TECHNIUM

Technium Vol. 26, pp.102-140 (2025)
ISSN: 2668-778X
www.techniumscience.com

Multimodal Recognition of Users States at Human-AI Interaction Adaptation

Izabella Krzemińska [ORCID: 0000-0002-6913-7393][1]

[1]Orange Innovation Poland, Warszawa, Al. Jerozolimskie 160

January 9, 2025

Abstract

The study investigates advancements in multimodal emotion recognition (MMER) within the evolving landscape of human-AI interaction. By synthesizing insights from psychology, computer science, and cognitive neuroscience, this paper examines the integration of multiple modalities—such as facial expressions, vocal tone, textual sentiment, and physiological signals—to achieve adaptive and emotionally intelligent AI systems. Leveraging a systematic literature review, it identifies state-of-the-art methodologies, including deep learning-based data fusion techniques, and their applications across sectors such as healthcare, education, and transportation. The review provides practical tools to support the implementation of MMER systems, such as a decision tree for selecting the most appropriate theoretical approach based on specific application needs and another for choosing the optimal fusion method. The proposed simplifications can help system designers and researchers solve key challenges, such as integrating different data types and real-time processing. The article will also address ethical issues such as privacy and anti-bias.

Keywords: Multimodal Emotion Recognition (MMER); Human-AI Interaction; Deep Learning; Adaptive AI Systems; AI Privacy & Ethics; Affective Computing; Decision Trees; Data Fusion.

1 Introduction

Multimodal emotion recognition (MMER) is a rapidly developing field that drives demand for adding emotional intelligence to AI-based creations. The ultimate goal is to improve human-computer interactions. By synthesizing data from various sources, such as facial expressions, tone of voice, text-based mood, and physiological signals, MMER systems more effectively interpret complex contexts and add emotions. This research area is primarily related to the design of adaptive, user-centric systems in line with the postulates of Industry 5.0. The theoretical framework of MMER is rooted in the psychology of emotions, including Ekman's Theory of Basic Emotions and Russell's Dimensional Model of Affect, which can help integrate multimodal data. Thanks to deep learning, technologies are rapidly developing and can increasingly combine implicit and explicit emotional signals, increasing prediction accuracy and expanding application areas. For example, combining EEG signals with facial recognition increases the efficiency of MMER systems despite the many challenges associated with collecting multimodal data.

Emotionally intelligent AI systems have transformational potential and can be used in many areas. Examples include applications already being implemented to support mental health or improve customer service quality, education or personalized healthcare. All of these applications are implemented with the requirement of real-time analysis, which currently poses the most significant challenge for analyzing large amounts of heterogeneous data types. Another challenge is undoubtedly ethical and privacy issues and regulatory requirements, including consent and information obligations. Solving these challenges requires a multidisciplinary approach integrating machine learning, psychology and user experience design.

Recent developments in combining modalities (e.g., EEG signals with facial data) have enhanced recognition accuracy by integrating implicit and overt cues despite practical challenges in collecting certain data types (Yan Wang *et al.*, 2022).

Deep learning has become central to MMER, enabling multimodal data integration and feature extraction from complex affective inputs. Studies have shown significant improvements in MMER

using neural networks, especially convolutional in MMER using neural networks, especially convolutional and recurrent models, which extract nuanced emotional information from multimodal data (Lian *et al.*, 2023).

Emotion theories continue to guide the development of MMER systems by explaining how emotions manifest across modalities and how they can be integrated to enhance accuracy in detection and response. For instance, works by (Yan Wang *et al.*, 2022) and (Zhao *et al.*, 2021) offer key insights into integrating multimodal emotion theories for enhanced user interaction.

In his book "The society of minds" Marvin Minsky emphasized the integral role of emotions in intelligence by positing that understanding and integrating emotions are crucial for developing brilliant machines (Minsky, 1988). The growing need to use MMER increases the importance of emotional intelligence in AI. Thanks to this, human-computer interactions can be more natural and more human. Machines endowed with emotional intelligence will provide personalized services, which is especially important for vulnerable groups such as the elderly, disabled and children (May *et al.*, 2017).

1.1 Purpose & Research Objectives

This review aims to explore the theoretical underpinnings of MMER and its evolution over time (1), examine the methodologies and algorithms used for multimodal data fusion (2), assess the application of MMER in virtual agents and human-AI interaction systems (3), identify current challenges, including ethical considerations and computational constraints (4), and propose future research directions for advancing MMER technologies (5). By addressing these objectives, this paper serves as a thorough reference for researchers and practitioners in the field, examining the potential and significance of emotionally aware AI systems in promoting empathetic and adaptive interactions. By addressing these objectives, this paper serves as a thorough reference for researchers and practitioners in the field, examining the potential and significance of emotionally aware AI systems in promoting empathetic and adaptive interactions.

RQ1. How can multimodal data improve the accuracy of user state recognition in AI-human interaction systems? This question examines methodologies and algorithms for integrating data from facial expressions, voice, text, and physiological signals to develop comprehensive and accurate models of user emotional and cognitive states.

RQ2. What role do individual differences (e.g., age, gender, personality, and cultural background) play in MMER performance? This inquiry investigates how user-specific factors influence the effectiveness of MMER systems and explore adaptive strategies to improve their responsiveness across diverse populations.

RQ3. Which modalities are most predictive of specific user states, and how do these modalities interact? This question evaluates the contribution of different modalities to recognizing specific emotional and cognitive states and examines how their integration optimizes recognition processes.

RQ4. How does real-time multimodal data processing affect a system's responsiveness and adaptability in dynamic environments? This question focuses on computational challenges and explores strategies to balance accuracy, speed, and resource efficiency in real-time emotion recognition.

RQ5. What ethical considerations arise from using multimodal data for user state recognition, and how can they be addressed? Key issues include data privacy, consent, potential biases, and the implications of deploying MMER systems in sensitive environments.

RQ6. To what extent can improved multimodal data processing techniques enhance user engagement and satisfaction? This question explores the impact of advanced data fusion methods on creating more intuitive and satisfying interactions in AI-driven interfaces.

RQ7. How can the optimal combination of multimodal data improve real-time user state recognition? This question goes beyond Q1 and focuses on identifying and optimizing combinations of modalities for high accuracy and efficiency in practical applications.

These questions will guide the systematic review of the literature and MMER application and methods for insights into the theoretical and practical dimensions of MMER research.

This study employs a systematic literature review to consolidate advancements in multimodal emotion recognition (MMER). The methodology follows a three-step process, as outlined by (Levy and Ellis, 2006; Okoli and Schabram, 2015):

- Search Strategy: A comprehensive search was conducted across major academic databases, including Google Scholar, SpringerLink, IEEE Xplore, ResearchGate, and ScienceDirect. Keywords such as "multimodal emotion recognition," "AI-human interaction," and "deep

learning" were used to identify relevant studies. Only peer-reviewed articles and conference papers published between 2018 and 2024 were included to ensure relevance and timeliness.

- Screening and Selection: Studies were screened based on predefined inclusion criteria, including relevance to MMER, emphasis on multimodal data integration, and practical applications in AI systems. Duplicate records and studies lacking methodological rigour were excluded.
- Data Extraction and Analysis: Key information, such as methodologies, datasets, algorithms, applications, and challenges, were extracted and categorized. This stage enabled the identification of trends, gaps, and emerging opportunities in MMER research.

The systematic approach ensures a comprehensive understanding of MMER advancements and lays the foundation for actionable recommendations in system design and future research.

2 Psychological Theory of Emotion Overview

Not only Paul Ekman's Basic Emotions Theory has served as an important milestone in psychology, but also underlining the universal nature of emotional expression. Ekman identified six primary emotions—happiness, sadness, fear, anger, surprise, and disgust—each associated with distinct and universally recognizable facial expressions. Rooted in empirical cross-cultural studies, this theory has significantly influenced diverse domains, including psychology, neuroscience, and artificial intelligence.

Applications of Ekman's theory in multimodal emotion recognition (MMER) include leveraging the Facial Action Coding System (FACS) to detect emotional states with high accuracy (Ekman and Friesen (1978), Ekman *et al.*, (2002)), what has facilitated the development of AI systems in security, customer service, and therapeutic tools. For example, MMER systems using FACS interpret subtle emotional expressions, what can be beneficial for naturalisation of human - AI interactions.

Despite its widespread influence, the theory faces criticism for oversimplifying the complexities of emotional expression. Critics argue that cultural and contextual differences significantly affect emotional recognition and perception. Studies such as those by Russell (1994) and Barrett (2006) challenge the universality of Ekman's framework, suggesting a more context-dependent and constructed nature of emotions. Nonetheless, Ekman's contributions remain pivotal, particularly in integrating facial recognition data into MMER systems for applications requiring rapid detection of fundamental emotions (Russell, 1994; Russell, 2003). Another researcher (Izard, 1994) similarly highlighted that although certain emotions are broadly recognized, the nuances of emotional expression differ depending on cultural norms.

To address these criticisms, Ekman and colleagues conducted extensive cross-cultural studies. Initial experiments, such as those involving the Fore people of Papua New Guinea (Ekman *et al.*, 1969), supported the hypothesis of universal emotional recognition. Further studies compared recognition rates across industrialized and non-industrialized cultures, including Japan, Brazil, and the United States, revealing high agreement in recognizing basic emotions (Ekman and Friesen, 1971). However, these studies also acknowledged that cultural factors can influence how emotions are displayed and interpreted, particularly regarding their intensity.

Furthermore, (Barrett, 2006) argued Ekman's theory that emotions are constructed from more basic psychological operations, which are mainly influenced by individual experiences and cultural contexts. Nevertheless, other research consistently showed high agreement across different groups, suggesting a biological basis for these emotional expressions (Ekman and Cordaro, 2011).

Ekman's foundational studies were expanded upon by subsequent research utilizing modern technology such as electromyography (EMG) to measure the muscle activity associated with specific facial expressions, thereby providing quantitative evidence supporting the theory (Hess and Thibault, 2009).

Moreover, neuro-scientific advancements have identified brain structures and pathways involved in processing basic emotions, linking specific emotions to distinct patterns of brain activity (Lindquist *et al.*, 2012).

Ekman's theory has been directly applied in developing effective computing systems used in customer service and security. Affectiva is a company known for specializing in emotion recognition, and its algorithms are using FACS to analyze facial expressions in real-time, enhancing customer interaction by providing immediate emotional feedback to service agents (McDuff *et al.*, 2014).

Another application is in security and surveillance. Systems like the one developed by iMotions use Ekman's FACS to detect suspicious behaviours by analyzing micro-expressions, thus aiding

in threat assessment and crime prevention (Lei *et al.*, 2017; Otamendi, 2022; Sham *et al.*, 2023). These applications of FACS are practical evidence of Ekman's theory implications and adress **RQ4**) by showcasing how real-time processing of facial expression data can enhance the responsiveness and adaptability of AI systems in dynamic interaction environments.

In addition, the development of high-speed cameras and advanced image processing technologies has allowed for the detailed study of micro-expressions—brief, involuntary facial expressions that manifest as concealed emotions. These studies have further validated Ekman's claims about the involuntary nature of emotional expressions and their potential use in various applications, from psychology to security (Porter and Ten Brinke, 2008).

2.1 The Dimensional Model of Affect (James Russell)

James Russell's Dimensional Model of Affect provides a continuous framework for understanding emotions, categorizing them along the axes of valence (pleasantness-unpleasantness) and arousal (activation-deactivation) (Russell, 1980; Russell *et al.*, 1989). This model enables the representation of emotions as points in a two-dimensional space, with extensions such as the dominance dimension adding further granularity to emotional classification. For instance, excitement and fear share high arousal levels but differ significantly in valence. The inclusion of a third dimension, **dominance** (the degree of control one feels over a situation), further refines the model, adding depth to its applications (Sutton *et al.*, 2019).

The Dimensional Model's flexibility makes it particularly suitable for applications requiring nuanced emotional analysis, such as virtual assistants, gaming environments, and educational technologies (Reuderink *et al.*, 2013). By mapping emotions on a continuous scale, AI systems can dynamically adjust their interactions to improve user engagement and satisfaction. However, critics argue that this model may oversimplify emotional experiences by reducing them to only a few dimensions. Additionally, cultural and individual variability in emotional expression poses challenges to its universal applicability.

Nevertheless, the Dimensional Model remains a cornerstone in MMER research, particularly in applications demanding real-time monitoring and adaptive responses. Recent advancements in deep learning have further enhanced its implementation, enabling multimodal data integration for improved accuracy in recognizing and predicting user states. The Dominance pertains to the degree of control and influence one feels over a situation or environment when experiencing an emotion. It ranges from feelings of dominance and empowerment to feelings of submission and powerlessness. Emotions such as pride and anger typically involve high dominance, while fear and helplessness are characterized by low dominance. (Fontaine *et al.*, 2013) explored how dominance, as a dimension of emotion, contributes to a more nuanced understanding of emotional experiences and their expression in social contexts.

The flexibility of this approach has led to its adoption in diverse fields. In virtual reality (VR), systems employing the dimensional model adjust environmental elements in real-time to match users' emotional states, as shown in studies like (Riva *et al.*, 2019). In educational technology, tools such as those studied by (D'Mello *et al.*, 2017), utilize valence-arousal tracking to tailor instruction to students' emotional states, boosting learning outcomes. These examples highlight the model's potential to enhance user satisfaction and performance across applications.

Critics of the Dimensional Model argue that it may oversimplify the complexities of emotional experiences by reducing them to just two or three dimensions. Research by Barrett (2017), emphasizes that emotions are often context-dependent and influenced by individual differences, which may not be fully captured by the valence-arousal-dominance framework. Furthermore, cultural and social factors can affect how emotions are experienced and expressed, challenging the universal applicability of this model.

Despite these criticisms, the Dimensional Model remains a cornerstone of emotion recognition research, particularly in systems requiring continuous monitoring of emotional states. Its adaptability, combined with its compatibility with multimodal data sources like voice, text, and facial expressions, makes it a powerful tool for advancing human-computer interaction.

Recent research, such as the work by (Buechel and Hahn, 2017), developed a model that integrates these three dimensions into sentiment analysis, showing improvements in the granularity and accuracy of emotional predictions.

The Dimensional Model of Affect is essential in answering **RQ3** as it helps identify which modalities (e.g., facial expressions, voice) are most predictive of specific user states. By mapping emotions along valence and arousal dimensions, this model facilitates a deeper understanding of how different modalities contribute to the overall emotional state.

2.2 The Appraisal Theory of Emotion (Richard Lazarus)

Richard Lazarus's Appraisal Theory emphasizes the cognitive evaluation—or appraisal—of events as the primary determinant of emotional responses (Moors *et al.*, 2013; Lazarus, 1991). Emotions, according to this theory, are not triggered solely by the events themselves but by the personal significance attributed to those events. This appraisal process is categorized into two types:

- Primary Appraisal: Determines whether an event is perceived as a threat, challenge, or benefit.
- Secondary Appraisal: Assesses the resources available to cope with the situation.

According to this perspective, different emotional responses can arise from the same situation depending on how an individual appraises the event, such as whether they perceive it as threatening, beneficial, or irrelevant to their goals (Lazarus, 1968). For example, encountering a dog may evoke fear in someone who appraises the dog as a threat but joy in another who appraises the encounter as an opportunity for companionship. According (Scherer *et al.*, 2001) appraisal involves the evaluation of various factors, including the potential for harm or benefit, the congruence with personal values and goals, and the perceived ability to cope with the consequences of the event. These appraisals can be immediate and automatic, or they can involve more deliberate processing.

This theory has significant implications for emotion recognition systems, as it highlights the variability in emotional responses based on personal and contextual factors. MMER systems integrating this model can analyze users' emotional reactions more effectively by considering individual goals, values, and situational contexts. For instance, in health services, systems employing appraisal theory can interpret patients' emotions within the context of their health concerns, tailoring responses to their specific needs. Adaptive tutoring systems, such as those developed by (Conati and Maclaren, 2009) use appraisal mechanisms to adjust educational strategies in response to students' emotions, improving learning outcomes.

Critics argue that the Appraisal Theory may overemphasize the cognitive aspects of emotion, potentially neglecting the role of automatic, unconscious processes. (Zajonc, 1984) contended that some emotional responses occur without conscious evaluation, challenging the theory's premise. However, experiments by (Smith and Lazarus, 1993) demonstrated that cognitive appraisals often predict emotional responses, providing robust support for the theory's relevance.

This theory also underscores the importance of individual differences in emotional expression, addressing how cultural backgrounds, personal history, and situational factors influence appraisals and is widely used by psychotherapists (Lisetti and Nasoz, 2002; Moors *et al.*, 2013), although in form of self-reports. This makes it particularly valuable for systems requiring a high degree of personalization, such as virtual assistants or educational tools for diverse populations.

By integrating the Appraisal Theory into MMER systems, researchers can develop applications capable of tailoring responses to the user's emotional context, enhancing adaptability and user satisfaction. However, implementing this theory in real-time systems remains challenging due to the computional modeling complexity of subjective appraisals.

2.3 Affective Processes Theory (Klaus Scherer)

Klaus Scherer's Affective Processes Theory, particularly his Component Process Model (CPM), proposes that emotions arise from the dynamic interaction of multiple subsystems. These include cognitive appraisals, physiological responses, motor expressions, and subjective feelings. Each component contributes to the emotional experience, operating on different timescales and intensities. Scherer emphasized that these processes are interdependent, with appraisals triggering physiological and behavioral responses (Scherer, 1982).

Key components of Scherer's CPM include:

- **Cognitive Appraisal**: Similar to Lazarus's theory, it involves evaluating the relevance of events to personal goals.
- **Physiological Responses**: Bodily changes such as heart rate, skin conductance, and neural activity.
- **Motor Expressions**: Observable behaviors like facial expressions, gestures, and posture.
- **Subjective Feelings**: The personal experience of emotion, often reported through self-assessment.

Magda B. Arnold is recognized for her pioneering work on the appraisal theory of emotion, which laid the groundwork for understanding the cognitive aspects of emotional processes (Arnold, 2013). Arnold's theories emphasized the role of personal judgment and evaluation in emotional responses, which influenced later Scherer's CPM.

