs. 5 and 6), emphasizing the significant role non-verbal cues play, notably physiological signals such as EEG, ECG, and EDA, either individually or in combination with other visual inputs. These signals can reveal subconscious emotional states that are often subtle to other

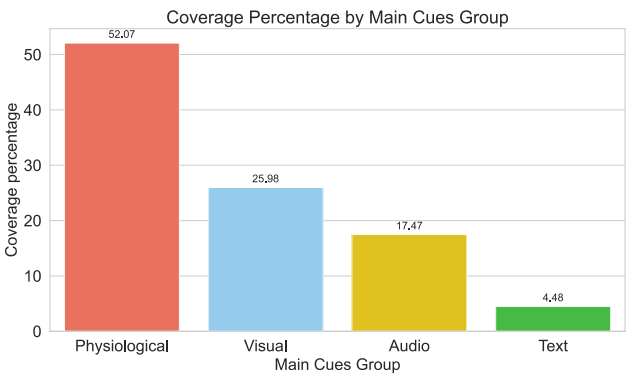

Fig. 5 Distribution of modalities in the study set. Physiological signals emerge as the predominant modality for emotion recognition, both when utilized individually and when combined with other modalities (see also Fig. 7)

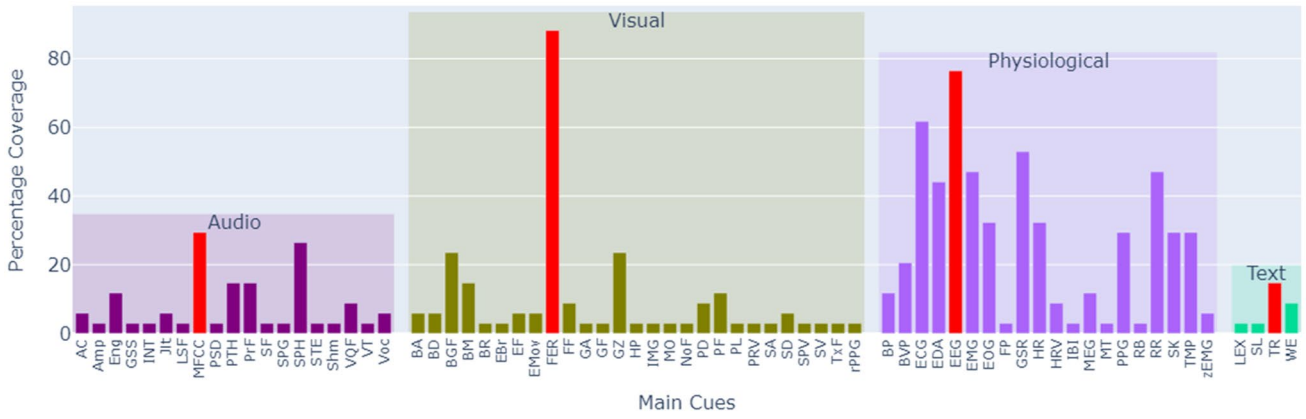

Fig. 6 The coverage of cues within the study set with cues highlighted in red indicating the highest coverage. Among the physiological signals group, EEG and ECG exhibit the most extensive coverage. FER emerges as the widely employed modality within the visual group. In the audio and text groups, MFCC, Text Reviews (TR), and Word Embeddings (WE) are the modalities that have been most frequently utilized

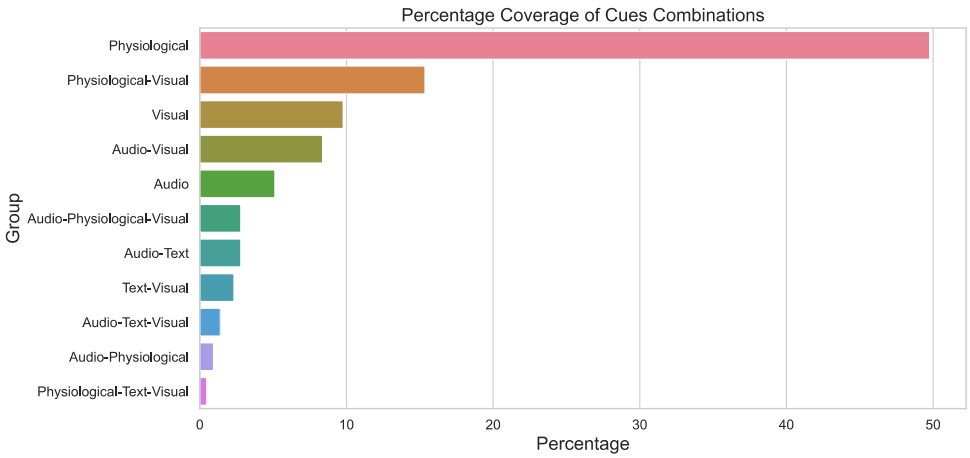

Fig. 7 The distribution of modality combinations within the study set, highlighting the prominence of various combinations. Physiological signal combinations are the most extensively employed choice, owing to their capability to accurately detect subtle emotional changes, relative immunity to deliberate manipulation, and provide continuous, real-time data, enabling a more comprehensive understanding of emotions

modalities. However, implementing physiological signal-based systems in the wild remains a challenge due to the aforementioned issues.

Our findings also highlight diverse combinations of modalities utilized in emotion recognition (see Fig. 7). In addition to the physiological and physiological–visual cues that are still largely required for the contact-based techniques, audio–visual modalities have been widely applied to MER, especially in the context of CMER. Some studies also explore the combination of audio, visual, and physiological modalities. Despite a smaller representation of text modality in our review, its effectiveness, particularly in conjunction with other modalities, is noteworthy. This is especially true in the context of conversational agents, where user input is largely text-based, emphasizing the necessity of an efficient emotion recognition system that incorporates audio, visual, text, and physiological modalities.

Within the scope of behavioral cues, Facial Emotion Recognition (FER) is highly employed, owing to its intuitive nature and the advancements in computer vision and machine learning that facilitate automated analysis. In addition, facial features provide profound insights into human emotion and behavior and are considered to be the strongest source of emotion cues [145, 146]. Paired with body gestures and movements, FER can provide a richer emotional context. Likewise, audio features such as Mel Frequency Cepstral Coefficients (MFCCs) and prosodic characteristics yield crucial emotional insights by dissecting spectral and rhythmic speech elements. Text modality, enhanced by natural language processing techniques like sentiment analysis, offers an added layer of emotional context, supporting the effectiveness of MER systems (see Fig. 6).

Figure 4 also categorizes the benchmarking metrics into two primary groups based on their coverage in review studies. Metrics in green are well-covered, representing a common pattern followed by most studies. In contrast, the metrics in orange indicate research gaps, as they have not been addressed by many studies. Figure 8 provides a

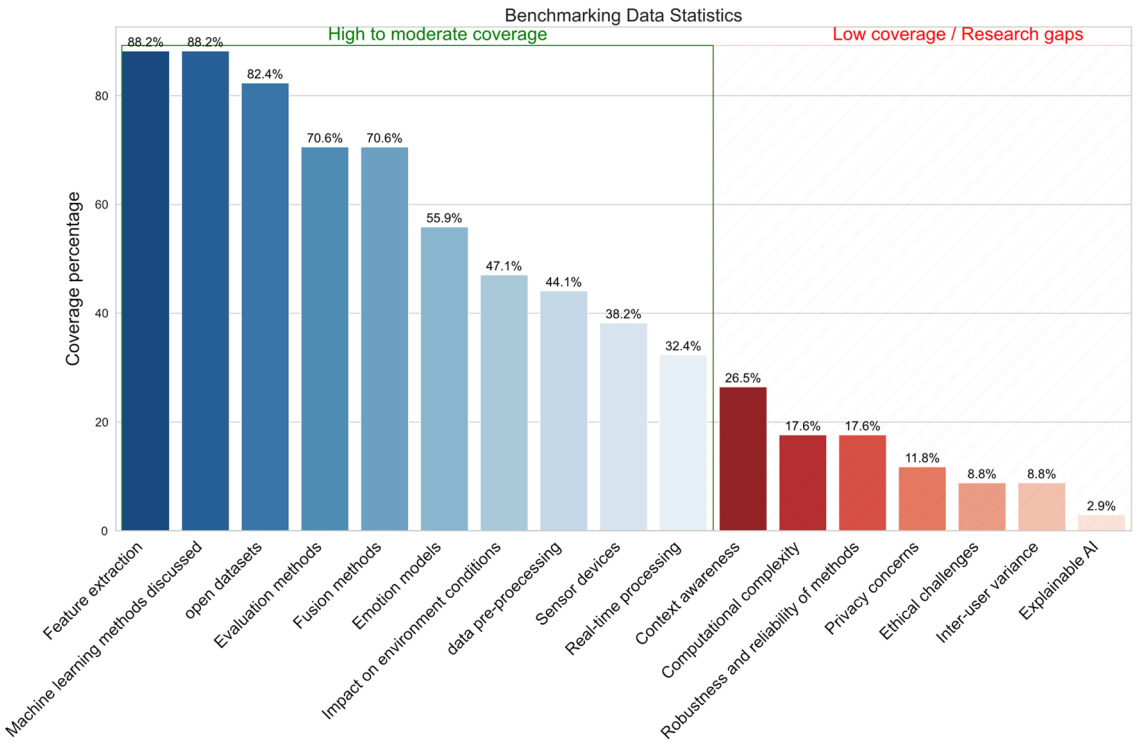

Fig. 8 The coverage of benchmarking metrics within the study group with metrics highlighted in red signifying the presence of under-addressed issues. Despite their significance for implementing MER in real scenarios, these metrics have not received adequate attention within the study group

clear illustration of the metrics that have been adequately addressed in the existing studies and those that have not received sufficient attention. In order to identify the research gaps, which are represented by the under-addressed metrics, we established a threshold of 30%.

One notable limitation in existing review studies is the lack of attention given to ethical and privacy concerns. With the rapid advancement of artificial intelligence, numerous ethical implications and privacy-preserving challenges have arisen. These concerns are crucial to address, as they involve respecting individuals' rights, ensuring fairness, and maintaining public trust in AI-enabled systems. We address ethical and privacy issues in detail in Sect. 7, where we also highlight the key principles for ensuring ethical and privacy compliance.

The computational complexity of MER systems is also overlooked in many review studies. The addition of more modalities to emotion recognition systems correlates with increased computational complexity [147], which can hinder real-time system deployment. Evaluating this complexity is vital in determining if these systems are feasible for real-time applications. There is no one-size-fits-all approach to managing computational complexity, but several strategies can enhance real-time performance. For example, we can select features that are highly representative of the emotional state [120], or apply dimensionality reduction to multimodal feature vectors [148] to simplify the model. In terms of deep learning models, the careful selection of parameters like convolution kernel size [149], the total number of model parameters [150], and the sizes of masks and weights in attention networks [151] can help manage complexity. Another strategy could be the utilization of the correct type of features for the task at hand. For instance, using virtual facial markers for facial emotion recognition (FER) may be more efficient than an approach based solely on image pixels [147].

The importance of robustness and reliability in emotion recognition (ER) systems cannot be overstated, however, these factors have received inadequate attention in review studies. Robustness characterizes the system's capacity to sustain performance in the face of external disturbances, and reliability represents its consistency in delivering accurate results over time. These qualities become particularly crucial in healthcare settings where lives hang in the balance. Take, for instance, an MER system employed for pain detection—it must possess high robustness to assess pain accurately. Several methods can be used to enhance the robustness and reliability of an MER system such as reducing the influence of environmental factors like noise [152], applying robust fusion techniques such as transformer-based cross-modality fusion [83], employing data augmentation in training

MER models [153], and adopting self-supervised learning approaches [154]. Recent studies have also indicated the high robustness of transformer-based multimodal ER models and deep canonical correlation analysis (DCCA) when combined with an attention-based fusion strategy [113, 155]. A more recent approach tackles the issue of robustness and reliability in MER systems by providing explanations for the predictions, thereby addressing the issue of label ambiguity due to the subjective nature of emotions. The model's predictions are deemed correct if the reasoning behind them is plausible [111].

Context awareness remains yet another significant research gap in the survey studies. While the integration of multiple modalities indeed enhances contextual comprehension, it only addresses one aspect of the overall context. The notion of context is vast, encompassing various elements like time, location, interaction, and ambiance, among others [156]. Recent studies have shown that self-attention transformer models are effective in capturing long-term contextual information [157]. These models can help in associating a modality cue—for instance, facial emotion recognition (FER)—with context cues such as the body and background [117]. Likewise, by employing a self-attention transformer, we can use background information and sociodynamic interaction as contextual cues in conjunction with other modality cues like FER and gait [59]. Therefore, these transformer models represent promising approaches for improving our understanding of the context in MER.

Although transparency and explainability are becoming increasingly significant in the field of AI, existing survey studies often overlook these aspects in emotion recognition. This oversight can largely be attributed to little focus on these attributes in the realm of emotion recognition. Earlier techniques mostly confined themselves to identifying influential modalities or cues [158, 159]. Nevertheless, a growing interest in these aspects has been observed recently. For instance, the authors in [160] sought to explain the predictions of MER systems using emotion embedding plots and intersection matrices. A recent study [161] implemented a more advanced technique. Their method provides explanations for its recognition results by leveraging situational knowledge, which included elements like location type, location attributes, and attribute-noun pair information, as well as the spatiotemporal distribution of emotions.

The issue of inter-subject data variance, or individual differences in emotion expression and perception, is often overlooked in emotion recognition research. Data related to specific modalities and emotions can vary significantly from person to person. For example, facial expressions, gait analysis, gestures, and physiological signals may not universally represent emotions, as individual responses and expressions can differ. More research about domain adaptation and developing attention-based models could help tackle inter-subject data variability [81].

6 Contactless multimodal emotion recognition (CMER)

In our second-tier investigation, we review the method papers that utilize MER techniques ([70, 85–89, 93–99, 162–212]), where modalities are obtained through sensors that are not in physical contact with the user's skin. A comprehensive analysis of contactless studies, exploring aspects like modality combination, types of emotion models, applications, and methods, is presented in Fig. 9. In our comparative analysis, we differentiate between verbal and nonverbal cues within the audio modality, treating them as separate entities. Verbal cues primarily pertain to the content of speech, while nonverbal cues encompass various nonlinguistic elements like laughter, cries, or prosodic features such as tone, pitch, and speech rate, amongst others. This distinction enables a clearer understanding of how different aspects of audio modalities have been explored. Similarly, we distinguish between RGB and thermal cues within the visual modalities to differentiate the roles of different visual cues in CMER. It is pertinent to mention that all the methods summarized in this section utilizing physiological modality use contactless techniques for data collection.

We identify the specific modality combinations used in CMER systems along with their percentage coverage, as shown in Fig. 10. The audio-visual emerges as the dominant combination of modalities. Specifically, verbal cues along with RGB visual cues remain dominant due to the fact that verbal communication is a key form of human interaction, while RGB visual cues provide critical context, enhancing communication effectiveness. Nonverbal and thermal cues of audio and visual modalities as augmented modalities remain under-explored despite their effectiveness. These cues have the potential to reveal subtle elements of emotional responses, engagement levels, and intentions, which are not explicitly conveyed through verbal and/or RGB visual cues. Similarly, the integration of text and contactless physiological signals as auxiliary modalities have been inadequately studied, despite their unique benefits (see Sects. 1.4.1, 1.4.4 and Sect. 5). Among other modality combinations, the audio-visual-physiological group has not been adequately explored, which could be promising.

Most of the studies have focused only on the categorical model, potentially due to its simplicity. The dimensional models, which provide a more complex but realistic representation of emotions as discussed in Sect. 3.1, have not been adequately addressed potentially due to the inherent complexity of dimensional models.

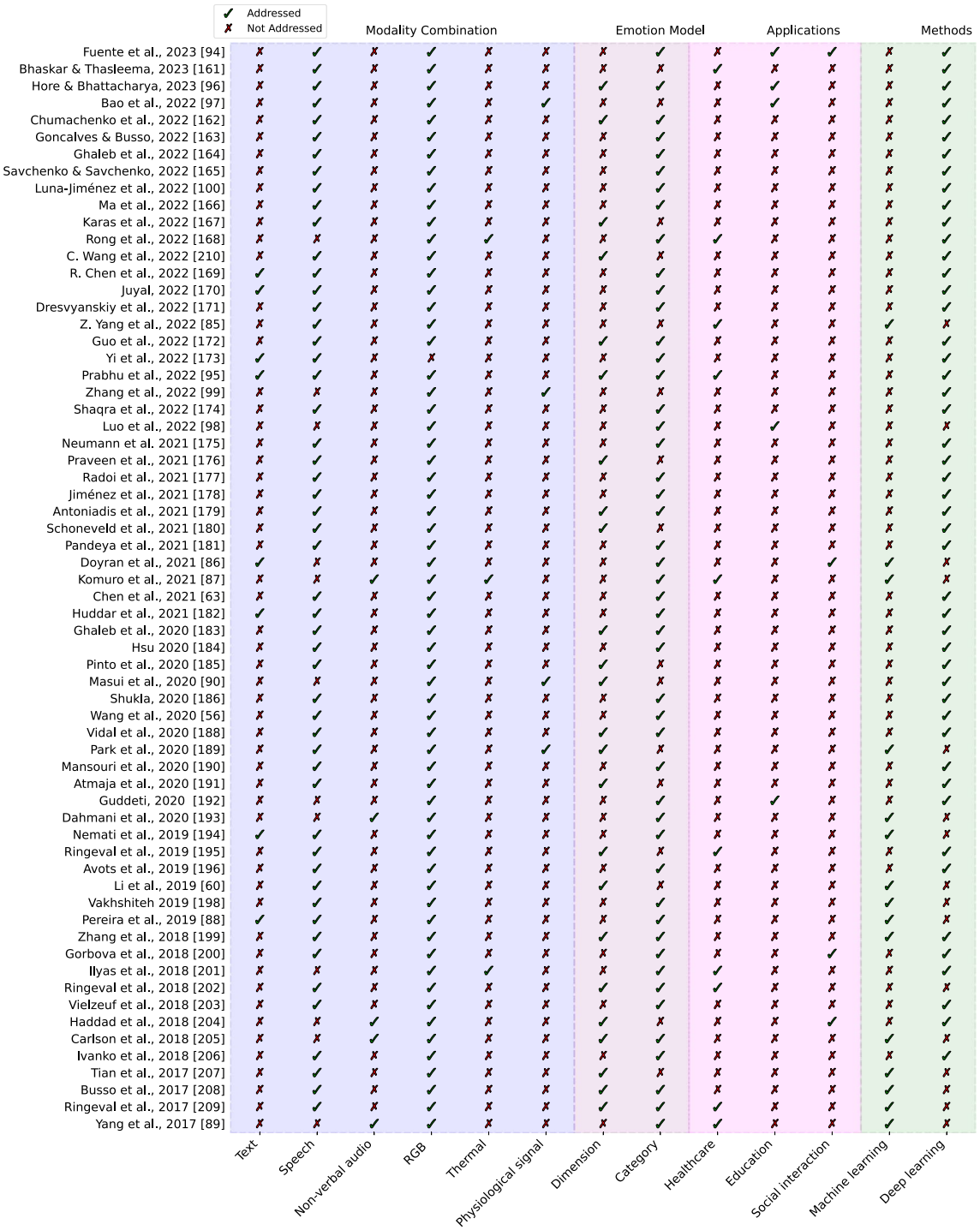

Fig. 9 A comparative analysis of the CMER methods with respect to modality combinations, types of emotion models, applications, and machine learning methods

We also analyze the applications of CMER systems in the studies in the context of three major fields: healthcare, education, and social interaction due to their intrinsic connection with human emotions. Despite the presence of studies discussing the application of existing CMER methods in one or more of these sectors, most research tends to address CMER in a general manner. We explore the implementation of the CMER system in these specific sectors along with other diverse scenarios in Sect. 6.

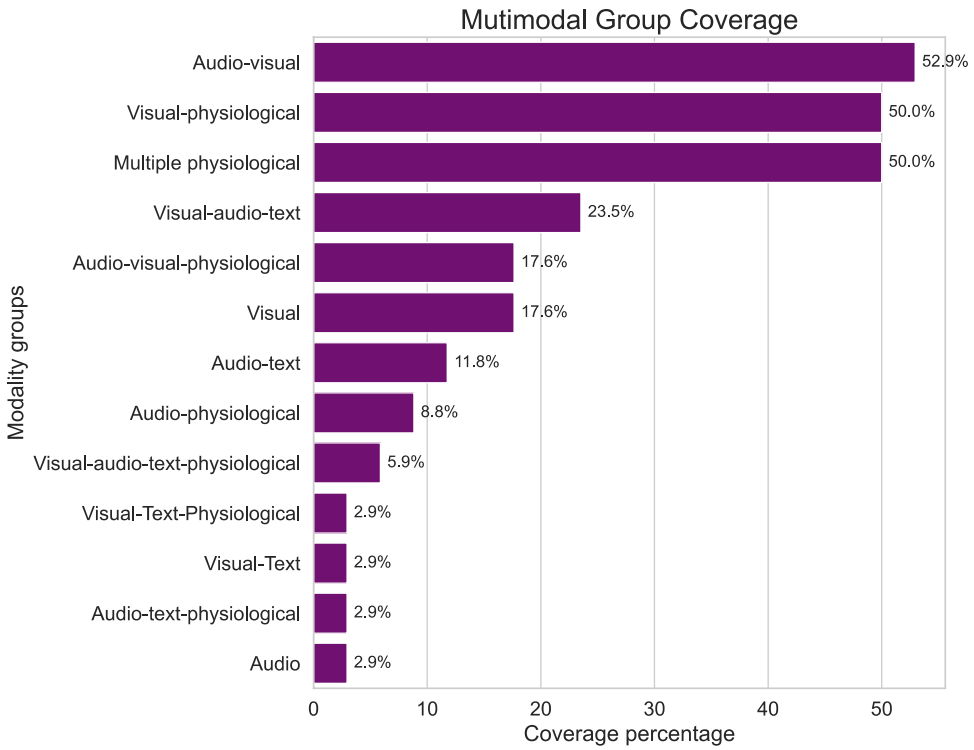

Fig. 10 Modality groups used in the listed papers

Due to the variation in modality groups and evaluation criteria across different studies, we do not draw performance comparisons between the CMER methods. However, we assess these studies based on the employed machine learning techniques: traditional methods like random forest, nearest-neighbors, and SVM [213], in contrast with deep learning methods. An upward examination of Fig. 9 reveals a progressive shift in CMER from traditional machine learning to deep learning. This shift can be attributed to deep learning's inherent ability to learn feature representations dynamically, which leads to enhanced accuracy and improved generalization. In addition, the proliferation of large-scale datasets and the surge in computational resources have further enhanced the accuracy of deep learning techniques in emotion recognition tasks.

