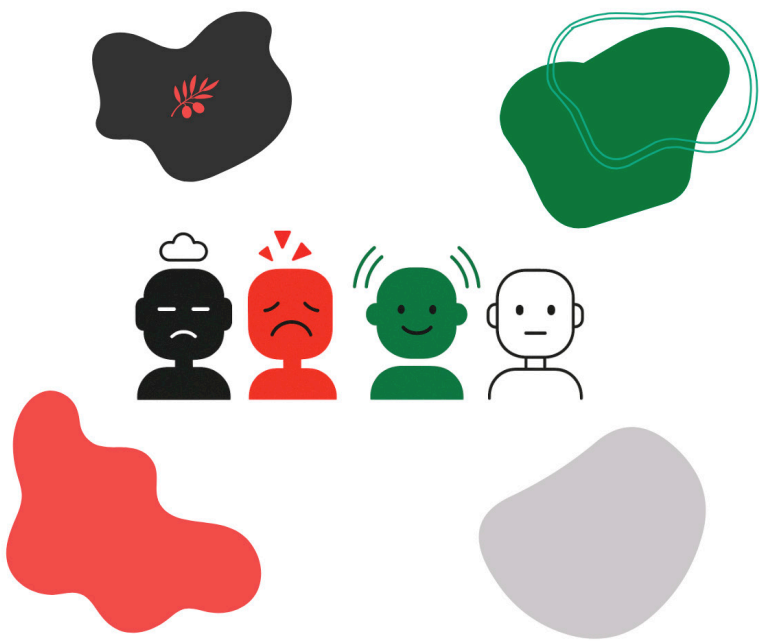
Doctoral Thesis in Computer Sceince

Multi-Modal Affective State Detection For Dyadic Interactions Using Thermal Imaging and Context

YOUSSEF MOHAMED

KTH ROYAL INSTITUTE OF TECHNOLOGY



Multi-Modal Affective State Detection For Dyadic Interactions Using Thermal Imaging and Context

YOUSSEF MOHAMED

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday the 26th of September 2025, at 1:00 p.m. in D37, Lindstedtvägen 9, Stockholm.

Doctoral Thesis in Computer Sceince
KTH Royal Institute of Technology
Stockholm, Sweden 2025

TRITA-EECS-AVL-2025:74
ISBN 978-91-8106-343-1

Printed by: Universitetsservice US-AB, Sweden 2025

Abstract

Until recently, most robotic systems have operated with limited emotional intelligence, primarily responding to pre-programmed cues rather than adapting to human emotional states. Thus, affect recognition in human-robot-interaction remains a significant challenge and is twofold: robots must not only detect emotional expressions but they also need to interpret them within their social context, requiring systems that are capable of collecting information from its surrounding, analyzing it and thereafter generalizing across different interaction scenarios and cultural contexts to handle more complex situations.

This thesis tackles affect recognition using multi-modal approaches that combine thermal imaging, facial expression analysis, and contextual understanding. Thermal imaging offers unique insights into physiological responses associated with emotional states, complementing traditional vision-based approaches while maintaining non-contact operation. The integration of thermal imaging, facial expression analysis, and contextual understanding creates a comprehensive multi-modal framework that addresses the key challenges in affect recognition, such as varying lighting conditions, occlusions, and ambiguous emotional expressions. This combination provides complementary information streams that enhance robustness in real-world environments, making it an effective case study for developing context-aware emotional intelligence in robotics.

We introduce a novel context-aware transformer architecture that processes multiple data streams while maintaining temporal relationships and contextual understanding. Each modality contributes complementary information about the user's emotional state, while the context processing ensures situation-appropriate interpretation. For instance, distinguishing between a smile indicating enjoyment during collaborative tasks versus one masking nervousness in stressful situations. This contextual awareness is crucial for appropriate robot responses in real-world deployments.

The research contributions span four areas: (1) developing robust thermal feature extraction techniques that capture subtle emotional responses (2) creating a transformer-based architecture for multi-modal fusion that effectively incorporates situational information, (3) implementing real-time processing pipelines that enable practical deployment in human-robot interaction scenarios, and (4) validating these approaches through extensive real-world interaction studies. Results show improved recognition accuracy from 77% using traditional approaches to 89% with our context-aware multi-modal system, demonstrating the ability to understand and appropriately respond to human emotions in dynamic social situations.

Sammanfattning

De flesta robotsystem har tidigare opererat med begränsad emotionell intelligens, där de främst reagerat på förprogrammerade signaler snarare än att anpassa sig till människors olika affektiva tillstånd. Detta har lett till mindre naturliga och effektiva interaktioner. Därav, affektiv igenkänning inom människa-robot interaktion förblir en stor utmaning och är tvåfaldig: robotar måste inte bara detektera känslouttryck utan även kunna förstå och tolka dem inom deras sociala kontext, vilket kräver system som kan samla in information från sin omgivning, analysera den och därefter generalisera över olika interaktions scenarier och kulturella sammanhang för att hantera mer komplexa situationer.

Denna avhandling syftar till att utforska multimodala metoder som kombinerar termisk avbildning, ansiktsuttryck, analys och kontextuell förståelse för att förbättra interaktionen mellan människor och robotar. Termisk avbildning ger insikter i fysiologiska responser associerade med känslotillstånd, vilket kompletterar traditionella visions baserade metoder samtidigt som beröringsfri operation bibehålls. Kombinationen av dessa modaliteter innefattar många av de generella utmaningarna inom affektiv igenkänning, vilket gör dem till effektiva fallstudier för utveckling av kontext medveten emotionell intelligens inom robotiken.

Genom detta arbete, introducerar vi en ny kontext medveten transformerarkitektur som behandlar olika typer av data samtidigt som temporala relationer och kontextuell förståelse upprätthålls. Varje modalitet bidrar med komplementerande information om användarens känslotillstånd, medan kontext bearbetningen säkerställer situationsanpassad tolkning. Systemet kan exempelvis skilja mellan ett leende som indikerar nöje under samarbetsuppgifter och ett leende som maskerar obehag i stressiga situationer. Denna kontextmedvetenhet är avgörande för lämpliga robot responser i verkliga tillämpningar.

Denna forskning bidrar till (1) utvecklingen av robusta termiska extraktions tekniker som fångar subtila känslomässiga responser med förmågan till miljöanpassning, (2) skapandet av en transformer baserad arkitektur för multimodal fusion som integrerar situationell information på ett effektivt sätt, (3) implementeringen av realtids bearbetning som möjliggör praktiska tillämpningar i människa-robot interaktion, och till sist (4) valideringen av dessa metoder genom omfattande interaktionsstudier i verkliga miljöer. Resultaten visar en förbättrad igenkännings noggrannhet från 77% med traditionella metoder till 89% med vårt kontext medvetna multimodala system, vilket demonstrerar betydande framsteg i robotars förmåga att förstå och till följd att lämpligt reagera på mänskliga känslor i dynamiska sociala situationer.

Acknowledgments

As I reflect on the journey that has led to this thesis, I am reminded of Mahmoud Darwish, the Palestinian poet's profound words:

وأنتَ تَصعدُ السّلّمَ، فكّرْ في مَنْ بناهُ مِنْ حَجَرِ التّعبِ.

محمود درويش

"As you climb the ladder, think of those who built it from the stone of fatigue." These words resonate deeply with me as I acknowledge that this academic achievement would not have been possible without the countless individuals who laid the foundation for my work, supported me through challenges, and contributed their knowledge, time, and care to my journey.

علَى هذِهِ الأَرضِ مَا يَسْتَحِقُّ الحَيَاةْ

محمود درويش

Darwish's verse—"On this land, there is what makes life worth living"—serves as a reminder to carry forward hope, even in moments of exhaustion. As I complete this chapter of my life, I hold onto the belief that this work, and the spirit behind it, can contribute to something greater than itself. This milestone is not an ending but a continuation—an invitation to imagine, to give back, and to build toward futures filled with meaning, justice, and shared purpose.

List of Papers

This thesis is based on the following papers:

A: **Youssef Mohamed**, Giulia Ballardini, Maria Teresa Parreira, Séverin Lemaignan, Iolanda Leite. *Automatic Frustration Detection Using Thermal Imaging. IEEE/RSJ International Conference on Human-Robot Interaction (HRI), 2022, pp. 451-459.*

B: **Youssef Mohamed**, Arzu Güneysu, Séverin Lemaignan, Iolanda Leite. *Multi-modal Affect Detection Using Thermal and Optical Imaging in a Gamified Robotic Exercise. International Journal of Social Robotics, 2023, pp. 981-997.*

C: **Youssef Mohamed**, Séverin Lemaignan, Arzu Güneysu, Patric Jensfelt, Christian Smith. *Context Matters: Understanding Socially Appropriate Affective Responses Via Sentence Embeddings. International Conference on Social Robotics, 2024.*

D: **Youssef Mohamed**, Séverin Lemaignan, Arzu Güneysu, Patric Jensfelt, Christian Smith. *Fusion in Context: A Multimodal Approach to Affective State Recognition. IEEE International Conference on Robot and Human Interactive Communication, 2025*

E: **Youssef Mohamed**, Séverin Lemaignan, Arzu Güneysu, Patric Jensfelt, Christian Smith. *Are You an Expert? Instruction Adaptation Using Multi-Modal Affect Detection with Thermal Imaging and Context. IEEE International Conference on Robot and Human Interactive Communication, 2025*

The author has contributed to the following article as well, although it is not included in this thesis:

F: Séverin Lemaignan, Antonio Andriella, Lorenzo Ferrini, Luka Juricic, **Youssef Mohamed**, Raquel Ros. *Social Embeddings: Concept and Initial Investigation*. In Open Research Europe, 2024

Contents

Part I

Overview

Chapter 1

Introduction

Over the past couple of decades, the field of Affective Computing in Robotics has gained momentum, striving to create robots that can perceive and respond to human emotions. Despite this, most current robotic systems operate with limited emotional intelligence, primarily responding to predefined inputs without understanding the subtle nuances of human emotional states [1] leading to systems that excel in controlled environments performing repetitive tasks, but when faced with more complex situations and environments they struggle to adapt to the nuanced emotional dynamics of human interaction. Therefore, as robots begin to spread into sectors like healthcare and education, the need for emotional intelligence has become even more important which has created new challenges in human-robot interaction.

Indeed, the ability to recognize and respond to human emotions is particularly important since many human-robot interaction tasks require understanding of emotional states. In healthcare settings, therapeutic robots need to identify patient distress and discomfort, yet current systems often miss these critical signs, reducing the effectiveness of treatment [2]. Similarly, in education, robots must adapt to varying levels of student engagement, but current approaches struggle to maintain consistent interaction due to their inability to detect subtle emotional cues [3]. Even in public spaces, where robots encounter people in diverse emotional states, the challenge is magnified by the considerable variation in how emotions are expressed across cultures and individuals [4]. These limitations stem from the inherent complexity of emotional expression and the context-dependent nature of appropriate responses. What might be a suitable reaction in one scenario could be inappropriate in another, making traditional, rigid approaches to affect recognition insufficient for real-world applications.

The challenge is further compounded by the technical limitations of current sensing approaches. Vision-based systems, while widely used, struggle with varying lighting conditions and occlusions common in real-world environments [5]. Physiological sensors like EEG [6] and GSR [7], though capable of providing detailed

emotional data, require physical contact with users, making them impractical for natural human-robot interaction. These technical constraints, combined with the need for real-time processing, create significant barriers to effective emotional recognition in robotic systems.