The CPM theory provides a comprehensive framework for understanding the complexity of emotions, making it well-suited for applications in multimodal emotion recognition (MMER). By integrating multiple data sources, such as physiological signals, facial expressions, and textual inputs, CPM-based systems can enhance the accuracy and depth of emotion recognition. For example, combining EEG data with facial recognition allows MMER systems to better capture user states in contexts like healthcare and human-computer interaction.

Despite its strengths, implementing CPM poses challenges due to its complexity. Real-time processing of multimodal data requires robust algorithms capable of synchronizing diverse inputs. Additionally, the subjectivity of emotional experiences complicates model validation. Scherer himself acknowledged the difficulty of achieving consistency in interpreting multimodal data (Scherer, 2009).

Scherer's theory remains influential in advancing MMER systems, especially in applications demanding detailed and nuanced emotional understanding (Scherer and Wallbott, 1994). Its emphasis on integrating cognitive, physiological, and behavioral data ensures a robust framework for developing responsive and empathetic AI.

Additionally, Nico H. Frijda's work on the action readiness theory provided insights into the motivational aspects of emotions, further enriching the affective processes framework (Frijda, 1986). Frijda's theories focused on how emotions prepare and motivate individuals to respond to environmental challenges, aligning with the idea that emotions involve multiple interrelated processes.

2.4 Other Theories of Emotion

Beyond traditional frameworks such as Ekman's Basic Emotions Theory and Russell's Dimensional Model, several other theories of emotion provide nuanced perspectives essential for advancing Multimodal Emotion Recognition (MMER) systems. These include Psychological Construction Theory, Social Constructivist Theory, and emerging approaches such as Dynamic Emotion Processes and Neurocognitive Models. Together, these frameworks address the complexities of emotional expression, variability, and context, enriching the design and adaptability of MMER technologies.

Psychological Construction Theory Lisa Feldman Barrett's Psychological Construction Theory redefines emotions as emergent phenomena constructed from core affect and conceptual knowledge shaped by individual experiences and cultural contexts (Barrett, 2017). This dynamic model challenges the universality of discrete emotions, emphasizing variability and context-dependence. For instance, sadness or happiness may manifest differently across cultural settings due to distinct interpretive frameworks.

This theory is particularly significant for cross-cultural and personalized MMER systems. Adaptive algorithms informed by Barrett's insights can tailor emotion recognition models to account for user-specific contexts and cultural norms, improving both accuracy and user satisfaction.

Social Constructivist Theory Lev Vygotsky's Social Constructivist Theory emphasizes the role of social and cultural factors in shaping emotional expression and experience. Emotions are not purely biological phenomena but are influenced by societal norms and interpersonal dynamics (Vygotsky and Cole, 1978; Bruner, 1990). For example, collectivistic cultures prioritize emotional expressions that maintain group harmony, while individualistic societies may encourage open displays of emotion (Mesquita *et al.*, 2016; Markus and Kitayama, 2010).

This theory is highly relevant for MMER systems designed for collaborative and culturally diverse environments, such as social robotics or virtual reality platforms. By incorporating social and contextual data, these systems achieve greater accuracy in interpreting emotions, enhancing interaction quality and cultural sensitivity (Hareli and Parkinson, 2008; Ghandeharioun *et al.*, 2019).

New Dynamic and Neurocognitive Models New theories such as Dynamic Emotion Processes and Neurocognitive Models add another aspect of variability in emotion recognition:

Researchers representing the dynamic trend view emotions as fluid and evolving rather than fixed states. This perspective benefits MMER systems that require real-time adaptation to changing emotional landscapes, such as games or healthcare (Kuppens *et al.*, 2010).

On the other hand, proponents of neurocognitive models like Pessoa (2013) investigate the brain mechanisms underlying emotional experiences. They do this using neurophysiological data such as EEG and fMRI (Dehghani *et al.*, 2023). According to research, neurocognitive models improve emotion recognition accuracy in clinical and psychological contexts (Dricu and Frühholz, 2020).

Incorporating these new theories into MMER systems allows for a multidimensional approach to emotion recognition, taking into account users' changing needs and contexts. However, they also seem challenging to implement due to the brain imaging techniques. If implemented in MMER, both approaches could enhance cultural and individual sensitivity by considering the variability of emotional expressions and increase contextual accuracy by integrating social and cultural factors in the interpretation of emotions. Dynamic and neurocognitive models provide a robust framework for real-time monitoring and advanced analysis, allowing systems to better adapt to the use case and thus be more versatile and scalable. However, this is not currently the approach used in MMER, probably due to the lack of available labelled training sets. The variability postulated by the models would require the preparation of a large number of data sets based on both intra- and inter-individual expression variability.

Table I: Comparative analysis of emotion theories

Theory	**Premise**	**Applications**	**Cultural Aspect**	**Critiques**
Basic Emotions Theory (Ekman)	Emotions are universal and biologically based, with distinct facial expressions for six core emotions.	Facial recognition, security systems, emotion-based AI applications.	Claims universality but limited by variability in expression across cultures.	Fails to address cultural influence on emotional perception.
Dimensional Model of Affect (Russell)	Emotions are represented along continuous dimensions of valence (positive-negative) and arousal (activation-deactivation).	User experience design, VR systems, sentiment analysis.	Framework is adaptable to different cultural interpretations of arousal and valence.	Oversimplifies complex emotional states, ignoring cultural subtleties.
Appraisal Theory (Lazarus)	Emotions arise from cognitive evaluations of events as threats, challenges, or benefits.	Health services, adaptive learning environments, personalized applications.	Cultural norms shape what is considered a threat or benefit, affecting appraisals.	Subjectivity complicates modeling across diverse cultural backgrounds.
Psychological Construction Theory (Barrett)	Emotions are constructed from core affect and conceptual knowledge, influenced by context and culture.	Cross-cultural AI systems, personalized emotion recognition models.	Directly addresses variability in emotional expression across cultures.	Complex computational implementation due to dynamic, contextual nature.
Social Constructivist Theory (Vygotsky)	Emotions are shaped by social norms and cultural practices rather than being purely biological.	Social robots, collaborative virtual environments, group-based AI interactions.	Highlights cultural specificity of emotional norms and expressions.	Neglects biological underpinnings, limiting physiological modeling.
Affective Processes Theory (Scherer)	Emotions result from the interaction of cognitive appraisals, physiological responses, motor expressions, and subjective feelings.	Multimodal AI systems, clinical emotion recognition, affective computing.	Applicable globally but requires cultural calibration of emotional components.	Challenging to integrate multimodal data in real-time systems.

A comparative analysis of the main features of selected emotion theories is presented in the table I. Each theory is assessed based on basic assumptions, cultural aspects, applications and criticism. Emotion theories, like most psychological concepts created from different perspectives

and premises, have a tangible goal of adapting these theories to specific applications and contexts. For example, Ekman's theory emphasizes the universality of emotions, which makes it better suited to systems requiring rapid detection of basic states, but it is limited by cultural dependence. In turn, the dimensional model of affect (Russell) provides a flexible framework for continuous tracking of valence and arousal but may oversimplify emotional nuances. Appraisal Theory (Lazarus) focuses on the role of cognitive evaluations, enabling personalized and context-aware applications like health services. However, it faces challenges in real-time implementation. Psychological Construction Theory (Barrett) strongly emphasizes individual and cultural differences, which may mean high computational requirements, difference from other theories, and encourage its use in cross-cultural AI applications. In turn, Social Constructivist Theory (Vygotsky) emphasizes the influence of social norms and may be better for implementation in Human-AI collaboration situations but less effective in capturing biological processes. Affective Processes Theory (Scherer) offers a comprehensive, multi-component approach that is valuable for detailed emotion recognition but demands computational resources and implementation.

These frameworks complement one another, offering diverse tools for MMER development depending on the requirements for universality, adaptability, or granularity in emotional recognition systems.

2.5 Decision tree for choosing best emotion theory for UC

Based on this literature review and the gathered knowledge, a simple decision tree was created to help identify which Theory of Emotion is more suitable for the deliberated use case. Choosing the appropriate emotional approach for MMER systems depends on the application's requirements and context, such as the need for cultural sensitivity, the complexity of emotions being recognized, and the technological context.

Simple Decision Tree: What Emotional Approach to Use? (own work)

- Q1. Do you need to recognize basic, universal emotions quickly and reliably?
  - Yes: Consider **Basic Emotions Theory (Ekman)** for straightforward applications requiring rapid identification of fundamental emotions.
  - No: Go to question 2
- Q2. Is your application focused on continuous monitoring and nuanced emotional states?
  - Yes: Consider the Dimensional Model of Affect (Russell) for applications requiring continuous tracking of emotional states across dimensions such as valence and arousal. Proceed also to the next question to explore the need for context-aware interpretation.
  - If the focus is not on continuous monitoring but rather on specific situational contexts, consider theories that allow for contextual evaluation. Proceed to the next question.
- Q3. Does your application require understanding how emotions are constructed from core affect and conceptual knowledge (e.g., influenced by specific contexts or cognitive evaluations)?
  - Yes: Consider Appraisal Theory for applications that need to evaluate the context or situation in which emotions arise and adapt responses based on this cognitive evaluation.
  - No: If the application does not require this level of cognitive appraisal and instead focuses on straightforward emotional recognition or tracking, consider the Dimensional Model of Affect or Basic Emotions Theory.
- Q4. Does your application require a combination of continuous monitoring and context-sensitive interpretation?
  - Yes: Consider combining Appraisal Theory with the Dimensional Model of Affect. This allows for continuous monitoring of emotional states with the flexibility to interpret those states based on contextual factors.
  - No: If only one aspect is needed (monitoring or contextual interpretation), choose the appropriate theory from the previous steps.
- Q5. Is real-time processing a critical component of your application?

- Yes: Depending on the previous answers, check whether the selected model (or combination of models) can be processed in real-time. The appraisal theory and dimensional affect model will probably require more computational resources than Ekman's classification models. The challenge will be to choose the appropriate time frame for class aggregation.
- No: If real-time processing is unnecessary, select the theory that best aligns with your application's needs, focusing on contextual understanding or emotion tracking.

Disclaimer: *Decision trees are designed to simplify complex decision-making processes by guiding users through a series of binary or categorical choices, that means that there is possibility that the decision tree may not perfectly align with all use cases.*

Example Application of Decision Tree for the Emotional Virtual Chatbot Based on the decision tree, let's discus the following use case: Emotional Virtual Chatbot as a Support Assistant at Hotel Reception.

Possible Scenario: The virtual chatbot is designed to assist hotel guests by providing required information, analyzing their emotions from video, text, and voice inputs during conversations in real-time (RT). The chatbot adapts its responses based on the customer's emotional state to improve service quality and assess its own performance.

Given the use case of an emotional virtual chatbot at a hotel reception, we can apply the decision tree for selecting an appropriate emotion theory model as follows:

- The chatbot operates within a Human-Computer Interaction (HCI) domain, specifically tailored for customer service in a hospitality setting. So we look for an emotion theory that supports real-time interaction and is sensitive to customer satisfaction.
- The chatbot processes video, text, and voice inputs. These multimodal data sources provide comprehensive information about the customer's emotional state. This could lead us to consider Psychological Construction Theory or Appraisal Theory due to their capacity to handle complex, multimodal data and adapt to the context.
- The chatbot needs to analyze emotions and adapt its responses in real-time. This requires an emotion theory that can support dynamic and continuous emotion assessment, which reinforces the selection of Appraisal Theory due to its focus on how individuals appraise events and how those appraisals influence emotional responses in real-time.
- The chatbot must handle a range of emotional expressions and adapt its responses dynamically, what require high flexibility and handling of complex interactions. The recommended theory can be Psychological Construction Theory, which is robust in integrating multiple modalities and adapting to varying emotional expressions.

Finally:

AD Q1 No, quick recognition of basic emotions is not the primary need; the chatbot requires nuanced emotional understanding.

AD Q2. Yes, the chatbot continuously monitors emotional states across multiple modalities (video, text, voice).

AD Q3. Yes, the chatbot needs to adapt its responses based on the specific context of the interaction.

AD Q4. Yes, both continuous monitoring and context aware interpretation are critical.

AD Q5. Yes, real-time processing is essential for effective interaction.

Final Model Selection: Combine Appraisal Theory with the Dimensional Model of Affect. Appraisal Theory will help the chatbot interpret emotions based on the context of the interaction. The Dimensional Model of Affect will allow the chatbot to continuously monitor and track the emotional states of users, providing a real-time understanding of changes in emotions.

Implementation Strategy should be based on two key elements: Real-Time integration & Customisation. Implement both models in tandem to ensure that the chatbot can continuously track emotional states while also adjusting its behavior based on cognitive appraisals of the situation. Fine-tune the models to prioritize either continuous monitoring or context-sensitive adaptation based on real-time demands and user interaction complexity.

A proposal would be to start with a Dimensional Model of Affect to monitor the user's emotional state in real time. The system will then process input from multiple modalities (e.g. video, text, voice, etc.) and map it onto dimensions of valence and arousal. This will allow for continuous tracking of the user's overall emotional state and detecting changes in emotions that are relevant to the interaction. Once an emotion is detected or a significant shift in the emotional state is

identified, the system applies Appraisal Theory to interpret the context of the emotion. Then the system evaluates the current situation, user interactions, and any relevant contextual factors (e.g., the content of the conversation, user history, specific events, etc) to determine the cause and meaning of the detected emotional state. For example, if the system detects a high-arousal, negative-valence state (such as frustration), it would then use appraisal mechanisms to determine if the frustration is due to a specific issue (e.r, a misunderstanding or a delay in service, etc).

Based on the combined results of the Dimensional Model of the Affect Theory and the Evaluation Theory models, the system generates an appropriate response. In turn, the classification according to the Dimensional Model provides the system with a continuous understanding of the user's emotional state, making it sensitive to changes over time.

This feedback loop ensures that the system becomes responsive and adaptive throughout the interaction, providing good context-dependent performance.

3 Application of Different models of Emotion

This section highlights how frameworks like Scherer's CPM, Russell's Dimensional Model of Affect, and Ekman's Basic Emotions Theory have been operationalized in real-world systems. For example, CPM integrating cognitive appraisals, physiological responses, and motor expressions, is a solid theoretical framework for understanding emotions in complex systems, essential for communication systems dedicated mainly to healthcare. Russell's Dimensional Model of Affect is a valuable framework for tracking emotional states continuously across dimensions, for example, in gaming, offering adjustments like in-game assistance during stressful situations. Ekman's Basic Emotions Theory focus on universal, discrete emotions, making it invaluable for quick and reliable emotion recognition.

Emotion recognition technologies enhance existing systems and catalyze innovation in personalized healthcare, adaptive learning, and virtual social spaces. For example, emotion-aware virtual assistants now respond to users more precisely by interpreting frustration or satisfaction from vocal and textual cues. Similarly, gaming platforms dynamically adjust challenges based on real-time emotional feedback, fostering engagement and reducing player frustration.

Ekman's theory has guided the development of algorithms for emotion recognition, which are now used in areas ranging from marketing to interactive gaming and mental health monitoring (Reardon *et al.*, 2019). These techniques can analyze facial expressions captured on video and can use them to assess the viewer's emotional state. Game designers incorporate this kind of data to adapt human-computer communication and the game's script. This data is used to adjust difficulty levels, narrative choices, and the game's dynamics. For example, a game that detects high arousal and negative valence (indicating stress or frustration) might offer in-game assistance or lower the difficulty to prevent the player from dropping out and getting discouraged. On the other hand, recognizing positive valence and high arousal might signal the game to increase the challenge to keep the player engaged while also potentially discouraging compulsive gamers with addictive tendencies. Studies by Hudlicka (2019) have shown that adaptive gaming systems that respond to player emotions can lead to more personalized and satisfying gaming experiences.

Also, the concept of metaverse, a virtual shared space, uses emotion recognition to deepen the effect of immersion and naturalness of the interactive environment. By reading the user's emotional states, metaverse platforms adapt to the virtual environment and create a digital experience for individual people. For instance, virtual social spaces can adjust the ambience and interactions based on the collective emotional state of participants, promoting positive social interactions and reducing harassment or harmful behaviour. Research by Liu *et al.*, (2024) and Singh and Kaunert (2024) highlights the potential of emotion-aware systems in the metaverse to create more inclusive and engaging virtual communities.

Recognizing and responding to user emotions, thus making interactions with machines more human, can improve user satisfaction in digitally delivered services such as e-commerce, online education, and health services. Mimicking human adaptive responses, such as detecting hesitation or confusion (negative valence, moderate arousal) while shopping or talking to a teacher, can prompt the system to offer additional information about the product or other assistance until satisfaction or understanding is achieved. Recognizing students' emotional states in online education can help educators adjust their teaching methods, provide timely support, and create more engaging learning experiences. Also, there are some applications in the transportation sector. For example, an emotional monitoring system in vehicles can detect drivers' frustration in real-time during a traffic jam and respond by playing soothing music, helping to calm the driver and reinforce safe driving practices (Zepf *et al.*, 2020).

Current research by Cambria *et al.*, (2024) explores and refines the integration of emotional dimensions in digital experiences. According to the authors, advances in machine learning and natural language processing are leading to the development of increasingly sophisticated and accurate emotion recognition systems. Methods based on deep learning can already analyze facial expressions, tone of voice, and text and use this to infer the emotional states of the interlocutor.

Implementing Klaus Scherer's complex theoretical component process model (CPM) in the MMER system is still challenging. The fundamental problem is still collecting and properly synchronizing data from different sensors, each with a different sampling rate and highly susceptible to noise and artefacts cites scherer2009dynamic. Precise time-stamping and preprocessing methods are essential for effective data alignment and cleaning. Integrating physiological, behavioural, and self-report sources is particularly difficult because these data types often come in different formats and levels of detail. Researchers continue to search for suitable, transient and universal techniques. To cope with complexity, advanced machine learning algorithms capable of combining heterogeneous data are used, such as autoencoders, Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs), Gaussian Mixture Models (GMMs) or Graph Neural Models (GNNs) (Picard, 2010).