For a comprehensive discussion of machine learning and deep learning emotion recognition techniques, we recommend recent surveys [214, 215], which explore this topic in greater detail.

7 Existing datasets for contactless multimodal emotion recognition

The availability of open datasets is of utmost importance in the advancement of MER systems, as they serve as the foundation for training AI models. The quality and diversity of these datasets directly impact the system's ability to generalize and accurately detect emotions in real-world scenarios. Open datasets not only facilitate the initial development and training of models but also contribute to their evaluation. By testing the system on separate data sets, developers can assess its performance and robustness against unseen instances, ensuring its reliability and validity. Furthermore, open datasets foster faster research progress by eliminating the need for individual data collection efforts. Conducting a review of existing open datasets also holds value in curating a selection of datasets specifically designed for contactless modalities. This analysis provides additional support for one of the metrics outlined in our comparative schema—data availability for a specific use-case (refer to Sect. 6).

Several open datasets have been published over the past 15 years, many of which have been previously examined in reviews such as [81, 124, 213]. In this section, we present a comprehensive list of relevant datasets, with a specific focus on CMER. Table 3 provides an overview of 33 datasets, each offering at least two contactless modalities. For these datasets, we select six key features that we deem most relevant when selecting data for model development in a specific use case or application. However, it is important to acknowledge that directly comparing these datasets is challenging due to their inherent heterogeneity. They were collected under varying conditions, utilizing different properties, and have distinct limitations. These factors will be discussed next.

In our dataset listings, we prioritize providing the number of separate subjects (N) instead of individual samples. This is crucial for ensuring model generalization to a broader population beyond the training sample. However,

Table 3 Collection of multimodal datasets for emotion recognition that includes more than one cue and contains emotion labels. Here, we use the following abbreviations: va = valence, ar = arousal, cat = category

Study	Database	N	Multimodal features	Condition	Type	Spon-taneous	Emotions
Dhall et al. [216]	AFEW	330	AVF	in-the-wild	media	no	7 cat
Kollias et al. [217]	Aff-wild2	458	AVF	in-the-wild	media	yes*	7 cat + va and ar
Sarkar et al. [218]	AVCAffe	106	AVF	controlled	paired	yes	va and ar
Valstar et al. [219]	AVEC	292	AVF	controlled	single	yes	va and ar
Zhalehpour et al. [220]	BAUM-1	31	AVF	controlled	single	yes	13 cat
Erdem et al. [221]	BAUM-2	286	AVF	in-the-wild	media	no	8 cat
Caridakis et al. [222]	CALLAS Expressivity Corpus	10	AVF	controlled	single	no	3 cat
Li et al. [223]	CAS(ME)3	247	AVF + CPS	controlled	single	yes	8 cat
Li et al. [224]	CHEAVD	238	AVF	in-the-wild	media	yes*	26 cat
Li et al. [225]	CHEAVD 2.0	527	AVF	in-the-wild	media	yes*	8 cat
Zadeh et al. [226]	CMU-MOSEI	1000	VATF	controlled*	media	yes*	6 cat + va
Cao et al. [227]	CREMA-D	91	AVF	controlled	single	no	6 cat
Ranganathan et al. [228]	emoFBVP	10	AVF + CPS	controlled	single	no	23 cat
Dhall et al. [229]	EmotiW 2017	1809*	AVF	in-the-wild	media	yes*	7 cat
Martin et al. [230]	eNTERFACE'05	42	AVF	controlled	single	no	6 cat
O'Reilly et al. [231]	EU Emotion Stimulus	19	AVF	controlled	single	no	20 cat
Bänziger et al. [232]	GEMEP	10	AVF	controlled	single	no	18 cat
Chen et al. [70]	HEU-part2	967	AVF	In-the-wild	media	yes*	10 cat
Douglas-Cowie et al. [233]	HUMAINE	48	AVF	In-the-wild	media	yes*	6 cat
Busso et al. [234]	IEMOCAP	10	VATF	controlled	paired	yes	10 cat + va and ar
Park et al. [191]	K-Emocon	32	AVF + CPS	controlled	paired	yes	18 cat + va and ar
Soleymani et al. [235]	MAHNOB-HCI	27	AVF + CPS	controlled	single	yes	9 cat + va and ar
Poria et al. [236]	MELD	13,708*	VATF	in-the-wild	media	no	10 cat
Shen et al. [237]	MemoR	8536*	VATF	in-the-wild	media	no	14 cat
Zhang et al. [159]	MMSE	140	VF + Th + CPS	controlled	single	yes	10 cat
Chou et al. [238]	NNIME	44	AVF + CPS	controlled	paired	no	6 cat + va and ar
Perepelkina et al. [239]	RAMAS	10	AVF + CPS	controlled	paired	no	9 cat
Livingstone and Russo [240]	RAVDESS	24	AVF	controlled	single	no	7 cat
Ringeval et al. [241]	RECOLA	46	AVF + CPS	controlled	paired	yes	va and ar
Clavens et al. [242]	SAFE	400	AVF	in-the-wild	media	no	4 cat + ar
McKeown et al. [243]	SEMAINE	150	VATF	controlled	paired	yes	13 cat + va and ar
Kossaifi et al. [244]	SEWA	398	VATF	controlled	single and paired	yes	va and ar
Metallinou et al. [245]	The USC CreativelT	16	AVF	controlled	paired	no	va and ar

*= ambiguous

for datasets derived from movies, TV series, and streaming services, the exact count of individual subjects is often not available. In such cases, we report the number of annotated samples instead. To maintain consistency and statistical significance, we have included only datasets with a minimum of ten subjects or samples. Datasets smaller than this threshold are typically inadequate for model development in practical applications.

Regarding annotated emotions, we report the number of distinct classes for categorical emotion models and valence and/or arousal for dimensional models, as discussed in Subsection 1.2. Some datasets offer both types of annotations. We acknowledge that certain properties are difficult to evaluate precisely and indicate them with an asterisk (*) symbol. For instance, the genuineness of emotions in TV series (e.g., reality TV) and online videos (e.g., YouTube blogs)

cannot be accurately assessed. Another point of ambiguity lies in the number of subjects. Notably, for datasets MELD and MEmor, we report 13,708 and 8,536 samples (clips), respectively. However, all these clips are derived from the TV series 'Friends' and 'Big Bang Theory,' predominantly featuring the main actors (5 and 7) from their respective shows.

Among the 33 datasets listed in Table 3, the distribution of modality groups is as follows: 18 datasets include audio-visual modalities (AVF), 7 datasets include audio-visual and physiological modalities (AVF + CPS), 6 datasets include visual-audio-text modalities (VATF), and 1 dataset includes visual, thermal, and physiological signals (VF + Th + CPS). While all aspects of training data are crucial for developing models applicable to real-life scenarios, we want to highlight three specific features in Table 3: condition, spontaneity, and emotions. Firstly, the condition under which the data is collected significantly impacts the generalizability of machine vision applications. In controlled environments, low-level features of the data remain relatively consistent, whereas in-the-wild data can exhibit significant variation. This variation affects aspects such as video resolution, occlusion, brightness, angles, ambient sounds, and noise levels. When utilizing the model in real-life applications, it must be capable of performing well under non-ideal conditions. Secondly, we emphasize whether the data captures genuine emotions or acted/imitated emotions. Genuine emotions are preferred to ensure the authenticity and real-life accuracy of the trained recognition model. Lastly, the choice of target labels determines the dataset's suitability for a specific use case. A significant distinction exists between datasets employing continuous models (valence/arousal) and those using discrete emotion models. The count of categories for discrete emotions varies widely, ranging from 3 to 26. Some datasets even include mental states such as boredom, confusion, interest, antipathy, and admiration. While certain applications may suffice with distinguishing between negative, neutral, and positive emotions (e.g., product reviews and customer satisfaction), other applications benefit from more nuanced annotations and require finer grained emotion distinctions.

Deep neural networks typically require a substantial amount of data for effective training. To address this, there are several strategies one can employ. Firstly, it is advisable to combine multiple existing datasets that are suitable for the task at hand. This approach helps increase the overall data volume and diversity, enhancing the model's ability to generalize. Additionally, collecting new data using the same setup of the application is essential to ensure that the model is trained on data that closely resembles what it will encounter during inference. Furthermore, data augmentation techniques can be leveraged to augment the existing dataset. Data augmentation has become a viable option, particularly with the introduction of Generative Adversarial Network (GAN) models and generative AI in general [246]. While image data augmentation is widely used, methods for augmenting audio-visual and physiological data also exist. For example, the authors in [179] proposed an augmentation method that considers temporal shifts between modalities and randomly selects audio-visual content. [168] developed GAN models specifically for audio-visual data augmentation. The authors in [174] demonstrated the effectiveness of data augmentation for physiological signals, improving emotion detection performance in deep learning models. Apart from expanding the training data, data augmentation can also help address potential biases in the dataset, such as gender or minority imbalances (see Sect. 7 for further discussion of ethical aspects).

8 Harnessing CMER for diverse use-cases

We explore the real-life scenarios in this section and explain how a CMER system can be effectively applied within the unique constraints and requirements of each use-case. The intention behind this detailed exploration is multifold. Primarily, it serves to highlight the versatility and adaptability of CMER systems, demonstrating their broad applicability across a range of scenarios with varying conditions and demands. Additionally, it provides a concrete, contextual understanding of how the selection and combination of modalities can be tailored to meet the specific needs of each use-case, thus adding a practical dimension to the theoretical understanding of these systems. By discussing the unique challenges inherent to each scenario and how these can be addressed within the framework of CMER, we aim to equip readers with a practical understanding that will inform and guide the future implementation of these technologies.

In our analysis, we identify a variety of modality combinations utilized in existing studies and propose a comparative schema comprising a wide range of metrics to evaluate the performance of these modalities. Figure 11 shows the CMER's comparative schema comprising various modality combinations such as audio-visual, physiological signals, and visual-audio-text, amongst others, and their benchmarks against the performance parameters such as real-time performance, in-the-wild feasibility, and context awareness, to name a few. Each cell in this comparative schema corresponds to the performance of a specific modality combination against a given metric. The process of constructing the comparative schema involved a detailed review and analysis of relevant literature, as well as an intuitive understanding derived from the studies. Each cell in the schema represents the performance of a specific modality combination against a given metric. The assignment of the values "Low", "Moderate", or "High" for each cell was carried out based on a consensus-driven analysis of the literature and the real-world application of these systems.

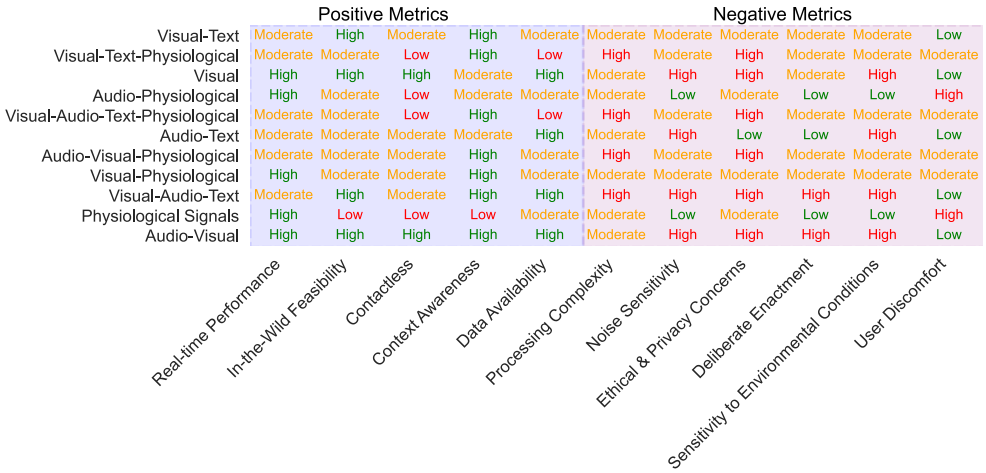

Fig. 11 CMER comparative schema for modality selection

The metrics are segmented into two distinct sets, each representing a unique aspect of the performance evaluation. The first set, highlighted in blue, represents 'positive metrics.' A higher value in this section, such as a 'High' rating in 'real-time performance,' is indicative of superior performance or desirable attributes. Conversely, the latter section, marked in orange, comprises 'negative metrics.' Within this section, an elevated value represents less favorable outcomes. For instance, a 'High' score in 'processing complexity' infers a significant computational demand, which is generally regarded as undesirable in the context of efficient and resourceful system design.

The term 'contactless' measures the system's capability to function with contactless interaction. An emotion recognition system that retrieves modality data from a smartphone, for instance, exhibits moderate contactless operation. Context awareness signifies the system's need for contextual data to accurately identify emotions. Systems functioning in a controlled environment, such as vital signs monitoring, require less context awareness than those operating in uncontrolled or 'wild' conditions. Data availability pertains to the accessibility of open datasets or the simplicity of gathering new data. Noise sensitivity involves the unintentional incorporation of undesired signals in a modality, which could occur due to external sounds in audio modalities, irrelevant objects in visual modalities, or electrical interference in physiological modalities.

Sensitivity to environmental conditions indicates how a system's performance might be affected by factors such as light intensity, weather conditions, or ambient temperature. Ethical and privacy concerns might range from low to high based on the application context. For example, a system designed for driver attention detection may carry fewer ethical and privacy implications than a system used in healthcare. Deliberate enactment refers to the faking of emotions by subjects. This factor could be moderate in a healthcare setting where patients might exaggerate symptoms, and higher in a law enforcement scenario where there's an incentive to deceive. User discomfort refers to any negative feelings users may have towards the system, whether due to its complexity, or any other factors that detract from a smooth user experience. Data availability refers to whether suitable open datasets are available in training a model that applies corresponding modalities simultaneously. This comparative schema aids in the selection of suitable modalities based on the requirements and constraints of a given use-case. It takes into account not only the functional attributes of each modality combination but also their implications on ethical and privacy concerns and user discomfort, ensuring a well-rounded and thorough evaluation.

When evaluating each combination of modalities, we initially take into account the inherent characteristics and limitations of the modalities involved. As an example, the 'audio-visual' modality combination receives a 'High' rating due to its lower computational demands and prompt feedback [90]. Additionally, this combination is also rated 'High' for context awareness, as it provides substantial contextual information [247]. Similarly, as another example, the 'video-audio-text-physiological' modality combination received a 'Low' rating for the "data availability" metric, as indicated in Table 3 of Sect. 5.

It is worth mentioning that within this comparative schema, physiological modalities are considered contact-based. This classification aligns with the relevant literature, as a majority of physiological cues are typically measured using contact-based methods. Despite the emergence of contactless methods to measure some physiological cues (such as heart rate and respiration rate [114], the table currently reflects the dominant paradigm in which these signals are captured and processed.

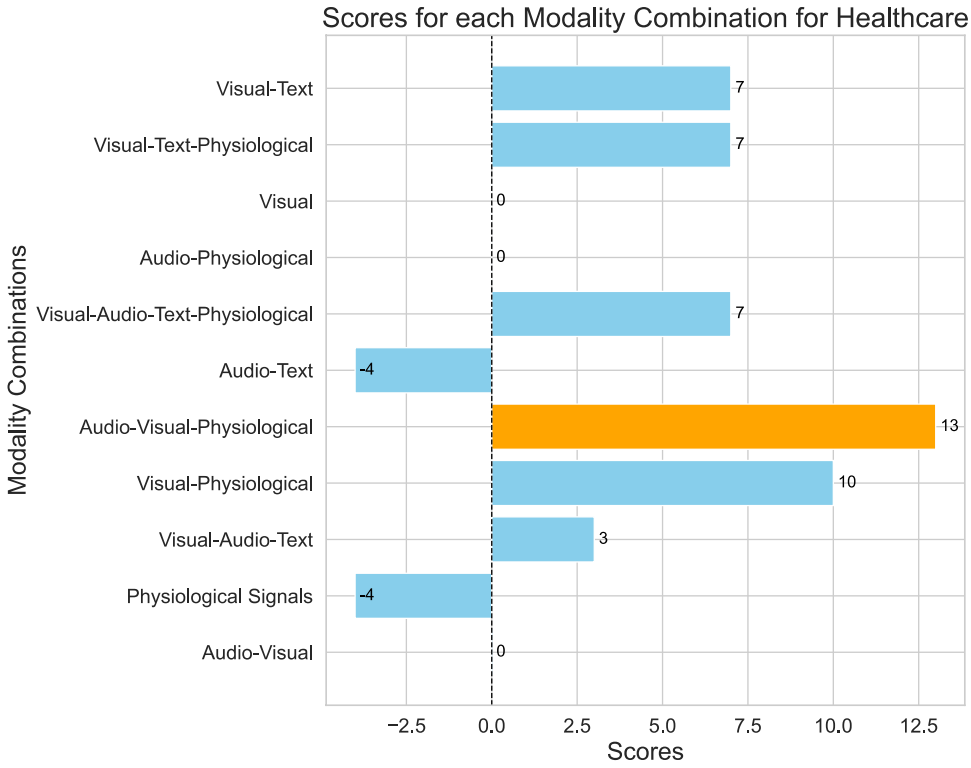

Fig. 12 Selection of modality combination for healthcare use-case

We envisage various scenarios for CMER and evaluate the suitability of various modality combinations for these use-cases. For this purpose, we first identify the requirements specific to a scenario and map these requirements to the respective metrics for each modality (Fig. 12). For each modality combination, the actual value of each metric is compared to our predefined requirements. A scoring mechanism is implemented, which rewards modalities that align with the requirements and penalizes those that do not. Through this process, we identify the modality combination that offers the most effective and feasible approach for feedback analysis in the hospitality industry. The algorithm for mapping the scenario requirements to the modality combinations is shown in Algorithm-I.

Algorithm-I: Scoring Modality Combination

```
Procedure SELECT_OPTIMAL_MODALITY(M)
    Initialize R - Define the requirements for the use case scenario
    for each modality combination m in M do
        Initialize score S(m)
        for each metric r and its required value v in R do
            Retrieve the value v' of the metric r for the current modality combination m from M
            if v' equals v then
                Increment S(m) by 2 #reward
            Else
                Decrement S(m) by the absolute difference between v' and v #penalize
            end if
        end for
        Assign the final calculated score S(m) to the current modality combination m
    end for
    Identify the modality combination m* with the highest score
    return m*
end procedure
```

8.1 Healthcare and wellbeing

In healthcare, contactless multimodal emotion analysis has the potential to enhance patient care and well-being. Existing studies have applied various contactless-based cues, including visual, audio, and textual cues, for mental health detection, specifically targeting depression [87, 94, 197, 211] and bipolar disorders [204]. These studies have identified indicators of depression, such as lack of eye contact, downward angling of the head, reduced frequency, intensity, and duration of smiles, variability in gestures, reduced speech variability and pitch, and increased tension in the vocal tract and vocal folds [248, 249]. In addition, it can be employed for real-time monitoring and providing

rehabilitation physiotherapy for traumatic brain-injured patients at the desired time for their recovery assistance [203]. It can also aid patients with special requirements, such as assisting in understanding emotions and facilitating communication among hearing-impaired individuals [163]. Moreover, by combining facial, vocal, and physiological cues from multiple modalities, contactless-based solutions have been successfully applied in supporting pain assessment [119].

As mentioned in Sect. 3, most current studies have primarily focused on utilizing audiovisual cues, and this also holds for healthcare applications. However, there is also potential in utilizing contactless solutions not only for facial and bodily expression-based emotional recognition but also for combining them with physiological signals obtained from contactless sources such as rPPG, radio, and WiFi. These contactless methods can be employed to monitor heartbeat and breathing in a smart home setting [250, 251]. Additionally, various elements of smart home architecture, such as lighting and music, can be utilized for emotion regulation [252].