Often, when faced with the challenge of emotion recognition, researchers have focused on single-modality approaches or controlled laboratory conditions [4]. While this leads to promising results in specific scenarios, it fails to address the complexity of real-world emotional expression in comparison to humans who can naturally interpret emotions across various contexts and situations. One goal of this research is therefore to create multi-modal affect recognition systems capable of understanding affect with accuracy approaching that of human perception. Contrary to traditional single-modality approaches intended for controlled environments, these systems must operate robustly in dynamic, real-world conditions while maintaining natural interaction capabilities [8].

1.1 The Critical Role of Context in Affect Understanding

Context plays a fundamental role in affect recognition systems, where physiological signals and facial expressions exhibit significant variability based on environmental, social, and cultural factors [9–11]. Traditional affect recognition systems often assume simplified contexts that don't reflect real-world complexity, similar to how early industrial robots were limited to controlled environments. Our research demonstrates this through thermal imaging data, where facial temperature patterns can indicate different emotional states depending on the setting - the same pattern might suggest cognitive load during an overwhelming task but emotional arousal while failing to win in a simple game [12–14].

While EEG has been the popular choice for detecting cognitive and emotional states [6], our comparative study has shown that thermal imaging can achieve comparable accuracy while being less intrusive (See Section 2.2). By combining thermal imaging with facial expression analysis, we managed to create a system that can perform well in varied environments even outside the lab, where the traditional usage of facial expressions analysis using RGB cameras struggles, such as varying lighting conditions or partial occlusions. This multimodal approach, enhanced by recent developments in large language models [15], enables a more nuanced interpretation of emotional states across different cultural and social settings, addressing the contextual challenges in both professional and social environments.

1.2 Thesis Motivation and Approach

The limitations of current affect recognition systems, combined with the growing deployment of social robots, create a pressing need for more reliable and context-aware approaches. Our work is motivated by the hypothesis that no single modality can capture the full complexity of human emotional expression. Instead, we pro-

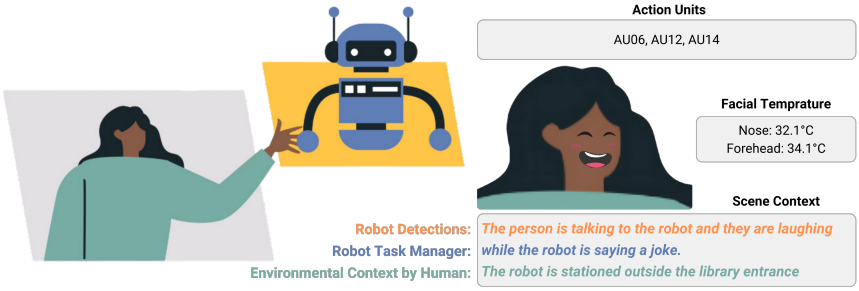


Figure 1.1: Overview of our proposed multi-modal affect recognition system combining thermal imaging, RGB analysis, and contextual information. The transformer-based architecture processes multiple data streams while maintaining temporal relationships and contextual understanding. Each modality contributes complementary information about the user's emotional state, while the context processing ensures situation-appropriate interpretation.

pose a multi-modal approach that combines thermal imaging, traditional vision-based analysis, and explicit context processing. We selected thermal imaging over alternatives like galvanic skin response (GSR) or EEG because it provides non-contact physiological measurements with comparable accuracy, enabling natural interactions while capturing autonomic nervous system responses through facial temperature variations - a critical advantage for real-world human-robot interaction scenarios where attaching sensors would disrupt natural behavior.

We extend the thermal imaging approach introduced by Abd Elrahman et al. [16] to capture physiological signals non-intrusively. This method detects subtle facial temperature variations that correlate with emotional states [17]. By integrating thermal data with RGB facial analysis and context processing, our system achieves more robust affect recognition than previously possible [18].

1.3 Scientific Contributions

This thesis presents four interconnected contributions to affect recognition in human-robot interaction, advancing from sensing methodologies to practical applications. Table 1.1 illustrates how these contributions build upon each other across Papers A through E.

Thermal Imaging for Non-Contact Affect Detection

Paper A establishes thermal imaging as a method for detecting human frustration in real-time interaction scenarios. The approach analyzes temperature patterns in specific facial regions (nose, forehead, and lower lip) to detect affective states [12, 19]. The contribution includes a region-of-interest tracking system that functions despite moderate head movements, developed through systematic feature selection to

Table 1.1: Contributions of Papers (A-E) to Affective Computing Challenges. This table maps how each included paper addresses specific research challenges across five domains. Papers with primary contributions and are highly central to the work (• • •) introduce novel solutions, while others have medium centrality and are implemented in the work but were not the main focus (• •) or mentioned as a concept from previous work (•).

Research Challenge	Specific Challenge	A	B	C	D	E
1. Non-intrusive Affect Recognition	Thermal imaging for detecting affective states	• • •	• •	•	•	•
	Differentiating cognitive and failure-induced frustration	• • •	• •			
	Identifying optimal facial regions for thermal analysis	• • •	• •		•	
2. Multi-modal Data Integration	Integration of thermal and action unit data		• • •		• •	
	Capturing temporal dynamics of affective states		• • •		• •	• •
	Handling heterogeneous modality characteristics		• • •		• •	
3. Contextual Understanding	Incorporating social context in affect recognition			• • •	• •	•
	Adapting to different environmental contexts			• • •	• •	•
	Measuring appropriateness of affective responses			• • •	• •	
4. Cross-modal Analysis	Modality-specific feature encoding				• • •	
5. Human-centered Application	Real-time affect detection in interaction scenarios		• •	• •		• • •
	Adaptive difficulty in gamified interactions		• •			• • •
	Testing in naturalistic environments			• •		• • •

identify the most informative facial regions. This approach addresses limitations of traditional physiological sensing methods by providing affect recognition without physical contact, making it suitable for naturalistic human-robot interaction scenarios where attaching sensors would disrupt normal behavior. The methodology is elaborated in Section 3.1 and fully detailed in Paper A.

Multi-Modal Temporal Integration

Paper B extends the foundation established in Paper A beyond frustration detection to capture four distinct affective states. By combining thermal imaging with facial expression analysis through LSTM networks, the system aims to distinguish between frustration, boredom, enjoyment, and neutral states [20, 21]. The paper introduces an LSTM-based architecture that models how emotions evolve over time, critical for interpreting emotional transitions that conventional single-frame methods miss. Based on ablation studies, Paper B identifies an optimal analysis window that balances temporal resolution with recognition accuracy, allowing sufficient time for thermal patterns to develop while maintaining system responsiveness. This multi-modal approach leverages complementary information from thermal and visual streams to improve robustness in varied environmental conditions, as detailed in Section 3.1.

Context-Aware Affect Recognition

Papers C and D address the context-dependent nature of emotional expressions by integrating contextual understanding into the affect recognition process. Paper C introduces a novel reference point technique using embedding spaces to compare social situations, while Paper D develops a transformer-based architecture that combines thermal features, facial expressions, and contextual information. This architecture employs modality-specific encoders that preserve the unique characteristics of each data stream, along with cross-modal attention mechanisms that enable bidirectional information flow between modalities. The context-aware approach aims to enhance disambiguation of similar expressions carrying different meanings across situations. Paper C demonstrates the importance of contextual understanding, while Paper D implements and validates the integrated architecture, with theoretical foundations provided in Section 3.2.

Validation in Human-Robot Interaction

Paper E validates the complete system through application in instruction delivery scenarios. The study examines how affect-based adaptation affects interaction quality and user experience in a practical human-robot task. The research investigates whether the system can reduce the need for manual interventions while maintaining high user satisfaction. Paper E explores how affect-based systems can implement

automatic level changes to create longer periods of stable interaction without sacrificing user experience. This validation aims to confirm the system's effectiveness in practical human-robot interaction, showing how affect recognition can support meaningful adaptation while preserving user agency. The broader implications of these findings for human-robot interaction are discussed in Section 3.3.

These five contributions collectively advance affect recognition from isolated sensing methodologies to integrated systems capable of performing in real-world interaction scenarios. The progression from Paper A through Paper E demonstrates how combining thermal imaging, multi-modal integration, and contextual understanding creates systems that can recognize and respond appropriately to human emotions in dynamic environments.

1.4 Delimitation and Scope

While this thesis makes contributions to multi-modal affect recognition, several important aspects remain outside its scope. Here we explicitly define the boundaries of our research to maintain focus on our core contributions.

The first limitation that has been set for our investigation is the number of specific affective states that we looked into, going into depth in only four: frustration, boredom, enjoyment, and neutral. While humans express a broader spectrum of emotions, these states were selected based on their prevalence and importance in human-robot interactions, as established in Papers A-C. This choice was also motivated by the availability of reliable ground truth data for these states, the ability to stimulate those states through HRI tasks, and their particular relevance to adaptive social robotics applications.

In addition, our work focuses exclusively on dyadic interactions between a human and a robot. While Paper E applies our system in collaborative tasks, it is limited to a one-on-one scenario. Group interactions, would introduce additional complexity in terms of multiple simultaneous affect recognition and inter-personal dynamics. This limitation allows us to thoroughly investigate the core challenges of context-aware affect recognition without the added complexity of multi-party interactions.

While we develop approaches for affect recognition using thermal imaging and traditional vision-based analysis, we do not attempt to create a comprehensive model of all possible contextual factors, but rather considering the immediate physical environment and social situation, as shown in Papers C and D. Other potentially relevant contextual elements, such as long-term interaction history or broader social relationships, are not addressed in this work.

Paper E shows the potential for using affect recognition in adaptive robot behavior, nonetheless, the exploration of different emotion regulation strategies and adaptations are not assessed in this work. Our research focuses on the challenge of accurate affect recognition rather than the subsequent problems of emotional response generation and behavior adaptation. The adaptive behaviors shown in our

experiments serve primarily to validate the practical utility of our affect recognition system.

Cultural effects on both facial expressions and thermal patterns, though important, would require a separate large-scale cross-cultural study beyond the scope of this work. Despite the fact that our data collection included participants from diverse cultural backgrounds, we do not explicitly model or analyze cultural variations in emotional expression. Our primary goal was to develop robust recognition methods that could work effectively across our participant population without explicitly accounting for cultural differences.

These limitations reflect deliberate choices to maintain a focused research agenda while thoroughly addressing the core challenges in multi-modal, context-aware affect recognition. Future work may build upon our foundations to address these additional dimensions of the problem.

1.5 Thesis Organization

The goal of this thesis is to advance affect recognition through thermal imaging and contextual understanding. The thesis is organized into six chapters, each addressing different aspects of the research:

Chapter 1 provides an introduction to the field of affective computing in robotics, highlighting the need for emotion recognition in human-robot interaction. It presents the motivation behind using multi-modal approaches, emphasizes the critical role of context in affect understanding, and outlines the scientific contributions and scope of the research.

Chapter 2 establishes the theoretical and technical foundations, starting with a historical overview of affect recognition. It examines various sensing modalities including thermal imaging, vision-based methods, and audio analysis, while discussing multi-modal integration challenges and solutions. This background information provides the necessary context for understanding the technical approaches developed in the thesis.