MMER requires real-time processing of large amounts of complex data to be functional. Much attention is currently being paid to optimizing available algorithms for their performance using edge computing techniques (Zeng *et al.*, 2007). Emotions are inherently subjective, variable, and dependent on individual differences, cultural background, and situational contexts. Therefore, self-adaptive models that learn and adapt based on contextual information are being created, increasing the accuracy of emotion recognition (Barrett, 2006). Another challenge is the reliability and accuracy of sensors over time because they require regular calibration and are susceptible to environmental factors (lighting) or user behaviour (face tilt angle). The solution may be integrating calibration procedures and using redundant sensors to ensure system continuity and integrity. Moreover, the collection and processing of sensitive data for emotion recognition, as well as the results of recognition models, require the highest levels and standards of data protection, such as encryption and anonymization, as well as obtaining informed consent to the processing and use of data along with ensuring transparency of their use (Scherer, 2009).

Explainable AI techniques provide additional insight, ensuring that even complex models remain understandable (Gunning and Aha, 2019). Despite these challenges, effective implementation of CPM in MMER systems promises a nuanced and comprehensive understanding of human emotions, crucial for applications in affective computing and psychological research (Zeng *et al.*, 2007).

The application of understanding valence, arousal, and dominance (James Russel model) extends beyond theoretical frameworks and sentiment analysis into practical domains, such as user state recognition in digital experiences. These include virtual assistants, gaming environments, the metaverse, and other digital services. In virtual gaming, emotion recognition personalizes game dynamics to increase user engagement and satisfaction. Games use emotional information to adjust, for example, difficulty level, narrative choices, and game pacing. For example, a game that detects a player's high arousal and negative valence (signifying stress or frustration) may offer in-game assistance or lower the difficulty to prevent the player from getting discouraged and failing. In contrast, recognition of positive valence and high arousal may signal to the game the need to increase challenge to maintain engagement. Studies by Hudlicka (2016), Akbar *et al.*, (2019), Frachi *et al.*, (2023), J. C. Lopes and R. P. Lopes (2022), Bălan *et al.*, (2020), Rezapour *et al.*, (2024), and Kadyr and Tolganay (2024) have shown that adaptive gaming systems that respond to player emotions can lead to a more personalized and satisfying gaming experience.

Virtual social spaces like metaverse can adjust the atmosphere and interactions for individuals and entire communities by analyzing the collective emotional state of participants. Such collective adjustment can enhance positive social interactions and reduce instances of harassment or harmful behaviour. Research by Pervez *et al.*, (2024) and Khalaf *et al.*, (2024) highlights the potential of emotion-aware systems in the metaverse to create more inclusive and engaging virtual communities. These applications address **RQ3** and **RQ6** by demonstrating how multimodal emotion recognition can enhance user engagement and satisfaction across various digital platforms.

Currently, increasingly sophisticated MMER systems that integrate data from multiple sources, based on CNNs and RNNs, have significantly enhanced the ability to process and understand complex emotional cues from various modalities as presented by LeCun *et al.*, (2015). Moreover, integrating multimodal data—combining audio, visual, textual, and physiological inputs—has proven more effective than unimodal approaches, leading to more reliable and nuanced emotion recognition systems (Zadeh *et al.*, 2017).

Kwon *et al.*, (2022) presents a framework for analyzing facial expressions to assess game experiences. The experimental setup and results, including accuracy and performance metrics, are

detailed, showing the potential for integrating this framework into game testing. The study delivers evidence that deep learning-based analysis of facial expressions can provide valuable insights into player experiences, which can be used to improve game design and testing processes.

In virtual assistants, recognizing and responding to user emotions is critical for creating more intuitive Virtual assistants, like Alexa, Siri and Google Assistant, benefit from models that incorporate valence, arousal, and dominance to interpret user queries better and provide appropriate responses. The possibility of detecting frustration during the conversation (negative valence, high arousal, and low dominance) in a user's voice can prompt the assistant to offer additional help or simplify instructions. Conversely, satisfaction recognition (positive valence, moderate arousal, and high dominance) allows the assistant to maintain its current interaction style. Research by Nasir *et al.*, (2022) has demonstrated that integrating emotional recognition into virtual assistants can significantly enhance user satisfaction and interaction quality.

Similarly, a study by A. Savchenko and LV Savchenko (2022) and Lyudmila Savchenko and V Savchenko (2021) demonstrated that audio-visual models provide a more comprehensive understanding of emotional states than those relying solely on visual or audio data. Similarly, (R. Chen *et al.*, 2022) applied a cross-modal auxiliary video network, finding that integrating visual and audio data improves accuracy in recognizing subtle emotions compared to unimodal systems.

The Dimensional Model of Affect is particularly conducive to MMER because it provides a flexible, continuous spectrum for emotion analysis, ideal for handling the variances in multimodal data. Emotions in this model are mapped based on valence (pleasant-unpleasant) and arousal (activated-deactivated), dimensions that can be effectively measured through algorithms analyzing physiological responses, voice tone, linguistic content, and facial expressions.

The usefulness of the Dimensional Model of emotion for MMER stems from its adaptive ability to interpret ambiguous or subtle emotional signals often present in multimodal input data. For instance, a video's combination of voice intonation and facial expressions can provide complementary information about the arousal and valence states, which can be integrated to predict more complex emotional states. This model's application in technology is extensively discussed in research that examines how different modalities contribute uniquely to understanding emotions (Russell, 2003; Mehu and Scherer, 2015).

Although highly relevant for understanding the cognitive processes behind emotions, the appraisal theory is less frequently used in MMER computational models. This is due to cognitive appraisals' subjective and intricate nature, which are difficult to quantify and model using current technologies. According to Scherer *et al.*, (2001), the theory's emphasis on individual differences and contextual factors presents additional challenges in creating generalized models that can accurately predict emotions across diverse users and contexts.

Emotional chatbot and virtual agent - EMMA - UC The article "EMMA: Emotion-Aware Wellbeing Chatbot" by Ghandeharioun *et al.*, (2019) examines the EMMA project, an emotionally intelligent chatbot for mental health interventions. EMMA has automated mood detection using data from smartphone sensors, including geolocation and activity. The system is based on Russell's two-dimensional model of emotion (valence and arousal) and machine learning to analyze patterns and attempt to provide personalized mental health support. It leverages multimodal data using classification models (e.g. logistic regression, random forest) to categorize mood into binary classes of valence and arousal. In turn, regression models (e.g. linear regression, support vector regression) predict the exact dimensions continuously. Regression models take into account individual differences by adjusting for personal reference points. The case of Emma shows that automated emotion detection is possible without sacrificing user perception. The key gap identified in the study is balancing automation with the need for user control to optimize engagement and effectiveness of the intervention and the long-term impact of using such a system. The conclusion is that using different emotion recognition models shows the transformative potential of integrating affective computing with human-AI interactions. Multimodality through the use of diverse data such as facial expressions, voice, text and physiological signals significantly increases accuracy, individualizes the experience towards naturalness and improves AI responsiveness. With a solid theoretical framework, different models, Scherer's Component Process Model, Russell's Dimensional Model of Affect and Ekman's Theory of Basic Emotions, researchers have significantly improved in many applications. Moreover, although real-time processing, data fusion, dealing with cultural and contextual variability, and ethical challenges remain challenges, the current rapid evolution of these models promises better results in creating empathetic, adaptive, and user-centric AI systems.

4 Multimodal vs Unimodal Emotion Recognition Models

Research on combining multiple data streams (intonation, facial expressions, and text content) highlights the importance of multimodality in improving emotion recognition techniques. Each modality brings unique information that, when integrated, provides a better understanding of user states and helps ensure performance stability despite significant individual differences. One notable advance is developing the Multimodal Transformer model that effectively combines data from different modalities. R. Wang *et al.*, (2024) demonstrated that their multimodal transformer outperforms unimodal models significantly in emotion detection tasks. The proposed models use artificial attention mechanisms to dynamically weigh different modalities' momentary importance, providing a more accurate interpretation of user states.

Speech analysis techniques provide increasingly nuanced data by examining prosodic features such as stress, intonation, rhythm, voice pitch (tone) and duration, speech rate, and pauses. Researchers like Choo *et al.*, (2023) have tested a deep learning model that analyzes prosody features in real-time, which improved the accuracy of assessing emotional states such as stress, happiness, and anger. The proposed by researcher model can be integrated with a virtual assistant system, providing new helpful information to customize user interactions by adapting the assistant's responses based on detected emotions.

Advances in computer vision and deep learning have enabled more accurate and real-time analysis of facial cues. Researchers Peng *et al.*, (2017) and Belaiche *et al.*, (2020) have developed a facial emotion recognition system that uses CNNs to detect microexpressions—subtle facial movements that reveal genuine emotions. The system has been integrated with various applications, from health care to virtual reality, providing immediate feedback on the user's engagement and emotional reactions. Recent advances in NLP have focused on transformer-based models such as BERT and GPT-3, demonstrating remarkable capabilities in understanding context and sentiment. The EmoBERTa model, proposed by Kim and Vossen (2021), fine-tunes BERT to detect emotions in text, achieving state-of-the-art results on multiple sentiment analysis benchmarks. EmoBERTa has also been validated for detecting depressive symptoms in text (Bucur *et al.*, 2023).

Enriching existing content-based models with voice and facial expression analysis will likely improve emotion recognition performance. The power of multimodal models is closely related to their improved ability to integrate different data sources. A pioneering study by Yong Zhang *et al.*, (2021) introduced a framework that combines voice, image, and text data to create a composite user profile. This approach uses a hierarchical fusion strategy, in which each modality is procesed individually before combining them into a unified representation. This approach allows the system to handle missing or noisy data more effectively, ensuring performance even in imperfect conditions. Experimental results shown by Akbar *et al.*, (2019) proof the accuracy and effectiveness of dynamically balancing game difficulty based on players' emotional states. Authors suggests future research to incorporate audio data and semi-supervised learning for even better performance. Additionally, it is a proof of concept for the simultaneous use of uni- and multimodal models.

Integration of multimodal models also has potential in various digital experiences (Butz, 2010). Hamdy and King (2018) proposes a Complex Multimodal Classification System used to recognize and respond to players' emotions in games with all its benefits. The conclusions highlight the potential of multimodal classification systems to significantly enhance gaming experiences by making them more responsive to players' emotions. Adapting the virtual assistant to the user, including counteracting or preventing negative emotions, is possible in general conditions (Rahman *et al.*, 2024) and in dedicated narrow scenarios in games (Hu *et al.*, 2024). Such personalization of the user experience increases engagement and, thus, ultimate satisfaction. Furthermore, by understanding the user's emotions through voice, facial expressions, and text, metaverse platforms will create more responsive and engaging virtual environments (Sayyed *et al.*, 2024). Recent studies like Kalateh *et al.*, (2024) have highlighted the importance of integrating multimodal emotion recognition approaches that use audio, visual, and textual data to improve the accuracy and reliability of systems used for patient monitoring (Khan *et al.*, 2024). Khan *et al.*, (2024) explores the potential of contactless multimodal emotion recognition (CMER) systems, highlighting the advantages of non-invasive sensors such as RGB, infrared, and Wi-Fi. These sensors capture physiological and behavioural cues without requiring direct skin contact, which improves user comfort and natural interaction. The development of multimodal non-contact techniques has excellent potential for use in the entertainment industry (gaming), healthcare (remote monitoring of the elderly), and work environments where the psychophysical state directly affects the quality of work performed, on which the safety of others depends.

Comparative analyses consistently show that multimodal emotion recognition models signifi-

cantly outperform unimodal models. For example, Mocanu *et al.*, (2023) explicitly demonstrated that audiovisual models provide a more comprehensive understanding of emotional states than those that rely solely on visual or auditory data. Similarly, He *et al.*, (2023) applied a video-based cross-modal auxiliary network, finding that integrating visual and auditory data improved the accuracy of recognizing emotion nuances compared to unimodal systems. Recently, the robustness of multimodal models under different conditions has also been studied. C. Kong *et al.*, (2024) emphasized that multimodal deep learning models maintain high performance despite environmental changes, such as lighting and background noise, which is less noticeable in unimodal models. Also Juyal (2022) supported these findings by noting that multimodal sentiment analysis models effectively manage and integrate different data sources, increasing their reliability and applicability in real-world conditions. Furthermore Yazhou Zhang *et al.*, (2021) discussed the application of multimodal models in detecting negative mood states using a robustness-centered fusion model that integrated data from multiple sensors.

In turn, Y. Yi *et al.*, (2023) introduced a dual-branch transformer model that significantly improved the capture of complex emotional cues from fused audiovisual data. Then Krishna (2021) further advanced this field by using large, pre-trained models with cross-modal attention mechanisms, which increased the model's ability to interpret and correlate information from different modalities.

In multimodality research, the challenge is the lack of large, appropriately labelled data sets available, so researchers use advanced data augmentation techniques. Ma *et al.*, (2022) used GANs to augment audiovisual data, improving the training process and, thus, the performance of MMER. The aforementioned CMER addresses the limitations of traditional unimodal systems that often fail to capture the complexity of human emotions (Khan *et al.*, 2024). However, implementing CMER systems in real-world scenarios requires super-efficient real-time data processing. Advances in edge computing and the development of lightweight models are key to meeting these requirements. Furthermore, emotional expressions vary across cultures and contexts, which affects the accuracy of emotion recognition systems (Lian *et al.*, 2023).

Multimodality through feature mining and person calibration is joining the trend of cross-cultural research and the search for models that consider context and person as a source of variability. Maintaining emotional data confidentiality is paramount, as is implementing data protection measures and obtaining informed user consent (Mamieva *et al.*, 2023). Addressing bias in emotion recognition algorithms is crucial to prevent discrimination and ensure equal treatment across populations. Ongoing research like Lian *et al.*, (2023) on new approaches, such as cross-modal attention mechanisms and graph neural networks, is expected to improve the integration of different modalities, leading to more advanced emotion recognition systems. Developing standard benchmarks, data sets, and assessment metrics, as well as fostering collaboration between computer scientists, psychologists, and ethcists, is crucial for the holistic development of MMER (Kalateh *et al.*, 2024).

5 Contactless vs. Contact Techniques of MMER

A recent review on CMER by Khan *et al.*, (2024) offers an in-depth look at different visual, auditory, textual, and physiological data modalities. Identified gaps for these systems are also effective integration methods, including cross-modal attention mechanisms, graph neural networks for multimodal data, and hierarchical fusion strategies, which could significantly improve the performance of CMER. Additionally, in practice, optimization of latency and computational efficiency is very important.

The comparison of non-contact and contact MMER techniques in Table II highlights the distinct advantages and challenges of each approach. CMER offers user comfort through non-invasiveness, which is why it is more widely accepted. These techniques are usually tailored for real-time operations, although their accuracy is variable and is a direct derivative of the application of newer deep learning techniques. Despite being non-invasive, their development raises ethical and privacy concerns, as well as data protection for both the data used for inference and the inference results. CMER is typically used in healthcare, education, social robotics, and customer service.

In contrast, non-contact techniques that use physiological sensors such as EEG and ECG, body-worn cameras, and tactile sensors have higher accuracy, especially for physiological data. However, these methods are often invasive and, therefore, uncomfortable, which results in moderate or low user acceptance. The authors note that contact techniques are less suitable for solutions requiring rapid decision-making in RT.

Contact techniques are typically used in controlled laboratory environments, healthcare, and some applications with wearable devices. The review by Khan *et al.*, (2024) does not discuss

Aspect	Contactless Techniques	Contact Techniques
Data Collection Methods	Visual (RGB cameras), auditory, text, infrared, radar	Physiological (EEG, ECG), visual (facial expression cameras), tactile
Sensor Types	RGB cameras, infrared cameras, radars, Wi-Fi, audio recorders	EEG sensors, ECG sensors, wearable cameras, tactile sensors
Accuracy	Varies, generally improving with deep learning advancements	High, especially for physiological data
User Acceptance	High, due to non-intrusive nature	Moderate to low, due to intrusiveness and discomfort
Real-time Capability	Good, dependent on processing power and optimization	Varies, can be slower due to the need for contact-based data collection
Cost	Moderate to high, depending on sensors used	High, due to specialized and often expensive sensors
Comfort and Intrusiveness	High comfort, non-intrusive	Moderate comfort, can be intrusive
Ethical and Privacy Concerns	Significant concerns about privacy, need robust data protection measures	Concerns about consent and user comfort, but generally better understood
Applications	Healthcare, education, social robotics, customer service	Healthcare, controlled lab environments, some wearable applications
Challenges	Environmental factors (lighting, occlusion), data privacy, real-time processing	User discomfort, high cost, invasive nature, data collection complexity

Table II: Comparison of Contactless and Contact Techniques of MMER, prepared based on Khan *et al.*, (2024)

how cultural differences and contextual variability affect CMER techniques—arguably, the topic is at such an early stage of development that research is yet unavailable. SO, there A thorough examination of cross-cultural studies and developing context-aware models could provide a more nuanced understanding of these effects. There is a lack of detailed case studies or practical implementations in various domains, which can enrich insight about the effectiveness and challenges of CMER systems. Another research gap is lack of longitudinal studies and analysis of emotional dynamics over time. Finally, the last but not least topic is the user experience and acceptance of CMER systems, investigating how users interact with these systems, their comfort level, and psychological effects of more emotional responsivness of virtual assistant. Technological parameters such as sensor accuracy, data processing limitations, and reliability under different conditions significantly impact the effectiveness of contact techniques. The paper also concludes that the studies need standardization of CMER studies, including benchmarks for standards and classifications, data sets, and evaluation metrics. The comparisons in Table III reveal varying levels of accuracy for different methods. Non-contact techniques, such as those using RGB cameras for facial recognition, show a range of accuracy from 75% to 90%. Infrared cameras used for thermal imaging typically achieve 70% to 85% accuracy rates. Radar systems that detect motion and heart rate show accuracy levels from 65% to 80%, while audio recorders analyzing speech present a broader accuracy range of 60% to 85%. Text analysis can achieve high accuracy rates between 70% and 90% using NLP techniques. There is a trend toward higher accuracy rates for contact techniques (especially those measuring physiological data). For example, EEG sensors, which measure brain activity, have accuracy ranges of 80% to 95%, while ECG sensors, which monitor heart rate, show a accuracy of 75% to 90%. Body-worn cameras used to analyze facial expressions have accuracy rates of 70% to 85%, as do tactile sensors measuring skin conductance. So, it is always a trade-off between user comfort and measurement precision and reliability.