In the health monitoring domain, the need for timely feedback and solutions entails high real-time performance. To meet the expectations of deployment in diverse environments, from homes and healthcare centers to outdoor settings through portable devices such as smartphones, the feasibility of contactless multimodal solutions is rated as moderately adaptable. Likewise, the priority for 'in-the-wild' applications is set at a moderate level. Accurate prediction and assessment are essential for health monitoring, which emphasizes the need for high context awareness. Furthermore, the ability to process complex information from coupled multimodal sensors results in high computational complexity. Noise sensitivity requirements could be moderate in facing both indoor and outdoor sceneries. Considering that the handled data mostly comes from patients, ethical and privacy concerns are highly significant. Since most of the scenarios may occur in relatively stable environments like smart homes or healthcare centers, sensitivity to the environment can be kept low. Though the possibility of deliberate enactment in a healthcare setup is relatively lower, it is still set to moderate to address the exaggeration of symptoms by some patients. As patients and individuals seeking health solutions are the primary targets in this use case, user discomfort can be compromised; hence, it can be kept at a low priority. The data availability is set to moderate primarily because while health monitoring systems can generate a substantial volume of data, access to this data may be constrained by privacy regulations and the personal preferences of the patients involved.

Figure 12 shows the scores assigned to each modality combination according to the specific requirements of the healthcare and well-being scenario. The audio-visual-physiological modality emerges as the most promising candidate. By merging visual cues including facial expressions and body movements, auditory signals such as voice tones, and contactless physiological cues such as heart rate, respiration rate, and electrodermal activity, a CMER system can be conceived that offers a multi-faceted evaluation of an individual's health status.

The determination of particular cues can be task specific. For example, consider a scenario of pain estimation in a healthcare setup. Although pain is not commonly classified as a basic emotion, it has been explored within the realm of emotion recognition [253], and a dedicated CMER system can be designed for automated pain detection. Existing pain evaluation tools that rely on observers, such as the PAINAD tool [254] for patients with cognitive impairments, could be automated through the audio-visual-physiological cues combination. The PAINAD scale quantifies pain levels in patients with dementia based on five behavioral indicators: breathing, negative vocalization, facial expression, body language, and consolability. A contactless CMER system, utilizing physiological cues for heart rate and breathing (rPPG), visual cues for facial expression, and body language, along with audio cues for negative vocalization, could potentially streamline and improve the accuracy of this process.

The major challenges in applying emotion recognition in the healthcare domain are ensuring the accuracy and robustness of CMER methods. Most current models and algorithms are based on healthy samples; therefore, factors such as variations in facial expressions from patients, lighting conditions, and environmental noise from the healthcare environment can all affect the reliability of the analysis. Therefore, it is necessary to build patient-based datasets that incorporate multimodal sensor data and improve existing algorithms to be used for diverse populations of patients. Developing user profiles and patient-tailored models would also help increase accuracy. However, this approach poses ethical challenges, especially regarding patient privacy. In this sense, implementing strict data protection measures, obtaining informed consent from both patients and caregivers, as well as adequately anonymizing and encrypting sensitive data, are highly prioritized and should be carefully considered in designing such kind of system.

8.2 Education

In the education domain, contactless multimodal emotion detection and recognition hold promise as a means to enhance learning performance and improve student engagement in the classroom. For example, the combination of facial expressions, hand gestures, and body language has been utilized to analyze students' affective states, assessing the engagement level [194], the interest in learning [97], and

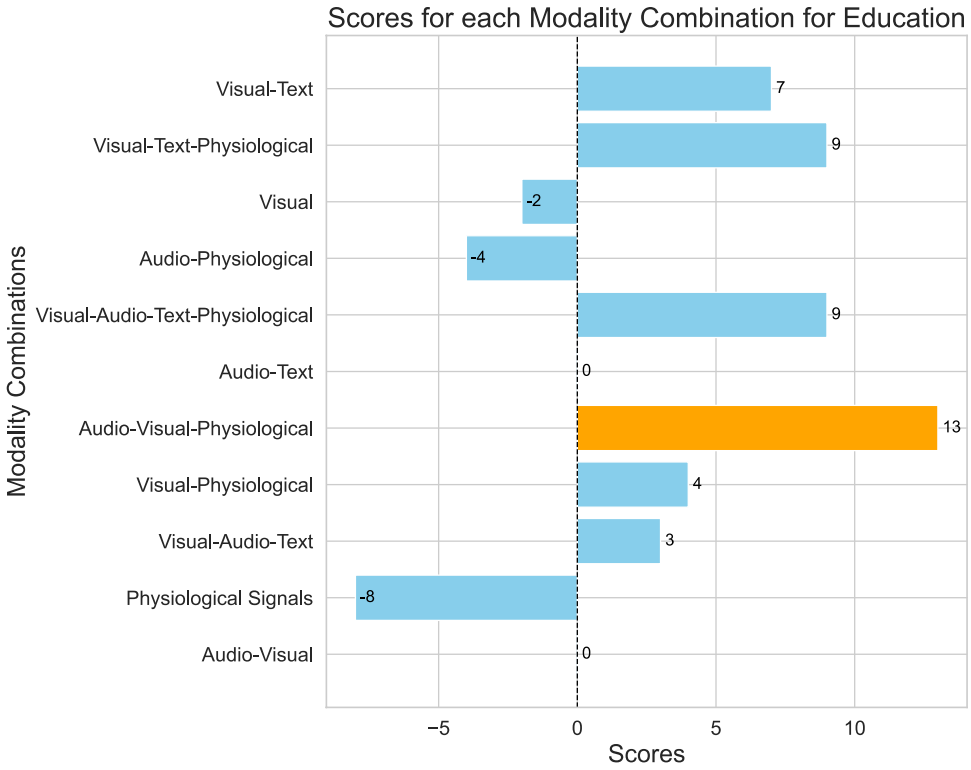

Fig. 13 Modality group selection for education use-case

the quality of collaboration [255] and students' frustration [93].

With the trend towards online or hybrid learning models for generation-Z, especially due to the significant impact of the COVID-19 pandemic, AI-powered contactless multimodal emotion analysis has gained traction. This approach offers great potential in assessing student engagement in real-time and providing feedback based on their behavioral and affective changes in the online learning environment. By integrating eye movements, audio, and video signals, the fusion of these features achieves an accuracy of as high as 81.9% in recognizing and distinguishing emotions such as confusion, boredom, and happiness [96]. In this sense, the adoption of contactless multimodal emotion analysis in education has the potential to create a feedback loop for study engagement, where real-time assessments of emotions can be used to provide timely feedback to students. This personalized approach can help educators tailor their teaching strategies and interventions to optimize student engagement and learning outcomes.

In an education use-case, the demands of real-time performance and in-the-wild feasibility of emotion recognition systems are moderate, because the analysis can be conducted either during the student learning or after the course. To get a comprehensive and effective result, the contactless setup, the context awareness, and the processing complexity are highly significant. We designate the noise sensitivity level as moderate to accommodate the system both within controlled environments and during outdoor events such as physical education classes. Given that our target audience in the education domain primarily comprises youth and teenagers, it is essential to address the ethical and privacy concerns associated with their participation. It is worth noting, however, that deliberate enactment of such concerns may be relatively low, as teenagers typically exhibit limited proficiency in concealing their emotions. Finally, the user discomfort should be low and considered for both teachers and learners, especially for underage students, so that they can concentrate on the educational process. We categorized data availability as moderate, considering that collecting data in the context of education, particularly in the post-pandemic era, is not overly challenging. However, similar to healthcare, acquiring this type of data may pose difficulties due to privacy concerns.

Figure 13 shows that the audio-visual-physiological modality garners the highest score when mapped against our requirement set. This combination of modalities can offer a comprehensive method of evaluating student interest and engagement. Visual cues, encompassing facial expressions, body language, and eye movements, combined with audio indicators like tone, speech rate, and pitch, are supplemented with contactless physiological data, such as heart and respiration rates. This multimodal approach promises to generate reliable feedback, proving invaluable to educators and virtual teaching aids.

Imagine a virtual classroom scenario. The CMER system integrates data from the students' webcams and microphones, to assess their affective states and engagement levels based on various cues. For instance, frequent yawning, slouched posture, or wandering eyes shown from the video data might indicate boredom or confusion. Furthermore, it can interpret slight color changes in students' faces to estimate their heart rates and respiration rates. For example, increased heart rates might denote excitement or stress. In parallel, the system analyzes audio signals to understand the emotional undertones in students' voices when they ask questions or participate in discussions. It examines the tone, speech rate, and pitch to discern sentiments. For instance, hesitant or stuttering speech might point to a lack of understanding, while an enthusiastic tone could imply interest. The students' engagement analyzed this way can help adjust teaching strategies accordingly.

The limitations of this system may include the variability in students' expressive behaviors and the subtlety of some emotional cues. The challenges include ensuring student privacy, maintaining a natural learning environment, and dealing with diverse lighting and noise conditions in online or hybrid learning settings. It also requires students to keep their webcams and microphones on to allow data collection. Potential solutions could involve using robust algorithms that can account for individual differences in expressions and employing noise-reduction technologies along with adaptive lighting adjustments for clear visual data capture. Furthermore, clear communication about the purpose and benefits of these features, along with strong assurances of data privacy and security, can also help mitigate any reservations students may have about keeping their cameras and microphones active.

8.3 Law enforcement and security

Emotion recognition technology has seen increasing applications in the field of law enforcement and security. The motivations behind these applications often stem from the ability of emotion recognition technology to analyze and interpret human states and feelings, potentially indicating criminal intentions or unlawful activities, and anticipate potentially dangerous situations beforehand [11, 12].

Visual features are the most commonly used modality via the wide availability of surveillance systems. Video and image analysis of facial expressions and body language can be used for security screening at airports, government buildings, and other public places to identify suspicious or agitated individuals. Various such systems are already in wide use, particularly in China (see Emotional Entanglement 2021 report [256]) and in the UK. Facial expression analysis is non-intrusive and can be performed at a distance, making it suitable for security applications. However, facial expression alone may not be reliable due to the possibility of deliberate masking or falsifying of emotions. Here, multimodal systems offer a more comprehensive approach.

Audio and speech emotion recognition from phone calls, interviews, interrogations, and negotiations could help detect deception, evaluate the mental state of suspects or distressed individuals, and assess the risk of violence [257]. However, speech-based emotion recognition is not precise, especially over the phone with low-quality audio. Physiological signals like galvanic skin response, heart rate, and blood pressure measured through wearable sensors could provide more reliable clues about a person's true emotional state. However, obtaining physiological data requires direct contact with the individual, which may not always be feasible or acceptable in security contexts.

By assessing individuals' emotional states in public spaces, such as airports or city streets, potential threats could be identified based on recognized patterns of fear, aggression, or stress. This provides an early warning system, enabling swift intervention before incidents escalate. Furthermore, it allows for better allocation of security resources by identifying areas of increased tension or fear. For example, police in China and the UK are using it to identify suspicious people based on their emotional states. It can also aid in the management of large gatherings or crowds. The authors in [258] developed a system that applied multi-modal data of human actions and facial expressions to obtain the emotional state, behavioral state, and behavioral semantic state of a crowd to predict potentially dangerous intentions. Detecting and monitoring collective emotional states at mass events, such as protests or sports games. This can aid in anticipating crowd behavior, enabling authorities to react more effectively to maintain safety and order.

The system must offer high real-time performance for immediate threat assessment with high in-the-wild feasibility due to varying conditions, angles, and poses of humans in public. Contactless operation and context awareness are highly prioritized. Processing complexity is set to moderate to ensure a real-time response. Noise sensitivity and sensitivity to environmental conditions are highly important due to the nature of the data. The need for ethical and privacy concerns is considered low when collecting data from strictly public spaces, which is allowed also for citizens in most countries. Deliberate enactment has a lower priority as spontaneous emotions are assumed. User discomfort should be minimal, with a focus on a non-intrusive and imperceptible experience despite necessary data collection. Data availability is moderate. Although there are lots of audio-visual datasets available, not many of those cover the demanding in-the-wild conditions (e.g., camera angles, poses, and noise levels) needed to analyze, e.g., surveillance feed.

Referring to Fig. 14, the comparative schema suggests that the audio-visual modality combination attains the

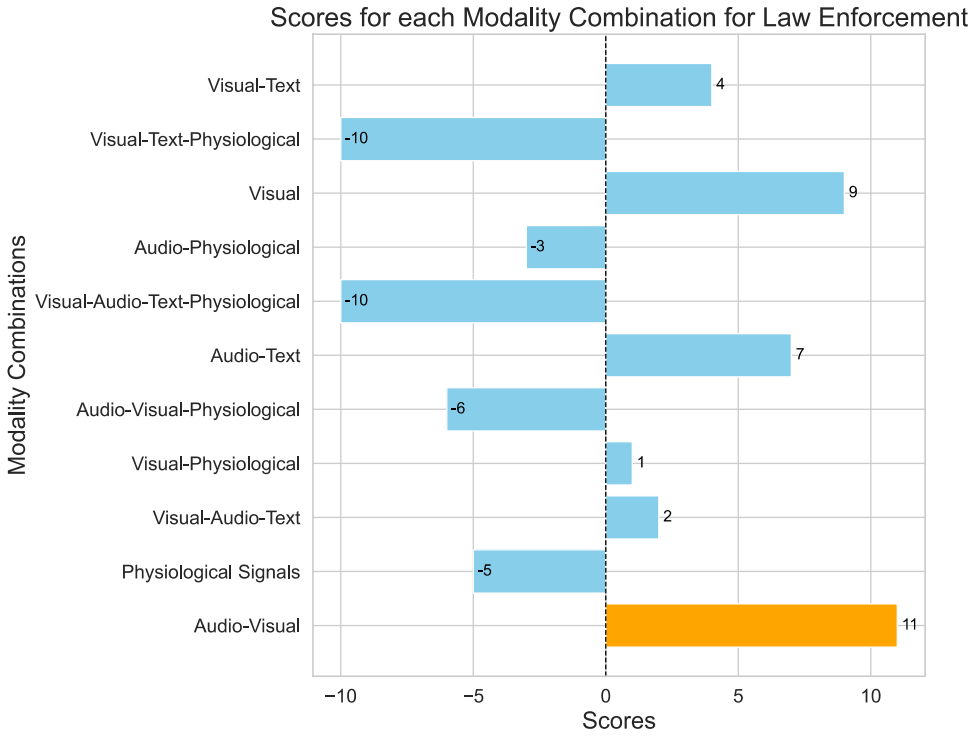

Fig. 14 Modality group selection for law enforcement and security

highest score when applied to law enforcement and security contexts. This combination is particularly effective due to its non-contact operation, sensitivity to environmental context, and adaptability to uncontrolled, real-world conditions. Consider a scenario of riot control and potential threat assessment in the busy downtown of a major city. The city's extensive network of surveillance cameras captures the visual cues of the crowd such as facial expressions and body gestures. Simultaneously, the wide array of microphones installed in appropriate locations collects audio cues such as the crowd's vocal intensity, speech rate, pitch, and tonal changes. By analyzing both the crowd's vocal patterns and visual behaviors in a synchronized manner, the system can recognize the collective emotional state of the crowd. See the studies [259, 260] for comprehensive details detecting the crowd's emotional states, datasets, opportunities, and prospects.

Implementing real-life solutions in law enforcement and security scenarios has several challenges. The quality of the data, especially visual data, is crucial for accurate emotion recognition. Monitoring a crowd, for instance, would require processing images of dozens or hundreds of people, demanding highly sophisticated and expensive camera technology. In addition to the high need for privacy, cultural and ethnic variations in emotional expression also pose challenges, as training datasets often predominantly feature one ethnic group, leading to potential biases in recognition accuracy [261].

8.4 Digital marketing and sales

Emotion recognition is becoming increasingly important in digital marketing and sales as it provides a deeper understanding of customer behavior and enables real-time personalization of user experiences. Emotions play a substantial role in consumer decision-making processes. If a business can recognize and respond to a customer's emotions in real time, it can significantly enhance the user experience, leading to improved customer engagement and potentially higher conversion rates.

We envisage a use case of personalized marketing that aims to elevate digital marketing and sales through CMER. It focuses on enhancing customer engagement via personalized experiences derived from real-time emotion detection. Customers interact with an e-commerce platform via a mobile application, navigating through a multitude of products and services. The online e-commerce platform aims to create a hyper-personalized shopping experience for its customers. To achieve this, they can integrate a CMER system into their platform to understand their emotional responses to different products or services.

The system must offer high real-time performance for immediate personalization based on users' emotions, thus boosting engagement and satisfaction. Moderate in-the-wild feasibility is acceptable due to the controlled nature of online browsing. Contactless operation is moderately prioritized, given the mobile app interface and the value of aligning with

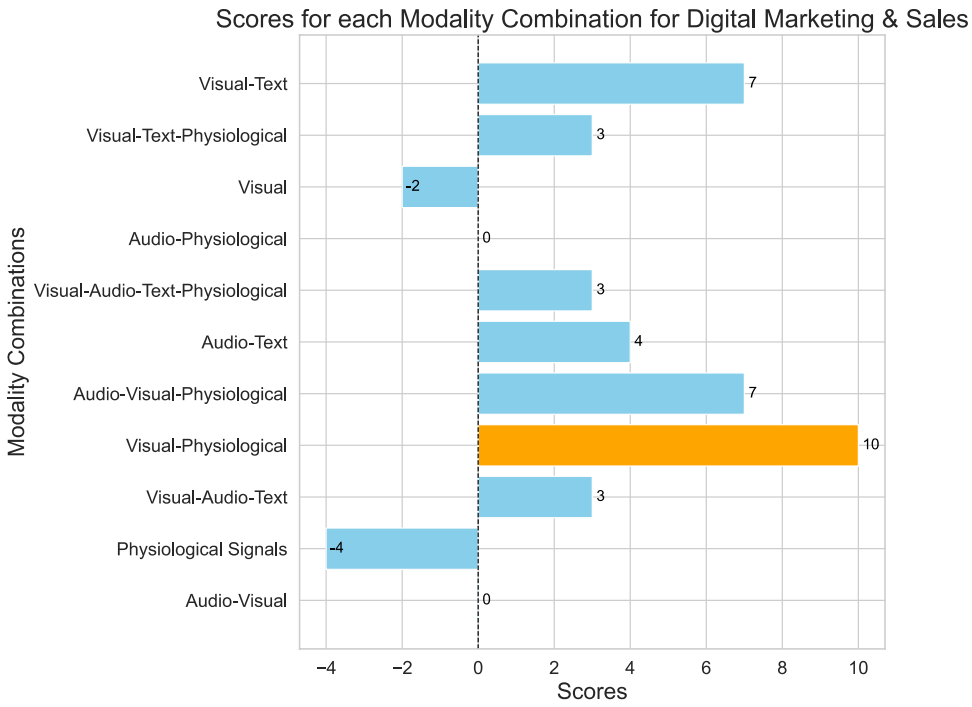

Fig. 15 Modality group selection for digital marketing and sales

user preferences. Context awareness is set high for capturing the interaction context. High processing complexity is required to integrate and interpret visual and physiological cues accurately, necessitating robust algorithms. Noise sensitivity, ethical, and privacy concerns, along with sensitivity to environmental conditions, are moderately important, considering the controlled browsing environment, user consent, and varying lighting conditions. Deliberate enactment has a lower priority as spontaneous emotions are preferred. User discomfort should be minimal, with a focus on a non-intrusive and aware user experience despite necessary data collection. Data availability is set to high because online marketing facilitates easy access to extensive user interaction data and existing datasets.

As per Fig. 15, these requirements align with the 'visual-physiological' modality combination. Facial expression recognition and eye tracking can be applied under the 'Visual' modality for emotion and attention detection, respectively. Under 'Physiological,' remote photoplethysmography (rPPG) can be used to infer emotional states from heart rate variability. Data collection can utilize the mobile device's front camera (by giving the option to enable their phone's camera), capturing visual data for expression and eye tracking and color changes for rPPG. Processing involves deep learning techniques for facial expressions, motion analysis for eye tracking, and signal processing for heart rate estimation from rPPG.

The demands of real-time performance and processing complexity for emotion interpretation necessitate high computational resources. Leveraging high-performance cloud computing and efficient algorithms can mitigate this. Context awareness poses challenges due to the intricacies of extensive user data analysis, addressable through AI models offering temporal understanding and adaptability. Moderate contactless operation challenges exist, particularly for rPPG data collection, and could be reduced by improvements in remote sensing technologies. Privacy concerns can be managed via robust encryption and user education on data usage. Noise and environmental condition sensitivity can be tackled with robust noise reduction algorithms and environmental adaptation techniques. User discomfort can be minimized by clear communication about data collection and its benefits.

8.5 Entertainment

In the entertainment industry, contactless multimodal emotion analysis has the potential to revolutionize user engagement and content delivery. By monitoring and analyzing users' emotions, such as interest or frustration [93], during activities such as gaming, movie-watching, or music-listening, the entertainment industry could obtain real-time user experiences and provide timely feedback. The content being offered could be adjusted and/or different recommendations based on the user's input could be provided. When combined with multiple-person analysis, contactless multimodal technology can enhance group-level entertainment activities, such as board games [86], by providing a more immersive experience that caters to the different skill levels of players, ensuring that everyone enjoys the game.