Chapter 3 presents the development and implementation of our multi-modal context-aware affect recognition system based on my five research papers (Papers A-E). This chapter details the methodological approaches and achievements of my research through three key aspects: (1) the multi-modal integration techniques I developed in Papers A and B that effectively combine thermal and visual data for improved affect detection, (2) the context understanding approaches introduced in Papers C and D that leverage embedding models to enhance social awareness, and (3) the impact of these technologies on human-robot interaction as validated in Paper E. This chapter connects the theoretical foundations to my practical implementations, demonstrating how each paper contributes to the overall system architecture and capabilities.

Chapter 4 presents a critical discussion of the research, focusing on both technical limitations (such as head pose tracking challenges and computational con-

straints) and ethical implications (including privacy concerns, representation issues, and questions of user agency).

Chapter 5 summarizes the five research papers that form the core contributions of the thesis. Each summary highlights the key innovations, methodologies, and findings, showing the progression from thermal imaging for frustration detection (Paper A) through multi-modal fusion (Paper B), context-aware affect recognition (Papers C and D), to practical application in instruction delivery (Paper E).

Chapter 6 concludes the thesis by synthesizing the empirical findings, discussing insights on human-robot collaboration, and identifying promising directions for future research. It reflects on how the contributions advance the field while acknowledging remaining challenges in creating affect-aware robotic systems that balance automation with human agency.

Chapter 2

Background

2.1 Historical Development and Evolution of Affect Recognition

The study of human emotions and their recognition has a rich history spanning over a century of research across multiple disciplines. This section traces the evolution of affect recognition from early theoretical foundations to modern computational approaches, highlighting key developments that have shaped our current understanding of affect detection and recognition systems.

Theoretical Foundations of Emotion Research

Early Evolutionary Perspectives

The scientific study of emotions began with Darwin's seminal work "The Expression of Emotions in Man and Animals" [22], which suggested that emotional expressions served an evolutionary purpose and were universal across cultures. This universality hypothesis laid the groundwork for modern affect recognition by suggesting that certain basic emotional expressions could be consistently identified regardless of cultural background. Building on Darwin's work, researchers in the mid-20th century started developing more systematic approaches to studying emotions.

Categorical Models of Emotion

Paul Ekman's research in the 1970s established the concept of basic emotions, identifying six emotional states: happiness, sadness, fear, disgust, surprise, and anger [23]. This categorical approach to emotions dominated the field for decades and continues to influence modern affect recognition systems. Ekman's work was important for developing the Facial Action Coding System (FACS) [24], which provided a standardized method for describing facial movements associated with emotional expressions.

Dimensional Approaches

The categorical approach to emotions faced criticism for being too simplistic and failing to capture the complexity of human emotional experiences. This led to the development of dimensional models of emotion, notably Russell's circumplex model [25], which proposed that emotions could be better understood along continuous dimensions of valence (pleasure-displeasure) and arousal (activation-deactivation). This shift from discrete categories to continuous dimensions opened up endless possibilities in affective computing, and how affect could be detected. But it also created a more challenging problem as there was no one definition of what affect looked like.

Computational Recognition of Affect

Early Automated Systems

The 1990s marked a major shift as researchers began applying computer vision and machine learning to transform theoretical models of emotion into automated recognition systems. The initial systems focused on geometric feature analysis, tracking facial landmarks to identify basic emotional states [26].

Advancement came in the late 1990s through appearance-based methods. Techniques such as Gabor wavelets [27] and Local Binary Patterns (LBP) [28] enabled analysis of facial texture patterns, providing more robust recognition compared to geometric approaches alone.

The early 2000s saw the integration of machine learning algorithms, particularly Support Vector Machines and Neural Networks, which enabled more sophisticated classification of emotional states. While these systems demonstrated high accuracy in laboratory conditions, they highlighted the persistent challenge of real-world generalization [29].

Deep Learning Revolution

Deep learning reshaped the landscape of affect recognition in the 2010s. In visual analysis, Convolutional Neural Networks achieved improvements in facial expression recognition [30]. The development of frameworks like OpenPose [31] and DensePose [32] expanded these capabilities by tracking body movements in both 2D and 3D space, offering new angle into how emotions manifest through physical expression.

The field also made progress in analyzing how emotions change over time. Recurrent Neural Networks (RNNs) improved the analysis of speech and video sequences [33], leading to better understanding of emotional dynamics. The introduction of transformer architectures brought further advances. In text analysis, models like BERT [34] and GPT [35] demonstrated sophisticated parsing of emotional content, while Vision Transformers [36] pushed forward the capabilities of visual emotion recognition.

Multi-modal Fusion Techniques

When combining multiple modalities, consideration must be given to the fusion strategy. Fusion techniques have evolved over time in multimodal affect recognition. Early fusion approaches primarily concatenated features at the input level, but these have been largely superseded by more sophisticated late fusion techniques that combine high-level representations [5]. As the field advanced, researchers explored attention-based fusion mechanisms that can dynamically weight different modalities based on their reliability and relevance to the current context [37, 38].

Data and Computational Challenges

A key challenge in deploying deep learning approaches for affect recognition remains the need for large annotated datasets [39]. While supervised learning has shown impressive results in controlled settings, there is growing interest in unsupervised and self-supervised approaches that can leverage unlabeled data [40]. Real-time processing requirements present additional challenges for affect recognition systems, driving research into optimized architectures [41] and model compression techniques [42].

Practical Challenges

Real-world Application Challenges

Affect recognition systems still face numerous challenges in real-world applications. These include:

1. **Individual Differences:** People express emotions differently based on personality, culture, and social context [43].

2. **Environmental Variability:** Real-world conditions introduce challenges like varying lighting, occlusions, and head poses [44].

3. **Temporal Dynamics:** Emotions evolve over time, making single-frame analysis insufficient [45].

4. **Context Dependency:** The same expression can have different meanings in different contexts [4].

Emerging Trends

Recent research has increasingly focused on multi-modal approaches that combine different types of information (visual, audio, physiological) to improve recognition accuracy [46]. There is also growing interest in considering contextual information and social situations when interpreting emotional expressions [4].

The integration of thermal imaging has emerged as a promising direction for affect recognition, offering advantages in detecting physiological responses associated with emotional states [47]. This non-contact approach provides valuable complementary information to traditional visual analysis, particularly in challenging lighting conditions or when subtle emotional changes need to be detected.

Hence, there is increasing recognition of the need for more nuanced approaches that can handle the complexity and variability of human emotions in real-world settings. This includes developing systems that can adapt to individual differences, account for cultural variations, and integrate contextual information effectively.

Sensing Modalities in Affect Recognition

Overview of Sensing Approaches

The evolution of affect recognition technology has progressed through multiple sensing modalities, each with distinct capabilities and limitations. Traditional approaches can be broadly categorized into three main streams: contact-based physiological sensing, vision-based methods, and audio/speech analysis. Physiological sensing systems gather emotional response data through direct contact sensors placed on the body, providing precise measurement of internal changes but limited by their intrusive nature. Vision-based approaches analyze facial expressions and body language through optical cameras, offering non-contact observation but susceptible to environmental conditions. Audio analysis examines speech patterns and vocal characteristics to detect emotional states, complementing visual information but constrained by linguistic variations. Each of these traditional approaches offers valuable insights while presenting specific challenges for practical affect recognition applications.

Contact-based Physiological Sensing

Brain activity measurements through EEG have provided valuable insights into the neural basis of emotions [48]. Different emotional states produce characteristic patterns across frequency bands that can be detected in EEG recordings [49]. The requirement for precise electrode placement on the scalp and high sensitivity to movement artifacts restrict its applicability outside controlled laboratory settings, which goes against the need for more real world settings.

Galvanic Skin Response (GSR) or Electrodermal Activity (EDA) measures changes in skin conductance associated with emotional arousal [50]. GSR has proven effective in detecting stress and anxiety levels, though it can be influenced by various non-emotional factors such as physical activity and environmental conditions.

Electromyography (EMG) measures muscle activity and has been useful in detecting subtle facial expressions [51]. Surface EMG sensors placed on facial muscles can detect activation patterns associated with different emotional expressions, even when these expressions are not visibly apparent. Even though this can advance our

understanding of facial movement, it still cannot be used in a natural setting and is considered intrusive.

Vision-based Methods

The field's progression mirrors broader advances in computer vision, with each new development bringing us closer to human-level perception capabilities.

The introduction of standardized frameworks, particularly the Facial Action Coding System (FACS), marked a milestone in the field. This systematic approach to describing facial movements provided researchers with a common language for emotion analysis. Building on this foundation, Bartlett et al. [52] demonstrated that machine learning algorithms could effectively detect action units in spontaneous behavior, achieving 93% accuracy across 20 distinct units. Modern implementations like OpenFace [5] have further refined these capabilities, offering toolkits that combine action unit detection with precise tracking of facial landmarks and gaze direction.

In addition, gaze tracking tools has evolved, moving away from hardware-dependent solutions toward more flexible appearance-based approaches [53]. Recent CNN architectures have pushed the boundaries of what's possible, with some systems achieving precision - error rates as low as 3.64 degrees in unconstrained environments [54]. This level of accuracy opens new possibilities for understanding attention and engagement during human-robot interactions.

The emergence of holistic frameworks that simultaneously track multiple bodily features represents another significant advancement. Systems like MediaPipe have increased our ability to capture the full range of physical expression, from subtle changes in head pose to complex hand gestures [55]. Yang et al. [56] demonstrated how these motion features could be effectively processed using hybrid architectures, achieving superior recognition rates while reducing data requirements compared to pure end-to-end approaches.

Despite these advances, vision-based systems still face fundamental challenges. Variations in lighting, partial occlusions, and diverse demographic representations continue to impact system reliability. The field has begun exploring self-supervised learning as a potential solution to the data scarcity problem, though questions remain about the generalizability of these approaches across different cultural contexts and interaction scenarios.

Audio and Speech Analysis

The landscape of speech-based affect recognition has shifted dramatically with the rise of deep learning. Moving beyond traditional analyses of pitch and speaking rate [57], modern approaches capture intricate patterns in voice quality and spectral characteristics that reveal emotional states [58].

Tools like OpenSmile [59] have transformed the field by standardizing feature extraction from speech signals. The platform's ability to generate rich statistical

features has made it invaluable for researchers, though the interpretation of these paralinguistic cues remains complex [60]. The field's progress owes much to initiatives like the Interspeech Computational Paralinguistics Challenge (ComParE), which has spurred innovations in analysis techniques.

Hybrid architectures combining RNNs with LSTM networks have shown particular promise in understanding emotional speech patterns. In a compelling demonstration, Tammewar et al. [61] achieved 63% accuracy in valence prediction using the Ulm State of Mind dataset. Schuller et al. [62] built on this foundation by integrating multiple deep learning toolkits, though their results - hovering around 50% accuracy for three-class tasks - underscore the field's current limitations.

The latest research explores end-to-end deep learning systems that process raw audio directly. While this approach could uncover subtle emotional indicators missed by traditional methods, it demands substantial computational resources and large datasets [63]. The scarcity of labeled emotional speech data remains a key obstacle to progress.

Speech analysis brings unique value to multi-modal affect recognition, offering crucial insights when facial cues prove unreliable. Yet the temporal nature of speech presents distinct challenges for system integration, particularly in synchronizing with other modalities. Success requires careful consideration of how speech patterns evolve alongside other emotional indicators.