The wide range of accuracy for different MMER techniques is attributed to various factors related to the performance of the measurement systems and algorithms. For example, face recognition accuracy using RGB cameras can be highly dependent on lighting conditions. At the same time, audio-based techniques can be sensitive to background noise and audio signal quality. Contact-based techniques such as EEG and ECG provide greater accuracy because they measure physiological responses directly related to emotional states.

These signals are also less susceptible to conscious control and hiding of emotions, offering a

Table III: Summary of Accuracy range for Contactless and Contact Techniques of MMER, prepared based on Khan *et al.*, (2024)

Technique Type	**Specific Techniques**	**Accuracy Range**
Contactless	RGB Cameras (facial recognition)	75% - 90%
Contactless	Infrared Cameras (thermal imaging)	70% - 85%
Contactless	Radar (motion and heart rate detection)	65% - 80%
Contactless	Audio Recorders (speech analysis)	60% - 85%
Contactless	Text Analysis (NLP)	70% - 90%
Contact	EEG Sensors (brain activity)	80% - 95%
Contact	ECG Sensors (heart rate)	75% - 90%
Contact	Wearable Cameras (facial expression)	70% - 85%
Contact	Tactile Sensors (skin conductance)	70% - 85%

more reliable indication of proper emotional responses. Although the authors do not mention this, the differences may also be due to individual differences in reactions between individuals, and even initial calibration may not be sufficient to account for these differences. Deep learning techniques such as neural networks typically provide higher accuracy because they automatically extract and learn complex patterns from data. Emotion recognition techniques must be carefully selected for specific use cases. When comparing run and multimodal models and contact and non-contact techniques, (A. Savchenko and LV Savchenko, 2022) conducted a comparative analysis, finding that multimodal models integrating audiovisual data outperformed unimodal models regarding emotion classification accuracy.

In turn, R. Chen *et al.*, (2022) introduced a cross-modal video auxiliary network to analyze visual and audio data, achieving an average classification accuracy of 92%. In their research, W. Kong (2024) showed that multimodal deep learning models maintain high performance under variable environmental conditions, suggesting more excellent reliability of multimodal approaches in practical applications. L. Yi and Mak (2019) used generative adversarial networks (GANs) to augment data, significantly improving multimodal emotion recognition systems' training process and performance.

The average effectiveness of contact-based techniques calculated from the analyzed studies is about 91.26%, and the average effectiveness of non-contact techniques is about 87.08%. Additionally, integrating multiple modalities, using advanced AI techniques such as SVM, CNN, and LSTM, using data augmentation with GANs, and providing real-time data processing and incorporating context awareness can further improve the performance and reliability of MMER.

6 Comparison of Fusion Methods

Choosing an appropriate fusion method for MMER is crucial for achieving efficient results. Various fusion techniques have been investigated and developed to integrate data from multiple modalities.

A systematic review by Gandhi *et al.*, (2023) provides a detailed review of MMER, highlighting its history, datasets, fusion methods, applications, and future directions. It discusses various fusion methods, including early, late, hybrid, model-level, tensor, hierarchical, bi-modal, attention-based, quantum-based, and word-level fusion. The review emphasizes the benefits of multi-source solutions to improve sentiment analysis accuracy. It also identifies key challenges in this research area, like the complexity of natural language, which requires more advanced machine-learning techniques. Focus on developing more sophisticated fusion methods and addressing ethical considerations will increase in time.

Tensor Fusion Network (TFN) for Multimodal Sentiment Analysis study by Zadeh *et al.*, (2017) introduces the TFN, which demonstrates state-of-the-art performance on the CMU-MOSI dataset by capturing complex interactions among different modalities. The authors demonstrated a more effective integration of intermodal dynamics compared to early fusion models. The presented examples nicely illustrated how TFN improves the accuracy of emotion state prediction. TFN effectively modeled higher-order interactions between modalities, which provides a good basis for recommending this solution.

A survey Multimodal Fusion of Visual Dialog by X. Chen *et al.*, (2020) covers advancements in Visual Dialog tasks, focusing on datasets, evaluation metrics, and the challenges of multimodal

fusion. It highlights progress in areas such as visual co-reference resolution and attention mechanisms and discusses the potential of graph neural networks (GNNs) for integrating multimodal features. The survey concludes that while significant progress has been made, ongoing challenges still require innovative solutions to improve the performance of visual dialogue systems.

Paper about Affective Computing as an introduction to Emotionally Intelligent Metaverse by Pervez *et al.*, (2024) explores the integration of affective computing in the metaverse, discussing potential applications in healthcare, education, gaming, and customer service. The key conclusion is that incorporating emotional intelligence into virtual environments can significantly enhance user interactions and experiences but must be done responsibly to address ethical issues (Pervez *et al.*, 2024; Akbar *et al.*, 2019).

Because based on the research review each fusion method offers distinct advantages and is dedicated to specific application contexts, the below is a list of found kinds of fusion with short description.

Key Fusion Methods

- **Early Fusion:** Combines features from different modalities at an initial stage before classification, simplifying implementation and speeding up processing but potentially losing modality-specific information.
- **Late Fusion:** Integrates the results from unimodal classifiers at a decision level, achieving high accuracy through specialized classifiers but at a higher computational cost.
- **Hybrid Fusion:** Utilizes both early and late fusion techniques to balance performance and flexibility, though it requires more computational resources.
- **Model-Level Fusion:** Integrates different models specialized in different modalities, offering improved accuracy through leveraging model strengths but is complex to implement.
- **Tensor Fusion:** Captures unimodal, bimodal, and trimodal interactions using tensor representations, providing high accuracy but demanding significant computational resources.
- **Hierarchical Fusion:** Structures the fusion process hierarchically, handling complex data effectively but may be slower to execute.
- **Attention-Based Fusion:** Employs attention mechanisms to focus on important parts of the input data, improving accuracy and noise reduction but requiring careful design and higher complexity.
- **Quantum-Based Fusion:** Applies principles from quantum computing to enhance the fusion process, promising higher computational efficiency but remains largely theoretical.
- **Word-Level Fusion:** Utilizes contextual information from neighboring utterances to improve emotional states recognition, offering excellent contextual understanding but requiring large datasets.

Summary of differences between various fusions metrics is presented at Table IV.

6.1 Comparative Analysis of Fusion Methods

In the MMER area, there are two popular techniques for achieving multimodal integration: **Tensor Fusion** and **Attention-Based Fusion**. Each of them offers distinct advantages and is dedicated to different problems, which makes the comparative analysis of recommended conditions can save disappointments in MMER implementations. Table V presents a comparison of these two methods, highlighting their key features, advantages and disadvantages, recent achievements and dedicated use cases. Tensor Fusion method consists in creating a multidimensional array (tensor) that represents interactions between different modalities, combining correlations between modalities into a structured tensor format. One of the greatest advantages of Tensor Fusion is its ability to encapsulate higher-order interactions, often leading to a more nuanced understanding of analyzed emotional states from complex data from different modalities. Another problematic (high computational power) is the use of this technique in the situation where real-time data is required. The growth of the tensor size with the addition of more modalities is exponential, which makes the scalability of such solutions difficult. In turn, Attention-Based Fusion methods use attention mechanisms to dynamically calculate weights for each modality based on its distinct importance for a specific task. These methods are flexible, so they can be applied to a wide range of contexts.

Method	Description	Pros	Cons	Dedicated Purpose	Performance
Early Fusion	Combines features from different modalities at an initial stage before classification.	Simpler implementation; reduces feature space; fast processing.	May lose modality-specific information; potential for data redundancy.	Simple applications where fast processing is needed.	Good for simple tasks, less effective for complex interactions.
Late Fusion	Integrates results from unimodal classifiers at a decision level.	Utilizes specialized classifiers for each modality; high accuracy.	Higher computational cost; complex integration process.	Applications requiring high accuracy and modality-specific insights.	High accuracy for specific tasks but computationally expensive.
Hybrid Fusion	Uses both early and late fusion techniques for improved performance.	Combines advantages of both early and late fusion; more flexible.	Increased complexity; requires more computational resources.	Flexible applications benefiting from both early and late fusion.	Balances performance and flexibility; suitable for diverse tasks.
Model-Level Fusion	Integrates different models specialized in various modalities.	Leverages strengths of different models; improves accuracy.	Complex implementation; requires large datasets.	Applications requiring strengths of multiple models.	Very effective if models are well-chosen; complex to implement.
Tensor Fusion	Captures unimodal, bimodal, and trimodal interactions using tensor representations.	Captures complex interactions; high accuracy.	High computational cost; requires large datasets.	High-accuracy applications needing detailed modality interactions.	Achieves high accuracy but requires significant resources.
Hierarchical Fusion	Structures the fusion process hierarchically, often using neural networks.	Organized structure can handle complex data; scalable.	Complex implementation; may be slower.	Scalable systems with complex data requirements.	Effective for large-scale, complex data but slower.
Attention-Based Fusion	Employs attention mechanisms to focus on important parts of the input data.	Focuses on relevant features; reduces noise; improves accuracy.	Requires careful design of attention mechanisms; higher complexity.	Applications needing high accuracy and noise reduction.	Achieves high accuracy by reducing noise and focusing on key features.
Quantum-Based Fusion	Applies principles from quantum computing to enhance fusion processes.	Potential for higher computational efficiency; novel approach.	Theoretical; less tested in practical applications.	Novel applications exploring advanced computational methods.	Potentially high efficiency but limited practical evidence.
Word-Level Fusion	Utilizes contextual information from neighboring utterances for improved sentiment analysis.	Improves contextual understanding; handles complex interactions.	Complex implementation; may require large datasets.	Applications needing deep contextual understanding.	Achieves excellent contextual understanding and accuracy.

Table IV: Performance of Fusion Methods Comparison & Summary

The dynamic nature of attention mechanisms makes it more adaptive and less demanding computationally compared to tensor operations, but also makes them more dependent on the quality and diversity of the training data. Moreover, although attention mechanisms improve the efficiency of MMER systems, they do not meet the requirements of interpretability, which makes it difficult to understand the decision-making process within the model. Table **??** summarizes the advantages and limitations of both methods in MMER. Tensor Fusion better captures higher-order interactions between modalities, but struggles with computational and scalability challenges. Attention-Based Fusion, on the other hand, provides a more adaptable and computationally efficient approach, although its effectiveness is dependent on the quality of the training data and is difficult to interpret. Perhaps future solutions dedicated to MMER will be built as hybrids combining the strengths of both methods while mitigating their weaknesses.

In MMER various fusion techniques—such as early fusion, late fusion, and hybrid fusion—are employed to integrate data from multiple modalities like text, audio, and visual inputs. These fusion methods are not mutually exclusive; they can coexist within a single system to leverage the strengths of each approach under specific conditions.

Early Fusion handles cross-modal correlations well, but the complexity of integrating heterogeneous data can potentially introduce errors.

Late Fusion allows for processing correlations for multiple modalities, but by matching across modalities, and can bypass complex cross-modal relationships through aggregation and generalization.

Hybridity involves, for example, performing early fusion on two modalities to capture their direct interactions, e.g. video and voice, and then combining these results with late fusion, with the output of the third modality, text. This strategy allows the system to effectively manage heterogeneous data that is analyzed in different time frames while preserving complex intermodal dynamics. Recent research supports the coexistence of different fusion types within MMER systems (Yong Zhang *et al.*, (2021), Cambria *et al.*, (2024), Tang *et al.*, (2024)). A studies by S. Zhang *et al.*, (2024) proposed a three-stage multimodal emotion recognition network that sequentially applies unimodal feature extraction, bimodal feature interaction (akin to early fusion), and multimodal fusion (similar to late fusion), demonstrating the integration of multiple fusion strategies within a single framework.

Table V: Comparison of Tensor Fusion and Attention-Based Fusion Methods.

Compared Aspect	**Tensor Fusion**	**Attention-Based Fusion**	**Key Differences**
Description	Captures high-order interactions using tensor representations.	Uses attention mechanisms to focus on important features.	Tensor Fusion emphasizes detailed, structured integration; Attention-Based focuses dynamically on relevant inputs.
Key Features	High-order interactions, rich representations, comprehensive integration.	Dynamic weights, noise reduction, flexibility.	Tensor Fusion prioritizes depth, while Attention-Based is agile in focusing on specific signals.
Pros	Detailed modeling, high accuracy, effective integration.	Improved accuracy, scalability, noise reduction.	Tensor excels in precision; Attention excels in adaptability.
Cons	High computational cost, requires large datasets.	Design complexity, computational overhead.	Tensor is resource-intensive; Attention demands careful architecture.
Recent Advancements	Low-rank tensor fusion, self-supervised learning.	Transformer-based models, sentiment knowledge-enhanced fusion.	Tensor sees advancements in data efficiency; Attention in model sophistication.
Best For	Applications needing detailed interaction modeling and high accuracy.	Real-time sentiment analysis, applications needing dynamic and focused feature integration.	Tensor for complex scenarios; Attention for real-time responsiveness.

Another example of effective combination of early and late fusion is the Joyful model described by D. Li *et al.*, (2023), which employs a joint modality fusion mechanism alongside with the graph contrastive learning for better performance.

6.2 Feature Alignment in MMER

Feature alignment (FA) in the context of MMER refers to the process of synchronizing and combining features from different modalities (e.g. text, audio, image) in such a way as to preserve the coherence of information and enable efficient data fusion for emotion recognition. The FA step precedes fusion, and its primary goals are to synchronize modalities, i.e. to align temporally and structurally data from different sources, to unify the feature space, i.e. to project features into a representation with a standard scale, dimensions, or semantics, and to reduce possible noise and redundancy in the data, i.e. to remove redundant or unnecessary data, which prevents model overload.

Without this step in obtaining user state scores, differences between modalities, such as differences in temporal resolution, semantic interpretation, and noise levels, can lead to errors and artefacts and undermine the reliability of MMER systems. In the analytical process, the FA provides a structured way to overcome time series inequalities, making it the cornerstone of all emotion analysis.

To appreciate the importance of FA, it is helpful to consider how different modalities feed into emotion recognition. For example, audio signals carry emotional cues via features such as rhythm, intonation and pitch. Video data captures facial expressions, body movement, and head posture. If an utterance accompanies this visual behaviour, semantic analysis is added, adding the explicit content and contextual subtleties. However, these modalities operate in different ways. Audio data exists as a continuous-time signal, video is typically broken down into discrete frames, and text is often processed as sequences of tokens. Analysts face data fragmentation and lack of synchronization without a proper mechanism for aligning these. FA reduces the potential for errors and incomplete or conflicting emotional interpretations. FA contravene this problem by ensuring all modalities are temporally synchronized and semantically aligned. For example, if a speaker's voice rises simultaneously as their facial expression shows a smile, FA ensures the two signals are

properly aligned. A unified framework for data analysis enables researchers and MMER systems to extract more relevant and holistic emotional insights that might be misinterpreted if the modalities were analyzed independently.

Various methods are used for FA. Dynamic Time Warping (DTW) is basic techniques useful for aligning temporal sequences. DTW adjusting the duration of one sequence to another and ensures that asynchronous data, such as audio and video, can be efficiently synchronized (Y. Chen *et al.*, 2024).

Another approach is to use cross-modal attention mechanisms, which allow for dynamic prioritization and alignment of the features based on their relevance to the analyzed emotional context (Shou *et al.*, 2024a). For example, suppose the system detects a strong emotional cue in the tone of a speaker's voice. In that case, it can adjust its attention to align corresponding visual features, such as changes in facial expression.

Shared latent space mapping is another way to implement FA (Suguitan *et al.*, 2024). This method involves projecting features from different modalities into a common representational space, where design enforces alignment. An example of such an approach is Semantic Alignment Networks (SANs). SANs use higher-order emotional representations to align features across modalities (X. Zhang *et al.*, 2024).

Adversarial learning frameworks also minimize noise, ensuring that only the most relevant features are aligned (Shou *et al.*, 2024b). These frameworks use techniques such as GANs to refine the feature alignment process. FAs directly improve the accuracy and robustness of emotion recognition systems. Studies such as Yusong Wang *et al.*, (2024) show how FAs can increase system performance by creating more coherent and interpretable multimodal representations.

Furthermore, FAs alleviate the problem of modality dominance, where one type of data—such as video—could otherwise overshadow others, such as text or audio. By artificially balancing the contributions of each modality, FA promotes fairer integration that allows systems to better generalize across contexts and emotional expressions (X. Zhang *et al.*, 2024). Recent research has focused on optimizing computational efficiency. Methods such as masked graph learning with recursive matching have shown promising potential in reducing the complexity of the matching process without sacrificing accuracy (Meng *et al.*, 2024).

But there are some situation when data alignment may not be necessary or even desirable. In some contexts (e.g., recognizing emotions in video recordings), the visual modality is more informative than the others (e.g., text or sound). This is because facial expressions or gestures often directly reflect emotions, while signals in text may be more ambiguous. If the data modalities naturally have different levels of importance for emotion recognition, it is worth considering models that can operate on non-synchronized features (e.g., using asynchronous attention mechanisms).

It is, therefore, necessary to design appropriate tests and experiments that will be able to show to what extent FA benefits MMER – and to what extent it is a stage generating spurious artefacts. In some situations, balancing the modalities, i.e. ignoring the natural differences in their informativeness about emotions, may cause bias at the level of the analysis design itself. This is the case for emotions subtly expressed through tone of voice (e.g. sarcasm), artificially "emphasizing" the visual modality could reduce the recognition efficiency of emotions. Modalities of lower quality (e.g. text in the case of an ironic speech recording) can introduce noise. If alignment artificially "amplifies" them in the fusion process, the risk that incorrect information from this modality will affect the final results increases. In such situations, hierarchical models are helpful, allowing for adaptive decisions about which modality dominates in a given context. Multimodal transformers with cross-modal attention mechanisms are also helpful because they "honestly" assess which modality is mostly informative at a given moment.