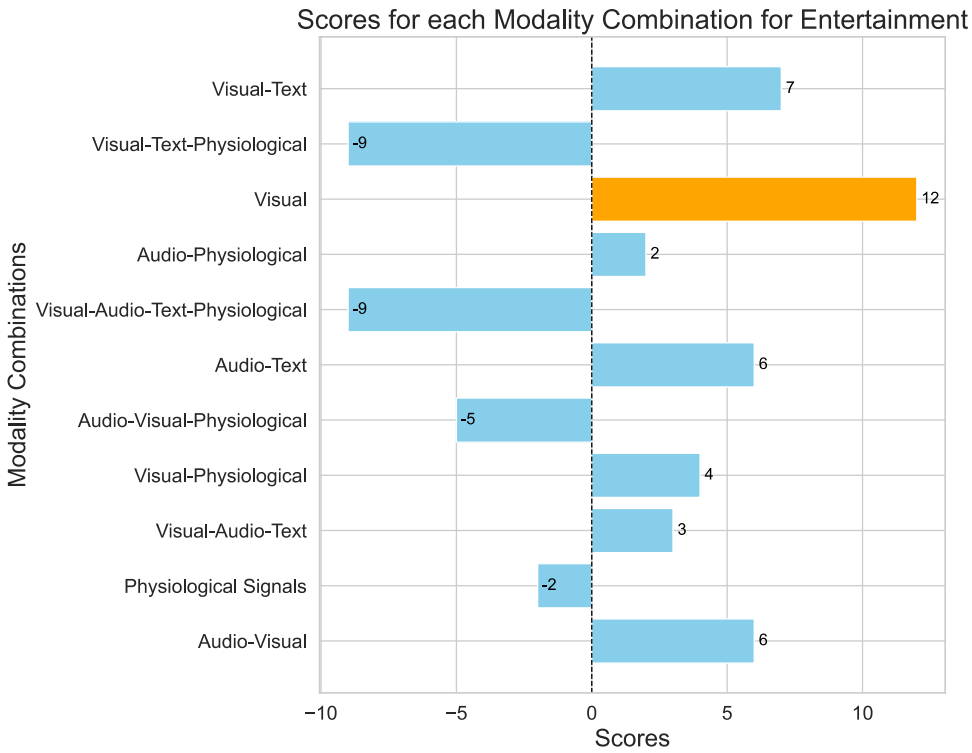

Fig. 16 Modality group selection for entertainment

Furthermore, in the Virtual Reality (VR) and/or Augmented Reality (AR) industry, contactless emotion analysis plays a crucial role that cannot be replaced by traditional contact-based multimodal emotion analysis. This technology contributes to a lighter VR/AR experience for users, allowing them to immerse themselves in a meta world without the need for physical interfaces or cables. Additionally, it enables personalized and real-time adjustments to the content based on the user's emotional feedback.

The employment of emotion recognition systems in entertainment applications is a new direction and thus has diverse requirements in different aspects. The real-time performance and in-the-wild feasibility of the system are crucial to effectively address the diverse needs of different users in a timely manner. We set the level of contactless interaction as moderate, considering that the audience in the entertainment domain can generally tolerate certain forms of contact-based devices such as earphones or VR/AR headsets. However, contactless solutions are more easily applicable and widespread in various conditions. The need for context awareness is moderate to provide accurate feedback in different scenarios, particularly in relatively stable situations. In the entertainment industry, the primary objective is to enhance immersion and enjoyment, resulting in relatively straightforward data requirements. Therefore, the processing complexity is set at a moderate level. On the other hand, noise sensitivity is set to a high level to accommodate different entertainment industries and accurately cater to customer needs. Ethical and privacy concerns are regarded as moderate, as this type of system is primarily used in relatively public settings, and the data collected is not as sensitive as in other domains like healthcare. The deliberate enactment is also considered moderate to ensure accurate predictions based on true requirements. Since most use cases occur in stable indoor environments, the sensitivity to the environment can be set to a low level. To enhance customer experience, minimizing user discomfort is prioritized, aiming for a low level of discomfort. Lastly, data availability is set high due to the pervasive use of digital media and interactive platforms in the entertainment industry, which provides a rich source of data for emotional analysis. Furthermore, the industry's forward-leaning stance towards adopting advanced technologies facilitates the collection and utilization of such data.

Figure 16 shows that visual modality emerges as the most suitable candidate for this use-case. Consider a scenario where an interactive movie experience is set up in a public space. Using high-resolution cameras equipped with emotion recognition algorithms, the system captures visual cues from the audience. These cues include facial expressions, body gestures, engagement patterns, eye movement, and even subtle changes in posture or orientation in real-time. The system detects fluctuations in emotions such as joy, fear, surprise, or boredom from the viewers and customizes the movie content dynamically. The idea of capturing emotions evoked by movies has been discussed in related literature [262]. The emergence of 'interactive cinema', where movies reshape their narratives based on the viewer's emotional

response, is making progress. The authors in [263] used an ECG headset and an eye tracker to measure viewers' emotional responses. However, a recent experiment [264, 265] relied solely on the audience's facial expressions and reactions to dynamically alter the movie content. The concepts of 'interactive cinema' and 'interactive movies' are still in their early stages. For a more in-depth understanding of interactive movies, consider referring to the study by Jinhao et al. (2023) [266].

In the mentioned scenario, environmental factors such as lighting conditions, crowd density, and diverse viewer orientations could complicate the task. Privacy and ethical concerns arise when collecting and analyzing data in public settings, especially without explicit consent. Potential solutions could involve addressing diverse environments and lighting conditions. Privacy-preserving techniques such as differential privacy could be used to anonymize data. Additionally, transparent privacy policies and opt-in consent could mitigate ethical concerns.

8.6 Traffic and transportation

The mental state and emotional condition of drivers impact their ability to safely navigate the complex and dynamic environment of road transportation systems [15]. In the realm of traffic and transportation management, AI-based emotion recognition technology is exhibiting transformative potential in enhancing safety and efficiency. One key application is real-time driver emotion monitoring, where the system can detect emotions such as stress, anger, or fatigue that often lead to accidents [15, 267, 268]. Problems in the field of traffic are often directly or indirectly related to emotions. Drivers experiencing intense emotions such as ecstasy, anger, or terror can significantly reduce self-control, making them prone to incorrect responses and potentially causing traffic accidents [11]. To balance that, vehicles need to be equipped with the capacity to monitor the state of the human user and to change their behavior in response [269]. By identifying emotional states, AI could initiate preventive measures, such as alerting the driver or temporarily assisting with vehicle control, thereby potentially reducing accident risks. Empathetic automation that responds to the driver's state could improve the experience and acceptance of self-driving vehicle drivers [270].

A promising avenue is the integration of emotion recognition technology within Advanced Driver Assistance Systems (ADAS; see, e.g., [271]). Such systems could adapt their level of intervention or automation based on the detected emotional state of the driver, thereby fostering a more responsive and intuitive driving environment. An intelligent vehicle could recognize and regulate drivers' emotions in various ways to improve driving safety and comfort [272]. Additionally, emotion recognition technology can be utilized to personalize in-vehicle information systems and cabin lighting, adjusting content or interface based on the driver's emotional state, ensuring optimal comfort, and minimizing distractions [273].

Several car manufacturers like BMW, Ford, GM, Tesla, Kia, and Subaru have implemented Driver Attention Warning (DAW) systems in their vehicles, utilizing sensors to caution fatigued drivers [274]. Exceptionally, Kia has introduced an emotion recognition system called R.E.A.D. to personalize the cabin experience based on a driver's emotional state by monitoring facial expressions, heart rate, and electrodermal activity. However, these systems lack emotion recognition. Studies highlight driver emotions as key influencers of behavior and accident risk [275]. Therefore, an emotion recognition-based system could enhance safety and personalization, surpassing fatigue-only detection. Besides drivers themselves, aggregated emotional data from drivers could inform infrastructure planning and traffic management strategies, highlighting areas or times of high emotional stress and suggesting potential interventions, and guiding urban planning [276]. Lastly, as autonomous vehicles advance, emotion recognition could serve in assessing the passengers' comfort and trust in the autonomous system, providing valuable feedback for system improvement.

In order to track the emotions of drivers and passengers, in most cases only contactless methods are feasible, although some contact-based sensors for EDA and HR could be embedded in the steering wheel (e.g., Kia Motors R.E.A.D concept). For drivers, facial expressions tend to be better for capturing changes in valence, whereas biosignals and speech tend to be better suited for detecting changes in arousal. Multi-modal approaches can significantly enhance emotion recognition performance [15].

The comfort of the driver is crucial. Real-time detection is very important as situations in traffic can change fast. Predictions made by the system can be limited to driver and car computers only, hence privacy issues are not a limiting factor for solutions. Finally, research has shown that dynamic driving and static life datasets were different in emotion recognition results [272], hence the context of data is important, which can limit the usefulness of existing datasets, setting the data availability to moderate. At present, there are no datasets available collected in spontaneous, real-life driving situations with annotations [15].

As depicted in Fig. 17, a combination of visual and physiological modalities aligns most closely with this requirement. A potential set of visual-physiological cues for this scenario is facial expressions, heart rate, and electrodermal activity. Regarding the measurement of electrodermal activity, two viable options exist. One option is to leverage the 'moderate' contactless approach and install sensors on the steering wheel to measure electrodermal activity, given that a driver's hands will consistently interact with the wheel

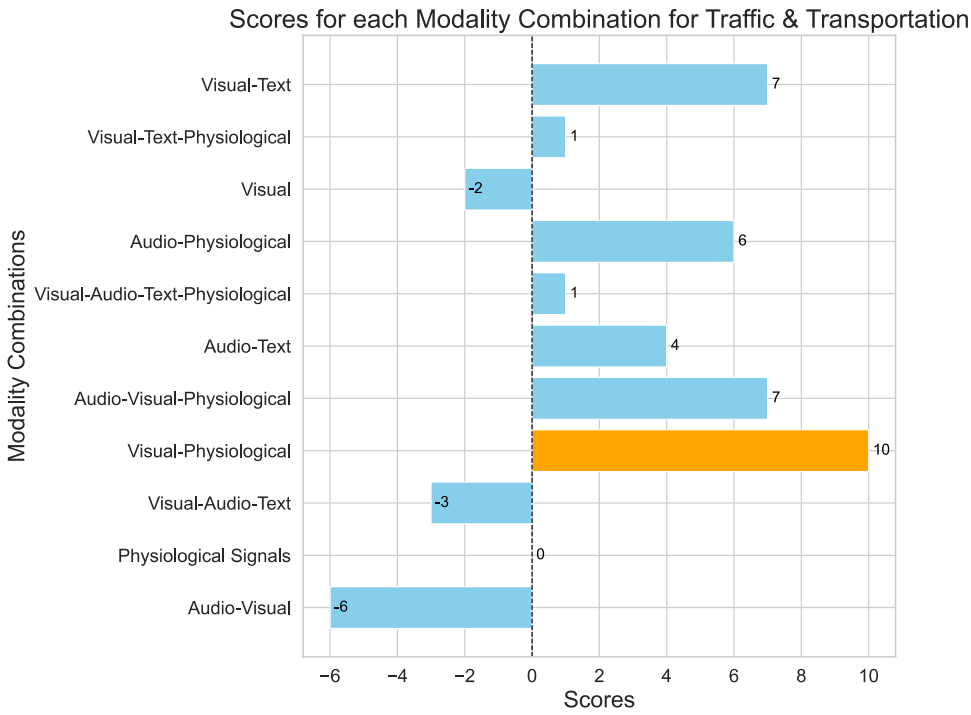

Fig. 17 Modality group selection for traffic and transportation

during driving. Alternatively, this physiological signal can also be measured through a fully contactless approach. For instance, a proof-of-concept study in [277] proposed the feasibility of measuring electrodermal activity using a camera.

Due to the necessity of hard real-time response, the processing complexity of the system is likely to increase. The system is intended to operate in a variety of daily conditions; therefore, while its susceptibility to noise is expected to be minimal in a closed environment, it must exhibit resilience against changing lighting conditions. To address the varying lighting conditions, near-infrared cameras could offer a viable solution. They are good at operating under poor lighting and provide high-quality images that can facilitate accurate facial emotion recognition as well as heart rate and electrodermal activity measurements.

8.7 Tourism and hospitality

Travel and hospitality experiences are based on visitor emotions, feelings, moods, satisfaction, loyalty, behavioral intentions, and other outcomes which have been studied in a plethora of studies [278]. There are a number of areas in tourism and hospitality in which emotion recognition can be used. FER has been of particular interest in tourism and hospitality, suggesting an enhanced value in the travel and tourism industry by providing personalized service delivery, special offers, value-based services, and optimized trip planning, based on recognizing customers' emotions [279]. Accommodation companies can use FER-based emotion recognition to identify which type of guest-room design elicits positive emotions in customers to improve customer experience [280]. Similarly, FER has also been used to analyze emotions and measure customer satisfaction during guided tours to measure tourists' emotions and satisfaction with the quality of service [281]. FER coupled with other cues can also be more effective for studying travelers' personal preferences and emotions. For example, screen-based eye tracking and FER can give clues about travelers' emotions during the booking process, such as navigating websites, understanding fees, and going through checkout, to understand their overall feelings about the booking process [164].

We discuss a few use-cases in tourism and hospitality where a contact-based approach with the relevant modalities and cues can be applied.

8.7.1 Improving destination experience

Measuring tourists' destination experiences necessitates examining both behavioral and physiological responses [282]. Given tourism's dynamic nature, moderate real-time performance and high 'in-the-wild' feasibility are crucial due to the unpredictable outdoor tourism environments. The high contactless operation, high context awareness, and moderate processing complexity balance computational efficiency with complex scenario handling. High noise sensitivity, high ethical and privacy considerations, moderate deliberate enactment capacity, high environmental condition sensitivity, and low user discomfort are prioritized. Data availability is set to moderate, owing to the nature of tourism environments, which allows for the collection of diverse

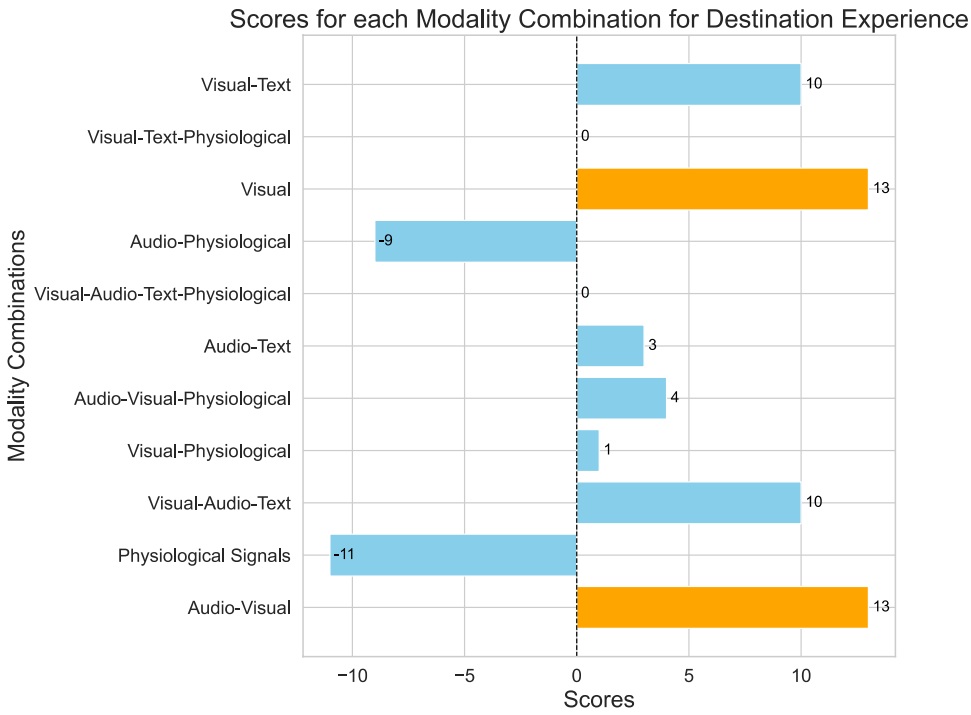

Fig. 18 Modality group selection for a destination experience

data, as well as the availability of existing datasets relevant to a wide array of possible tourist experiences.

As per Fig. 18, visual modalities are selected. A typical tourist experience often involves the visitor approaching monuments or significant attractions for a closer examination, creating an opportunity to capture several visual cues including physiological signals. Considering this, the selection of visual cues leans heavily towards behavioral responses such as FER and gestures. These cues are crucial to the understanding of the visitor's experience and can be efficiently gauged using traditional contactless methods [282]. Physiological responses such as heart rate, assessable via rPPG add depth to emotion recognition [283].

Ensuring real-time performance, crucial in dynamic tourist settings, is challenging due to the computational demands of processing visual cues and physiological signals like heart rate. This could be potentially mitigated with edge IoT devices for localized data processing and optimized machine learning algorithms. Unpredictable outdoor environments test the 'in-the-wild' feasibility, particularly the analysis of facial expressions and body gestures. However, robust computer vision techniques and diverse machine learning models could enhance performance. The requirement for contactless operation and context awareness increases system complexity, coupled with ethical issues regarding data privacy. Potential solutions include data anonymization, informed consent, strict privacy protocols, and public awareness initiatives to reduce resistance and promote system acceptability.

8.7.2 Personalized recommendations

Personalized recommendations based on tourists' emotions can enhance their tour experience [284]. A moderate real-time performance is set for personalization across diverse environments, thus establishing the need for high in-the-wild feasibility. A balance between contactless data collection and accurate emotion recognition is achieved with moderate contactless operation. Furthermore, high context awareness is crucial to understand the interplay between tourists' emotions, their undertaken activities, and the surrounding factors. A moderate level of processing complexity is deemed sufficient for effective data analysis. High noise sensitivity can better account for ambient noises typically present in tourist environments. High ethical and privacy considerations ensure user consent and data protection. While deliberate enactment remains low to facilitate natural behavior, high sensitivity to environmental conditions is crucial to adapt to various settings. Ensuring a low user discomfort level contributes to a pleasant and non-intrusive user experience. Data availability is set to low, indicating the need for more specific and varied datasets for the tourism domain.

Based on the suggested modality combination in Fig. 19 (visual-audio-text), the set of specific modality cues could be text reviews, facial expressions, voice sentiments, and location data from mobile devices. Data can be sourced from tourist posts on social media and location tracking from mobile phones. This location data provides context to the physical activities of the tourist, where changes in movement patterns can indicate different emotional states [285]. The

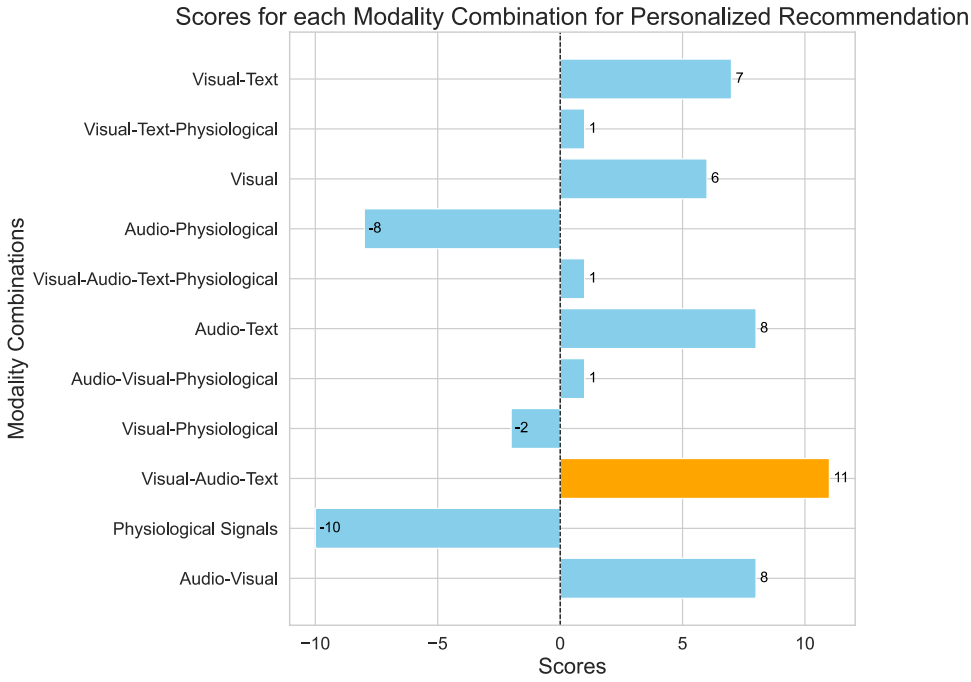

Fig. 19 Modality suggestion for personalized recommendations

personalized recommendation could be further enhanced by collecting data directly from tourists and creating a personal tourism profile in a digital travel companion app. This profile, storing tourist preferences and past experiences, enables the system to make specific recommendations tailored to individual likes and interests. These specific recommendations encourage tourists to provide regular updates, including location, activities, feedback, and mood changes. These continual updates refine the recommendation algorithm, promoting a cycle of continuous improvement that further enhances the user experience.

Local processing on mobile phones, while addressing latency and privacy issues associated with remote servers, faces energy constraints. Efficient algorithms can be the key to overcoming these challenges. Privacy concerns necessitate stringent data privacy measures, user consent, and data anonymization. Diverse usage environments might affect performance; however, robust models trained on diverse datasets can provide consistency. User discomfort can be minimized with transparent communication, user-friendly interfaces, and user-controlled data collection. Encouraging frequent data updates through gamification or incentives can ensure system accuracy. Context misinterpretation can be managed by integrating various contextual cues and efficient algorithms for precise recommendation generation.