Multi-modal Integration and Fusion

Benefits of Multi-modal Approaches

Recognition of human affect often benefits from the integration of multiple modalities, as different sensing channels can provide complementary information about emotional states, where these systems have proved to have higher recognition accuracy by leveraging complementary information from the different modalities [46]. For example, facial expressions alone might indicate happiness while physiological signals could reveal underlying stress, providing a more complete picture of the emotional state.

Different modalities also offer varying levels of robustness to different types of noise and interference. In the case that one modality is compromised (e.g., poor lighting affecting visual analysis), other modalities can help maintain system performance [5]. This complementary nature of different modalities has been evident in real-world deployments, where environmental conditions can impact individual sensing channels.

Fusion Strategies

Multimodal fusion strategies can be broadly categorized into three main approaches: early fusion (feature-level), late fusion (decision-level), and hybrid/model-level fusion, each with distinct characteristics and applications. Early fusion directly

merges features from different modalities before performing classification [64]. This approach enables learning rich joint representations but introduces vulnerability to noisy or missing sensor data. Building on these observations, Psaltis et al. [37] demonstrated how attention mechanisms could dynamically adjust the influence of each modality, improving robustness while preserving the benefits of early fusion.

Decision-level fusion, in contrast, processes each modality independently and combines their outputs [65]. This late fusion approach offers greater robustness to missing data and allows for modality-specific processing optimizations. Alonso-Martin et al. [38] showed how decision-level fusion could effectively combine facial expressions, speech, and body posture for more robust emotion detection in human-robot interactions.

Model-level fusion represents a middle ground, using specialized architectures to learn joint representations while maintaining some degree of modality independence [66]. These approaches often employ parallel processing streams with cross-modal attention mechanisms to capture interactions between modalities. Recent advances in transformer architectures have enabled more sophisticated fusion approaches that can dynamically weight different modalities based on their reliability and relevance [67].

Real-time Processing Challenges

The integration of multiple modalities for affect recognition introduces processing challenges in maintaining temporal synchronization and real-time performance. Our experimental work showed that thermal imaging and facial expressions operate at different timescales - facial expressions can change within 1-3 seconds while thermal patterns evolve over several seconds [12]. As shown by Chen et al. [68], these temporal mismatches create practical constraints on system responsiveness. Their speech-based cognitive load measurements required 2-3 seconds of active speech, while linguistic analysis demanded complete sentences or phrases, highlighting how different modalities impose varying temporal requirements.

The processing pipeline in multi-modal affect recognition systems must handle the different temporal dynamics across modalities, with facial expressions changing rapidly while thermal patterns evolve more slowly. Sliding window approaches have been proposed in the literature to address this challenge, typically processing several seconds of data with partial overlap between consecutive windows [68]. This technique balances the need for thermal pattern development against overall system responsiveness. Chen et al. [68] demonstrated in their work on cognitive load assessment that synchronizing multiple behavioral indices through temporal windows can achieve good performance, despite the inherent differences in timescales between various sensing modalities.

Missing or corrupted data presents another challenge in real-world deployments [65]. Environmental factors can degrade sensor data quality - changing lighting conditions affect RGB cameras while air currents can impact thermal measurements [65]. That is why the aspects discussed in 2.1 are essential in solving this

issue, as multi-modal systems can operate by relying on other sources of data if one data stream is corrupted.

The computational demands of processing multiple synchronized data streams create additional constraints for real-world operation [69]. Each modality requires its own processing pipeline - thermal feature extraction, facial expression analysis, and context embedding generation all compete for computational resources.

These challenges highlight the importance of considering temporal dynamics and processing efficiency in multi-modal affect recognition system design. In addition, the integration of contextual understanding further improves system performance but requires careful consideration of processing overhead and temporal alignment [69].

2.2 Thermal Imaging for Affect Detection

We propose thermal imaging for affect detection as a promising non-contact approach to measuring physiological responses associated with emotional states. This section examines the physiological basis for our thermal imaging approach, its advantages compared to traditional sensing methods, and key technical considerations for implementation.

Physiological Basis

The relationship between emotional states and temperature changes in specific facial regions is rooted in the activity of the autonomic nervous system (ANS). When a person experiences different emotional states, the ANS triggers changes in blood flow patterns throughout the body, including facial blood vessels [70]. These blood flow changes manifest as subtle but detectable temperature variations that can be measured through thermal imaging [71].

The sympathetic division of the ANS plays an important role in emotional responses, causing vasoconstriction or vasodilation of blood vessels, which directly affects skin temperature [7]. For example, during states of increased arousal or stress, vasoconstriction in peripheral blood vessels can lead to decreased skin temperature in regions like the nose and cheeks, while increased muscle activation may cause temperature increases in other facial regions [72].

Through extensive research examining thermal responses during emotional states, several key facial regions of interest (ROIs) have been identified as reliable indicators of affect: the nose region, forehead, periorbital region, and maxillary area [73]. Each region exhibits characteristic temperature patterns:

- The nose region typically shows temperature decreases during states of heightened arousal or stress due to vasoconstriction
- The forehead temperature often increases during cognitive load and emotional activation due to increased blood flow

- The periorbital area shows rapid temperature changes linked to sympathetic nervous system activation
- The maxillary (cheek) region can reflect emotional valence through temperature variations

Advantages over Traditional Methods

Thermal imaging offers several key advantages compared to conventional physiological sensing approaches used in affect detection:

Comparison with EEG

While EEG provides direct measurement of neural activity and has been widely used for emotion recognition, it requires complex setup procedures and is highly sensitive to motion artifacts [74].

Benefits over GSR

Galvanic skin response has traditionally been used to measure emotional arousal through changes in skin conductance [75]. However, GSR requires direct skin contact through electrodes which can interfere with natural interaction. Thermal imaging provides similar insights into autonomic nervous system activation without any physical contact. Additionally, GSR measurements can be affected by environmental factors like temperature and humidity, while thermal imaging remains relatively robust to these variations when properly calibrated.

Advantages versus EMG

Electromyography effectively captures muscle activation patterns associated with facial expressions but requires precise sensor placement on the face [76]. Thermal imaging can detect similar patterns of facial muscle activation through temperature changes while allowing completely unrestricted facial movement. This is important for natural human-robot interaction scenarios where participants need to speak and express themselves freely.

Environmental Robustness

Unlike RGB cameras which depend heavily on ambient lighting conditions, thermal cameras can operate effectively in various lighting environments since they detect infrared radiation emitted by the subject [77]. This makes thermal imaging suitable for real-world applications where lighting conditions may be variable or sub-optimal.

Technical Considerations

Thermal imaging provides a non-invasive method for capturing physiological signals relevant to affect detection. To effectively translate these thermal measurements into meaningful affect predictions, our pipeline consists of distinct yet interconnected stages. This pipeline begins with image acquisition through specialized hardware, followed by signal processing and feature extraction. Each stage presents unique requirements and challenges that must be addressed for successful affect detection.

Hardware Requirements

The first component in our thermal imaging pipeline is the acquisition hardware. Accurate affect detection requires thermal cameras with specific capabilities. Based on previous research [70], the camera must have a spatial resolution of at least 320×240 pixels to capture sufficient facial detail for region-of-interest analysis. Temporal dynamics of facial temperature variations necessitate a minimum frame rate of 10Hz, ensuring that rapid emotional responses are not missed. Additionally, the measurement range must span 20-40°C to accommodate baseline facial temperatures while capturing subtle emotion-related temperature variations. Proper calibration of these devices is essential to ensure measurement accuracy across varied environmental conditions and subject characteristics.

2.3 Context in Affect Recognition

Traditional affect recognition systems have focused primarily on analyzing facial expressions or physiological signals in isolation, but research has proved that emotional expressions must be interpreted within their broader social and environmental context [4]. This section examines the theoretical foundations and computational approaches for incorporating contextual information in affect recognition systems, as well as the implementation challenges that arise in real-world applications.

Theoretical Foundations

Social context plays a crucial role in interpreting emotional expressions. As shown by Hoang et al. [13], identical physiological responses can signal different emotional states - for instance, elevated heart rate may indicate either excitement or anxiety depending on the situation. Barrett et al. [4] demonstrate that emotion perception emerges from the integration of sensory inputs with contextual knowledge, rather than from sensory data alone.

Cultural influences play a crucial role in both the expression and interpretation of emotions. Studies have shown variations in emotional display rules across cultures [43], affecting both the intensity and manner of emotional expression. These cultural differences extend to the interpretation of emotional signals, with the same

expression potentially carrying different meanings across cultural contexts [78]. For instance, research has proved that Eastern and Western cultures differ systematically in their interpretation of certain facial expressions [43].

Situational factors further modulate emotional expression and interpretation [79]. Professional settings typically require more restrained emotional displays compared to casual social interactions. Environmental conditions, such as the presence of others or the physical setting, can also influence both the expression and perception of emotions. Recent work has highlighted how these situational factors can lead to systematic variations in emotional expression that must be accounted for in affect recognition systems [14].

The temporal aspect of context adds another layer of complexity [45]. Emotions evolve over time, and their interpretation often depends on understanding the sequence of events that preceded them. This temporal context is important in social interactions, where the history of the interaction can influence the interpretation of current emotional displays. According to research in affective computing considering temporal dynamics can substantially improve recognition accuracy [45].

Implementation Challenges

Extracting context in real-time remains a significant challenge for practical systems [18]. Natural interaction requires synchronized processing of visual, thermal, and contextual data streams while meeting strict timing requirements.

The integration of contextual information with affect recognition poses both theoretical and practical challenges [14]. The system must weigh the relative importance of different contextual factors and determine how they should influence the interpretation of emotional signals. This becomes challenging in situations where contextual cues might conflict with observed emotional expressions.

Handling ambiguous situations requires sophisticated decision-making mechanisms. Our research showed that even state-of-the-art embedding models can struggle with complex social scenarios where multiple interpretations are possible. We found that using reference points in the embedding space can help resolve some of these ambiguities, though perfect disambiguation remains challenging [20].

2.4 Transformer Architectures

The emergence of transformer architectures has revolutionized multi-modal fusion and context integration in affect recognition systems [80]. This section examines the key aspects of transformer-based approaches and their applications in our work.

Multi-modal Fusion

Cross-modal attention mechanisms form the foundation of transformer-based fusion approaches [80]. These mechanisms allow the system to dynamically weight

different modalities based on their relevance to the current context. Our architecture employs self-attention mechanisms across modalities, enabling it to learn complex relationships between thermal patterns, visual features, and contextual information.

The integration of multi-modal data streams demands robust feature representation and alignment mechanisms [5,67]. Our architecture implements this through modality-specific encoders paired with cross-modal attention layers, which preserve the characteristics of each modality while enabling bidirectional information flow between thermal, visual, and contextual streams. The transformer-based design extends beyond static feature fusion by incorporating temporal attention mechanisms that capture both fine-grained and long-term dependencies in the data sequence. This temporal modeling proves important for affect recognition tasks, where emotional expressions manifest as dynamic patterns that must be interpreted within their temporal context, rather than as isolated observations.

2.5 Limitations and Solutions

This section examines challenges that affect recognition faces and presents our solutions, supported by empirical results from Papers A-E.

Sensing Modality Limitations

Traditional vision-based approaches suffer from sensitivity to lighting conditions and occlusions, while contact-based physiological sensors are impractical for natural interactions. Our thermal imaging approach, introduced in Paper A and validated in Paper B, provides an alternative that is immune to lighting variations while enabling non-contact physiological measurement.