In summary, feature matching is not just a technical necessity but a fundamental factor in the design and development of MMER. It bridges the gaps between different modalities, enabling synergy effects. Multiple studies emphasize the importance of FA, treating this element as a central pillar of innovation in multimodal analytic (Srivastava *et al.*, 2024; S. Li *et al.*, 2024).

6.3 Summary of Fusion Methods

The analyzed literature allows for the formulation of several important generalizations. The first is undoubtedly that this area is in a phase of muscular development and exploration, where tensor fusion and attention-based mechanisms in capturing complex interactions in modalities are currently the leading techniques. Since the results of these activities remain in the area of challenges due to the quality of the results, further development and search for better solutions are undoubtedly necessary. The need to use or at least explore the potential of MMER for emotion recognition is multi-application, which increases its attractiveness in the face of potential benefits.

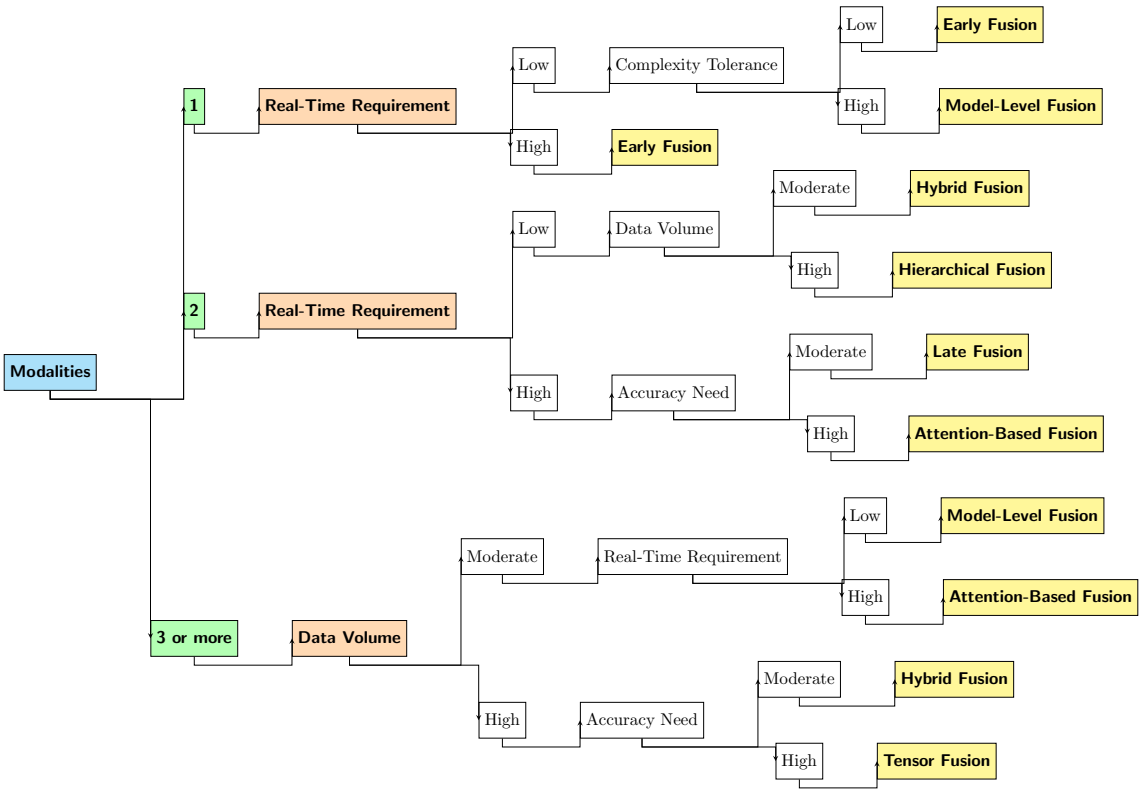
The main challenges are data quality, computational efficiency and incorporating social and cultural contexts to exploit these technologies' potential fully. Fusion techniques allow for achieving higher efficiency through effective modelling of higher-order interactions, and the analysis's complexity is joined with the goal of data usability and solving problems related to quality (missing data) and different timing for different modalities. Another key challenge is to balance the analytical complexity and computational requirements in advanced fusion methods since they are rarely performed after the fact and are mostly performed in real time.

The race is to develop more efficient algorithms that provide high-quality analysis results without compromising processing speed. The third challenge is the growing recognition of the importance of social and cultural contexts in emotion research. Research emphasizing the key role of emotions in interactions, their role in building the individuality of interactions and their impact on interaction evaluation will be a strong motivator for developing multimodal data integration techniques. This indicates an important trend towards more context-aware and socially intelligent AI systems.

These key elements can also be summarized in a simplified diagram 6.3 in the form of a tree diagram classifying the types of Fusion. It can help select the most appropriate fusion method for MMER based on key criteria such as the number of modalities involved, the volume of data, and the needs of a given use case regarding prediction accuracy, real-time processing, complexity, and availability of computational resources. This domain development does not give grounds to assume that the decision tree created for this review will be accurate in the long term, so approach it with caution and common sense, remembering that everything depends on the use case created.

Today, Tensor Fusion is recommended for high-accuracy scenarios with multiple modalities and large volumes of data. Attention-Based Fusion is ideal for applications requiring real-time analysis with a dynamic focus on features. Hybrid Fusion balances performance and flexibility, while Early Fusion and Late Fusion address simpler and specific needs for high accuracy. Hierarchical and Model-Level Fusion are suitable for complex, large-scale applications and do not exclude using the previously mentioned fusions. On the other hand, Quantum-Based Fusion offers potential performance gains for increasingly complex MMER systems.

Graph 5. Decision Tree for Choosing the Best Fusion Method



Attention-Based Fusion is generally more scalable than Tensor Fusion, especially when optimized for efficiency. It is better suited for applications where real-time processing is crucial, such as sentiment analysis in social media streams and interactive systems like virtual assistants. Both methods involve complex implementation processes. Tensor Fusion requires sophisticated tensor operations, while Attention-Based Fusion demands intricate attention mechanisms. The choice be-

tween these methods may depend on the application requirements and the available computational resources.

7 Ethical Considerations in Multimodal Emotion Recognition

Because of sensitive personal data use, which are protected MMER requires special attention and efforts concerning responsibility of AI and AI ethics. The collection and processing of this data raise significant privacy issues, mainly if the data is used without users' explicit consent or if it is stored insecurely.

The potential misuse of sensitive data, especially in facial recognition or physiological monitoring cases, could lead to unauthorized surveillance or unintended disclosure of private information (C.-K. Zhang *et al.*, 2017). In addition, the information about the user that is created as a result of the MMER algorithms should also be protected as private data. In addition, there is a potential risk of using information about emotional states to manipulate the user.

Ensuring informed consent is critical when deploying AI systems that collect and analyze multimodal data. Regulation and simple ethical rules aimed at autonomy and privacy protection postulate making sure that the user is informed what data is being collected, what the purpose is, how it will be used, and who will use it.

Transparency is vital for maintaining user trust and ensuring that the whole process of MMER practices complies with legal and ethical standards. Implementing a clear and straightforward consent mechanism can be an element of formal information requirements and help in aware decision (Binns *et al.*, 2018).

Bias in AI models is another critical ethical issue, particularly in systems that utilize multimodal data. Different modalities can introduce more varying biases, resulting in unwanted inequality effects. For example, facial recognition algorithms perform differently across demographics due to cultural differences and can lead to even unrecognized discrimination. To address these biases, it is essential to use diverse and representative datasets and to implement regular audits to detect and mitigate bias (Buolamwini and Gebru, 2018; Hutchinson and M. Mitchell, 2019; S. Mitchell *et al.*, 2021). To mitigate these ethical concerns, several general strategies are recommended to be implemented, like data minimization practices (collecting only the data necessary), anonymizing data, or the "privacy by design" approach, executed by local data processing or federated learning methods. Also, differential privacy can help safeguard individual data while allowing for aggregate analysis (Dwork, Roth, *et al.*, 2014). Bias mitigation strategies, such as using fair algorithms and conducting regular model evaluations, are also crucial for ensuring equitable outcomes in AI systems (Mehrabi *et al.*, 2021).

In summary, although using MMER in AI-human interactions presents several ethical challenges, there are some general strategies such as data minimization, transparent consent mechanisms, bias mitigation, and adherence to regulatory and ethical guidelines, which can minimize the potential risk.

- **Data Minimization and Anonymization:** One effective strategy to protect privacy is data minimization, which involves collecting only the data necessary for a specific purpose and anonymizing it to prevent the identification of individuals. Techniques like differential privacy can add noise to data to protect privacy while allowing for aggregate analysis (Dwork, Roth, *et al.*, 2014). These approaches help to reduce the risk of data breaches and unauthorized access.

- **Transparent Consent Mechanisms:** Transparent and straightforward consent mechanisms ensure users are well informed about the data collection process. The whole process should be a process that does not favour mechanical consent but is balanced with naturalness and non-intrusiveness. Users should be aware of what data is collected, what purpose will be used, and what potential risks are involved. Providing users with control over their data, such as options to opt-out or delete their data, can further enhance trust and compliance with ethical standards (Nissenbaum, 2011).

- **Bias Mitigation Techniques:** To address this bias, AI developers should use diverse and representative datasets that reflect the population's demographics using the technology. There are also algorithmic fairness techniques, such as re-weighting training data or

incorporating fairness constraints into model training, helpful in reducing bias. Regular audits and testing for bias should be an integral part of the development process to ensure ongoing fairness and accountability (Mehrabi *et al.*, 2021).

- **Regulatory Compliance and Ethical Guidelines:** The General Data Protection Regulations GDPR (2016) and the Artificial Intelligence Act (AI act) [1] (Parliament, 2021) are available and in force, which outline the directions for creating new solutions. Additionally, multiple ethical standards have been adopted to case frameworks promoted by international organizations such as OECD, IEEE, Partnership on AI, or GPAI. These institutions also offer expert assistance in developing new standards for new cases and technologies. These regulations emphasize user privacy, consent, and the right to explanation, which are crucial for maintaining ethical standards.

Finally, it is worth mentioning that only some countries respect such high ethical standards and approach responsibility with the EU at the forefront. This disproportion between regulations is disadvantageous both for creators (higher entry threshold for creating technology in a responsible manner), but also for users. In most of the world, the inalienable rights to privacy and autonomy are not sufficiently respected.

8 Research Questions, Research Gaps and Unanswered Questions

The following summary of conclusions organized according Research Questions were prepared after review joined with domain and practical knowledge. Unanswered questions and research Gaps can help other researcher to find the need for new knowledge.

RQ1. How can multimodal data be effectively integrated to improve user state recognition accuracy in AI-human interaction systems? Currently, data integration methods such as Tensor Fusion and Attention Based -Fusion are being researched, developed and considered - their diversity stems from the need to match the strengths of different types of data, such as facial expressions, tone of voice and text, to create a holistic picture of the user's emotional and cognitive states.

Based on the review presented in Durante *et al.*, (2024), it can be concluded that integrating multimodal data effectively improves user state recognition in AI-human interaction systems. Currently, ready-made complex models such as large language models (LLM) and vision-language models (VLM) are often used, which have multimodal data fusion integrated. The option of interactive learning allows AI systems to adapt through real-time feedback. Finally, enabling contextual memory increases the adaptability and continuity of user interactions. In the current review, a data fusion method selection scheme is proposed to improve the performance of MMER for different conditions and use cases.

RQ2.What are the impacts of individual differences (such as age, gender, personality (BIG5), cultural background or others) on the effectiveness of multi-modal user state recognition systems? This question investigates whether and how individual differences affect the performance of multi-modal systems in recognizing user states, and how systems can be adapted to accommodate these variations.

Recent studies from 2023 and 2024 have provided valuable insights into the impacts of individual differences on the effectiveness of multimodal user state recognition systems. These differences include factors such as age, gender, personality traits (e.g. the Big Five personality model), and cultural background. Understanding and accommodating these variations is crucial for developing robust and inclusive AI systems. Here, I summarize the findings from the latest research on this topic.

The impact of individual differences, such as age, gender, personality traits (like those described by the Big Five model), and cultural background, on the effectiveness of multi-modal user state recognition systems is significant. These factors can influence how different modalities (e.g., facial expressions, speech, text) are interpreted and how accurately a system can recognize the user's state.

Younger and older individuals may express emotions differently. For instance, facial expressions and vocal intonations might vary with age, affecting the accuracy of emotion recognition systems. Research indicates that age-related changes in muscle tone and speech patterns need to

[1] UE AI ACT: https://artificialintelligenceact.eu/

be considered to enhance system performance across age groups (Lian *et al.*, 2023; Xiaoming *et al.*, 2022).

Men and women often display and interpret emotional cues differently. For example, studies have shown that women might use more expressive facial and vocal cues compared to men. Recognizing these differences can help tailor the algorithms to better understand gender-specific expressions (Simić *et al.*, 2024).

The Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, neuroticism) can significantly impact how emotions are expressed and perceived. For instance, extraverts might exhibit more pronounced facial expressions and gestures, while individuals high in neuroticism may show more varied and intense emotional responses. Adapting systems to these traits can improve recognition accuracy (R. Wang *et al.*, 2024; Chavez and Heatherton, 2015).

Based on primary psychological emotion theories it's obvious that cultural context (and differences) has impact on emotions expression and interpretation. For example, certain facial expressions and gestures might be common in one culture but not in another. Thus, emotion recognition systems need to be trained on diverse datasets to cover a wide range of cultural expressions to avoid biases and inaccuracies.

To accommodate these variations, MMER systems can be personalized based on the user's demographic and psychographic information, what can help in tailoring the recognition process. For example, incorporating user-specific calibration sessions can refine the system's accuracy.

This approach allows the system to learn from data distributed across multiple users without compromising privacy. It helps in creating models that are generalizable yet adaptable to individual differences by continuously learning from diverse user inputs (Simić *et al.*, 2024). Using advanced techniques like cross-modal transformers can help in effectively integrating and interpreting data from various modalities, ensuring that the system captures nuanced differences across individuals (R. Wang *et al.*, 2024).

RQ3. Which modalities are most predictive of specific user states and how do these modalities interact in AI-human interaction? This study focuses on understanding the contribution of each modality to the overall prediction of user states and explores interactions between modalities that may enhance or hinder the recognition process. A series of studies in 2023 and 2024 have provided new knowledge on which modalities (facial expressions, tone of voice, and text) are most predictive of specific user states and how these modalities interact in AI-human interaction systems. The review suggests that the strength of a modality can be individual and situation-dependent, but that having multiple modalities certainly yields more accurate results. Here are the main findings:

- Facial expressions: Facial expressions are highly predictive of emotional states and provide immediate, visible cues. CNNs and RNNs are particularly effective in analyzing facial expressions. For example, research by Chuangao Tang and Yuan Zong highlights the importance of deep learning approaches in multimodal emotion recognition, particularly emphasizing the role of facial expressions in understanding emotional states (Lian *et al.*, 2023).

- Voice: Tone of voice is crucial for emotion recognition through changes in pitch, volume, and tempo of speech, but can be misleading . Combining audio features with facial expressions can significantly improve the accuracy of emotion recognition systems. A study by Simić *et al.*, (2024) discusses the integration of audio and visual data, demonstrating improved performance with a federated learning approach that preserves privacy while increasing detection accuracy (Simić *et al.*, 2024).

- Text: Text data provides context and additional emotional cues, enriching the understanding of user states. Natural language processing (NLP) techniques, especially those using large language models (LLMs), are crucial for decoding sentiment and emotional tone from text. Wei Xu and Zaifeng Gao's work on human-centric AI emphasizes the integration of text data with other modalities to create more comprehensive user state recognition systems (Xu and Z. Gao, 2023).

- Interaction of modalities: Multimodal systems that integrate facial expressions, tone of voice, and text are more effective at predicting user states. The interaction between these modalities enables comprehensive analysis in which each modality compensates for the limitations of the others. This synergistic approach is emphasized in various studies, including a Google Research study that examines the integration of predictive user interfaces and multimodal emotion recognition systems (*Web article:Human-Computer Interaction and Visualization* n.d.).

- Cross-modal attention mechanisms: Advanced techniques such as cross-modal transformers facilitate the interaction between different modalities. These models use attention mechanisms to dynamically weight the importance of each modality's input, improving prediction by focusing on the most salient features of each modality. This approach has been detailed in recent research on multimodal transformers for human state recognition (R. Wang *et al.*, 2024).

- Personalization and adaptation: Individual differences such as age, gender, and cultural background affect how modalities are expressed and perceived. Personalizing models to account for these differences and using federated learning approaches can create adaptive systems that learn from different user interactions while preserving privacy. Microsoft Research Guidelines for Human-AI Interaction emphasize the importance of adapting AI systems to account for these individual differences (Amershi *et al.*, 2019).

RQ4. How does real-time multimodal data processing affect the responsiveness and adaptivity of AI systems in dynamic interaction environments? Real-time multimodal data processing is currently a trend in building responsive and adaptive AI systems. A well-functioning MMER system adds a layer of emotional intelligence and makes dynamic adaptation of human-machine interactions possible.

Over the past 2 years, researchers have been trying to adjust the technological and computational requirements necessary to obtain timely and user-adaptive responses. The use of MMER will soon enable smooth adaptation. The challenges are mainly related to the speed of real-time multimodal data processing. Each modality — be it audio, visual or text — requires specific processing techniques, which, when combined, can significantly increase the computational load.

Real-time data processing allows AI systems to quickly interpret and respond to a variety of inputs such as audio, video, and text, providing timely responses in dynamic scenarios such as traffic management and healthcare emergencies, which increases the overall safety, efficiency, and effectiveness of these systems in critical applications (Zepf *et al.*, 2020; Lisetti and Nasoz, 2002). Continuous real-time processing also allows systems to adapt to changing conditions and user needs. Edge computing minimizes latency by processing data locally (Schmidt *et al.*, 2019), while cloud servers handle more complex calculations, providing a balance between speed and computing power (Zhao *et al.*, 2021).

The aforementioned computing requirements are being addressed with new, more efficient data preprocessing techniques. Often, edge devices and cloud solutions are required to handle the volume of data and the complexity associated with multimodal integration (Schmidt *et al.*, 2019; Yan Wang *et al.*, 2022).