8.7.3 Feedback analysis

Emotion recognition technology in hospitality, including resorts, guest houses, and hotels, provides an innovative way to understand customer experiences. Traditional feedback methods, like questionnaires, have limitations, including a narrow scope and lack of automated analytics, making systematic, long-term analysis difficult. A multimodal, contactless emotion recognition system could overcome these constraints, providing richer insights into customer experiences.

In this context, moderate real-time performance suffices, allowing efficient emotion data processing without constant immediacy. The system requires high context awareness to interpret emotions in relation to the environment and activities. High processing complexity caters to diverse emotional states and data nature, while high noise sensitivity acknowledges outdoor ambient noise. Upholding stringent ethical and privacy standards is crucial. The system's requirement for deliberate enactment is high, accommodating spontaneous and deliberate customer behaviors. Moderate environmental sensitivity ensures system robustness in varying outdoor settings, and low user discomfort maintains a pleasant customer experience. Data availability is set to low due to the scarcity of publicly available emotion databases in the hospitality sector, and the potential sensitivity and privacy concerns surrounding the collection of such data.

Unlike the 'improving destination experience' case, here, capturing physiological signals through the rPPG method suits destination settings where tourists approach specific sites. With visual-audio-text modalities (Fig. 20), feedback analysis combines data on facial expressions, body language, voice sentiment, and text reviews. However, stringent privacy rules make data collection in private spaces, like hotel rooms, impracticable. Outdoor environments offer a viable alternative. Accommodation providers can install smart digital billboards with audio-visual sensors in places like

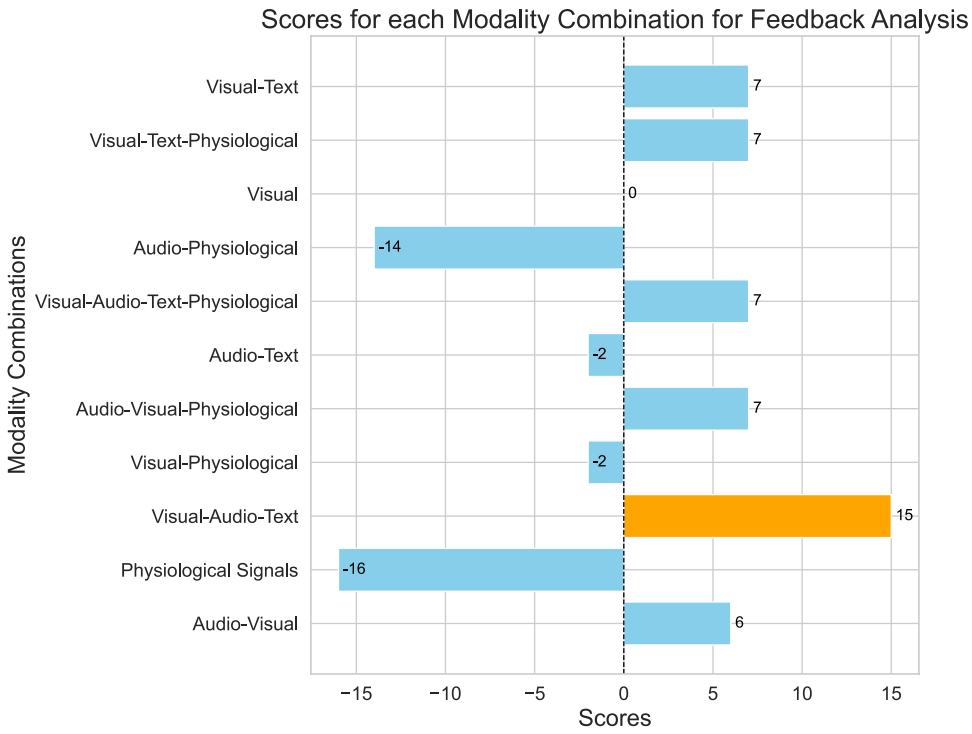

Fig. 20 Modality selection for customer feedback analysis

gardens and pathways. These billboards, displaying promotional content, can concurrently gather emotion-related data by analyzing facial expressions, body language, and voice sentiments. A similar method can be employed at activity check-in kiosks, adding text reviews from guests to assess satisfaction levels. This data integration provides a comprehensive view of the customer's experience. For example, if a guest's positive review aligns with their emotional data, it confirms a good experience. However, discrepancies between review and emotion data warrant further staff investigation.

Interpreting emotional signals in relation to diverse, unpredictable environments demands complex context-aware understanding. This can be mitigated by integrating various context-aware signals and robust algorithms. The system's high noise sensitivity, potentially causing false positives or negatives in noisy environments, necessitates efficient noise reduction methods. Strict ethical and privacy standards impose limitations on data collection and processing, requiring stringent data protection measures, anonymization, and informed user consent. The system's potential to cause user discomfort due to continuous data collection can be offset by user-friendly interfaces, transparent data usage communication, and user control over data collection. In addition, the system's dependency on accurate data can lead to inaccuracies if data sources are unreliable or absent, though this can be managed by using redundant sensors and integrating multiple modalities for a comprehensive understanding of the user's emotional state and context.

9 Ethical and privacy considerations

The deployment of CMER systems necessitates robust attention to ethical and privacy considerations. By integrating multiple signals from diverse data sources, these systems can capture the subtleties of emotions more effectively. However, this multimodal approach enhances the complexity of ethical and privacy issues as it facilitates a more detailed representation of an individual compared to a single data source. The ethical and privacy concerns are intensified in the CMER context due to the unobtrusive nature of data collection, which often occurs without explicit user knowledge. This implicit methodology underscores the urgency for comprehensive and stringent ethical and privacy frameworks.

9.1 Ethical issues

The deployment of CMER systems in environments where people work, interact, and collaborate, presents several ethical considerations [286]. While existing ethical guidelines, such as those proposed by the High-Level Expert Group on Artificial Intelligence [287] and the World Health Organization [288], can provide a foundational framework for instituting key ethical principles in CMER, user acceptance of such systems remains a unique challenge. The inherent limitations of AI models' furnishing explainability and transparency could be one of the major reasons for embracing CMER systems in various fields, particularly when deployed in high-stakes scenarios such as healthcare [289]. This absence of

interpretability hinders users' comprehension of the underlying decision-making process, thereby inhibiting their willingness to adopt these systems extensively. For instance, consider a CMER system designed for pain detection using FER and body movement analysis. Compared to existing self-report or observer-based pain detection methods, such a system offers the potential for more timely interventions. Despite the evident efficacy, patients and caregivers may harbor ethical concerns, rooted in skepticism and limited knowledge about the technology, which could constitute a significant impediment to the successful deployment of this system [290].

In other critical domains, such as law enforcement, the necessity for explainability and transparency becomes pivotal. For instance, a CMER system deployed for detecting potential threats or suspicious behaviors in public spaces must exhibit high levels of transparency and explainability. Authorities making decisions based on the CMER output should be able to understand and explain the criteria used by the system to flag an individual as a potential threat. Without this understanding, it would be difficult to justify the system's decisions, which could potentially lead to false accusations and erosion of public trust in the system.

Apart from the need for transparency and explainability, several other ethical challenges emerge when deploying CMER systems. These encompass potential emotional manipulation, intrusions into personal boundaries, and apprehensions around possible punitive actions based on the emotion recognition results. For instance, in education, the use of technology to probe into a student's emotional state may be seen as a violation of their right to emotional autonomy and space, even in an academic setting. Similarly, in the sphere of digital marketing and sales, the question of emotional manipulation arises. If marketers have detailed insight into a consumer's emotional state, it could be used to influence purchasing decisions in a way that exploits emotional vulnerabilities. This leads to concerns about whether such strategies might overstep ethical boundaries, leveraging emotions to encourage consumer behavior that may not be in their best interest.

In the scenario of traffic and transportation, while the primary objective of CMER systems may be to enhance safety, there can be concerns about the potential for these systems to be used for disciplinary purposes. For instance, the systems could be used to issue traffic violations based on perceived driver stress or distraction, potentially leading to concerns about fairness and proportionality. Furthermore, ethical questions might arise about the consent process. Given the ubiquity of transportation and the critical nature of these services, it is essential to ensure that users are not merely presented with a binary choice of using the service with CMER or not using it at all. There needs to be a clear, informed, and accessible mechanism for users to choose such monitoring without sacrificing their ability to use the transportation service.

When developing CMER applications, one must need to pay attention to potential biases in the system that might put individuals in unequal positions based on, e.g., gender, age, or skin color. One must verify that the data used in training the system contains samples coming from people of various nationalities, groups, and minorities. Systemic and implicit biases such as racism and other forms of discrimination can inadvertently manifest in AI through the data used in training, as well as through the institutional policies and practices underlying how AI is commissioned, developed, deployed, and used [291]. This must be taken into account when planning data collection or when applying open datasets, e.g., those listed in Sect. 5.

9.2 Privacy issues

Addressing privacy concerns is crucial in the deployment of CMER systems, especially in sensitive sectors like healthcare. Patient apprehensions regarding privacy can be influenced by numerous factors, including demographic traits, data type, collection context, and institutional reputation, which can all elevate privacy concerns [292–294]. Past experiences with medical research can also shape privacy perceptions [295]. Notably, the perceived benefits of data collection and the patient's health status significantly influence privacy attitudes [296, 297]. Patients might exhibit an increased willingness to consent to data use if they understand the potential benefits, such as improved diagnosis or treatment planning, derived from CMER system analytics.

Taking the example of personalized recommendations in tourism, collecting FER data at various attractions, recording vocal sentiments during interactions, and tracking physical movements across locations carry the potential to infringe upon a tourist's privacy. Visitors may express apprehensions about how their emotional data is being used, stored, or potentially shared with third-party entities. Further, the lack of transparency around the AI decision-making process in providing personalized recommendations may add to these concerns.

Thus, it is essential to strike a balance between personal data privacy and perceived benefits when crafting data collection protocols and consent forms. By elucidating the benefits, we can foster trust, encourage data sharing, and promote broader acceptance of CMER systems in healthcare, aligning with the premise that user benefits support more extensive data usage [297]. In the European Union, the requirements of the General Data Protection Regulation (GDPR) and regulation on artificial intelligence (AI Act) must be followed in designing and implementing MER and CMER applications.

Table 4 Key ethical principles for CMER systems

Key ethical and privacy principles	Meaning for CMER systems
Human Autonomy	Participants retain the right to seek information about research, care, or study procedures. They have the right to determine their level of participation in processes involving AI-based systems
Human Well-being	Participants should feel comfortable and anticipate potential benefits when participating in processes employing AI-based multimodal methods. These benefits may be direct or contribute to research outcomes that can benefit others
Fairness	Research, care, or study settings, along with collected data, should not instigate discrimination, biased results, or participant stigmatization. Training data should be representative of real-life data. This principle remains crucial when results are disseminated or published
Privacy, Consent, and Perceived Benefits	Participants are required to provide informed consent for the use, sale, or sharing of their data, while also understanding the potential benefits of CMER systems. They should feel their privacy is preserved and that their data, along with the AI systems in use, are responsibly managed. This balance between privacy concerns and perceived benefits fosters greater trust and acceptance of CMER systems
Technological Trustworthiness	Data management procedures and AI systems should promote transparency, explainability, and responsibility to the maximum extent technologically possible. Practitioners and researchers involved in the planning, management, and dissemination of research or studies must foster transparency and responsibility by openly describing the procedures and devices employed to participants

9.3 Ethics and privacy principles for CMER

Based on the ethics and privacy discussion in several domains [287, 288, 290, 298], we highlight five key ethics and privacy principles to be considered while planning and deploying CMER systems. These principles are summarized in Table 4.

10 Summary, challenges and future directions

A predominant observation in the existing studies on multimodal recognition is the emphasis on non-verbal cues, particularly physiological signals such as EEG, ECG, and EDA. Although these modalities offer comprehensive and accurate emotional data, their practical implementation outside a controlled environment remains challenging due to technical limitations and practical inconveniences. In the realm of behavior cues, facial expression recognition emerges as a popular choice due to its natural and intuitive nature, supported by advancements in computer vision and machine learning. This is complemented by the inclusion of body gestures, audio features, textual analysis, and even contactless physiological signals which enhance emotion recognition capabilities by providing additional, contextually rich data.

Ethical and privacy concerns that are linked to the rapid progress of artificial intelligence are often overlooked. Additionally, computational complexity, a significant factor affecting the real-time performance of emotion recognition systems, remains largely unexplored in most studies. Another significant gap lies in the insufficient attention to system robustness, reliability, and context awareness. These factors are especially pivotal for systems deployed in critical applications such as healthcare, where consistent performance, context sensitivity, and resilience to external disturbances are crucial. Moreover, transparency and explainability are also inadequately addressed in existing MER studies. Understanding how these systems reach certain decisions is crucial, particularly for applications in sensitive domains. In addition, the diverse combinations of modalities employed in emotion recognition, particularly the combination of audio-visual-text-physiological data, are promising but necessitate further exploration. While the text modality has lower coverage, its integration with other modalities, especially in the context of advanced conversational agents, should not be dismissed.

Audio-visual modality combination is dominant in CMER systems. The accessibility of these modalities and the sophistication of their analysis techniques, thanks to the evolution of machine learning and computer vision, contribute to their widespread adoption in CMER. However, this observation also underlines a notable imbalance in the exploration of other modality combinations. In particular, the incorporation of text and physiological signals in CMER systems remains substantially under-researched. This stands in contrast to their potential to enrich the emotion recognition process significantly by contributing additional layers of context and depth to the interpretation of emotional cues. The text modality, for example, is particularly relevant in contemporary scenarios where human–computer interaction is becoming increasingly conversational, and much of the user's input is text-based. As for physiological signals, despite the challenges associated with contactless acquisition, our use-case discussions highlighted their potential

value in CMER systems. We examined the possibilities of harnessing non-contact methods for extracting physiological signals such as heart rate and respiratory rate.

There is a significant reliance on categorical and dimensional models to define emotions, with most studies utilizing a single description model. However, the potential benefits of a more comprehensive analysis, incorporating both models, remain largely unexplored. This finding emphasizes the need for a broader, multi-model approach in future studies to capture the subtle complexities of human emotions more effectively.

The CMER's comparative schema, constructed upon analysis and benchmarking of extant studies, stands as a decision-support tool, aiding researchers and practitioners in making informed choices about modality combinations based on specific application scenarios. The rich discussions centered around implementing CMER with an assessment model in various use-cases provide practical insights into the intricacies of real-world deployment and the importance of this framework. The detailed analyses are designed to equip researchers with a solid understanding of how to tailor modality combinations to meet unique situational demands. The formulation of the CMER framework is the first step towards a more capable, context-aware, and integrated platform that systematically optimizes the usage of multimodalities for emotion recognition and is fine-tuned to cater to the distinct necessities of varying application scenarios.

Despite the transformative potential of CMER, ethical and privacy concerns pose significant challenges to its deployment, especially in sensitive sectors such as healthcare, law enforcement, education, and digital marketing. The necessity for transparency and explainability in AI systems, especially in critical scenarios, surfaces as a critical finding. Concerns around emotional manipulation, intrusions into personal boundaries, and potential punitive actions based on emotion recognition outcomes entail ethical and privacy principles. With these considerations, five pivotal ethics and privacy principles emerge for the deployment of CMER systems. These principles, encompassing human autonomy, human well-being, fairness, privacy, consent, perceived benefits, and technological trustworthiness, offer an invaluable compass for subsequent research and applications. By adhering to these principles, the evolution and implementation of CMER systems can be ensured to be ethically robust and respectful of user privacy.

The development of an efficient CMER system entails training effective AI models, necessitating rich and diverse data. To encompass a broad emotional spectrum, work efficiently in non-ideal conditions, and maintain robust functionality under varying environmental conditions, it is essential to secure datasets marked by a rich prevalence of genuine emotional instances. This should include instances of complex emotional states, culturally diverse expressions, and emotions under situational influences, amongst others. Open datasets hold the potential to significantly alleviate the challenges posed by data acquisition. As revealed by our in-depth analysis, the relevant open datasets do have a wide range of emotions and a higher number of combinations of contactless modalities, offering a substantial foundation for training CMER systems. The integration of AI generative models could further improve these datasets by enabling data augmentation to amplify their utility in developing more robust and generalizable emotion recognition models.

10.1 Potential challenges and suggestions

The following sections provide an analysis of potential risks associated with CMER systems and the potential mitigation strategies.

10.1.1 Data heterogeneity

CMER systems, due to their use of diverse sensing technologies, modalities, and cues, inherently face heterogeneity in data collection. Such diversity necessitates sophisticated fusion methods capable of seamlessly integrating multimodal data, considering temporal alignment and cross-modality influences, and accounting for the temporal dynamics of emotions and the synchrony between different modalities to ensure accurate emotion recognition [299].

Multimodal fusion faces the issue of certain modalities occasionally being absent or having poor quality, and the unpredictability of which modality might be missing adds to the complexity [164, 300]. There could be several reasons for missing modalities such as obscured or damaged sensors [301], errors in data collection [302], and absence of verbal and/or textual cues due to the inactivity of users [70]. Several advanced techniques for addressing the missing and/or uncertain modalities can be adopted such as meta-sampling [302], self-attention fusion [164], exploiting modality-invariant features [301], and missing modality imagination network [300].

10.1.2 Inter-individual and cultural variations

Emotion expression and perception are significantly influenced by individual characteristics and cultural backgrounds [303] which can potentially affect CMER system performance. To counter these variations, demographic and cultural metadata should be incorporated during system training to support adaptability across diverse users and cultural contexts. Personalized adaptation techniques should also be embedded into CMER system design, enabling systems to learn from individual user interactions and progressively fine-tune emotion recognition capabilities. Such personalized tuning aids in bridging the gap between generic models

and individual-specific emotional patterns, improving system efficacy and user experience.

10.1.3 Misinterpretation of emotional states

The potential for misinterpretation of emotional states presents a considerable challenge in the application of CMER systems. There are several reasons leading to misclassifications of emotions such as noisy or missing modality and inaccurate fusion [148], insufficiently labeled data, and complexity of emotions [304]. Incorrectly classified emotional states can lead to improper responses, which could have dire consequences, especially in sensitive fields like healthcare and law enforcement. For instance, in a scenario where a CMER system is used for depression detection, misinterpreting emotional cues like sadness or apathy as normal could lead to overlooked or postponed depression diagnoses. Sources of such misinterpretations could include inadequate training data, inherent algorithmic biases, or the inability of the system to fully comprehend emotional context.

It is crucial to ensure the CMER systems are trained on comprehensive and diverse datasets to reduce the likelihood of misinterpretations. Additionally, the development of algorithms should take into account possible biases and aim to minimize their impact. Moreover, integrating an understanding of the context into the CMER systems could offer a better interpretation of emotions. This might involve leveraging auxiliary information or employing more sophisticated models capable of understanding the intricate dynamics of human emotions.

10.1.4 Algorithmic bias

Algorithmic bias [305] poses a significant risk to the integrity of CMER systems. This bias can stem from either skewed training data or inherent biases in the design of the machine learning algorithms themselves. The resultant bias could potentially cause unfair or discriminatory outcomes, undermining trust in the CMER systems and possibly leading to detrimental consequences. For example, consider a CMER system developed for personalized tourism recommendations in a multicultural city like New York, but trained primarily on data collected from Western tourists. When deployed to assist a visitor from Japan, it might not optimally align its suggestions to the cultural preferences or interests of this tourist. For instance, it could overemphasize recommendations for steakhouses while failing to highlight the city's extensive offerings of sushi bars.

The training datasets should be representative of the diverse population that the system Is expected to serve, including different demographics and cultural backgrounds. This could reduce the likelihood of bias arising from unrepresentative training data. Furthermore, techniques like bias auditing [306], where the outcomes of an algorithm are examined for potential disparities, and bias mitigation methods, such as preprocessing the data or postprocessing the model outputs, could be employed to tackle algorithmic bias.

10.1.5 Lack of generalization and misclassification risks

The risk of lack of generalization leading to the misclassification of emotional states presents another significant challenge. Given the high individual and cultural variability in emotional expressions, CMER systems might, in their attempts to establish generalized models of emotion recognition, overlook these subtleties, leading to inaccurate predictions. For example, In the educational realm, a CMER system might misinterpret subtle signs of student engagement, such as attentive listening, as disengagement if it is predominantly trained on overt engagement signals like active participation or positive facial expressions. This could lead to inappropriate interventions. Similarly, in telehealth consultations, a CMER system trained primarily on clear signs of discomfort like grimacing or restlessness may overlook patients who exhibit discomfort less explicitly. The lack of generalizations primarily stems from the scarcity of real training data or the data all [304]. With ample data acquired under uncontrolled conditions, CMER systems can better discern spontaneous behaviors, leading to improved generalization for unseen data.

10.1.6 Cybersecurity risks

The deployment of CMER systems inherently invites cybersecurity threats that could jeopardize system integrity, user privacy, and misuse of the system for malicious activities. Potential risks include data breaches, unauthorized system access, and performance tampering. Robust security measures, including strong encryption protocols, stringent access controls, and comprehensive threat monitoring systems, need to be integral parts of CMER deployment to mitigate these risks. Additionally, regular system audits and security updates can further strengthen the resilience of CMER systems against emerging cybersecurity threats.

10.2 Future directions

Rapid advancements in sensing technologies, machine learning algorithms, fusion methods, and data acquisition techniques are paving the way for ubiquitous CMER systems. The advancement in deep learning is significantly enhancing the efficiency of vision algorithms. These improvements enable more precise measurement of physiological signals, such as respiratory rate, by extracting respiration waveforms from the RGB camera of the chest area [307], thereby marking a significant leap in the evolution of CMER systems.