The challenge of real-time processing and synchronization across multiple modalities, highlighted by Chen et al. [68], is addressed through our efficient processing pipeline detailed in Paper C.

Context Integration Challenges

The interpretation of affect signals is heavily dependent on context, yet many existing systems treat emotional expressions in isolation. Our context-aware architecture, presented in Paper C, explicitly incorporates situational information using a sentence embedding approach, revealing the importance of contextual understanding.

Papers D and E further explored context integration in practical scenarios, addressing the challenge of real-time context extraction in dynamic environments. Paper E confirmed successful context integration in human-robot collaborative tasks.

Cultural and Individual Variations

A major challenge in developing robust affect recognition systems lies in accommodating the rich diversity of human emotional expression. Through cross-cultural research, Jack et al. [43] uncovered systematic differences between Eastern and Western emotional expressions and interpretations. Their finding that Eastern participants predominantly utilize eye regions for emotion recognition, while Western participants integrate information from the entire face, highlights the complexity of creating universal recognition systems. This cultural variation directly influenced our approach in Paper A, where we carefully evaluated multiple action units before identifying the most important ones as features that are consistently reliable indicators across our diverse participant pool.

The impact of cultural differences extends into the physiological domain, particularly affecting thermal imaging applications. As reviewed by Al Qudah et al. [47], thermal patterns associated with emotional responses can vary significantly across different cultural and ethnic backgrounds. These variations are observed not only in baseline facial temperatures but also in the magnitude and regions of temperature changes during emotional states. Such cultural and individual differences create challenges for thermal-based affect recognition systems, requiring adaptable approaches that can account for this variability. Our LSTM-based architecture in Paper B addressed this challenge by dynamically adapting to individual baseline differences, enabling consistent performance across participants from diverse backgrounds despite the physiological variations documented in the thermal imaging literature.

Recent research has explored different strategies for handling cultural variations in affect recognition. Zhang et al. [46] developed a culture-aware system that dynamically adjusts its interpretation based on detected cultural contexts, achieving a 15% accuracy improvement compared to conventional approaches. However, their method required extensive culture-specific training data - a significant practical limitation. Our work in Paper C took an alternative path by leveraging contextual embeddings, enabling cultural and social adaptation without explicitly modeling cultural differences. This approach proved more flexible and generalizable across different cultural contexts.

Beyond cultural considerations, affect recognition systems must also contend with substantial individual variations in emotional expression. Barrett et al. [4] documented remarkable person-to-person differences in both the experience and manifestation of emotions, even within culturally homogeneous groups. Their meta-analysis revealed that identical emotional states often manifest through distinctly different combinations of facial action units. Paper D addressed this challenge through our transformer architecture's attention mechanisms, which dynamically learned to weight different modalities based on their reliability for each individual participant.

Practical Deployment Challenges

The transition from laboratory prototypes to real-world affect recognition systems introduces a complex set of practical challenges. While controlled environments allow for careful calibration and optimal sensor placement, real-world conditions present dynamic and often unpredictable variables. Wang et al. [41] conducted extensive deployment studies that revealed how real-world conditions could degrade recognition accuracy by up to 40% compared to laboratory results, identifying three fundamental challenges: maintaining reliable sensing, managing computational constraints, and handling missing or corrupted data streams.

A critical factor affecting system performance is the precise placement and calibration of sensors. Through careful experimentation, Nguyen et al. [81] demonstrated that even minor deviations from optimal thermal camera positioning could reduce temperature measurement accuracy by 0.3°C, enough to mask the subtle thermal patterns associated with emotional responses. Similarly, variations in lighting conditions and head poses significantly impact visual feature tracking, with Baltrusaitis et al. [5] reporting drops in action unit detection accuracy of up to 30% under poor lighting conditions. Our work in Paper A addressed these challenges through a ROI detection approach within the thermal processing pipeline, maintaining performance even with moderate head movement.

The integration of multiple sensing modalities introduces significant computational demands that must be carefully managed for real-time operation. While Chen et al. [68] achieved 10Hz processing rates, their implementation required substantial computational resources. Through careful optimization of our feature extraction pipeline, Papers A and B demonstrated more efficient processing that maintained 15Hz operation. The computational complexity grows further with modern deep learning approaches - Hazarika et al. [67] highlighted how transformer architectures often require specific optimization for real-time applications. Paper D tackled this challenge through a streamlined architecture that carefully balanced computational efficiency against recognition accuracy.

Recent research has proposed various strategies to address these deployment challenges. Wagner et al. [65] developed robust fusion approaches that could maintain 85% of full-system accuracy even when some modalities became unavailable. However, as comprehensively analyzed by Mittal et al. [14], maintaining consistent system performance across diverse deployment scenarios remains an open challenge, with real-world performance typically falling 15-30% below laboratory results. The successful validation of our system in Paper E's practical instructional setting represents an important step forward, demonstrating how careful system design can help bridge this gap between laboratory and real-world performance.

Chapter 3

Multi-Modal Context-Aware Affect Recognition System Development and Implementation

This thesis develops approaches for affect recognition in human-robot interaction, with contributions spanning sensing methods, contextual understanding, and real-world applications. Our research began by investigating single-modality thermal sensing and progressively expanded to incorporate contextual information and multiple sensing streams to increase affective state detection robustness and adaptive detections. A key insight emerging across Papers A through E aligns with recent theoretical frameworks in human-AI collaboration [82, 83] that emphasize moving beyond simple binary choices between human and automated control. Instead of pursuing full automation, our work demonstrates the value of hybrid decision-making systems that can fluidly share control between automated affect-based adaptations and human oversight. As Leitão et al. [82] argue, optimal human-AI collaboration emerges from flexible interaction patterns rather than rigid delegation rules, a principle that guided our development from sensing technologies through to practical implementations. Building on Punzi et al.'s [83] framework for hybrid decision-making systems, our papers collectively demonstrate how combining sensing modalities with contextual awareness not only enables more accurate emotion recognition across different environments but also supports more natural and effective human-robot collaboration through balanced control-sharing approaches. This chapter analyzes these developments while examining their current limitations.

3.1 Multi-modal Integration

In Papers A and B, we hypothesized that thermal imaging complements traditional visual features by capturing physiological indicators invisible to standard cameras [70]. The papers concluded that thermal-only system reached 87% F1-

score for frustration-only detection, the combination with visual features through LSTM networks enabled reliable classification of four distinct emotional states with 77% F1-score. These findings align with previous work by Ioannou et al. [70] on thermal indicators while addressing their key limitations of head movement sensitivity and single-modality constraints through our fusion of complementary RGB and thermal data streams with real-time processing.

As we have progressed and piloted our system in different scenarios whether it was in the wild or in the lab environment, we have realized that even perfect sensing cannot resolve inherent ambiguities that can emerge depending on the environment [4] and becomes unstable if tested with new context, considering that same facial expression or thermal pattern could indicate different emotional states depending on context. This observation led to our second hypothesis: Context-aware fusion is necessary for more robust real-world affect recognition. Papers C and D provided strong validation, with context integration improving accuracy from 77% to 89%. This 12% improvement quantifies the impact of contextual understanding.

Multi-modal data streams are harder to use as the synced streams need to align their temporal data to achieve optimal results. Therefore, we did an ablation study with different window sizes in Paper A and confirmed in Paper B, showing that larger window sizes are optimal for thermal imaging and facial expressions. This aligns with established research by Ekman et al. [24] showing that physiological signals usually take 5-15 seconds to take effect fully. In addition, maintaining the information across the window sizes is also important, as demonstrated by [84].

In our studies, a 7-second analysis rolling window proved optimal, balancing temporal resolution against recognition accuracy. This timeframe was decided based on our ablation studies in Papers A and B, allowing sufficient time for thermal patterns to develop while maintaining responsive system performance. For the fusion architecture, we implemented a transformer-based approach [67] similar to Hazarika's MISA work , which enabled effective integration of multiple modalities while preserving their unique characteristics [67]. The transformer architecture's ability to maintain cross-modal relationships was particularly valuable for our application. We further enhanced temporal coherence by implementing a rolling window technique to preserve information continuity across consecutive analysis frames, ensuring smooth transitions in affect detection.

Importantly, Paper E showcased the practical usage of the whole system and its ability to work in real-time. This validates our core premise that multi-modal, context-aware affect recognition enables a natural and effective human-robot interaction.

Going one step beyond incorporating multiple sensing modalities, the key to meaningful progress is better contextual understanding. The fusion pipeline developed in this work is a practical foundation for this step forward, enabling a more contextualized affect recognition while navigating real-world implementation constraints.

The current implementation assumes a relatively stable head pose and requires all modalities to be available for optimal performance. Processing speed remains

another challenge - while the 1Hz rate used in Paper E proved sufficient for our instructional applications, more dynamic interaction scenarios might require higher temporal resolution. Encouragingly, given the rapid advancement of LLMs and embedding models, achieving faster processing rates appears increasingly feasible in the near future.

3.2 Context Understanding

Our exploration of context understanding in affect recognition builds upon our recent advances in social embeddings by Lemaignan et al. [85], where we contributed to developing methods for encoding social situations using pre-trained language models. In this collaborative work, we introduced social embeddings as compact vector representations of social scenes, demonstrating how textual descriptions of social environments could be transformed into meaningful numerical embeddings using large language models [85]. The approach achieved robust results using both Meta's Llama2 and OpenAI's text embedding models, with particular success in maintaining semantic consistency across different pragmatic variations of scene descriptions. Building on this foundation, we extended these insights in Papers C-E to show that the relationship between physical expressions and emotional states is intertwined with contextual information. Papers C-E demonstrate that context is not just an additional input but rather a fundamental component of affect recognition. This finding supports Barrett's theory of constructed emotion [29], where emotional expressions derive their meaning from situational context. Our embedding-based approach significantly outperformed random chance, achieving 68% accuracy in detecting socially appropriate interactions compared to the 33% baseline, the results highlight current limitations in computational social understanding. Modern language models can grasp basic social context but still fall short of human-level comprehension of nuanced social situations [86].

The Social IQA benchmark [87] provides a standardized way to evaluate social reasoning capabilities. Our experiments with this dataset demonstrate that embedding models, particularly OpenAI's ada-2, can effectively distinguish between socially appropriate and inappropriate interactions with 68% accuracy.

The effectiveness of different embedding models in capturing social context was evaluated through a reference point technique that measures semantic distances in embedding space. This approach introduces two polar reference sentences - one indicative of a positive interaction and one of a negative interaction. The premise is to associate the semantic proximity of a sentence to these references with the sentiment it conveys. For instance, if a test sentence embeds closer to "This is a positive interaction" than "This is a negative interaction", it is inferred to depict a positive situation.

To identify optimal reference sentences, a systematic experimental series calculated distances between correct answer averages and positive references, and separately for incorrect answers against negative references, as shown in Table 3.1.