The interaction of multiple data modalities (e.g. visual, auditory, textual) provides comprehensive analysis, leveraging the strengths of each modality to increase the overall performance and reliability of the system in real time (Zadeh *et al.*, 2017; Poria *et al.*, 2017).

Continuous learning is also enhanced by real-time processing; it allows AI systems to adaptively learn from incoming data, refine models, and increase accuracy. Feedback mechanisms such as reinforcement learning help systems adapt to new scenarios and user behaviors, providing continuous optimization (Zhao *et al.*, 2021; Lian *et al.*, 2023).

RQ5. What ethical issues arise from the use of MMER in AI-human interactions and how can they be resolved?

The use of MMER in AI-human interactions is associated with several ethical issues that should always be carefully considered, not only to maintain ethical standards, but above all to increase and maintain user trust in new AI-based technologies. The law actively supports these standards with regulations, enforcing adequate attention from both MMER researchers and MMER solution developers. Fundamental and unquestionable ethical issues include the protection of privacy and autonomy, informed consent and information obligation, transparency and lack of bias. All of them pose serious challenges when implementing MMER technologies.

Multimodal data are usually treated as protected data, so any collection, storage and processing of this data can lead to potential privacy violations. These risks can only be counteracted. In addition, facial recognition technologies potentially expose individuals to unauthorized surveillance, and physiological data can reveal health information that users did not intend to share (C.-K. Zhang *et al.*, 2017). Therefore, regulators have placed most of these systems on high-risk lists - especially in the workplace.

Since too often users do not understand what information is being collected, for what purpose it is being used, or who has access to it, obtaining their informed consent to use MMER is potentially problematic. There are of course users who unwisely agree to everything without reading, but there are also those who are overly suspicious. Such different attitudes can result in users unwittingly

agreeing to extensive data collection and use, or too hastily blocking the operation of a system that is safe for them, generalizing a lack of trust. Also ensuring that users are fully informed about the nature and purpose of data collection can be difficult when dealing with less educated users. The consent process must be simple and transparent enough for most people, not just a select few. Users must be aware of their rights and the potential uses of their data (Binns *et al.*, 2018).

The potential for MMER systems to be manipulated, and the potential bias in AI systems, is a serious ethical issue, and the more complex and opaque the system. For example, facial recognition systems have been shown to perform differently based on race and gender, leading to unequal treatment and potential discrimination. This side effect can be exacerbated in multimodal systems, where biases in one modality can be compounded by biases in others (Buolamwini and Gebru, 2018).

To address this, the entire MMER development process needs to be overseen from the very beginning and training data used through to the final solution - to minimize potential biases as much as possible and ensure the highest level of fairness. AI models should also be regularly audited and updated to identify and mitigate any biases that may emerge over time (Hutchinson and M. Mitchell, 2019; S. Mitchell *et al.*, 2021).

RQ6. To what extent can improvements in MMER improve user engagement and satisfaction in AI-based interfaces?

Although the main goal of MMER is to improve the quality of human-machine interaction and increase the naturalness of conversation by making it more human, this is not yet obvious and identical to the satisfaction of recipients. However previous studies confirm that adding an emotional layer increases engagement, it is not certain whether this will change in the future.

Improvements in MMER go towards obtaining increasingly detailed information about the user; the systems are increasingly aware of the context in terms of the individual relationship and are, therefore, increasingly responsive. Integration of multiple sources adds a holistic perspective similar to human intuition. The benefits are obvious but not without potential errors. After all, we train algorithms based on labels assigned by competent judges or different types of voting between available algorithms.

The most reliable systems seem to be those trained on contact data - physiological combined with non-contact data. The presented in the review studies has shown that the improvement in the quality of emotion recognition can lead to significant improvements in user experience and increase satisfaction and engagement.

User satisfaction increases as interactions become more engaging and empathetic, and the bond between users and AI assistants is also strengthened (Chatterjee *et al.*, 2024; S. Ghosh *et al.*, 2023; A. Ghosh *et al.*, 2023). In games, adaptability can create more engaging and engaging experiences, making the experience more personalized and enjoyable (Barthet *et al.*, 2024). Customer service platforms also use MMER, enabling more efficient and satisfactory handling of the customer's case. The systems often have greater potential than a human employee, faster analyzing the content and tone of the customer's query and historical data. This improves the user experience and increases the efficiency of customer service operations, leading to higher satisfaction rates among users (Shahin *et al.*, 2024). These systems can offer more nuanced and effective service by understanding and responding to the full range of human communication—beyond text or voice. The potential challenge can be over-satisfaction and emotional dependence on a virtual assistant or game.

RQ7. How does the integration of multiple modalities (such as facial expressions, vocal intonations, text, and physiological signals) impact the effectiveness and accuracy of emotion recognition systems in different application contexts?.

Based on review it was documented that each modality provides unique insights into the user's emotional state, and always their combination offers a more holistic understanding of human state. For example, facial expressions provide direct visual indicators of emotion, while vocal intonations add auditory cues that capture nuances in tone and pitch, and text analysis offers insights into sentiment and intent through language (Poria *et al.*, 2017; Y. Yi *et al.*, 2023).

Multimodal emotion recognition systems outperform unimodal systems by capturing a richer set of emotional cues, as demonstrated in studies comparing audio-visual models with those relying solely on visual or auditory data. This comprehensive approach allows these systems to maintain high performance even under varying environmental conditions, such as changes in lighting or background noise, which can significantly impact the accuracy of unimodal systems (Mocanu *et al.*, 2023; Y. Yi *et al.*, 2023).

In virtual reality and gaming environments, MMER more and more change user experience by enabling real-time adaptation to players' emotional states, creating more immersive and engaging experiences. For instance, games can dynamically adjust difficulty levels, narratives, and inter-

actions based on real-time emotional feedback, enhancing player engagement and satisfaction by making games more enjoyable and reducing the likelihood of frustration-induced dropout (Akbar *et al.*, 2019; Hamdy and King, 2018).

In healthcare, integrating modalities like facial, vocal, and physiological data allows for precise monitoring of patients' emotional well-being, which is crucial for timely interventions. Systems that combine EEG data with facial expression and speech analysis have been shown to detect early signs of mental health issues more accurately, thus improving patient care by providing timely interventions and support (Schmidt *et al.*, 2019).

Overall, the integration of MMER systems provides a richer and more detailed understanding of user emotions, enhancing system performance across diverse contexts. This holistic approach not only improves the reliability of emotion detection but also enables more personalized and context-aware interactions, which are essential for creating meaningful and effective user experiences in fields such as gaming, virtual reality, healthcare, and more (Yong Zhang *et al.*, 2021).

9 Identified Resarch Gaps and Unanswered Questions

Research on MMER, particularly on contactless techniques, has advanced significantly in recent years. However, several research gaps and unanswered questions remain. Identifying these gaps is valuable for guiding future research and developing of MMER.

Based on the current literature, including the recent comprehensive review by Khan *et al.*, (2024), some key research gaps and unanswered questions still require deep investigation. For the prioritization purpose, the unanswered questions were divided into three classes:

- **High Priority (Currently Active Research Areas)**. Topics that are well-established in the field with a substantial amount of ongoing research. These areas are critical for immediate advancements, and there is a rich literature supporting them.
- **Medium Priority (Emerging Research Areas):** Areas gaining research interest and have a moderate amount of literature. These topics are important for developing more advanced systems but are not as thoroughly explored as high-priority areas.
- **Low Priority (Virgin Research Areas):** Topics with limited research or that are relatively unexplored. These represent new frontiers or complex challenges that require foundational research and innovation to address effectively.

High Priority (Currently Active Research Areas).These are research gaps that are receiving significant attention from the academic community and are considered critical for the advancement of multimodal emotion recognition systems. Researchers on these issues is ongoing, and there is a substantial amount of literature available.

- **Ethical and Privacy Concerns**: Addressing ethical and privacy issues, particularly in terms of data collection, storage, and usage, is continuously high-priority area. Introducing regulations like AI Act in EU with many limitation possibly increased substantiality of research activity aimed at developing frameworks and guidelines for ethical use of emotion recognition technologies. How can we address ethical and privacy issues associated with the deployment of multimodal emotion recognition systems, particularly in terms of data collection, storage, and usage? What frameworks or guidelines should be established to ensure the ethical use of emotion recognition technologies in various applications?
- **Integration with Emerging Technologies**. The integration of MMER with emerging technologies, including AR, VR, and IoT, is a high priority due to its business potential to enhance user experiences in these domains. This area is actively researched because of the rapid development, using additional data sources, and application of these technologies in various fields. How can multimodal emotion recognition be effectively integrated with emerging technologies such as augmented reality (AR), virtual reality (VR), and the Internet of Things (IoT) to enhance user experiences in these domains?
- **Improving Real-Time Processing Capabilities**: Research on reducing the computational complexity of multimodal emotion recognition systems to enable real-time processing without compromising accuracy is crucial and ongoing. Improvements can be achieved by hard-wear and soft-wear as well. Advances in computational efficiency and the development of new algorithms are key focus area. How can the computational complexity of multimodal emotion recognition systems be reduced to enable real-time processing without compromising accuracy and robustness?

- **Evaluation Metrics and Benchmarking**. Developing and appropriate evaluation metrics, especially with standardized bench-marking frameworks is a priority (as they are lacking) to ensure consistent performance assessment across different multimodal emotion recognition systems. This foundational research is necessary for advancing the field and comparing systems effectively. What are the most appropriate evaluation metrics for assessing the performance of multimodal emotion recognition systems, and how can standardized bench-marking frameworks be developed?

- **Multimodal data fusion**: Developing standardized methods for effectively fusing data from multiple modalities to ensure robustness and accuracy across different applications and environments is a heavily researched topic due to its foundational importance in emotion recognition systems. How can we develop standardized methods for effectively fusing data from multiple modalities to ensure robustness and accuracy across different applications and environments?

Medium Priority (Emerging Research Areas). These research gaps are gaining attention but are not as extensively covered as the high-priority topics. These areas have a growing body of literature and are beginning to attract more research efforts, especially as foundational problems are becoming understood and technology progresses.

- **Cross-Cultural Generalization**. Designing systems that generalize across different cultural contexts is increasingly important as these technologies are deployed globally. Understanding cultural differences in emotional expressions is crucial for building fair and effective systems, making this an emerging priority. Since emotional expression can vary significantly between cultures, how can multimodal emotion recognition systems be designed to generalize across different cultural contexts?

- **Handling Multimodal Data Synchronization**: Best practices for ensuring temporal synchronization between different modalities are being explored, especially for real-time emotion recognition systems. This area is increasingly gaining relevance as multimodal systems are becoming more sophisticated. What are the best practices for ensuring temporal synchronization between different modalities, such as audio, video, and physiological signals, in real-time emotion recognition systems?

- **Real-World Deployment and User Acceptance**. As MMER is increasingly implemented, a better understanding of the factors that influence user acceptance and trust is important. Research in this area is focused on improving system design to meet user expectations and ensure widespread adoption. What factors influence user acceptance and trust in emotion recognition systems, and how can these systems be designed to meet user expectations and requirements in real-world deployments?

- **Addressing Ambiguities in Emotion Recognition**. Improving systems to handle ambiguous or mixed emotions is an emerging research area. Although complex, recognizing multiple emotional states simultaneously is vital for real-world applications and is beginning to receive more attention. How can systems be improved to handle ambiguous or mixed emotions, where an individual may exhibit multiple emotional states simultaneously?

- **Context-Aware Emotion Recognition**: New approaches are needed to incorporate contextual information from interactions in real-world scenarios. These new ideas will be important for making MMER more applicable and effective in diverse environments. To improve MMER accuracy and applicability in real-world scenarios, what approaches can be developed to incorporate contextual information from interaction (e.g., environmental factors, social context) into emotion recognition models?

- **Robustness to Environmental Variability**: Making MMER more robust to environmental variations such as lighting conditions and background noise is becoming an important area of focus as these systems are deployed in more diverse and uncontrolled settings. How can multimodal emotion recognition systems be more robust to environmental variations such as lighting conditions, background noise, and occlusions?

- **Explainability and Transparency of Models**. Enhancing the explainability and transparency of deep learning models used in emotion recognition is a newer area of research. As the adoption of these systems grows, the demand for understanding how these models make decisions will likely increase, but current research is still preliminary. How can we enhance

the explainability and transparency of deep learning models used in emotion recognition to build trust and understanding among users and stakeholders?

- **Development of Large-Scale, High-Quality Datasets**. Creating and maintaining diverse and representative multimodal datasets is a critical gap yet to be fully addressed. The development of such datasets is necessary for training and validating robust emotion recognition systems but is still in its early stages. What are the effective strategies for creating and maintaining large-scale, high-quality multimodal datasets representative of diverse populations and various emotional states? How can we ensure that these datasets are annotated accurately and consistently?

Low Priority (Virgin Research Areas) These research questions are less explored, often due to their complexity, the nascent state of the technology, or the need for more foundational research before they can be fully addressed. These are "virgin" areas with limited literature and represent opportunities for groundbreaking research.

- **Addressing Individual Differences in Emotional Expression**. MMER systems that can take into account inter-item variability are something that businesses are counting on. While there is a common awareness of the importance of this problem, comprehensive solutions are still lacking. How can emotion recognition systems be designed to account for inter-subject variability, such as differences in how individuals express emotions through facial expressions, voice, and physiological responses?

- **Longitudinal Studies and Temporal Dynamics**. Conducting longitudinal studies to understand how emotional expressions and recognition accuracy evolve is relatively unexplored. This area is important for evaluating long-term user interactions and system performance, but logistical challenges have limited research efforts. What are the best approaches for conducting longitudinal studies to understand how emotional expressions and recognition accuracy evolve in real-world applications?

- **Impact of Physiological Signals on Emotion Recognition** Using physiological signals measured by wearable in MMER is an emerging area with significant potential. However, this field is still in its early stages, and more foundational research is needed to understand its impact and utility fully.

- **Mitigating Bias and Ensuring Fairness** Addressing bias and ensuring fairness in emotion recognition systems is essential but has received less focus than other ethical considerations. When speaking about bias, we are talking not only about identified cultural or gender differences, but also those resulting from the technologies themselves (such as feature alignment) or other methodological factors influencing recognition errors. This area will likely grow in importance as awareness of AI biases increases.

- **Enhancing User Engagement and Satisfaction** While improving user engagement and satisfaction through multimodal emotion recognition is important, it is less researched than other topics. There is a growing interest in this area, because it can help in building trust and consent for using advanced technology, but it remains a lower priority due to its broader focus on user experience.

- **Application-Specific Customization**. Customizing emotion recognition systems for specific applications is a relatively new area. Research here focuses on adapting systems to different use cases, but it is still developing and not yet a primary focus for many researchers.

- **Easy to use, low or even no code and responsible by design frameworks** - in the MMER area there is a lack of ready-made off-the-shelf frameworks for easy creation of state/emotion analysis systems that would already contain elements enforcing responsible use "by design".

MMER is becoming a multidisciplinary field; therefore, solving the listed research gaps will be based on innovative interdisciplinary research combining achievements in computer science, psychology, neuroscience, ethics and other fields.

Based on the above topics classification, the visualization of Technology Lifecycle analysis for MMER was prepared (1) organized by priority used at Technology Lifecycle Stages (Gartner Hype Cycle):

The Standard Gartner Hype Cycle model by (Gregory *et al.*, 2013) is based on stages that represent the typical progression of a technology or research area from inception to maturity and

Technium Vol. 26, pp.102-140 (2025)
ISSN: 2668-778X
www.techniumscience.com

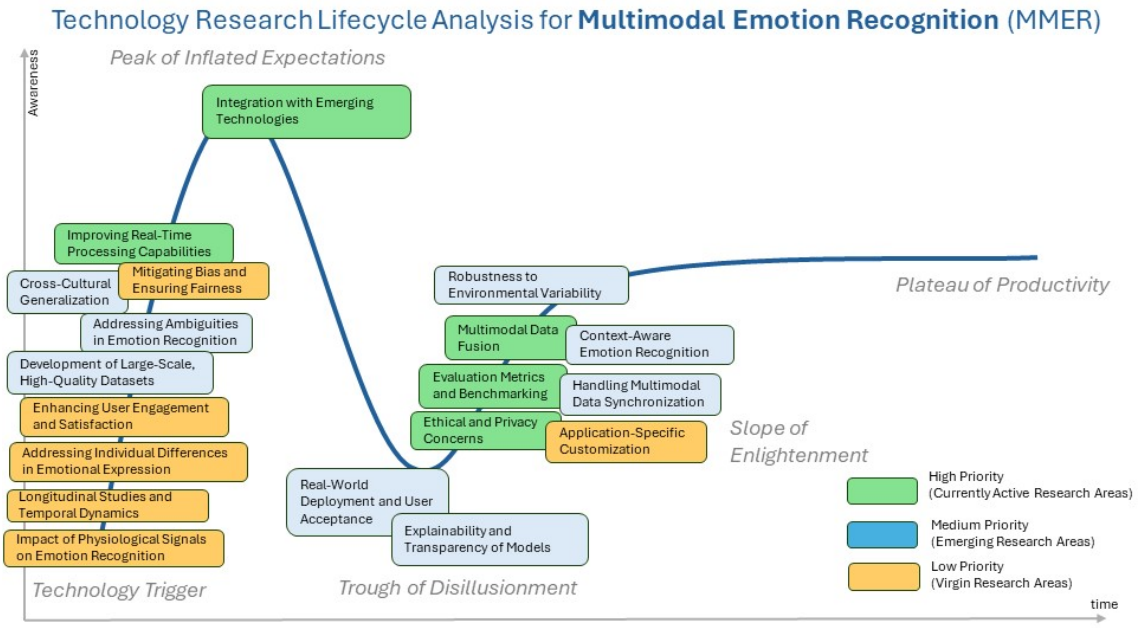


Figure 1: Technology Research Lifecycle Analysis for Multimodal Emotion Recognition (MMER) - own work

widespread adoption. The stages are (names and definitions taken from the source (Gregory *et al.*, 2013; Colett, 2017):

- Innovation Trigger: A technology breakthrough or significant innovation gains interest. Often, it is still in the lab or early prototyping stage.
- Peak of Inflated Expectations: Early publicity produces many success stories—often accompanied by scores of failures. Some companies take action; many do not.
- Trough of Disillusionment: Interest wanes as experiments and implementations fail to deliver. Producers of the technology shake out or fail. Investments will continue only if the surviving providers improve their products to the satisfaction of early adopters.
- Slope of Enlightenment: More instances of how technology can benefit the enterprise start crystallizing and becoming more widely understood. Second- and third-generation products appear from technology providers.
- Plateau of Productivity: Mainstream adoption starts to take off. Criteria for assessing provider viability are more clearly defined. The technology's broad market applicability and relevance are paying off.