The rPPG technique, notably, enables contactless and precise measurement of physiological signals previously only accessible through invasive methods. Recent research has seen rPPG successfully employed with everyday devices like smartphone cameras to measure heart rate, breathing rate, heart rate variability, and SpO2 levels [308]. To compensate for poor lighting, thermal sensors, multimodal sensors, and near-infrared sensors with near-infrared illumination have also been recommended for measuring physiological signals such as heart rate [309]. These metrics, representing crucial emotion-related cues within the visual modality, hold the potential to considerably augment the precision and efficacy of CMER systems. Parallel to this, the evolution of deep learning techniques is transforming signal estimation. Their application in estimating rPPG signals is notably improving accuracy [310], thus boosting CMER system performance. These advancements offer a promising trajectory for future development.

Radar technology's application in physiological sensing is an expanding field of research, with continuous efforts dedicated to enhancing the technology's precision and dependability. Various radar technologies are being harnessed to gauge physiological signals, including continuous-wave radars for respiratory rate measurement [115], millimeter-wave radar for heart rate and respiratory rate [311], impulse radio ultra-wideband radar for chest movement tracking [312], and Doppler radar for assessing respiration rate, heartbeat, and body movements [313]. Moreover, a reconfigurable intelligent surface-based 4D radar has been developed to measure respiratory and cardiac signals [314]. These methods, along with the recent strides in capturing advanced physiological signals like ECG and EEG using electric potential sensors [315], albeit with current range limitations, present an encouraging trajectory for CMER systems. Another study has presented a proof-of-concept for measuring electrodermal activity using an RGB camera [277]. These non-contact sensing methods promise significant improvements in emotion recognition.

Emerging technologies are expanding the scope of contactless methods for capturing behavioral and non-behavioral cues. For instance, Laser Doppler Vibrometry has demonstrated progress in heartbeat detection [316] and the measurement of facial muscle activity, suggesting potential alternatives to physiological signals like EMG [317]. Concurrently, research on WiFi signals is uncovering their potential for capturing diverse behavioral and non-behavioral cues, including gestures [318], respiration rate [319], and patterns of activity and gait [70, 320]. While these methodologies are not currently emotion-specific, they represent promising pathways for future integration into CMER systems.

Progress in sensing technologies, increased computational capacity, the ubiquity of computing devices, significant progress in machine learning, and the vast range of information signals within our environment collectively chart a promising path forward for CMER systems. The forthcoming era sees the collection of multimodal data as a commonplace procedure, akin to harnessing off-the-shelf components. The recent advancements in artificial intelligence over the past decade have substantially enhanced human–computer interaction, prompting the necessity of CMER systems and predicting their inevitable ubiquity. The advent of Large Language Models (LLMs), the emergence of conversational agents, AI-empowered search engines, and a plethora of AI tools have further redefined the interaction landscape, forging a more immersive human–computer interface that holds immense potential for future growth.

Given the current pace of advancements in artificial intelligence, the dynamic evolution of AI toolkits, and the forthcoming multimodality of conversational agents, it is possible to predict the trajectory of human–computer interaction and the enhanced contextual awareness of machines. With context awareness closely linked to emotion recognition, CMER systems will be instrumental in enabling machines to comprehend and respond to human emotions more intuitively. These systems will provide personalized and customized user experiences, driving the next wave of innovation in AI and human–computer interaction.

Future research in CMER systems should focus on the incorporation of emerging sensing technologies to provide avenues for the integration of more contactless modalities. This would significantly enhance the robustness, reliability, accuracy, and context-awareness of CMER systems, especially in critical applications such as healthcare or law enforcement. The acquisition of modality data from off-the-shelf and readily available sources necessitates exploration as it provides a more practical and cost-effective way to collect diverse and continuous emotional cues, thus facilitating in-the-wild applications of CMER systems.

The development of methods to mitigate environmental impacts, including ambient noise, lighting and weather conditions, occlusions, and user motion artifacts, is paramount to ensure the robustness and versatility of CMER systems. Such advancements would enable CMER systems to operate effectively in varied and unconstrained environments, expanding their application range and improving their practical usability.

Given the inherent heterogeneity in multimodal data and the possible absence of certain modalities due to environmental or sensor issues, developing efficient fusion methods becomes essential. These methods would enable the seamless integration of diverse and incomplete data sources, improving the robustness and resilience of CMER systems under challenging conditions.

Further exploration and enhancement of the CMER's comparative schema introduced in this paper are encouraged. As a pioneering attempt towards a more generic and

context-aware framework, it lays a foundation that can be built upon to accommodate more sophisticated applications and complex environments, thereby achieving a truly ubiquitous emotion recognition system.

As the context awareness and ubiquity of CMER systems increase, ethical and privacy issues will inevitably surface with more complexity. The five principles of ethics and privacy proposed in this paper serve as an initial roadmap, but further investigations are required to identify and incorporate additional factors, ensuring comprehensive protection of user privacy and ethical integrity.

Last but not least, the realm of explainability and transparency in MER, a largely under-explored area, warrants active research. As we entrust machines with the delicate task of emotion recognition, the ability to understand and scrutinize the decision-making process of these systems becomes crucial. This would ensure their accountability, increase user trust, and allow for continual improvements based on discernible insights.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00530-024-01302-2.

Acknowledgements This research was primarily supported by the AI Forum project (Grant no. OKM/116/523/2020), funded by the Ministry of Education and Culture, Finland. Additionally, the research was supported in part by the following projects: i) AI Driver project (Grant no. OKM/108/523/2021), funded by the Ministry of Education and Culture, Finland, ii) the Finnish Cultural Foundation for North Ostrobothnia Regional Fund (Grant no. 60231712), and iii) the Instrumentarium Science Foundation (Grant no. 240016).

Author contributions Umair Ali Khan contributed to writing and shaping the major part of the manuscript. Qianru Xu and Yang Liu focused on evaluating contactless emotion recognition studies, contributing to Sect. 5 and multiple subsections in Sect. 6. Altti Lagstedt provided expertise in analyzing emotion recognition studies and developed a benchmarking framework that supports our research. Ari Alamäki led the discussion on ethical and privacy concerns, crafting a compelling narrative around these aspects. Janne Kauttonen conducted a comprehensive analysis of open datasets, strengthening the paper's empirical foundation. Additionally, he contributed to the exploration of use-cases in Sect. 6, adding depth and practical relevance.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Cabanac, M.: What is emotion? Behav. Processes **60**, 69–83 (2002)
2. Feidakis, M., Daradoumis, T., Caballé, S.: Emotion measurement in intelligent tutoring systems: what, when and how to measure. In: Third International Conference on Intelligent Networking and Collaborative Systems. pp 807–812 (2011)
3. Damasio, A.R.: Emotion in the perspective of an integrated nervous system. Brain Res. Rev. **26**, 83–86 (1998)
4. Scherer, K.R.: What are emotions? And how can they be measured? Soc. Sci. Inf. **44**, 695–729 (2005)
5. Gonçalves, V.P., et al.: Assessing users' emotion at interaction time: a multimodal approach with multiple sensors. Soft. Comput. **21**, 5309–5323 (2017)
6. Szwoch, M., Szwoch, W.: Emotion recognition for affect aware video games. In: Image Processing and Communications Challenges 6. pp 11–20 (2015)
7. Liu, H., e. al.: Review on emotion recognition based on electroencephalography. Front. Comput. Neurosci. **15**, 84 (2021)
8. Wang, W., et al.: Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. Complexity **2020**, 1–9 (2020)
9. Tanko, D., et al.: Shoelace pattern-based speech emotion recognition of the lecturers in distance education: ShoePat23. Appl. Acoust. **190**, 108637 (2022)
10. Hasnul, M.A., Ab Aziz, N.A., Alelyani, S., Mohana, M., Abd Aziz, A.: Electrocardiogram-based emotion recognition systems and their applications in healthcare-a review. Sensors (2021). https://doi.org/10.3390/s21155015
11. Xia, H., Wu, J., Shen, X., Yang, F.: The Application of Artificial Intelligence in Emotion Recognition. In: Proceedings of the 2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI). pp 62–65 (2020)
12. Shoumy, N.J., Ang, L.-M., Seng, K.P., Rahaman, D.M.M., Zia, T.: Multimodal big data affective analytics: a comprehensive survey using text, audio, visual and physiological signals. J. Netw. Comput. Appl. **149**, 102447 (2020). https://doi.org/10.1016/j.jnca.2019.102447
13. Ribeiro, B., et al.: Deep learning in digital marketing: brand detection and emotion recognition. Int. J. Mach. Intell. Sens. Signal Process **2**, 32–50 (2017)
14. Cao, W., et al.: Video emotion analysis enhanced by recognizing emotion in video comments. Int J Data Sci Anal **14**, 175–189 (2022)
15. Zepf, S., Hernandez, J., Schmitt, A., Minker, W., Picard, R.W., et al.: Driver emotion recognition for intelligent vehicles: a survey. ACM Comput. Surv. **53**, 1–30 (2020)
16. Sini, J. et al. Passengers' emotions recognition to improve social acceptance of autonomous driving vehicles. In: Progresses in Artificial Intelligence and Neural Systems. pp. 25–32 (2020)
17. Tan, L., et al.: Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space–air–ground integrated intelligent transportation system. IEEE Trans. Intell. Transp. Syst. **23**, 2830–2842 (2021)
18. Chatterjee, et al.: Real-time speech emotion analysis for smart home assistants. IEEE Trans. Consum. Electron. **67**, 68–76 (2021)
19. Santhoshkumar, R., Geetha, M.K.: Deep learning approach: emotion recognition from human body movements. J. Mech. Contin. Math. Sci. **14**, 182–195 (2019)

20. Tsiourti, C., et al.: Multimodal integration of emotional signals from voice, body, and context: effects of (in) congruence on emotion recognition and attitudes towards robots. Int. J. Soc. Robot. **11**, 555–573 (2019)
21. Muhammad, G., Hossain, M.S.: Emotion recognition for cognitive edge computing using deep learning. IEEE Internet Things J. **8**, 16894–16901 (2021)
22. Research, G.V.: Emotion Detection And Recognition Market Size Report, 2030 (2021)
23. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion **37**, 98–125 (2017). https://doi.org/10.1016/j.inffus.2017.02.003
24. Li, Z., et al.: Fundamentals of multimedia. Pearson Prentice Hall (2004)
25. Dzedzickis, A., Kaklauskas, A., Bucinskas, V.: Human emotion recognition: review of sensors and methods. Sensors **20**, 592 (2020)
26. Houssein, E.H., Hammad, A., Ali, A.A.: Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. Neural Comput. Appl. **34**, 12527–12557 (2022)
27. Hinkle, L.B., Roudposhti, K.K., Metsis, V.: Physiological measurement for emotion recognition in virtual reality. In: 2019 2nd International Conference on Data Intelligence and Security (ICDIS). pp 136–143 (2019)
28. Du, G., et al.: A noncontact emotion recognition method based on complexion and heart rate. iEEE Trans. Instrum. Meas. **71**, 1–14 (2022)
29. Zhang, L., et al.: Non-contact dual-modality emotion recognition system by CW radar and RGB camera. IEEE Sens. J. **21**, 23198–23212 (2021)
30. Fragopanagos, N., Taylor, J.G.: Emotion recognition in human–computer interaction. Neural Netw. **18**, 389–405 (2005)
31. Ahmed, N., Al, A.Z., Girija, S.: A systematic survey on multimodal emotion recognition using learning algorithms. Intell. Syst. Appl. **17**, 200171 (2023). https://doi.org/10.1016/j.iswa.2022.200171
32. Gallo, L.C., Matthews, K.A.: Understanding the association between socioeconomic status and physical health: do negative emotions play a role? Psychol. Bull. **129**, 10 (2003)
33. Richman, L.S., et al.: Positive emotion and health: going beyond the negative. Heal Psychol. **24**, 422 (2005)
34. Fredrickson, B.L.: The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. Am. Psychol. **56**, 218 (2001)
35. Sreeja, P.S., Mahalakshmi, G.: Emotion models: a review. Int. J. Control Theory Appl. **10**, 651–657 (2017)
36. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(169), 200 (1992)
37. Bruna, O., Avetisyan, H., Holub, J.: Emotion models for textual emotion classification. J. Phys. Conf. Ser. **772**, 12023 (2016)
38. Plutchik, R.: The emotions: facts, theories, and a new model. Random House (1962)
39. Yannakakis, G.N., Cowie, R., Busso, C.: The ordinal nature of emotions: an emerging approach. IEEE Trans. Affect. Comput. **12**, 16–35 (2018)
40. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image Vis. Comput. **31**, 120–136 (2013)
41. Kossaifi, J., et al.: AFEW-VA database for valence and arousal estimation in-the-wild. Image Vis. Comput. **65**, 23–36 (2017)
42. Jaimes, A., Sebe, N.: Multimodal human–computer interaction: a survey. Comput. Vis. Image Underst. **108**, 116–134 (2007)
43. Karray, F., Alemzadeh, M., Saleh, J.A., Arab, M.N.: Human-computer interaction: overview on state of the art. Int J smart Sens Intell Syst **1**, 137–159 (2008)
44. Su, B., Peng, J.: Sentiment analysis of comment texts on online courses based on hierarchical attention mechanism. Appl. Sci. **13**, 4204 (2023)
45. Nandwani, P., Verma, R.: A review on sentiment analysis and emotion detection from text. Soc. Netw. Anal. Min. **11**, 81 (2021)
46. Jiang, Y., et al.: A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. Inf. Fusion **53**, 209–221 (2020)
47. Shaheen, S., El-Hajj, W., Hajj, H., Elbassuoni, S.: Emotion recognition from text based on automatically generated rules. In: 2014 IEEE International Conference on Data Mining Workshop. Shenzhen, China, pp 383–392 (2014)
48. Yoon, S., Byun, S., Jung, K.: Multimodal Speech Emotion Recognition Using Audio and Text. In: 2018 IEEE Spoken Language Technology Workshop, pp. 112–118. Greece, Athens (2018)
49. Park, S.-H., Bae, B.-C., Cheong, Y.-G.: Emotion recognition from text stories using an emotion embedding model. In: 2020 IEEE International Conference on Big Data and Smart Computing. Busan, Korea (South), pp 579–583 (2020)
50. Adoma, F., Henry, N.-M., Chen, W.: Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition. In: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing. Chengdu, China, pp 117–121 (2020)
51. Alvarez-Gonzalez, N., Kaltenbrunner, A., Gómez, V.: Uncovering the limits of text-based emotion detection. arXiv Prepr arXiv210901900 (2021)
52. Murthy, A.R., Kumar, K.M.A.: A review of different approaches for detecting emotion from text. IOP Conf. Ser. Mater. Sci. Eng. **1110**, 12023 (2021)
53. Heredia, J., et al.: Adaptive multimodal emotion detection architecture for social robots. IEEE Access **10**, 20727–20744 (2022)
54. Tarnowski, P., et al.: Emotion recognition using facial expressions. Procedia Comput. Sci. **108**, 1175–1184 (2017)
55. Abramson, et al.: Social interaction context shapes emotion recognition through body language, not facial expressions. Emotion **21**, 557 (2021)
56. Shen, Z. et al.: Emotion recognition based on multi-view body gestures. In: 2019 IEEE International Conference on Image Processing. Taipei, Taiwan, pp 1–5 (2019)
57. Lim, J.Z., Mountstephens, J., Teo, J.: Emotion recognition using eye-tracking: taxonomy, review and current challenges. Sensors **20**, 2384 (2020)
58. Kosti, R., et al.: Context based emotion recognition using EMOTIC dataset. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 2755–2766 (2020)
59. Mittal, T. et al.: Emoticon: Context-aware multimodal emotion recognition using Frege's principle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 14234–14243 (2020)
60. Zhang, S., Tang, C., Guan, C.: Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition. Pattern Recognit. **130**, 108833 (2022)
61. Domínguez-Jiménez, J.A., et al.: A machine learning model for emotion recognition from physiological signals. Biomed. Signal Process. Control **55**, 101646 (2020)
62. Wu, C.-H., Liang, W.-B.: Speech emotion recognition using acoustic-prosodic information and semantic labels. IEEE Trans. Affect. Comput. **2**, 10–21 (2010)
63. Wang, J., Xia, M., Li, H., Chen, X.: Speech emotion recognition with dual-sequence LSTM architecture. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, pp. 7314–7318. Speech and Signal Processing. Barcelona, Spain (2020)

64. Issa, D., Demirci, M.F., Yazici, A.: Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control **59**, 101894 (2020)
65. Kwon, S.: A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors **20**, 183 (2019)
66. Huang, K.-Y. et al.: Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, pp 6885–6889 (2019)
67. Li, R. et al.: Dilated Residual Network with Multi-head Self-attention for Speech Emotion Recognition. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, pp 6875–6879 (2019)
68. Abdelhamid, A., et al.: Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm. IEEE Access **10**, 49265–49284 (2022)
69. Amjad, A., Khan, L., Chang, H.-T.: Effect on speech emotion classification of a feature selection approach using a convolutional neural network. PeerJ Comput Sci **7**, e766 (2021)
70. Chen, J., Wang, C., Wang, K., Yin, C., Zhao, C., Xu, T., Zhang, X., Huang, Z., Liu, M., Yang, T.: HEU emotion: a large-scale database for multimodal emotion recognition in the wild. Neural Comput. Appl. **33**, 8669–8685 (2021)
71. Khalil, R.A., et al.: Speech emotion recognition using deep learning techniques: a review. IEEE Access **7**, 117327–117345 (2019)
72. Li, Y., Zhao, T., Kawahara, T.: Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In: Interspeech (2019)
73. Dai, D., Wu, Z., Li, R., Wu, X., Jia, J., Meng, H.: Learning discriminative features from spectrograms using center loss for speech emotion recognition. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK, pp 7405–7409 (2019)
74. Lee, M., et al.: Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features. Appl. Sci. **10**, 3501 (2020)
75. Ozdemir, M.A., et al.: EEG-based emotion recognition with deep convolutional neural networks. Biomed Eng Tech **66**, 43–57 (2021)
76. Salankar, N., Mishra, P., Garg, L.: Emotion recognition from EEG signals using empirical mode decomposition and second-order difference plot. Biomed. Signal Process. Control **65**, 102389 (2021)
77. Vazquez-Rodriguez, J. et al.: Transformer-Based Self-Supervised Learning for Emotion Recognition. In: 2022 26th International Conference on Pattern Recognition (ICPR). Montreal, QC, Canada (2022)
78. Subasi, A., et al.: EEG-based emotion recognition using tunable Q wavelet transform and rotation forest ensemble classifier. Biomed. Signal Process. Control **68**, 102648 (2021)
79. Liu, Y., Fu, G.: Emotion recognition by deeply learned multi-channel textual and EEG features. Futur. Gener. Comput. Syst. **119**, 1–6 (2021)
80. Shu, L., et al.: A review of emotion recognition using physiological signals. Sensors **18**, 2074 (2018)
81. Ahmad, Z., Khan, N.: A survey on physiological signal-based emotion recognition. Bioengineering **9**, 688 (2022)
82. Zhang, S., et al.: Learning deep multimodal affective features for spontaneous speech emotion recognition. Speech Commun. **127**, 73–81 (2021)
83. Xie, B., Sidulova, M., Park, C.H.: Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. Sensors **21**, 4913 (2021)
84. Tan, Y., et al.: A Multimodal emotion recognition method based on facial expressions and electroencephalography. Biomed. Signal Process. Control **70**, 103029 (2021)
85. Yang, Z., Nayan, K., Fan, Z., Cao, H.: Multimodal Emotion Recognition with Surgical and Fabric Masks. In: Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing. pp 4678–4682 (2022)
86. Doyran, M., Schimmel, A., Baki, P., Ergin, K., Türkmen, B., Salah, A.A., Bakkes, S.C.J., Kaya, H., Poppe, R., Salah, A.A.: MUMBAI: multi-person, multimodal board game affect and interaction analysis dataset. J. Multimodal User Interfaces **15**, 373–391 (2021)
87. Yang, T.H., Wu, C.H., Huang, K.Y., Su, M.H.: Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. J. Ambient. Intell. Humaniz. Comput. **8**, 895–906 (2017)
88. Komuro, N., Hashiguchi, T., Hirai, K., Ichikawa, M.: Predicting individual emotion from perception-based non-contact sensor big data. Sci. Rep. **11**, 1–9 (2021)
89. Masui, K., Nagasawa, T., Tsumura, N., et al.: Continuous estimation of emotional change using multimodal affective responses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp 290–291 (2020)
90. Hussain, T. et al.: Deep Learning for Audio Visual Emotion Recognition. In: 25th International Conference on Information Fusion (FUSION). pp 1–8 (2022)
91. Harár, P., Burget, R., Dutta, M.K.: Speech emotion recognition with deep learning. In: 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN). pp 137–140 (2017)
92. Mamieva, D., Abdusalomov, A.B., Kutlimuratov, A., Muminov, B., Whangbo, T.K.: Multimodal emotion detection via attention-based fusion of extracted facial and speech features. Sensors **23**, 5475 (2023)
93. Fuente, C., Castellanos, F.J., Valero-Mas, J.J., Calvo-Zaragoza, J., de la Fuente, C., Castellanos, F.J., Valero-Mas, J.J., Calvo-Zaragoza, J.: Multimodal recognition of frustration during game-play with deep neural networks. Multimed. Tools Appl. **82**, 13617–13636 (2023). https://doi.org/10.1007/s11042-022-13762-7
94. Prabhu, S., Mittal, H., Varagani, R., Jha, S., Singh, S.: Harnessing emotions for depression detection. Pattern Anal. Appl. **25**, 537–547 (2022)
95. Hore, S., Bhattacharya, T.: Impact of lockdown on Generation-Z: a fuzzy based multimodal emotion recognition approach using CNN. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-14543-6
96. Bao, J., Tao, X., Zhou, Y.: An emotion recognition method based on eye movement and audiovisual features in MOOC learning environment. IEEE Trans. Comput. Soc. Syst. (2022). https://doi.org/10.1109/TCSS.2022.3221128
97. Luo, Z., Zheng, C., Gong, J., Chen, S., Luo, Y., Yi, Y.: 3DLIM: Intelligent analysis of students' learning interest by using multimodal fusion technology. Educ. Inf. Technol. 1–21 (2022)
98. Zhang, R., He, N., Liu, S., Wu, Y., Yan, K., He, Y., Lu, K.: Your heart rate betrays you: multimodal learning with spatio-temporal fusion networks for micro-expression recognition. Int J Multimed Inf Retr **11**, 553–566 (2022)
99. Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M., Fernández-Martínez, F.: A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset. Appl. Sci. **12**, 327 (2022)
100. Do, L.-N., Yang, H.-J., Nguyen, H.-D., Kim, S.-H., Lee, G.-S., Na, I.-S.: Deep neural network-based fusion model for emotion recognition using visual data. J. Supercomput., 1–18 (2021)
101. Venkatakrishnan, R., Goodarzi, M., Canbaz, M.A.: Exploring large language models' emotion detection abilities: use cases