Table 3.1: Distances between averages of correct and incorrect answers for reference sentence pairs

Positive Reference	Negative Reference	Distance
Socially appropriate behavior	Socially inappropriate behavior	0.006
Socially appropriate interaction	Socially inappropriate interaction	0.007
A positive scenario	A negative scenario	0.008
A positive interaction	A negative interaction	0.010
A socially appropriate scenario	A socially inappropriate scenario	0.005
A socially acceptable scenario	A socially unacceptable scenario	0.006

Table 3.2: Accuracy and Embedding Dimensions of Language Models in Detecting Social Situations

Model	Parameters	Dimensions	Accuracy (%)
OpenAI text-3-small	N/A	1536	67
OpenAI Ada-2	N/A	1536	68
OpenAI text-3-large	N/A	3072	68
Instruction XL	N/A	4096	67
LLaMA-2 7B	7B	3204	61
LLaMA-2 13B	13B	3204	63
Phi-2 2.7B	2.7B	2048	62
MixTRaL 8x7B	46.7B	8192	57
Random Chance	N/A	N/A	33

The reference pair yielding maximum separation was selected as providing the most distinguishable boundary between appropriate and inappropriate interactions. The distance metric was computed using:

$$\text{Ref. Distance} = \frac{\sum(\text{Neg. Ref} - \text{Corr. Ans})}{n} - \frac{\sum(\text{Pos. Ref} - \text{Incorr. Ans})}{m} \quad (3.1)$$

The methodology for evaluating social context understanding involved reformulating Social IQA dataset examples into uniform sentence structures. For example, the question "How would you describe Quinn?" with context "Quinn spilt their guts with raucous jokes" was reformulated into complete sentences like "Quinn is a very funny person after spilling their guts with raucous jokes" for the correct answer. This reformulation was used for consistent embedding and comparison across different social scenarios. In addition, facial expressions captured through GPT-4(V) were combined with context embeddings to provide a more complete picture of social situations.

Our research showed that combining thermal imaging data with traditional visual analysis and context improved F1-scores from 77% to 89% [21]. This improvement validates the value of considering multiple streams of contextual information in affect recognition systems.

The reference point technique proved effective for comparing social situations. Instead of treating context as a static background, this method enabled dynamic comparison between situations through their relative positions in embedding space. This addressed a crucial limitation regarding the difficulty of comparing social contexts across different scenarios. Results indicated that even complex social situations could be effectively mapped and compared using carefully selected reference points in the embedding space.

An interesting finding emerged from our model comparison studies, as presented in Table 3.2. Larger models did not necessarily perform better at social reasoning [88]. OpenAI's Ada-2 and Instruction XL achieved comparable performance (68% and 67%, respectively), while the much larger MixTRaL 8x7B model performed worse at 57% accuracy despite its 46.7B parameters. This counterintuitive result suggests that current approaches to scaling language models may not automatically improve their social reasoning capabilities, supporting recent findings by Yuan et al. [88] on the limitations of scale for social norm understanding.

Paper C introduces a reference point technique that leverages the linear properties of embedding spaces demonstrated by Mikolov et al. [89] in their work on word vector compositionality. By establishing contrasting reference points like "This is an appropriate interaction" versus "This is an inappropriate interaction" in the embedding space, the method creates a measurable scale for social appropriateness that adapts to different contexts. This builds on Reimers and Gurevych's [90] findings on the effectiveness of cosine distances for semantic comparison in sentence embeddings. This also matches with the approach used by Anwar et al. [91] showing how embedding-based comparison methods can effectively capture context-dependent social relationships. By representing situations as points relative to these references in embedding space, the method enables direct comparison between different social scenarios while maintaining context sensitivity, addressing a key challenge in social robotics identified by Mittal et al. [14] regarding the integration of contextual factors in emotion perception.

The real-world validation in Paper E revealed another important insight: context understanding must operate at multiple timescales. Immediate context affects emotional interpretation, longer-term contextual factors like user expertise also play crucial roles. This multi-scale nature of context has been largely overlooked in previous work, which typically focuses on immediate situational factors. Our results suggest that comprehensive context understanding must integrate both immediate and evolving contextual elements.

In our transformer-based fusion architecture (Paper D), we demonstrated that cross-modal attention mechanisms could improve performance. While Zhang et al. [92] emphasized the need for specialized architectures that can handle the unique challenges of social computing, our work extends this argument by showing both the limitations of scaling existing architectures and the benefits of purpose-built fusion mechanisms. The success of our multi-scale context integration approach, particularly in handling both immediate physical responses and longer-term interaction patterns, provides empirical support for Zhang's theoretical framework.

Current embedding models still struggle with highly nuanced social situations, notably when multiple contextual factors interact in complex ways. The computational requirements of large language models present challenges for real-time operation, though our optimized implementation maintained acceptable performance. Another question that remains unsolved is how to represent and reason about social context in a computationally tractable way.

Looking forward, several promising directions emerge from our work. The success of the reference point technique suggests that relative rather than absolute context representation may be more effective. The highlighted importance of multiple timescales shows that hierarchical context modeling could improve performance. Most importantly, the strong interaction between context understanding and physical sensing suggests that integrated architectures treating these as linked rather than separate components may be necessary for the progression in affect recognition.

3.3 Impact on Human-Robot Interaction

Our research shows how multi-modal sensing and contextual awareness can create robots that adapt their behavior during human interactions. Building on Mittal et al.'s [14] work in emotion perception, we translated these capabilities to real-world instructional environments. Through our studies, we discovered that successful adaptation isn't just about accurate affect detection, it requires an understanding of when and how to respond. In Paper E during instruction sessions, we saw manual interventions drop from 2.0 to 0.7 times per session, all while maintaining the high user satisfaction levels typically associated with human-guided adaptation. These findings suggest that our system can effectively read and respond to users' needs while maintaining the natural flow of interaction that is so crucial for human-robot engagement.

When robots adapted instruction detail based on detected affect, participants found it as natural as manual adaptation, evidenced by the lack of significant differences in questionnaire responses between conditions. Paper A's distinction between cognitive-load and failure-induced frustration provides important nuance for adaptive systems. The experimental validation in Paper E shows how this distinction could potentially be leveraged. This approach to adaptation aligns with recent findings by Hoang et al. [13] on the importance of context-specific responses in emotion recognition systems.

However, important limitations emerged through our studies. We found that while affect-based adaptation reduced the need for manual intervention, some participants still occasionally adjusted instruction levels manually. This suggests that perfect affect detection and response may not be achievable or even desirable - instead, systems should aim to reduce user burden while maintaining user control when desired.

These results collectively indicate that effective human-robot interaction requires a balance between proactive adaptation and user autonomy. Our findings align with and extend Leitão et al.'s [82] framework, which challenges the traditional binary view of either full automation or complete human control. Through Papers A-E, particularly in Paper E's instructional setting, we observed that successful human-robot interaction emerges from fluid patterns of shared control rather than rigid delegation rules. The continuous but subtle adjustments achieved through our multi-modal affect detection system exemplify what Punzi et al. [83] describe as "hybrid decision-making paradigms" - where systems and humans collaboratively shape interaction outcomes. Their theoretical framework emphasizes that optimal human-AI collaboration occurs when systems maintain user agency while leveraging AI capabilities, which our work validates empirically in the context of affect-aware robotics. This is particularly evident in Paper E, where the system's ability to modulate instruction detail based on detected affect while preserving manual override options demonstrates the practical implementation of these hybrid decision-making principles. Rather than attempting to eliminate manual control entirely, our research shows that affect-aware systems should strive for a complementary relationship where automated adaptations reduce user burden while maintaining meaningful human agency - a finding that bridges the theoretical insights of recent human-AI collaboration research with practical implementations in human-robot interaction.

Chapter 4

Discussion

4.1 Technical and Practical Limitations

The development of multi-modal affect recognition systems revealed several challenges. Based on our findings across Papers A-E, we identify three primary categories of limitations: technical constraints, data requirements, and generalization challenges.

The transition from controlled laboratory settings to real-world applications revealed important limitations of our system. A critical challenge is maintaining stable head pose tracking, as rapid head movements can disrupt the entire analysis pipeline. While our thermal processing approach shows improvements compared to Mittal et al. [14], two main constraints affect real-time applications: the required 7-second analysis window and the current processing rate. These limitations are particularly relevant for time-critical scenarios such as emergency response, where rapid system feedback is essential.

Another limitation concerns the availability and synchronization of multiple modalities [14]. Our transformer architecture assumes access to thermal, visual, and contextual information streams. When any modality becomes unavailable or degraded, system performance may suffer. This extends beyond our specific implementation - as Mittal et al. [14] note, multi-modal systems inherently face increased vulnerability to sensor or processing failures compared to single-modality approaches. The real-world validation in Paper E highlighted this challenge, as occasional sensor glitches required fallback mechanisms.

Despite their capabilities in social understanding, our embedding-based context model faces several key limitations when dealing with cultural and situational factors. By treating context as static within each 7-second interaction window, we potentially miss rapidly evolving social dynamics - a fundamental challenge that Barrett et al. [4] highlighted in their work on emotional context.

Computational demands pose another significant hurdle in our current implementation. Running at 1Hz in paper E proved sufficient for basic single-person

interactions, yet this processing rate would fall short in dynamic environments involving multiple people. The contextual processing pipeline, heavily dependent on large language models, requires substantial computational resources. To make these systems more practical and scalable, we'll need to develop both more efficient architectures and smarter processing optimizations.

The most pressing challenge for our system lies in its data requirements when expanding to new domains. Papers C and D demonstrated strong performance on existing datasets, but moving into new contexts isn't straightforward - it demands extensive effort in collecting and annotating fresh data. This is particularly obvious in affective computing where we need perfectly synchronized data across multiple modalities, making dataset creation far more complex than traditional single-modality approaches. Our between-subjects study in Paper E brought this issue into focus, showing how different users' behaviors and interaction patterns could push our system to its generalization limits. These findings point to a critical need: future work must prioritize collecting more diverse training data to build robust and adaptable systems.

4.2 Ethical and Societal Implications

The deployment of affect-aware robotic systems raises ethical challenges that demand attention far beyond technical performance metrics. These systems, which recognize and influence aspects of human life related to emotion, introduce capabilities that carry significant potential for both benefit and harm [93]. As affect pervades human life and grounds rationality rather than opposing it [94], the ethical implications of systems that engage with human affect are particularly profound.

Privacy and Data Collection

Privacy emerges as a fundamental concern in continuous affect monitoring. While thermal imaging offers advantages over contact sensors in terms of physical intrusion, it still captures intimate physiological responses that could reveal more about a person's internal state than they might wish to share. As Devillers and Cowie note, these systems enable "the massive collection of sensitive personal data in an invisible and unaccountable manner" [93], raising concerns about surveillance and potential misuse.

The pervasive nature of emotional monitoring presents unique challenges distinct from traditional robotics systems. Floridi's information ethics theory [95] highlights how the social deployment of AI creates ethical questions not well handled by traditional formulations. This is particularly relevant for affective computing, where the technology's ability to detect and influence emotional states raises novel privacy concerns.

Oversimplification and Representation

A critical ethical concern lies in how affective computing systems may oversimplify human emotional states. Contemporary understanding indicates that emotions involve complex, far-reaching patterns of activity, including evaluative perception, memory evocation, and behavioral preparation [96]. Current representations often reduce these complex emotional experiences to simplified categories or dimensions, potentially missing crucial nuances in human affect.

The challenge extends to the way these systems learn to recognize emotions. Training data often relies on acted or simplified emotional expressions, which may not accurately reflect the complexity of genuine emotional experiences in real-world settings [97]. This raises ethical questions about the validity and reliability of affect recognition systems, particularly when they inform important decisions or interactions.