The typical Gartner cycle is based on interest, awareness and visibility of technology concept which change in time. Classification of priority indicate the current focus or urgency in addressing specific **research areas or challenges**, often driven by external demands, immediate benefits, or foundational gaps that need filling.

This slight difference is a reason why *Innovation Trigger* is not directly Correlated with *Virgin Research Areas*. *Innovation Trigger* refers to the beginning of the lifecycle for technologies that may have been around for some time but are now gaining interest because of a recent innovation or breakthrough. It does not necessarily mean the research is brand new; it might just be a new application or renewed interest due to technological advancements. While *Virgin Research Areas*, classified as low priority, are those that are new or exploratory and are not yet seen as critical. These areas are more speculative and foundational. While they may coincide with early stages of the technology lifecycle, they are categorized as *low priority* because they are not yet a focus for active development or investment.

So in general, class of maturity is not highly correlated with technology lifecycle stages. Maturity classifications (High, Medium, Low Priority) focus on the urgency and importance of research areas from a practical perspective:

- High-priority areas often correspond to stages like the *Trough of Disillusionment* or *Slope of Enlightenment* where there are active attempts to address gaps, failures, or ethical concerns to push the technology forward.
- Medium-priority areas align with stages like *Slope of Enlightenment*, where the technology is gaining a clearer understanding, but there's less urgency compared to high-priority issues.
- Low-priority areas are often still in the *Innovation Trigger* stage, where exploration is happening but without immediate pressure or a clear path to practical application.

Otherwise, Technology Lifecycle Stages are about the journey of a technology's development and market adoption over time. A technology can be at the Innovation Trigger phase but not be considered a high priority if it doesn't immediately impact critical areas or have clear, short-term applications.

So, High Priority Areas in *Trough of Disillusionment* are issues like ethical concerns have surfaced due to real-world deployments failing to consider these aspects. Thus, they are high-priority to fix, even if they are in a "trough" where expectations have been adjusted due to these failures. In Medium Priority Areas in *Innovation Trigger* there topics like "Cross-Cultural Generalization" are just beginning to be explored (*Innovation Trigger*) but are not seen as immediately crucial, hence medium priority. And within class of Low Priority Areas in *Slope of Enlightenment* are topics like "Application-Specific Customization" are starting to gain clarity (*Slope of Enlightenment*) as specific applications of the technology are understood better but are not seen as pressing needs currently.

It's important to add that this classification will change in time quickly so should be actualised regularly and treated as knowledge that quickly becomes outdated, although it is valuable at a given point in time as a determinant of research priorities in a commercial company.

For readers of this review convenience, the challenges identified are briefly summarized in the Table VI:

Challenge	**Key Issues and Observations**
Integration of Data	Lack of standard multimodal data fusion methods, need for advanced real-time signal processing, and insufficient research on effective data integration algorithms.
Real-Time Multimodal Data Processing	Need for low-latency algorithms, optimized edge computing methods, and better integration of real-time processing with existing systems.
Data Quality and Reliability	Noise or inaccuracies in one modality affect overall performance; scalability issues arise with increased modalities, and fusion techniques need to maintain system efficiency.
Integration of Modalities	Fusion requires sophisticated algorithms and significant computational resources; real-time processing challenges include responsiveness and adaptation in dynamic environments.
Impact of Individual Differences	Limited datasets reflecting demographic and cultural diversity, need for adaptable models, and insufficient analysis of cultural impacts on emotion recognition.
Best Predictive Modalities	Insufficient comparative research on modalities, lack of understanding of their interactions in dynamic settings, and need for improved feature extraction techniques.
Ethical Considerations	Lack of privacy protection guidelines, need to address biases in data/models, and insufficient analysis of ethical implications in various contexts.
User Engagement and Satisfaction	Few empirical studies on multimodal emotion recognition's impact on satisfaction, need for long-term engagement analysis, and optimization of interfaces using emotion data.
Transparency and Interpretability	Difficulty in understanding modality contributions; improving transparency is critical for ethical use and building trust in systems.
Privacy Concerns	Significant challenges in managing sensitive data like facial expressions and physiological signals; critical need for privacy-preserving techniques and compliance with regulations.

Table VI: Identified Challenges in Multimodal Emotion Recognition

Finally, the most important thing is **ensuring the quality and reliability of data from different modalities**. Any interference or inaccuracy in one modality can significantly impact

the overall performance of the emotion recognition system. Maintaining high-quality data interpretation in all modalities is essential for accurate emotion detection. This involves a large amount of work for multiple cross-validated tests.

With the expected increase in the number of modalities, the complexity of data fusion increases exponentially. Due to scalability issues, problems will arise from the balance between the accuracy and practicality of MMER systems, especially in large-scale implementations. Therefore, **developing efficient fusion techniques** that can handle multiple data sources without compromising performance is a serious challenge.

Another paradox - trap is that multimodal systems increase accuracy. However, as they increase in complexity and operate on artefacts inaccessible to human evaluation, the possibilities of their reliable interpretation will decrease. **Understanding how different modalities contribute to the final emotion recognition decision can be difficult, directly impacting the explanation of the system's decision. Therefore, improving the transparency of these systems and their interpretability is necessary to build trust and ensure ethical use.**

Collecting and processing sensitive personal data, such as facial expressions and physiological signals, raises significant privacy issues. Ensuring compliance with regulations like the AI Act is crucial to protecting user data and maintaining public trust. Implementing robust privacy-preserving techniques is essential to address these concerns. **And the fundamental question is: Do we need to collect any personnel for MMER in adaptive systems?** Is it possible to process the sensitive data, use them in real-time and not store them at all? That is also an important challenge.

10 Final Conclusions

This review highlights the transformative potential of MMER in developing increasingly natural human-AI interactions. Thanks to using different modalities, MMER systems are becoming more intuitive, adaptive, and emotionally intelligent year by year. The main goal of the review was to identify attractive gaps in the current state of knowledge so that researchers can invest their efforts in potential innovations and perhaps even inventions.

This article provides a helpful overview of basic emotion theories for MMER researchers and developers—from Ekman's Basic Emotions to Barrett's Psychological Construct—along with assessing their applicability to different MMER use cases. It also offers several conclusions on modelling human emotions, which is especially helpful for MMER developers and designers without a psychology background. The most popular classification of basic emotions in the peer-reviewed literature, Ekman's (most popular only because of the popularity of labelled training sets), is not necessarily the best for adaptive systems using MMER.

The speed of development of deep learning and multimodal fusion techniques is constantly increasing the accuracy of emotion recognition and the chances of their use in adaptive real-time systems, and the decision tree included in the proposed one simplifies the choice of appropriate theoretical approaches and fusion methods for system designers.

It is also proposed to evaluate the current knowledge base and works based on the Hype Cycle, dividing technical components into five development phases and assessing their level of advancement. This analysis shows that, as researchers, we are at the beginning of a gigantic field of knowledge which aims to realize the imitation of human emotional intelligence.

Technical challenges like handling heterogeneous data and increasing computational requirements are combined with ethical challenges and concerns, such as manipulation, misrecognition or addiction. MMER is mainly based on the interdisciplinary cooperation of researchers and developers.

The future of MMER primarily involves developing data fusion techniques, cultural adaptability, and personal adaptability. A desirable direction will be to develop hybrid models that can integrate multiple modalities more effectively and operate even without single-modality data. MMER systems will also be developed to self-adapt cultural differences and individual characteristics, such as personality, to improve the personalization of user interactions.

Undoubtedly, the range of real-time processing capabilities will improve, with a particular focus on optimizing deep learning models for use in resource-constrained environments, where the constant requirement is to minimize latency.

As MMER evolves, ethical considerations will become increasingly important, focusing on ensuring privacy and mitigating bias in MMER applications. Techniques such as differential privacy and federated learning can help address privacy concerns. At the same time, mitigation strategies

will be integrated into MMER by design systems to ensure appropriate assessment of user states and fair treatment of different user groups. In addition, it will be necessary to develop explainable AI (XAI) models as this will be a key factor in improving the transparency and trust of users, enabling them to understand how the system assesses their emotional states.

Another direction of MMER development is to adapt them to the specific needs of emerging new interpersonal, entertainment, work and business technologies, such as the metaverse, autonomous vehicles or virtual personal assistants. The continued development of MMER for mental health applications, adaptive learning (educational systems), personalized services and customer care or human capital will also drive future research. A broad exploration of MMER applications will make systems increasingly efficient, adaptive and ethically sound, enabling the creation of emotionally intelligent AI systems.

Current research on MMER systems suggests that future advances may also prioritize integrating contactless and contact approaches to enhance user comfort while maintaining high accuracy. Researchers will address challenges such as improving sensor robustness and creating culturally adaptable algorithms to ensure reliable performance across different user populations. Moreover, ongoing research on various multimodal fusion approaches, especially tensor and hierarchical models, emphasizes the need to dynamically combine modalities, even in cases where individual modalities are unavailable or inconsistent. The review proposes a classification of fusion for use cases.

Intricate hierarchical modality fusion will enhance the system's resilience and efficiency, particularly in uncertain real-world situations. On the other hand, development of cross-modal attention techniques hold the potential to further enhance data integration, and increasing MMER's capability to grasp the subtleties of emotional states with minimizing computational demands.

As MMER systems mature, their application in highly interactive and immersive environments such as virtual reality and the metaverse will redefine how humans interact with technology. Emotional data can dynamically adapt to virtual environments, creating deeply personalized experiences that adapt in real time. More and more there will be a strong need for creating ethical guidelines to prevent any possible misuse or violation when changing user conditions.

In conclusion, MMER offers a unique chance to connect the intricacies of human emotions with machine intelligence. As studies advance the achievement of this domain will rely on promoting cooperation across disciplines.

It is also important to note that the current approach to obtaining MMER is focused primarily on reproducing another person's ability to assess the other person's emotional state. An interesting future direction for research would be to develop methods that allow for assessing the actual emotional state rather than just predicting the state resulting from the other person's assessment. Automation in the creation of labelling training sets for MMER training could significantly change the pace of development in this field.

ACKNOWLEDGMENTS: Preparation of this review was financed by Orange Innovation Research departments and was technically prepared by the AI Skills Center department at Orange Innovation Poland under the direction of Paweł Tuszyński. The author would like to thank Michał Szczerbak for his supervision.

Izabella Krzemińska is a Technology Leader in Orange Innovation Poland. She received her M. Sc in psychology and psychometric from University of Warsaw and PhD in Data Science in Business in the Faculty of Computer Science and Electronic Economy/Poznan University of Economics. She is a researcher and data analyst with almost 20 years of work experience (both research institutes, consulting and client-side). She manages R&D projects concerning creating new technologies and services based on digital data and creating additional value based on it, especially in the psychology area (cyber-personality, emotion recognition). Currently she works on Responsible AI and AI supporting business in Industry 5.0 era.

References

Akbar, M Taufik *et al.*, (2019). "Enhancing game experience with facial expression recognition as dynamic balancing", *Procedia Computer Science*, Vol. 157, pp. 388–395.

Amershi, Saleema *et al.*, (2019). "Guidelines for human-AI interaction", *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13.

Arnold, Magda B (2013). *Feelings and emotions: The Loyola symposium*, Vol. 7. Academic Press.

Bălan, Oana *et al.*, (2020). "An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy", *Sensors*, Vol. 20 No. 2, p. 496.

Barrett, Lisa Feldman (2006). "Solving the emotion paradox: Categorization and the experience of emotion", *Personality and social psychology review*, Vol. 10 No. 1, pp. 20–46.

— (2017). *How emotions are made: The secret life of the brain*, Pan Macmillan.

Barthet, Matthew *et al.*, (2024). "Closing the Affective Loop via Experience-Driven Reinforcement Learning Designers", *arXiv preprint arXiv:2408.06346*,

Belaiche, Reda *et al.*, (2020). "Cost-effective CNNs for real-time micro-expression recognition", *Applied Sciences*, Vol. 10 No. 14, p. 4959.

Binns, Reuben *et al.*, (2018). "'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions", *Proceedings of the 2018 Chi conference on human factors in computing systems*, pp. 1–14.

Bruner, Jerome (1990). *Acts of meaning: Four lectures on mind and culture*, Vol. 3. Harvard university press.

Bucur, Ana-Maria *et al.*, (2023). "It's just a matter of time: Detecting depression with time-enriched multimodal transformers", *European Conference on Information Retrieval*. Springer, pp. 200–215.

Buechel, Sven and Hahn, Udo (2017). "A flexible mapping scheme for discrete and dimensional emotion representations: Evidence from textual stimuli", *CogSci 2017—Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pp. 180–185.

Buolamwini, Joy and Gebru, Timnit (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification", *Conference on fairness, accountability and transparency*. PMLR, pp. 77–91.

Butz, Andreas (2010). "User interfaces and HCI for ambient intelligence and smart environments", *Handbook of ambient intelligence and smart environments*. Springer, pp. 535–558.

Cambria, Erik *et al.*, (2024). "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing", *International Conference on Human-Computer Interaction (HCII)*.

Chatterjee, Chintan *et al.*, (2024). "A Survey on Multi-modal Emotion Detection Techniques",

Chavez, Robert S and Heatherton, Todd F (2015). "Multimodal frontostriatal connectivity underlies individual differences in self-esteem", *Social cognitive and affective neuroscience*, Vol. 10 No. 3, pp. 364–370.

Chen, Rongfei *et al.*, (2022). "Video-based cross-modal auxiliary network for multimodal sentiment analysis", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 32 No. 12, pp. 8703–8716.

Chen, Xiaofan, Lao, Songyang, and Duan, Ting (2020). "Multimodal fusion of visual dialog: A survey", *Proceedings of the 2020 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence*, pp. 302–308.

Chen, Yiyuan *et al.*, (2024). "EEG emotion recognition based on Ordinary Differential Equation Graph Convolutional Networks and Dynamic Time Wrapping", *Applied Soft Computing*, Vol. 152, p. 111181.

Choo, Sanghyun *et al.*, (2023). "Effectiveness of multi-task deep learning framework for EEG-based emotion and context recognition", *Expert Systems with Applications*, Vol. 227, p. 120348.

Colett, Alexis (2017). *The Art of Hype*, PhD thesis. Berklee College of Music.

Conati, Cristina and Maclaren, Heather (2009). "Empirically building and evaluating a probabilistic model of user affect", *User Modeling and User-Adapted Interaction*, Vol. 19, pp. 267–303.

D'Mello, Sidney, Dieterle, Ed, and Duckworth, Angela (2017). "Advanced, analytic, automated (AAA) measurement of engagement during learning", *Educational psychologist*, Vol. 52 No. 2, pp. 104–123.

Dehghani, Amin, Soltanian-Zadeh, Hamid, and Hossein-Zadeh, Gholam-Ali (2023). "Probing fMRI brain connectivity and activity changes during emotion regulation by EEG neurofeedback", *Frontiers in Human Neuroscience*, Vol. 16, p. 988890.

Dricu, Mihai and Frühholz, Sascha (2020). "A neurocognitive model of perceptual decision-making on emotional signals", *Human Brain Mapping*, Vol. 41 No. 6, pp. 1532–1556.

Durante, Zane *et al.*, (2024). "Agent ai: Surveying the horizons of multimodal interaction", *arXiv preprint arXiv:2401.03568*,

Dwork, Cynthia, Roth, Aaron, *et al.*, (2014). "The algorithmic foundations of differential privacy", *Foundations and Trends® in Theoretical Computer Science*, Vol. 9 No. 3–4, pp. 211–407.

Ekman, Paul and Cordaro, Daniel (2011). "What is meant by calling emotions basic", *Emotion review*, Vol. 3 No. 4, pp. 364–370.

Ekman, Paul, Friesen, W v, and Hager, J (2002). "Facial action coding system: Research Nexus", *Network Research Information, Salt Lake City, UT*, Vol. 1.

Ekman, Paul and Friesen, Wallace V (1971). "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, Vol. 17 No. 2, p. 124.

— (1978). "Facial action coding system", *Environmental Psychology & Nonverbal Behavior*,

Ekman, Paul, Sorenson, E Richard, and Friesen, Wallace V (1969). "Pan-cultural elements in facial displays of emotion", *Science*, Vol. 164 No. 3875, pp. 86–88.

Fontaine, Johnny JR, Scherer, Klaus R, and Soriano, Cristina (2013). *Components of emotional meaning: A sourcebook*, OUP Oxford.

Frachi, Yann, Chanel, Guillaume, and Barthet, Mathieu (2023). "Affective gaming using adaptive speed controlled by biofeedback", *Companion Publication of the 25th International Conference on Multimodal Interaction*, pp. 238–246.

Frijda, Nico H (1986). *The emotions*, Cambridge University Press.

Gandhi, Ankita *et al.*, (2023). "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions", *Information Fusion*, Vol. 91, pp. 424–444.

GDPR, General Data Protection Regulation (2016). "General data protection regulation", *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*,

Ghandeharioun, Asma *et al.*, (2019). "Emma: An emotion-aware wellbeing chatbot", *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 1–7.

Ghosh, Anay *et al.*, (2023). "A multimodal sentiment analysis system for recognizing person aggressiveness in pain based on textual and visual information", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 14 No. 4, pp. 4489–4501.

Ghosh, Soumitra *et al.*, (2023). "Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data", *Knowledge-Based Systems*, Vol. 260, p. 110182.