from the middle east. In: 2023 IEEE Conference on Artificial Intelligence (CAI). pp 241–244 (2023)
102. Zhao, Z., Wang, Y., Wang, Y.: Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition. arXiv Prepr arXiv220704697 (2022)
103. Krishna, D.N.: Using large pre-trained models with cross-modal attention for multi-modal emotion recognition. arXiv Prepr arXiv210809669 (2021)
104. Yi, Y., Tian, Y., He, C., Fan, Y., Hu, X., Xu, Y.: DBT: multimodal emotion recognition based on dual-branch transformer. J. Supercomput. **79**, 8611–8633 (2023)
105. Zhang, H., Gou, R., Shang, J., Shen, F., Wu, Y., Dai, G.: Pre-trained deep convolution neural network model with attention for speech emotion recognition. Front. Physiol. **12**, 643202 (2021)
106. Lee, S., Han, D.K., Ko, H.: Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification. IEEE Access **9**, 94557–94572 (2021)
107. Tran, M., Soleymani, M.: A pre-trained audio-visual transformer for emotion recognition. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 4698–4702 (2022)
108. Padi, S., Sadjadi, S.O., Manocha, D., Sriram, R.D.: Multimodal emotion recognition using transfer learning from speaker recognition and bert-based models. arXiv Prepr arXiv220208974 (2022)
109. Sun, L., Xu, M., Lian, Z., Liu, B., Tao, J., Wang, M., Cheng, Y.: Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model. In: Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge. pp 15–20 (2021)
110. Zhao, J., Li, R., Jin, Q., Wang, X., Li, H.: Memobert: pre-training model with prompt-based learning for multimodal emotion recognition. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp 4703–4707 (2022)
111. Lian, Z. et al.: Explainable Multimodal Emotion Reasoning. arXiv Prepr arXiv230615401 (2023)
112. Lu, X.: Deep learning based emotion recognition and visualization of figural representation. Front. Psychol. **12**, 818833 (2022)
113. Liu, D., et al.: Multi-modal fusion emotion recognition method of speech expression based on deep learning. Front. Neurorobot. **15**, 697634 (2021)
114. Finotti, G., Serwadda, A., Elhoseiny, M., Menti, E., Bicci, D., Zafeiriou, S., Cristani, M.: Remote photoplethysmography (rPPG) in the wild: Remote heart rate imaging via online webcams. arXiv Prepr arXiv220212024 (2022)
115. Gouveia, C., et al.: Study on the usage feasibility of continuous-wave radar for emotion recognition. Biomed. Signal Process. Control **58**, 101835 (2020)
116. Keele, S. (2007) Guidelines for performing systematic literature reviews in software engineering
117. Wang, D.D., Zhao, X.M.: Affective video recommender systems: a survey. Front. Neurosci. (2022). https://doi.org/10.3389/fnins.2022.984404
118. Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Murugappan, M., Chakole, Y., Acharya, U.R.: Automated emotion recognition: Current trends and future perspectives. Comput. Methods Progr. Biomed. (2022). https://doi.org/10.1016/j.cmpb.2022.106646
119. Werner, P., et al.: Automatic recognition methods supporting pain assessment: a survey. IEEE Trans. Affect. Comput. **13**, 530–552 (2022)
120. He, Z.P., Li, Z.N., Yang, F.Z., Wang, L., Li, J.C., Zhou, C.J., Pan, J.H.: Advances in multimodal emotion recognition based on brain-computer interfaces. BRAIN Sci. (2020). https://doi.org/10.3390/brainsci10100687
121. Skaramagkas, V., et al.: Review of eye tracking metrics involved in emotional and cognitive processes. IEEE Rev. Biomed. Eng. **16**, 260–277 (2021)
122. Seng, J.K.P., Ang, K.L.-M.: Multimodal emotion and sentiment modeling from unstructured big data: challenges, architecture, and techniques. IEEE Access **7**, 90982–90998 (2019)
123. Chaturvedi, V., et al.: Music mood and human emotion recognition based on physiological signals: a systematic review. Multimed. Syst. **28**, 21–44 (2022)
124. Siddiqui, M.F.H., Dhakal, P., Yang, X., Javaid, A.Y.: A survey on databases for multimodal emotion recognition and an introduction to the VIRI (visible and infrared image) database. Multimodal Technol. Interact **6**, 47 (2022). https://doi.org/10.3390/mti6060047
125. Rouast, P.V., Adam, M.T.P., Chiong, R.: Deep learning for human affect recognition: Insights and new developments. IEEE Trans. Affect. Comput. **12**, 524–543 (2021)
126. Karimah, S.N., Hasegawa, S.: Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. Smart Learn. Environ. **9**, 1–48 (2022)
127. Landowska, A., Karpus, A., Zawadzka, T., Robins, B., Barkana, D.E., Kose, H., Zorcec, T., Cummins, N.: Automatic emotion recognition in children with autism: a systematic literature review. Sensors **22**, 1649 (2022). https://doi.org/10.3390/s22041649
128. Dhelim, S., et al.: Artificial intelligence for suicide assessment using audiovisual cues: a review. Artif. Intell. Rev. **56**(6), 5591–5618 (2023)
129. Koromilas, P., Giannakopoulos, T.: Deep multimodal emotion recognition on human speech: a review. Appl. Sci. **11**, 7962 (2021). https://doi.org/10.3390/app11177962
130. Singh, J., Hamid, M.A.: Cognitive computing in mental healthcare: a review of methods and technologies for detection of mental disorders. Cognit. Comput. **14**, 2169–2186 (2022)
131. Gu, X., Shen, Y., Xu, J.: Multimodal emotion recognition in deep learning: a survey. In: International Conference on Culture-oriented Science and Technology (ICCST). pp 77–82 (2021)
132. Karani, R., Desai, S.: Review on multimodal fusion techniques for human emotion recognition. Int. J. Adv. Comput. Sci. Appl. **13**, 287–296 (2022). https://doi.org/10.14569/IJACSA.2022.0131035
133. Spezialetti, M., Placidi, G., Rossi, S.: Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. Front. Robot. AI (2020). https://doi.org/10.3389/frobt.2020.532279
134. Nandi, A. et al.: A survey on multimodal data stream mining for e-learner's emotion recognition. In: 2020 International Conference on Omni-layer Intelligent Systems. pp 1–6, (2020)
135. Krishna, S., Anju, J.: Different approaches in depression analysis: a review. In: International Conference on Computational Performance Evaluation. pp 407–414 (2020)
136. Song, X., Chen, H., Wang, Q., Chen, Y., Tian, M., Tang, H.: A review of audio-visual fusion with machine learning. J. Phys. Conf. Ser. (2019). https://doi.org/10.1088/1742-6596/1237/2/022144
137. Vankudre, G. et al.: A survey on infant emotion recognition through video clips. In: 2021 International Conference on Computational Intelligence and Knowledge Economy. pp 296–300 (2021)
138. Yadegaridehkordi, E., Noor, N., Bin Ayub, M.N., Affal, H.B., Hussin, N.B.: Affective computing in education: a systematic review and future research. Comput. Educ. (2019). https://doi.org/10.1016/j.compedu.2019.103649

139. Giuntini, F., et al.: A review on recognizing depression in social networks: challenges and opportunities. J. Ambient. Intell. Humaniz. Comput. **11**, 4713–4729 (2020)
140. Garg, D., Verma, G.K., Singh, A.K.: A review of deep learning based methods for affect analysis using physiological signals. Multimed. Tools Appl. (2023). https://doi.org/10.1007/s11042-023-14354-9
141. Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D.W., Li, X.L., Gao, S.Y., Sun, Y.X., Ge, W.F., Zhang, W., Zhang, W.Q.: A systematic review on affective computing: emotion models, databases, and recent advances. Inf. Fusion **83**, 19–52 (2022). https://doi.org/10.1016/j.inffus.2022.03.009
142. Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., Dario, P.: Emotion modelling for social robotics applications: a review. J. Bionic Eng. **15**, 185–203 (2018)
143. Lin, W.Q., Li, C.: Review of studies on emotion recognition and judgment based on physiological signals. Appl. Sci. (2023). https://doi.org/10.3390/app13042573
144. Schmidt, P., Reiss, A., Durichen, R., Van Laerhoven, K.: Wearable-based affect recognition-a review. Sensors (2019). https://doi.org/10.3390/s19194079
145. Dalvi, M.R., Patil, S.S., Gite, S.P., Kotecha, K.: A survey of ai-based facial emotion recognition: features, ML \& DL techniques, age-wise datasets and future directions. IEEE Access **9**, 165806–165840 (2021)
146. Zloteanu, M., et al.: Veracity judgment, not accuracy: reconsidering the role of facial expressions, empathy, and emotion recognition training on deception detection. Q. J. Exp. Psychol. **74**, 910–927 (2021)
147. Hassouneh, A., Mutawa, A.M., Murugappan, M.: Development of a real-time emotion recognition system using facial expressions and EEG based on machine learning and deep neural network methods. Inf. Med. Unlocked **20**, 100372 (2020)
148. Razzaq, M.A., et al.: A hybrid multimodal emotion recognition framework for UX evaluation using generalized mixture functions. Sensors **23**, 4373 (2023)
149. Liu, W., et al.: Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Trans. Cogn. Dev. Syst. **14**, 715–729 (2021)
150. Middya, A.I., Nag, B., Roy, S.: Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowl. Based Syst. **244**, 108580 (2022)
151. Rasendrasoa, S. et al.: Real-Time Multimodal emotion recognition in conversation for multi-party interactions. In: Proceedings of the 2022 International Conference on Multimodal Interaction. pp 395–403 (2022)
152. Huang, Y., et al.: Research on robustness of emotion recognition under environmental noise conditions. IEEE Access **7**, 142009–142021 (2019)
153. Lévêque, L., et al.: Comparing the robustness of humans and deep neural networks on facial expression recognition. Electronics **11**, 4030 (2022)
154. Yoon, Y.C.: Can we exploit all datasets? Multimodal emotion recognition using cross-modal translation. IEEE Access **10**, 64516–64524 (2022)
155. Cohen, D., et al.: Masking important information to assess the robustness of a multimodal classifier for emotion recognition. Front. Artif. Intell. **6**, 1091443 (2023)
156. Dey, A.K.: Understanding and using context. Pers. Ubiquitous Comput. **5**, 4–7 (2001)
157. Song, Q., Sun, B., Li, S.: Multimodal sparse transformer network for audio-visual speech recognition. IEEE Trans. Neural. Netw. Learn. Syst. 1–11 (2022)
158. Lin, J. et al.: An explainable deep fusion network for affect recognition using physiological signals. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp 2069–2072 (2019)
159. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H. et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp 3438–3446 (2016)
160. Kumar, P., Kaushik, V., Raman, B.: Towards the explainability of multimodal speech emotion recognition. In: Interspeech. pp 1748–1752 (2021)
161. Palash, M., Bhargava, B.: EMERSK–explainable multimodal emotion recognition with situational knowledge. arXiv Prepr arXiv230608657 (2023)
162. Pereira, M.H.R., Pádua, F.L.C., Dalip, D.H., Benevenuto, F., Pereira, A.C.M., Lacerda, A.M.: Multimodal approach for tension levels estimation in news videos. Multimed. Tools Appl. **78**, 23783–23808 (2019)
163. Bhaskar, S., Thasleema, T.M.: LSTM model for visual speech recognition through facial expressions. Multimed. Tools Appl. **82**, 5455–5472 (2023)
164. Chumachenko, K., Iosifidis, A., GabboujIEEE, M.: Self-attention fusion for audiovisual emotion recognition with incomplete data. Institute of Electrical and Electronics Engineers Inc., Tampere Univ, Dept Comp Sci, Tampere, Finland (2022)
165. Goncalves, L., Busso, C.: Robust audiovisual emotion recognition: aligning modalities, capturing temporal information, and handling missing features. IEEE Trans. Affect. Comput. **13**, 2156–2170 (2022). https://doi.org/10.1109/TAFFC.2022.3216993
166. Ghaleb, E., Niehues, J., Asteriadis, S.: Joint modelling of audio-visual cues using attention mechanisms for emotion recognition. Multimed. Tools Appl. **82**, 11239–11264 (2022)
167. Savchenko, A.V., Savchenko, L.V.: Audio-visual continuous recognition of emotional state in a multi-user system based on personalized representation of facial expressions and voice. Pattern Recognit Image Anal. **32**, 665–671 (2022). https://doi.org/10.1134/S1054661822030397
168. Ma, F., Li, Y., Ni, S., Huang, S.-L., Zhang, L.: Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN. Appl. Sci. **12**, 527 (2022)
169. Karas, V., Tellamekala, M.K., Mallol-Ragolta, A., Valstar, M., Schuller, B.W.: Time-continuous audiovisual fusion with recurrence vs attention for in-the-wild affect recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp 2381–2390 (2022)
170. Rong, Q., Ding, S., Yue, Z., Wang, Y., Wang, L., Zheng, X., Li, Y.: Non-contact negative mood state detection using reliability-focused multi-modal fusion model. IEEE J. Biomed. Heal. Inf. **26**, 4691–4701 (2022)
171. Chen, R., Zhou, W., Li, Y., Zhou, H.: Video-based cross-modal auxiliary network for multimodal sentiment analysis. IEEE Trans. Circuits Syst. Video Technol. **32**, 8703–8716 (2022)
172. Juyal, P.: Multi-modal sentiment analysis of audio and visual context of the data using machine learning. In: 3rd International Conference on Smart Electronics and Communication. pp 1198–1205, (2022)
173. Dresvyanskiy, D., Ryumina, E., Kaya, H., Markitantov, M., Karpov, A., Minker, W.: End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild. Multimodal Technol. Interact (2022). https://doi.org/10.3390/mti6020011
174. Guo, P., Chen, Z., Li, Y., Liu, H.: Audio-visual fusion network based on conformer for multimodal emotion recognition. In: Artificial Intelligence, CICAI 2022, PT II. Springer, pp 315–326 (2022)

175. Yi, Y., Tian, Y., He, C., Fan, Y., Hu, X., Xu, Y.: DBT: multimodal emotion recognition based on dual-branch transformer. J Supercomput 0123456789 (2022)
176. Abu Shaqra, F., Duwairi, R., Al-Ayyoub, M.: A multi-modal deep learning system for Arabic emotion recognition. Int. J. Speech Technol. 123–139 (2022)
177. Neumann, M., Vu, N.T., IEEE: Investigations on audiovisual emotion recognition in noisy conditions. 2021 IEEE Spok. Lang. Technol. Work. 358–364 (2021)
178. Praveen, R.G., Granger, E.: Cardinal P cross attentional audio-visual fusion for dimensional emotion recognition. 2021 16TH IEEE Int. Conf. Autom. FACE GESTURE Recognit. (FG 2021)
179. Radoi, A., Birhala, A., Ristea, N.C., Dutu, L.C.: An end-to-end emotion recognition framework based on temporal aggregation of multimodal information. IEEE Access **9**, 135559–135570 (2021). https://doi.org/10.1109/ACCESS.2021.3116530
180. Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J.M., Fernández-Martínez, F.: Multimodal emotion recognition on RAVDESS dataset using transfer learning. Sensors **21**, 7665 (2021)
181. Antoniadis, P., Pikoulis, I., Filntisis, P.P., Maragos, P.: An audio-visual and contextual approach for categorical and continuous emotion recognition in-the-wild. 2021-Octob 3638–3644, (2021)
182. Schoneveld, L., Othmani, A., Abdelkawy, H.: Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recognit. Lett. **146**, 1–7 (2021). https://doi.org/10.1016/j.patrec.2021.03.007
183. Pandeya, Y.R., Bhattarai, B., Lee, J.: Music video emotion classification using slow–fast audio–video network and unsupervised feature representation. Sci. Rep. **11**, 1–14 (2021)
184. Huddar, M.G., Sannakki, S.S., Rajpurohit, V.S.: Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. Multimed. Tools Appl. **80**, 13059–13076 (2021)
185. Ghaleb, E., Niehues, J., Asteriadis, S.: Multimodal attention-mechanism for temporal emotion recognition. In: 2020 IEEE International Conference On Image Processing (ICIP). Maastricht Univ, Maastricht, Netherlands, pp 251–255 (2020)
186. Hsu, J.-H., Wu, C.-H.: Attentively-coupled long short-term memory for audio-visual emotion recognition. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. pp 1048–1053 (2020)
187. Pinto, J.R., Goncalves, T., Pinto, C., Sanhudo, L., Fonseca, J., Goncalves, F., Carvalho, P., Cardoso, J.S.: Audiovisual classification of group emotion valence using activity recognition networks. In: 2020 IEEE International Conference on Image Processing, Applications and Systems. pp 114–119 (2020)
188. Shukla, A.: Learning self-supervised multimodal representations of human behaviour. In: MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, Inc, pp 4748–4751 (2020)
189. Wang, Y., Wu, J., Heracleous, P., Wada, S., Kimura, R., Kurihara, S.: Implicit knowledge injectable cross attention audiovisual model for group emotion recognition. In: ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction. Association for Computing Machinery, Inc, pp 827–834 (2020)
190. Vidal, A., Salman, A., Lin, W.C., Busso, C.: MSP-Face Corpus: a natural audiovisual emotional database. In: International Conference on Multimodal Interaction. pp 397–405 (2020)
191. Park, C.Y., Cha, N., Kang, S., Kim, A., Khandoker, A.H., Hadjileontiadis, L., Oh, A., Jeong, Y., Lee, U.: K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. Sci. Data (2020). https://doi.org/10.1038/s41597-020-00630-y
192. Mansouri-Benssassi, E., Ye, J., Intelligence AAA (2020) Synch-graph: multisensory emotion recognition through neural synchrony via graph convolutional networks. Thirty-Fourth AAAI Conf. Artif. Intell. THIRTY-SECOND Innov. Appl. Artif. Intell. Conf. TENTH AAAI Symp. Educ. Adv. Artif. Intell. 34:1351–1358
193. Atmaja, B.T., Akagi, M., IEEE (2020) Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. In: 2020 IEEE Int. Conf. Acoust. Speech, Signal Process. 4482–4486
194. Ashwin, T.S., Guddeti, R.M.R.: Generative adversarial nets in classroom environment using hybrid convolutional neural networks. Educ. Inf. Technol. **25**, 1387–1415 (2020)
195. Dahmani, S., Colotte, V., Ouni, S.: Some consideration on expressive audiovisual speech corpus acquisition using a multimodal platform. Lang. Resour. Eval. **54**, 943–974 (2020)
196. Nemati, S., Rohani, R., Basiri, M.E., Abdar, M., Yen, N.Y., Makarenkov, V.: A hybrid latent space data fusion method for multimodal emotion recognition. IEEE Access **7**, 172948–172964 (2019)
197. Ringeval, F. et al.: AVEC workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In: Proceedings of the 9th International Workshop on Audio/Visual Emotion Challenge. pp 3–12 (2019)
198. Avots, E., Sapinski, T., Bachmann, M., Kaminska, D.: Audiovisual emotion recognition in wild. Mach. Vis. Appl. **30**, 975–985 (2019). https://doi.org/10.1007/s00138-018-0960-9
199. Li, X., Lu, G.M., Yan, J.J., Li, H.B., Zhang, Z.Y., Sun, N., Xie, S.P.: Incomplete cholesky decomposition based kernel cross modal factor analysis for audiovisual continuous dimensional emotion recognition. KSII Trans. Internet Inf. Syst. **13**, 810–831 (2019). https://doi.org/10.3837/tiis.2019.02.0018
200. Vakhshiteh, F., Almasganj, F.: Exploration of properly combined audiovisual representation with the entropy measure in audiovisual speech recognition. Circuits Syst Signal Process **38**, 2523–2543 (2019)
201. Zhang, Z., Han, J., Deng, J., Xu, X., Ringeval, F., Schuller, B.: Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning. IEEE Access **6**, 22196–22209 (2018). https://doi.org/10.1109/ACCESS.2018.2821192
202. Gorbova, J., Avots, E., Lusi, I., Fishel, M., Escalera, S., Anbarjafari, G.: Integrating vision and language for first-impression personality analysis. IEEE Multimed. **25**, 24–33 (2018)
203. Ilyas, C.M.A., Nasrollahi, K., Rehm, M., Moeslund, T.B.: Rehabilitation of Traumatic Brain Injured Patients: Patient Mood Analysis from Multimodal Video. In: 2018 25th IEEE International Conference on Image Processing. IEEE, pp 2291–2295 (2018)
204. Ringeval, F. et al.: AVEC Workshop and challenge: bipolar disorder and cross-cultural affect recognition. In: Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. pp 3–13 (2018)
205. Vielzeuf, V., Kervadec, C., Pateux, S., Lechervy, A., Jurie, F.: An Occam's Razor view on learning audiovisual emotion recognition with small training sets. In: ACM Conference on International Conference on Multimedia Retrieval. pp 589–593 (2018)
206. El Haddad, K., Rizk, Y., Heron, L., Hajj, N., Zhao, Y., Kim, J., Trung, N.T., Lee, M., Doumit, M., Lin, P., Kim, Y., Cakmak, H.: End-to-end listening agent for audiovisual emotional and naturalistic interactions. J. Sci. Technol. ARTS **10**, 49–61 (2018). https://doi.org/10.7559/citarj.v10i2.424
207. Carlson, J.M., Conger, S., Sterr, J.: Auditory distress signals potentiate attentional bias to fearful faces: evidence for multimodal facilitation of spatial attention by emotion. J. Nonverbal Behav. **42**, 417–426 (2018)