User Agency and Autonomy

The question of user agency and control has emerged as perhaps the most practically significant ethical concern. While automated systems can reduce the need for manual intervention, maintaining appropriate user autonomy is crucial. Research has shown that people approach interactions with artificial agents similarly to human interactions [98], which can lead to overestimation of system capabilities and inappropriate emotional attachment.

Floridi's concept of "faultless responsibility" [99] becomes particularly relevant here, as affect-aware systems often involve networks of agents - both human and artificial - whose local interactions may have significant moral implications. This distributed nature of responsibility raises questions about who bears ethical accountability for system outcomes.

Cultural and Individual Differences

Cultural differences in emotional expression create a particularly complex challenge at the intersection of ethics and technical capability. Research has demonstrated systematic variations in how emotions are expressed and interpreted across cultures [43, 100]. These differences extend beyond mere expression to fundamental variations in emotional experience and interpretation across societies.

Individual differences add another layer of complexity. Barr et al. [101] identified seven dimensions of variation in emotional expression, highlighting the challenge of creating systems that can appropriately respond to diverse emotional styles. Systems that fail to account for these differences risk imposing culturally biased interpretations of emotional behavior.

Manipulation and Emotional Authenticity

A significant ethical concern involves the potential for emotional manipulation. As demonstrated in studies of artificial companions [102], these systems can create emotional dependencies that may not be well-grounded in reality. The ability to influence emotional states through "nudging" techniques [103] raises critical questions about autonomy and authentic emotional experience.

The risk of emotional manipulation becomes particularly acute in therapeutic applications. While systems like Woebot show promise in delivering therapeutic interventions [104], they also raise questions about the nature of therapeutic relationships and the role of authentic human connection in emotional well-being.

Addiction and Engagement

The risk of addiction presents another significant ethical challenge. As recognized in DSM-5 [105], internet gaming disorder and similar technology-related conditions demonstrate how emotionally engaging systems can lead to problematic usage patterns. This risk is particularly relevant for affective computing systems designed to create positive emotional experiences [106].

Future Considerations

Our experience developing and deploying these systems suggests several key areas for future ethical consideration:

- Development of transparent data collection policies aligned with information ethics principles [95]
- Integration of cultural sensitivity in system responses [43]
- Implementation of flexible control mechanisms that maintain meaningful user agency
- Establishment of clear guidelines for deployment in sensitive domains
- Creation of frameworks for assessing long-term societal impact

The validation studies provide initial insights into these challenges, but broader deployment will require more extensive ethical consideration. Moving forward, we must integrate ethical considerations into the core design process of affect-aware systems, recognizing that their impact extends far beyond immediate human-robot interactions to shape broader patterns of emotional expression and social interaction [93].

Chapter 5

Summary of Publications

This chapter summarizes five papers that form the core of this thesis. Together, these works develop a framework for multi-modal affect detection by combining thermal imaging, facial expressions, and contextual understanding. The papers highlight a systematic progression from thermal imaging through multi-modal fusion and context understanding to practical applications. The author was the main contributor in all the Papers as seen in Table 5.1. These contributions included designing the studies, creating the architectures, testing them and then writing the manuscript.

Table 5.1: Comprehensive Analysis of Author Contributions Across Research Papers

Paper	Technical Design	Implementation	Validation	Writing
A	●●●	●●●	●●	●●
	Thermal Imaging pipeline	Real-time processing	18 participants	Core draft
B	●●●	●●●	●●●	●●●
	Multi-modal fusion	LSTM architecture	4-state validation	Lead author
C	●●●	●●●	●●●	●●●
	Context framework	Embedding system	collecting in-the-wild data and testing	Lead author
D	●●●	●●●	●●●	●●●
	Transformer design	Real-time fusion	Accuracy metrics	Lead author
E	●●●	●●●	●●●	●●●
	Adaptive system	Full deployment	40-user study	Lead author

Legend: ●●● Primary contribution (>80%); ●● Significant contribution (50-80%); ● Supporting contribution (<50%)

Paper A

Automatic Frustration Detection Using Thermal Imaging

Youssef Mohamed, Maria Teresa Parreira, Giulia Ballardini, Séverin Lemaignan, Iolanda Leite

Published in 2022 IEEE/RSJ International Conference on Human-Robot Interaction (HRI)

This paper investigates the use of thermal imaging for detecting two types of frustration in human-robot interaction scenarios: cognitive load-induced and failure-induced. The work develops a processing pipeline combining thermal and RGB cameras to track facial regions of interest while operating at 15Hz. When both frustration cases were combined and used as training input, the model reached an accuracy of 89% with RGB features, 87% using only thermal features and 84% using EDA.

Through experimental validation with 18 participants, the study revealed that three facial regions (nose, forehead, and lower lip) provide reliable indicators of frustration. Feature selection analysis showed that the cheek region could be discarded while maintaining performance. The thermal processing pipeline incorporates ROI tracking techniques that address key technical challenges in thermal-based affect detection.

The findings highlight that thermal imaging can effectively detect different types of frustration, providing a viable non-contact sensing modality for affect recognition in human-robot interaction. The developed methods formed the foundation for the multi-modal integration explored in subsequent papers. The real-time processing capabilities and high accuracy show the potential for practical deployment in interactive robotic systems.

Contributions

The author developed the thermal processing pipeline, implemented the integrated thermal-RGB tracking system, collected experimental data from 18 participants, devised the feature selection methodology, and created the real-time ROI tracking techniques. This core technical work established the foundation for multi-modal affect detection in human-robot interaction scenarios.

Paper B

Multi-modal Affect Detection Using Thermal and Optical Imaging in a Gamified Robotic Exercise

Youssef Mohamed, Arzu Güneysu, Séverin Lemaignan, Iolanda Leite

Published in International Journal of Social Robotics, 2023

Building on Paper A, this work extends affect detection to multiple modalities and states. The system combines thermal imaging with visual action unit analysis to detect four states: frustration, boredom, enjoyment and neutral. Using LSTM networks for temporal modeling, the system achieves 77% accuracy across all states, proving that thermal and visual features provide complementary information.

The LSTM-based approach captures affect dynamics across 7-second windows, which proves useful for distinguishing between similar expressions that evolve differently over time. Ablation studies analyze the contribution of different features and modalities to overall system performance.

Integration with a gamified exercise platform enables the collection of natural interaction data. The system shows good performance in detecting transitions between affective states, leveraging both instantaneous features and temporal patterns. The LSTM architecture captures these temporal dependencies while maintaining real-time operation.

Contributions

The author conceived and implemented the multi-modal fusion architecture and LSTM-based temporal modeling approach. He conducted comprehensive ablation studies, integrated the system with a gamified platform, collected and analyzed the experimental data, and developed the real-time implementation that maintains temporal dependencies.

Paper C

Context Matters: Understanding Socially Appropriate Affective Responses Via Sentence Embeddings

Youssef Mohamed, Séverin Lemaignan, Arzu Güneysu, Patric Jensfelt, Christian Smith

Published in International Conference on Social Robotics, 2024

This paper examines the role of context in affect recognition through an embedding-based approach. The system uses GPT-4(V) to generate facial expression descriptions and ada-2 embeddings to detect socially inappropriate interactions. Tests on benchmark datasets and real-world deployments show improved recognition accuracy when incorporating contextual understanding.

The paper presents a method using sentence embeddings to capture social context in affect recognition, including a reference point technique for comparing social situations. The system achieves 68% accuracy in social appropriateness detection during in-the-wild deployment. Analysis of different embedding models reveals trade-offs between model size, embedding dimensionality, and context understanding capabilities.

The use of GPT-4(V) for facial expression description enables richer representation compared to action unit analysis alone. This approach helps capture subtle variations in expression that carry social meaning beyond basic emotional categories.

Contributions

The author developed the novel embedding-based context framework, implemented the GPT-4(V) integration, created the reference point technique for social situation comparison, and conducted the comparative analysis of embedding models. He designed and executed the in-the-wild validation experiments demonstrating the system's effectiveness in real-world scenarios.

Paper D

Fusion in Context: A Multimodal Approach to Affective State Recognition

Youssef Mohamed, Séverin Lemaignan, Arzu Güneysu, Patric Jensfelt, Christian Smith

Published in International Conference on Robot and Human Interactive Communication, 2025

This paper presents a transformer-based architecture for context-aware multimodal fusion in affect recognition. The system uses modality-specific encoders for thermal, visual, and contextual data, combined through additive fusion. Evaluation shows improved recognition accuracy through context-aware fusion, achieving 89% accuracy compared to 77% for previous approaches.

The transformer architecture enables the integration of multiple data streams while maintaining temporal relationships. Modality-specific encoding helps maintain the characteristics of each data stream during fusion. The system performs well in ambiguous situations where context helps disambiguate similar expressions.

Contributions

The author architected the transformer-based fusion system, developed the modality-specific encoders, implemented the additive fusion mechanism, and optimized the system for real-time operation. His work established a new approach for integrating multiple data streams while preserving their temporal relationships and unique characteristics.

Paper E

Are You an Expert? Instruction Adaptation Using Multi-Modal Affect Detection with Thermal Imaging and Context

Youssef Mohamed, Séverin Lemaignan, Arzu Güneysu, Patric Jensfelt, Christian Smith

Published in International Conference on Robot and Human Interactive Communication, 2025

The final paper demonstrates the practical application of multi-modal affect recognition in adaptive instruction delivery. The system uses thermal imaging, facial expressions, and contextual information to dynamically adjust robot instruction detail during assembly tasks. A pilot study (N=21) validated that manipulating instruction detail successfully induced target affective states (57.1% for boredom, 50.0% for frustration).

In the main between-subjects study (N=40), the adaptive system significantly reduced manual interventions to 0.73 adjustments per session compared to 2.0 in the manual condition, while automatically implementing 2.62 level changes per session. This created longer periods of stability (3.65 vs. 2.58 minutes) without sacrificing user experience, as shown by comparable interest/enjoyment measures (M=0.472 vs. M=0.397) and frustration levels (M=0.343 vs. M=0.381) between conditions.

This work demonstrates how affect-aware systems can balance automation with user control, enabling fluid patterns of shared decision-making rather than rigid automation. The successful real-world deployment validates the practical viability of thermal-based affect detection in instructional settings, particularly when coupled with contextual understanding to guide adaptation decisions.

Contributions

The author designed and implemented the adaptive instruction system, developed the real-time affect-based adaptation mechanisms, and integrated multi-modal affect detection with expertise estimation. He designed the experimental flow and conditions, led the 40-participant user study, analyzed the experimental results, and validated the system's effectiveness in real-world instruction scenarios.

Chapter 6

Conclusion

6.1 Overview

This thesis has advanced the field of affect recognition in human-robot interaction through several contributions spanning technical approaches, empirical validations, and practical applications. Our work demonstrates that combining thermal imaging, visual analysis, and contextual understanding (as discussed in Section 3.1) creates affect recognition systems that support collaborative human-robot interactions.

6.2 Empirical Findings

The empirical findings from our research validate several hypotheses. First, we established thermal imaging as a method for detecting emotional states, providing a non-contact alternative to traditional physiological sensors (as detailed in Section 3.1). Our results demonstrated the role of context in affect recognition (Section 3.2), with accuracy improving from 77% to 89% when incorporating contextual understanding. We identified specific facial regions for thermal measurement, creating a focused detection approach. Real-world deployment studies provided performance benchmarks in non-laboratory conditions (addressing challenges outlined in Section 4.1).