Gregory, Sue, Tavares-Jones, Nancy, and Jerry, Paul (2013). *The Hype Cycle Upswing: The Resurgence of Virtual Worlds*,

Gunning, David and Aha, David (2019). "DARPA's explainable artificial intelligence (XAI) program", *AI magazine*, Vol. 40 No. 2, pp. 44–58.

Hamdy, Salma and King, David (2018). "Affective games: a multimodal classification system", *19th annual European GAME-ON Conference (GAME-ON'2018) on Simulation and AI in Computer Games*. EUROSIS.

Hareli, Shlomo and Parkinson, Brian (2008). "What's social about social emotions?", *Journal for the theory of social behaviour*, Vol. 38 No. 2, pp. 131–156.

He, Ping *et al.*, (2023). "Cross-Modal Sentiment Analysis of Text and Video Based on Bi-GRU Cyclic Network and Correlation Enhancement", *Applied Sciences*, Vol. 13 No. 13, p. 7489.

Hess, Ursula and Thibault, Pascal (2009). "Why the same expression may not mean the same when shown on different faces or seen by different people", *Affective information processing*. Springer, pp. 145–158.

Hu, Guimin *et al.*, (2024). "Recent Trends of Multimodal Affective Computing: A Survey from NLP Perspective", *arXiv preprint arXiv:2409.07388*,

Hudlicka, Eva (2016). "Virtual affective agents and therapeutic games", *Artificial intelligence in behavioral and mental health care*. Elsevier, pp. 81–115.

— (2019). "Modeling cognition–emotion interactions in symbolic agent architectures: Examples of research and applied models", *Cognitive Architectures*, pp. 129–143.

Hutchinson, Ben and Mitchell, Margaret (2019). "50 years of test (un) fairness: Lessons for machine learning", *Proceedings of the conference on fairness, accountability, and transparency*, pp. 49–58.

Izard, Carroll E (1994). "Innate and universal facial expressions: evidence from developmental and cross-cultural research."

Juyal, Prachi (2022). "Multi-modal sentiment analysis of audio and visual context of the data using machine learning", *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, pp. 1198–1205.

Kadyr, Sarsenbek and Tolganay, Chinibayeva (2024). "Affective computing methods for simulation of action scenarios in video games", *Procedia Computer Science*, Vol. 231, pp. 341–346.

Kalateh, Sepideh *et al.*, (2024). "A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges", *IEEE Access*,

Khalaf, Oshamah Ibrahim *et al.*, (2024). "Elevating metaverse virtual reality experiences through network-integrated neuro-fuzzy emotion recognition and adaptive content generation algorithms", *Engineering Reports*, e12894.

Khan, Umair Ali *et al.*, (2024). "Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects", *Multimedia Systems*, Vol. 30 No. 3, pp. 1–48.

Kim, Taewoon and Vossen, Piek (2021). "Emoberta: Speaker-aware emotion recognition in conversation with roberta", *arXiv preprint arXiv:2108.12009*,

Kong, Chenqi *et al.*, (2024). "M ˆ{3} FAS: An Accurate and Robust MultiModal Mobile Face Anti-Spoofing System", *IEEE Transactions on Dependable and Secure Computing*,

Kong, Weiqi (2024). "Research Advanced in Multimodal Emotion Recognition Based on Deep Learning", *Highlights in Science, Engineering and Technology*, Vol. 85, pp. 602–608.

Krishna, DN (2021). "Using large pre-trained models with cross-modal attention for multi-modal emotion recognition", *arXiv preprint arXiv:2108.09669*, Vol. 2.

Kuppens, Peter, Oravecz, Zita, and Tuerlinckx, Francis (2010). "Feelings change: accounting for individual differences in the temporal dynamics of affect." *Journal of personality and social psychology*, Vol. 99 No. 6, p. 1042.

Kwon, Seungjin *et al.*, (2022). "Analytical framework for facial expression on game experience test", *IEEE Access*, Vol. 10, pp. 104486–104497.

Lazarus, Richard S (1968). "Emotions and adaptation: Conceptual and empirical relations." *Nebraska symposium on motivation*. University of Nebraska Press.

— (1991). "Progress on a cognitive-motivational-relational theory of emotion." *American psychologist*, Vol. 46 No. 8, p. 819.

LeCun, Yann, Bengio, Yoshua, and Hinton, Geoffrey (2015). "Deep learning", *nature*, Vol. 521 No. 7553, pp. 436–444.

Lei, Jing, Sala, Johannan, and Jasra, Shashi K (2017). "Identifying correlation between facial expression and heart rate and skin conductance with iMotions biometric platform", *Journal of Emerging Forensic Sciences Research*, Vol. 2 No. 2, pp. 53–83.

Levy, Yair and Ellis, Timothy J (2006). "A systems approach to conduct an effective literature review in support of information systems research." *Informing Science*, Vol. 9.

Li, Dongyuan *et al.*, (2023). "Joyful: Joint Modality Fusion and Graph Contrastive Learning for Multimoda Emotion Recognition", *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Bouamor, Houda, Pino, Juan, and Bali, Kalika. Association for Computational Linguistics: Singapore, pp. 16051–16069. DOI: 10.18653/v1/2023.emnlp-main.996. **available at**: https://aclanthology.org/2023.emnlp-main.996.

Li, Shaokai *et al.*, (2024). "Feature distribution Adaptation Network for Speech Emotion Recognition", *arXiv preprint arXiv:2410.22023*,

Lian, Hailun *et al.*, (2023). "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face", *Entropy*, Vol. 25 No. 10, p. 1440.

Lindquist, Kristen A *et al.*, (2012). "The brain basis of emotion: a meta-analytic review", *Behavioral and brain sciences*, Vol. 35 No. 3, pp. 121–143.

Lisetti, Christine L and Nasoz, Fatma (2002). "MAUI: a multimodal affective user interface", *Proceedings of the tenth ACM international conference on Multimedia*, pp. 161–170.

Liu, Chang *et al.*, (2024). "EmoFace: Audio-driven Emotional 3D Face Animation", *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, pp. 387–397.

Lopes, Júlio Castro and Lopes, Rui Pedro (2022). "A review of dynamic difficulty adjustment methods for serious games", *International Conference on Optimization, Learning Algorithms and Applications*. Springer, pp. 144–159.

Ma, Fei *et al.*, (2022). "Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN", *Applied Sciences*, Vol. 12 No. 1, p. 527.
Mamieva, Dilnoza *et al.*, (2023). "Multimodal emotion detection via attention-based fusion of extracted facial and speech features", *Sensors*, Vol. 23 No. 12, p. 5475.
Markus, Hazel Rose and Kitayama, Shinobu (2010). "Cultures and selves: A cycle of mutual constitution", *Perspectives on psychological science*, Vol. 5 No. 4, pp. 420–430.
May, Alyxander David *et al.*, (2017). "Human emotional understanding for empathetic companion robots", *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*. Springer, pp. 277–285.
McDuff, Daniel *et al.*, (2014). "Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads", *IEEE Transactions on Affective Computing*, Vol. 6 No. 3, pp. 223–235.
Mehrabi, Ninareh *et al.*, (2021). "A survey on bias and fairness in machine learning", *ACM computing surveys (CSUR)*, Vol. 54 No. 6, pp. 1–35.
Mehu, Marc and Scherer, Klaus R (2015). "Emotion categories and dimensions in the facial communication of affect: An integrated approach." *Emotion*, Vol. 15 No. 6, p. 798.
Meng, Tao *et al.*, (2024). "Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*,
Mesquita, Batja, Boiger, Michael, and De Leersnyder, Jozefien (2016). "The cultural construction of emotions", *Current opinion in psychology*, Vol. 8, pp. 31–36.
Minsky, Marvin (1988). *Society of mind*, Simon and Schuster.
Mitchell, Shira *et al.*, (2021). "Algorithmic fairness: Choices, assumptions, and definitions", *Annual review of statistics and its application*, Vol. 8 No. 1, pp. 141–163.
Mocanu, Bogdan, Tapu, Ruxandra, and Zaharia, Titus (2023). "Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning", *Image and Vision Computing*, Vol. 133, p. 104676.
Moors, Agnes *et al.*, (2013). "Appraisal theories of emotion: State of the art and future development", *Emotion review*, Vol. 5 No. 2, pp. 119–124.
Nasir, Jauwairia *et al.*, (2022). "What if social robots look for productive engagement? Automated assessment of goal-centric engagement in learning applications", *International Journal of Social Robotics*, Vol. 14 No. 1, pp. 55–71.
Nissenbaum, Helen (2011). "A contextual approach to privacy online", *Daedalus*, Vol. 140 No. 4, pp. 32–48.
Okoli, Chitu and Schabram, Kira (2015). "A guide to conducting a systematic literature review of information systems research",
Otamendi, F Javier (2022). "Statistical emotion control: Comparing intensity and duration of emotional reactions based on facial expressions", *Expert Systems with Applications*, Vol. 200, p. 117074.
Parliament, European (2021). "Artificial intelligence act", *European Parliament: European Parliamentary Research Service*,
Peng, Min *et al.*, (2017). "Dual temporal scale convolutional neural network for micro-expression recognition", *Frontiers in psychology*, Vol. 8, p. 273835.
Pervez, Farrukh *et al.*, (2024). "Affective Computing and the Road to an Emotionally Intelligent Metaverse", *IEEE Open Journal of the Computer Society*,
Pessoa, Luiz (2013). *The cognitive-emotional brain: From interactions to integration*, MIT press.
Picard, Rosalind W (2010). "Affective computing: from laughter to IEEE", *IEEE transactions on affective computing*, Vol. 1 No. 1, pp. 11–17.
Poria, Soujanya *et al.*, (2017). "Context-dependent sentiment analysis in user-generated videos", *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883.
Porter, Stephen and Ten Brinke, Leanne (2008). "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions", *Psychological science*, Vol. 19 No. 5, pp. 508–514.
Rahman, MD *et al.*, (2024). *A comprehensive NLP-based voice assistant system for streamlined information retrieval in metro rail services of Bangladesh*, PhD thesis. Brac University.
Reardon, Claudia L *et al.*, (2019). "Mental health in elite athletes: International Olympic Committee consensus statement (2019)", *British journal of sports medicine*, Vol. 53 No. 11, pp. 667–699.

Reuderink, Boris, Mühl, Christian, and Poel, Mannes (2013). "Valence, arousal and dominance in the EEG during game play", *International journal of autonomous and adaptive communications systems*, Vol. 6 No. 1, pp. 45–62.

Rezapour, Mohammad Mahdi, Fatemi, Afsaneh, and Nematbakhsh, Mohammad Ali (2024). "A methodology for using players' chat content for dynamic difficulty adjustment in metaverse multiplayer games", *Applied Soft Computing*, Vol. 156, p. 111497.

Riva, Giuseppe, Wiederhold, Brenda K, and Mantovani, Fabrizia (2019). "Neuroscience of virtual reality: from virtual exposure to embodied medicine", *Cyberpsychology, behavior, and social networking*, Vol. 22 No. 1, pp. 82–96.

Russell, James A (1980). "A circumplex model of affect." *Journal of personality and social psychology*, Vol. 39 No. 6, p. 1161.

— (1994). "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies." *Psychological bulletin*, Vol. 115 No. 1, p. 102.

— (2003). "Core affect and the psychological construction of emotion." *Psychological review*, Vol. 110 No. 1, p. 145.

Russell, James A, Lewicka, Maria, and Niit, Toomas (1989). "A cross-cultural study of a circumplex model of affect." *Journal of personality and social psychology*, Vol. 57 No. 5, p. 848.

Savchenko, AV and Savchenko, LV (2022). "Audio-visual continuous recognition of emotional state in a multi-user system based on personalized representation of facial expressions and voice", *Pattern Recognition and Image Analysis*, Vol. 32 No. 3, pp. 665–671.

Savchenko, Lyudmila and V Savchenko, Andrey (2021). "Speaker-aware training of speech emotion classifier with speaker recognition", *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*. Springer, pp. 614–625.

Sayyed, Mudassar *et al.*, (2024). "Human-Machine Interaction in the Metaverse: A Comprehensive Review and Proposed Framework", *Impact and Potential of Machine Learning in the Metaverse*, pp. 1–28.

Scherer, Klaus R (1982). *The nature and function of emotion*,

— (2009). "The dynamic architecture of emotion: Evidence for the component process model", *Cognition and emotion*, Vol. 23 No. 7, pp. 1307–1351.

Scherer, Klaus R, Schorr, Angela, and Johnstone, Tom (2001). *Appraisal processes in emotion: Theory, methods, research*, Oxford University Press.

Scherer, Klaus R and Wallbott, Harald G (1994). "Evidence for universality and cultural variation of differential emotion response patterning." *Journal of personality and social psychology*, Vol. 66 No. 2, p. 310.

Schmidt, Philip *et al.*, (2019). "Wearable-based affect recognition—A review", *Sensors*, Vol. 19 No. 19, p. 4079.

Shahin, Mohammad, Chen, F Frank, and Hosseinzadeh, Ali (2024). "Harnessing customized AI to create voice of customer via GPT3. 5", *Advanced Engineering Informatics*, Vol. 61, p. 102462.

Sham, Abdallah Hussein *et al.*, (2023). "Towards context-aware facial emotion reaction database for dyadic interaction settings", *Sensors*, Vol. 23 No. 1, p. 458.

Shou, Yuntao *et al.*, (2024a). "A low-rank matching attention based cross-modal feature fusion method for conversational emotion recognition", *IEEE Transactions on Affective Computing*,

Shou, Yuntao *et al.*, (2024b). "Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations", *Information Fusion*, Vol. 112, p. 102590.

Simić, Nikola *et al.*, (2024). "Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks", *Applied Sciences*, Vol. 14 No. 4, p. 1325.

Singh, Bhupinder and Kaunert, Christian (2024). "Augmented Reality and Virtual Reality Modules for Mindfulness: Boosting Emotional Intelligence and Mental Wellness", *Applications of Virtual and Augmented Reality for Health and Wellbeing*. IGI Global, pp. 111–128.

Smith, Craig A and Lazarus, Richard S (1993). "Appraisal components, core relational themes, and the emotions", *Cognition & emotion*, Vol. 7 No. 3-4, pp. 233–269.

Srivastava, Arun Pratap *et al.*, (2024). "Bridging the Gap Between Modalities with Cross-Modal Generative AI and Large Model", *2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, pp. 965–971.

Suguitan, Michael *et al.*, (2024). "Face2Gesture: Translating facial expressions into robot movements through shared latent space neural networks", *ACM Transactions on Human-Robot Interaction*, Vol. 13 No. 3, pp. 1–18.

Sutton, Tina M, Herbert, Andrew M, and Clark, Dailyn Q (2019). "Valence, arousal, and dominance ratings for facial stimuli", *Quarterly Journal of Experimental Psychology*, Vol. 72 No. 8, pp. 2046–2055.

Tang, Jiehao *et al.*, (2024). "Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment", *Information Fusion*, Vol. 103, p. 102129.

Vygotsky, Lev Semenovich and Cole, Michael (1978). *Mind in society: Development of higher psychological processes*, Harvard university press.

Wang, Ruiqi *et al.*, (2024). "Husformer: A multi-modal transformer for multi-modal human state recognition", *IEEE Transactions on Cognitive and Developmental Systems*,

Wang, Yan *et al.*, (2022). "A systematic review on affective computing: Emotion models, databases, and recent advances", *Information Fusion*, Vol. 83, pp. 19–52.

Wang, Yusong, Li, Dongyuan, and Shen, Jialun (2024). "Inter-Modality and Intra-Sample Alignment for Multi-Modal Emotion Recognition", *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8301–8305.

Web article:Human-Computer Interaction and Visualization, (n.d.). https://research.google/research-areas/human-computer-interaction-and-visualization/.

Xiaoming, ZHAO, Yijiao, YANG, and Shiqing, ZHANG (2022). "Survey of deep learning based multimodal emotion recognition", *Journal of Frontiers of Computer Science & Technology*, Vol. 16 No. 7, p. 1479.

Xu, Wei and Gao, Zaifeng (2023). "Enabling Human-Centered AI: A Methodological Perspective", *arXiv preprint arXiv:2311.06703*,

Yi, Lu and Mak, Man-Wai (2019). "Adversarial data augmentation network for speech emotion recognition", *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 529–534.

Yi, Yufan *et al.*, (2023). "DBT: multimodal emotion recognition based on dual-branch transformer", *The Journal of Supercomputing*, Vol. 79 No. 8, pp. 8611–8633.

Zadeh, Amir *et al.*, (2017). "Tensor fusion network for multimodal sentiment analysis", *arXiv preprint arXiv:1707.07250*,

Zajonc, Robert B (1984). "On the primacy of affect."

Zeng, Zhihong *et al.*, (2007). "A survey of affect recognition methods: audio, visual and spontaneous expressions", *Proceedings of the 9th international conference on Multimodal interfaces*, pp. 126–133.

Zepf, Sebastian *et al.*, (2020). "Driver emotion recognition for intelligent vehicles: A survey", *ACM Computing Surveys (CSUR)*, Vol. 53 No. 3, pp. 1–30.

Zhang, Chuan-Ke *et al.*, (2017). "An extended reciprocally convex matrix inequality for stability analysis of systems with time-varying delay", *Automatica*, Vol. 85, pp. 481–485.

Zhang, Shiqing *et al.*, (2024). "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects", *Expert Systems with Applications*, Vol. 237, p. 121692.

Zhang, Xiaoheng *et al.*, (2024). "A Multi-Level Alignment and Cross-Modal Unified Semantic Graph Refinement Network for Conversational Emotion Recognition", *IEEE Transactions on Affective Computing*,

Zhang, Yazhou *et al.*, (2021). "Multi-task learning for jointly detecting depression and emotion", *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 3142–3149.

Zhang, Yong, Cheng, Cheng, and Zhang, Yidie (2021). "Multimodal emotion recognition using a hierarchical fusion convolutional neural network", *IEEE access*, Vol. 9, pp. 7943–7951.

Zhao, Sicheng *et al.*, (2021). "Emotion recognition from multiple modalities: Fundamentals and methodologies", *IEEE Signal Processing Magazine*, Vol. 38 No. 6, pp. 59–73.