208. Ivanko, D., Karpov, A., Fedotov, D., Kipyatkova, I., Ryumin, D., Ivanko, D., Minker, W., Zelezny, M.: Multimodal speech recognition: increasing accuracy using high speed video data. J. Multimodal User Interfaces **12**, 319–328 (2018)
209. Tian, L., Muszynski, M., Lai, C., Moore, J.D., Kostoulas, T., Lombardo, P., Pun, T., Chanel, G.: Recognizing Induced emotions of movie audiences: are induced and perceived emotions the same? In: Seventh International Conference on Affective Computing and Intelligent Interaction. pp 28–35 (2017)
210. Busso, C., Parthasarathy, S., Burmania, A., Abdelwahab, M., Sadoughi, N., Provost, E.M.: MSP-IMPROV: an acted corpus of dyadic interactions to study emotion perception. IEEE Trans. Affect. Comput. **8**, 67–80 (2017)
211. Ringeval, F., Gratch, J., Mozgai, S., Schuller, B., Cowie, R., Cummins, N., Pantic, M., Valstar, M., Scherer, S., Schmitt, M.: AVEC—Real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction. pp 3–9 (2017)
212. Wang, C., Zhang, J., Gan, L., Jiang, W.: a prediction method for dimensional sentiment analysis of the movie and tv drama based on variable-length sequence Input. In: 2022 International Conference on Culture-Oriented Science and Technology (CoST). pp 1–5 (2022)
213. Tawsif, K., Aziz, N.A.A., Raja, J.E., Hossen, J., Jesmeen, M.Z.H.: A systematic review on emotion recognition system using physiological signals: data acquisition and methodology. Emerg. Sci. J. **6**, 1167–1198 (2022)
214. Li, Y., Wei, J., Liu, Y., Kauttonen, J., Zhao, G.: Deep learning for micro-expression recognition: a survey. IEEE Trans. Affect. Comput. **13**(4), 2028 (2022)
215. Liu, Y., Zhang, X., Li, Y., Zhou, J., Li, X., Zhao, G.: Graph-based facial affect analysis: A review. IEEE Trans. Affect. Comput. **14**(4), 2657–2677 (2022)
216. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed. **9**, 34–41 (2012)
217. Kollias, D., Zafeiriou, S.: Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace (2019)
218. Sarkar, P., Posen, A., Etemad, A.: AVCAffe: a large scale audio-visual dataset of cognitive load and affect for remote work. AAAI (2022). https://doi.org/10.1609/aaai.v37i1.25078
219. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.: Avec 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. pp 3–10 (2014)
220. Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E.: BAUM-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans. Affect. Comput. **8**, 300–313 (2017)
221. Erdem, C.E., Turan, C., Aydin, Z.: BAUM-2: a multilingual audio-visual affective face database. Multimed. Tools Appl. **74**, 7429–7459 (2015)
222. Caridakis, G., Wagner, J., Raouzaiou, A., Lingenfelser, F., Karpouzis, K., Andre, E.: A cross-cultural, multimodal, affective corpus for gesture expressivity analysis. J. Multimodal User Interfaces **7**, 121–134 (2013)
223. Li, J., Dong, Z., Lu, S., Wang, S.J., Yan, W.J., Ma, Y., Fu, X.: CAS (ME) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. IEEE Trans. Pattern Anal. Mach. Intell. **45**, 2782–2800 (2022)
224. Li, Y., Tao, J., Chao, L., Bao, W., Liu, Y.: CHEAVD: A Chinese natural emotional audio–visual database. J. Ambient. Intell. Humaniz. Comput. **8**, 913–924 (2017)
225. Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., Jia, J.: Mec 2017: Multimodal emotion recognition challenge. In: Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia). pp 1–5 (2018)
226. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp 2236–2246 (2018)
227. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. **5**, 377–390 (2014)
228. Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp 1–9 (2016)
229. Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T.: From individual to group-level emotion recognition: Emotiw 5.0. In: Proceedings of the 19th ACM international conference on multimodal interaction. pp 524–528 (2017)
230. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The enterface'05 audio-visual emotion database. In: Proceedings of the 22nd International Conference on Data Engineering Workshops. p 8 (2006)
231. O'Reilly, H., Pigat, D., Fridenson, S., Berggren, S., Tal, S., Golan, O., Bölte, S., Baron-Cohen, S., Lundqvist, D.: The EU-emotion stimulus set: a validation study. Behav. Res. Methods **48**, 567–576 (2016)
232. Bänziger, T., Mortillaro, M., Scherer, K.R.: Introducing the geneva multimodal expression corpus for experimental research on emotion perception. Emotion **12**, 1161 (2012)
233. Douglas-Cowie, E., Cox, C., Martin, J.C., Devillers, L., Cowie, R., Sneddon, I. et al.: The HUMAINE database. In: Emotion-oriented systems: The Humaine handbook. pp 243–284 (2011)
234. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**, 335 (2008)
235. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. **3**, 42–55 (2012)
236. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: a multimodal multi-party dataset for emotion recognition in conversations (2018)
237. Shen, G., Wang, X., Duan, X., Li, H., Zhu, W.: Memor: a dataset for multimodal emotion reasoning in videos. In: Proceedings of the 28th ACM International Conference on Multimedia. pp 493–502 (2020)
238. Chou, H.C., Lin, W.C., Chang, L.C., Li, C.C., Ma, H.P., Lee, C.C.: NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In: Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp 292–298 (2017)
239. Perepelkina, O., Kazimirova, E., Konstantinova, M.: RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. In: Proceedings of the International Conference on Speech and Computer. pp 501–510 (2018)
240. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE **13**, e0196391 (2018)
241. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). pp 1–8 (2013)

242. Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T., Sedogbo, C.: The SAFE Corpus: illustrating extreme emotions in dynamic situations. In: First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006)). Genoa, Italy, pp 76–79 (2006)
243. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME). pp 1079–1084 (2010)
244. Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., et al.: Sewa db: a rich database for audio-visual emotion and sentiment research in the wild. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 1022–1040 (2019)
245. Metallinou, A., Yang, Z., Lee, C.C., Busso, C., Carnicke, S., Narayanan, S.: The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. Lang. Resour. Eval. **50**, 497–521 (2016)
246. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems. pp 2672–2680 (2014)
247. Cheng, H., Tie, Y., Qi, L., Jin, C.: Context-aware based visual-audio feature fusion for emotion recognition. In: IEEE International Joint Conference on Neural Networks (IJCNN). pp 1–8 (2021)
248. He, L., Niu, M., Tiwari, P., Marttinen, P., Su, R., Jiang, J., Guo, C., Wang, H., Ding, S., Wang, Z., et al.: Deep learning for depression recognition with audiovisual cues (2022)
249. Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Morency, L.-P., et al.: Automatic audiovisual behavior descriptors for psychological disorder analysis. Image Vis. Comput. **32**, 648–658 (2014)
250. Adib, F., Mao, H., Kabelac, Z., Katabi, D., Miller, R.C.: Smart homes that monitor breathing and heart rate. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp 837–846 (2015)
251. Liu, X., Cao, J., Tang, S., Wen, J., Guo, P.: Contactless respiration monitoring via off-the-shelf WiFi devices. IEEE Trans. Mob. Comput. **15**, 2466–2479 (2015)
252. Fernández-Caballero, A., Martinez-Rodrigo, A., Pastor, J.M., Castillo, J.C., Lozano-Monasor, E., López, M.T., Zangróniz, R., Latorre, J.M., Fernández-Sotos, A.: Smart environment architecture for emotion detection and regulation. J. Biomed. Inform. **64**, 55–73 (2016)
253. Cao, S. et al.: Can AI detect pain and express pain empathy? A review from emotion recognition and a human-centered AI perspective (2021). arXiv Prepr arXiv211004249
254. Dunford, E., West, E., Sampson, E.L.: Psychometric evaluation of the pain assessment in advanced dementia scale in an acute general hospital setting. Int. J. Geriatr. Psychiatry **37**, 1–10 (2022)
255. Li, Y., Liu, Y., Nguyen, K., Shi, H., Vuorenmaa, E., Jarvela, S., Zhao, G.: Exploring Interactions and Regulations in Collaborative Learning: An Interdisciplinary Multimodal Dataset (2022). arXiv Prepr arXiv221005419
256. Emotional Entanglement: China's emotion recognition market and its implications for human rights (2021)
257. Deschamps-Berger, T., Lamel, L., Devillers, L.: End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction. pp 1–8 (2021)
258. Miao, Y., Yang, J., Alzahrani, B., Lv, G., Alafif, T., Barnawi, A., Chen, M.: Abnormal behavior learning based on edge computing toward a crowd monitoring system. IEEE Netw. **36**, 90–96 (2022)
259. Kuppusamy, P., Bharathi, V.C.: Human abnormal behavior detection using CNNs in crowded and uncrowded surveillance–A survey. Meas Sensors **24**, 100510 (2022)
260. Sanchez, F.L., et al.: Revisiting crowd behaviour analysis through deep learning: taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. Inf Fusion **64**, 318–335 (2020)
261. North-Samardzic, A.: Biometric technology and ethics: beyond security applications. J. Bus. Ethics **167**, 433–450 (2020)
262. Hayat, H., Ventura, C., Lapedriza, A.: Recognizing Emotions evoked by Movies using Multitask Learning. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII) (2021)
263. Cohendet, R. et al.: Emotional interactive movie: adjusting the scenario according to the emotional response of the viewer. EAI Endorsed Trans Creat Technol 4 (2017)
264. News, N.: Emotionally responsive interactive movie developed (2023)
265. Forum, W.E.: Interactive cinema: How films could alter plotlines in real time by responding to viewers' emotions (2023)
266. Li, J., Liu, J., Jahng, S.G.: Research and dynamic analysis of interactive methods of film in the context of new media. In: SHS Web of Conferences (2023)
267. Perello-March, J.R., Burns, C.G., Birrell, S.A., Woodman, R., Elliott, M.T.: Physiological measures of risk perception in highly automated driving. IEEE Trans. Intell. Transp. Syst. **23**, 4811–4822 (2022)
268. Muhlbacher-Karrer, S., Mosa, A.H., Faller, L.M., Ali, M., Hamid, R., Zangl, H., Kyamakya, K.: A driver state detection system: combining a capacitive hand detection sensor with physiological sensors. IEEE Trans. Instrum. Meas. **66**, 624–636 (2017)
269. Izquierdo-Reyes, J., Ramirez-Mendoza, R.A., Bustamante-Bello, M.R., Pons-Rovira, J.L., Gonzalez-Vargas, J.E.: Emotion recognition for semi-autonomous vehicles framework. Int. J. Interact. Des. Manuf. **12**, 1447–1454 (2018)
270. Alsaid, A., Lee, J.D., Noejovich, S.I., Chehade, A.: The effect of vehicle automation styles on drivers' emotional state. IEEE Trans. Intell. Transp. Syst. **24**, 3963–3973 (2023)
271. Antony, M.M., Whenish, R.: Advanced driver assistance systems (ADAS). In: Automotive Embedded Systems: Key Technologies, Innovations, and Applications. Springer International Publishing, pp 165–181 (2021)
272. Li, W., Cui, Y., Ma, Y., Chen, X., Li, G., Zeng, G., Guo, G,. Cao, D.: A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: emotions triggered by video-audio clips in driving scenarios. IEEE Trans. Affect. Comput. (2021)
273. Kim, T., Kim, Y., Jeon, H., Choi, C.S., Suk, H.J.: Emotional response to in-car dynamic lighting. Int. J. Automot. Technol. **22**, 1035–1043 (2021)
274. Reports C: Driver monitoring systems can help you be safer on the road (2022)
275. Sukhavasi, S.B., et al.: A hybrid model for driver emotion detection using feature fusion approach. Int. J. Environ. Res. Public Health **19**, 3085 (2022)
276. Resch, B., Puetz, I., Bluemke, M., Kyriakou, K., Miksch, J.: An interdisciplinary mixed-methods approach to analyzing urban spaces: the case of urban walkability and bikeability. Int. J. Environ. Res. Public Health **17**, 6994 (2020)
277. Bhamborae, M.J., Flotho, P., Mai, A., Schneider, E.N., Francis, A.L., Strauss, D.J.: Towards contactless estimation of electrodermal activity correlates. In: 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp 1799–1802 (2020)

278. Godovykh, M., Tasci, A.D.A.: Emotions, feelings, and moods in tourism and hospitality research: conceptual and methodological differences. Tour. Hosp. Res. **22**, 247–253 (2022)
279. Gupta, S. et al.: The future is yesterday: use of AI-driven facial recognition to enhance value in the travel and tourism industry. Inf Syst Front, 1–17 (2022)
280. Iván, A.L., Begoña, J.N., Yoon, S.Y.: Identifying customer's emotional responses towards guest-room design by using facial expression recognition, in hotel's virtual and real environments. J. Indones. Tour. Hosp. Recreat. **2**, 104–118 (2019)
281. González-Rodríguez, M.R., Díaz-Fernández, M.C., Pacheco Gómez, C.: Facial-expression recognition: an emergent approach to the measurement of tourist satisfaction through emotions. Telemat. Inf. **51**, 101404 (2020)
282. Kim, J., Fesenmaier, D.R.: Measuring emotions in real time: implications for tourism experience design. J. Travel Res. **54**, 419–429 (2015)
283. Cai, Y., Li, X., Li, J.: Emotion recognition using different sensors, emotion models, methods and datasets: a comprehensive review. Sensors **23**, 2455 (2023)
284. Santamaria-Granados, L., et al.: Tourist experiences recommender system based on emotion recognition with wearable data. Sensors **21**, 7854 (2021)
285. Sheikh, M., Qassem, M., Kyriacou, P.A.: Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. Front. Digit. Heal. **3**, 662811 (2021)
286. Austin, W.: The ethics of everyday practice: healthcare environments as moral communities. Adv. Nurs. Sci. **30**, 81–88 (2007)
287. On Artificial Intelligence H-LEG: Ethics guidelines for trustworthy AI (2019)
288. Organization WH: Ethics and governance of artificial intelligence for health: WHO guidance (2021)
289. Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., Floridi, L.: The ethics of algorithms: key problems and solutions. In: Mittelstadt, B., Floridi, L., Taddeo, M. (eds.) Ethics Governance and Policies in Artificial Intelligence, pp. 97–123. Springer (2021)
290. Saheb, T., Saheb, T., Carpenter, D.O.: Mapping research strands of ethics of artificial intelligence in healthcare: a bibliometric and content analysis. Comput. Biol. Med. **135**, 104660 (2021)
291. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P.: Towards a standard for identifying and managing bias in artificial intelligence. NIST Spec. Publ. **1270**, 1–77 (2022)
292. Vassilakopoulou, P., Aanestad, M.: Communal data work: data sharing and re-use in clinical genetics. Health Inf. J. **25**, 511–525 (2019)
293. Kroes, S., Janssen, M., Groenwold, R., van Leeuwen, M.: Evaluating privacy of individuals in medical data. Health Informatics J. **27**, 1460458220983398 (2021)
294. Wilkowska, W., Ziefle, M.: Privacy and data security in E-health: requirements from the user's perspective. Health Inf. J. **18**, 191–201 (2012)
295. Milne, R., Morley, K.I., Howard, H.C., Niemiec, E., Nicol, D., Critchley, C., Prainsack, B.: Trust in genomic data sharing among members of the general public in the UK, USA, Canada and Australia. Hum. Genet. **138**, 1237–1246 (2019)
296. Lafky, D., Horan, T.: Personal health records: Consumer attitudes toward privacy and security of their personal health information. Health Inf. J. **17**, 63–71 (2011)
297. Parvinen, L., Alamäki, A., Hallikainen, H., Mäki, M.: Exploring the challenges of and solutions to sharing personal genomic data for use in healthcare. Health Informatics J. 29 (2023)
298. Adams, C., Pente, P., Lemermeyer, G., Rockwell, G.: Artificial intelligence ethics guidelines for K-12 Education: a review of the global landscape. In: et al. IR (ed) AIED 2021, LNAI 12749. pp 24–28 (2021)
299. Kumar, P., Malik, S., Raman, B.: Interpretable Multimodal Emotion Recognition using Hybrid Fusion of Speech and Image Data (2022). arXiv Prepr arXiv220811868
300. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, (Volume 1: Long Papers) (2021)
301. Zuo, H. et al.: Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1–5 (2023)
302. Chi, H. et al.: Missing Modality meets Meta Sampling (M3S): An Efficient Universal Approach for Multimodal Sentiment Analysis with Missing Modality (2023). arXiv Prepr arXiv221003428
303. Soto, J.A., Levenson, R.W.: Emotion recognition across cultures: the influence of ethnicity on empathic accuracy and physiological linkage. Emotion **9**, 874 (2009)
304. Aguilera, A., Mellado, D., Rojas, F.: An assessment of in-the-wild datasets for multimodal emotion recognition. Sensors **23**, 5184 (2023)
305. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Comput. Surv. **55**, 1–44 (2022)
306. Pagano, T.P. et al.: Bias and unfairness in machine learning models: a systematic literature review (2022). arXiv Prepr arXiv220208176
307. Liu, Z. et al.: Contactless Respiratory Rate Monitoring for ICU Patients Based on Unsupervised Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6004–6013 (2023)
308. Qayyum, A. et al.: Assessment of physiological states from contactless face video: a sparse representation approach. Computing, pp. 1–21 (2022)
309. Zhang, X., et al.: Recent progress of optical imaging approaches for noncontact physiological signal measurement: a review. Adv. Intell. Syst. (2023). https://doi.org/10.1002/aisy.202200345
310. Li, B., et al.: Non-contact PPG signal and heart rate estimation with multi-hierarchical convolutional network. Pattern Recognit. **139**, 109421 (2023)
311. Dang, X., Chen, Z., Hao, Z.: Emotion recognition method using millimetre wave radar based on deep learning. IET Radar Sonar Navig. **16**, 1796–1808 (2022)
312. Siddiqui, H.U.R., et al.: Respiration based non-invasive approach for emotion recognition using impulse radio ultra-wide band radar and machine learning. Sensors **21**, 8336 (2021)
313. Islam, S.M.M.: Radar-based remote physiological sensing: progress, challenges, and opportunities. Front. Physiol. **13**, 2135 (2022)
314. Li, Z. et al.: MetaPhys: contactless Physiological Sensing of Multiple Subjects Using RIS-based 4D Radar. IEEE Internet Things J, (2023)
315. Tang, X., Chen, W., Mandal, S., Bi, K., Özdemir, T.: High-sensitivity electric potential sensors for non-contact monitoring of physiological signals. IEEE Access **10**, 19096–19111 (2022)
316. Abonga, C.: Evaluation of a model to detect vital signs of a subject trapped in hard-to-reach environment using a laser doppler vibrometry technique, (2022)
317. Casaccia, S., et al.: Facial muscle activity: High-sensitivity noncontact measurement using laser Doppler vibrometry. IEEE Trans. Instrum. Meas. **70**, 1–10 (2021)
318. Zhang, Y., et al.: Widar3.0: zero-effort cross-domain gesture recognition with wi-fi. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 8671–8688 (2022)
319. Bao, N., Du, J., Wu, C., Hong, D., Chen, J., Nowak, R., Lv, Z.: Wi-breath: A WiFi-based contactless and real-time respiration

monitoring scheme for remote healthcare. IEEE J. Biomed. Heal. Inf. (2022)
320. Hao, Z., et al.: Wi-CAS: a contactless method for continuous indoor human activity sensing using Wi-Fi devices. Sensors **21**, 8404 (2021)