6.3 Insights on Human-Robot Collaboration

This research demonstrates that human-robot interaction emerges not from purely automated responses, but from patterns of shared control (as explored in Section 3.3). Our system's ability to reduce manual interventions while maintaining user autonomy demonstrates hybrid decision-making approaches where humans and robots collaboratively shape interaction outcomes. The balance between automated

adaptation and user control supports interactions, aligning with theoretical frameworks in human-AI collaboration (with considerations for implications discussed in Section 4.2).

6.4 Future Research Directions

Technical Opportunities

Looking ahead, several promising research directions emerge from this work. From a technical perspective, we see opportunities for more sophisticated context understanding models, such as hierarchical attention mechanisms [107] and multi-scale temporal transformers [108], that can better capture the nuances of human emotional expression and support more nuanced collaborative decision-making. Improvements in real-time processing methods could enable faster system responses, while enhanced fusion strategies like uncertainty-aware weighting and adversarial feature alignment might better handle missing or corrupted sensor data.

Validation Challenges

The validation landscape also presents important challenges to address. There is a pressing need for large-scale datasets that combine thermal, visual, and contextual information. Cross-cultural validation studies would help ensure these systems work effectively across different societies and contexts [109, 110]. Additionally, long-term deployment analyses could reveal how these adaptive systems perform over extended periods of real-world use and how the balance of human-robot control evolves over time.

Application Domains

In terms of practical applications, we envision several exciting possibilities. Integration with diverse robotic platforms could expand the reach of affect-aware systems into new domains [111]. Adapting these technologies for group interactions presents both challenges and opportunities, particularly in developing frameworks for multi-party collaborative decision-making. Domain-specific implementations could address unique requirements in fields like healthcare or education, where the balance between automation and human oversight is especially crucial.

Research Opportunities

Further research opportunities lie in developing cultural adaptation mechanisms that allow systems to adjust their interpretation, response patterns and even their roles based on cultural context [112]. Continuous learning systems could enable ongoing improvement through real-world interactions, while context prediction models might anticipate and prepare for upcoming emotional states. Importantly, these

advances must maintain the principle of shared control, ensuring that increased system capability enhances rather than replaces meaningful human agency.

6.5 Final Remarks

As social robots become more common in our daily lives, the importance of robust affect recognition and adaptive interaction will only grow. Our work demonstrates that combining multiple sensing modalities with contextual understanding can create systems that interact more naturally and effectively with humans while maintaining appropriate boundaries between automated and human-guided behavior - specifically, preserving the user's agency to override automated decisions while reducing cognitive burden through context-aware adaptations. Challenges remain, particularly in areas like cross-cultural adaptation and real-time processing, the path forward is clear: future systems must continue to integrate multiple streams of information while maintaining awareness of the broader context in which human-robot interactions occur.

The contributions presented in this thesis establish a foundation for more sophisticated affect-aware robotic systems that can collaborate with their human partners. As technology advances and our understanding of human emotions deepens, we expect continued progress toward robots that can understand and appropriately respond to human emotional states while maintaining the delicate balance between automation and human agency. This evolution will be crucial for creating robot companions and assistants that can interact with humans in natural and meaningful ways while supporting, rather than supplanting, human decision-making capabilities.

Bibliography

[1] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[2] A. Tapus, M. Maja, and B. Scassellati, "Socially assistive robotics [grand challenges of robotics]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 35–42, 2007.

[3] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[4] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[6] N. S. Suhaimi, J. Mountstephens, and J. Teo, "Eeg-based emotion recognition: a state-of-the-art review of current trends and opportunities," *Computational intelligence and neuroscience*, vol. 2020, no. 1, p. 8875426, 2020.

[7] D. Shastri, A. Merla, P. Tsiamyrtzis, and I. Pavlidis, "Imaging facial signs of neurophysiological responses," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 477–484, 2009.

[8] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[9] S. L. Stewart, A. Schepman, M. Haigh, R. McHugh, and A. J. Stewart, "Affective theory of mind inferences contextually influence the recognition of emotional facial expressions," *Cognition and Emotion*, vol. 33, no. 2, pp. 272–287, 2019.

[10] M. E. Milanak and H. Berenbaum, "The effects of context on facial affect recognition," *Motivation and Emotion*, vol. 38, pp. 560–568, 2014.

[11] C. De Sousa, S. Morgado, J. Ferreira, S. Tukaiev, and R. Fonseca, "The impact of clinical context on the recognition of facial expressions," *European Psychiatry*, vol. 67, no. S1, pp. S114–S114, 2024.

[12] Y. Mohamed, G. Ballardini, M. T. Parreira, S. Lemaignan, and I. Leite, "Automatic frustration detection using thermal imaging," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 451–459, IEEE, 2022.

[13] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "Context-aware emotion recognition based on visual relationship detection," *IEEE Access*, vol. 9, pp. 90465–90474, 2021.

[14] T. Mittal, A. Bera, and D. Manocha, "Multimodal and context-aware emotion perception model with multiplicative fusion," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2021.

[15] H. Lu, X. Niu, J. Wang, Y. Wang, Q. Hu, J. Tang, Y. Zhang, K. Yuan, B. Huang, Z. Yu, *et al.*, "Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing," *arXiv preprint arXiv:2403.05916*, 2024.

[16] Y. Abdelrahman, E. Velloso, T. Dingler, A. Schmidt, and F. Vetere, "Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–20, 2017.

[17] R. Mehta, S. Saha, S.-B. Cho, and P. K. Kalra, "A survey on facial thermal image analysis: Past, present and future," *Information Fusion*, vol. 91, pp. 73–99, 2023.

[18] H. Wang, A. Mehra, C. N. Duong, W. Guo, Y. Jin, X. Liu, and G. Z. Li, "Multimodal emotion recognition: A survey of recent advances and trends," *IEEE Transactions on Affective Computing*, 2023.

[19] Y. Mohamed, A. Güneysu, S. Lemaignan, and I. Leite, "Multi-modal affect detection using thermal and optical imaging in a gamified robotic exercise," pp. 981–997, Springer, 2023.

[20] Y. Mohamed, S. Lemaignan, A. Güneysu, P. Jensfelt, and C. Smith, "Context matters: Understanding socially appropriate affective responses via sentence embeddings," in *International Conference on Social Robotics*, pp. 78–91, Springer, 2024.

[21] Y. Mohamed, S. Lemaignan, A. Guneysu, P. Jensfelt, and C. Smith, "Fusion in context: A multimodal approach to affective state recognition," *arXiv preprint arXiv:2409.11906*, 2024.

[22] C. Darwin, *The expression of the emotions in man and animals*. John Murray, 1872.

[23] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[24] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Consulting Psychologists Press*, 1978.

[25] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.

[26] K. Mase, "Recognition of facial expression from optical flow," *IEICE Transactions on Information and Systems*, vol. 74, no. 10, pp. 3474–3483, 1991.

[27] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205, 1998.

[28] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[29] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," *Computer Vision and Pattern Recognition Workshop*, 2003.

[30] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," *IEEE International Conference on Computer Vision*, pp. 2983–2991, 2015.

[31] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[32] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306, 2018.

[33] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[35] T. B. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[36] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[37] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal affective state recognition in serious games applications," in *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 435–439, 2016.

[38] F. Alonso-Martin, M. Malfaz, J. Sequeira, J. F. Gorostiza, and M. A. Salichs, "A multimodal emotion detection system during human–robot interaction," *Sensors*, vol. 13, no. 11, pp. 15549–15581, 2013.

[39] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2019.

[40] A. Taylor, D. M. Chan, and L. D. Riek, "Robot-centric perception of human groups," *ACM Transactions on Human-Robot Interaction*, vol. 9, no. 3, pp. 1–21, 2020.

[41] Y. Wang, M. Li, S. Liu, and M. Li, "Context-aware emotion recognition networks," *arXiv preprint arXiv:1908.05913*, 2019.

[42] L. Guerdan and H. Gunes, "Federated continual learning for socially aware robotics," *arXiv preprint arXiv:2303.xxxx*, 2023.

[43] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[44] S. Li and W. Deng, "Deep learning for facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[45] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Learning to recognize dynamic facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1739–1755, 2019.

[46] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Multimodal intelligence: Representation learning, information fusion, and applications," in *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[47] M. M. Al Qudah, A. S. Mohamed, and S. L. Lutfi, "Affective state recognition using thermal-based imaging: A survey.," *Computer Systems Science & Engineering*, vol. 37, no. 1, 2021.

[48] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, 2017.

[49] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, 2018.

[50] H. D. Critchley, "Electrodermal responses: what happens in the brain," *The Neuroscientist*, 2002.

[51] M. Hamedi, S.-H. Salleh, T. S. Tan, and K. Ismail, "Emg-based facial gesture recognition through versatile elliptic basis function neural network," *Biomedical Engineering Online*, 2011.

[52] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully automatic facial action recognition in spontaneous behavior," in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 223–230, IEEE, 2006.

[53] H. Chennamma and X. Yuan, "A survey on eye-gaze tracking techniques," *arXiv preprint arXiv:1312.6410*, 2013.

[54] J. Lemley, A. Kar, A. Drimbarean, and P. Corcoran, "Efficient cnn implementation for eye-gaze estimation on low-power/low-quality consumer imaging systems," *arXiv preprint arXiv:1806.10890*, 2018.

[55] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[56] Z. Yang, A. Kay, Y. Li, W. Cross, and J. Luo, "Pose-based body language recognition for emotion and psychiatric symptom interpretation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 294–301, 2021.

[57] B. W. Schuller, "Speech emotion recognition: State of the art and challenges," *IEEE Signal Processing Magazine*, 2018.

[58] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[59] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[60] B. W. Schuller, "The computational paralinguistics challenge [social sciences]," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, 2012.

[61] A. Tammewar, A. Cervone, E.-M. Messner, and G. Riccardi, "Modeling user context for valence prediction from narratives," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 1–10, 2019.

[62] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings of INTERSPEECH*, pp. 2042–2046, 2020.

[63] Q. Mao, X. Zhao, Z. Huang, and Y. Zhan, "Speaker-independent speech emotion recognition by fusion of functional and accompanying paralanguage features," *Journal of Zhejiang University SCIENCE C*, vol. 14, no. 7, pp. 573–582, 2013.

[64] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[65] J. Wagner, E. André, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 206–218, 2011.

[66] Z. Liu, Y. Shen, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2247–2256, 2018.

[67] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

[68] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, and D. Conway, "Multimodal behavior and interaction as indicators of cognitive load," *ACM Transactions on Interactive Intelligent Systems*, vol. 7, no. 3, pp. 1–26, 2017.

[69] J. Zhang, K. Zhao, Y. Xu, Y. Zhao, Y. Xue, Y. Zhou, and J. Feng, "Context-aware attention network for image-emotion classification," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1830–1842, 2019.

[70] S. Ioannou, V. Gallese, and A. Merla, "Thermal infrared imaging in psychophysiology: Potentialities and limits," *Psychophysiology*, vol. 51, no. 10, pp. 951–963, 2014.

[71] I. Pavlidis, N. L. Eberhardt, and J. A. Levine, "Seeing through the face of deception," *Nature*, vol. 415, no. 6867, pp. 35–35, 2002.

[72] H. Genno, K. Ishikawa, O. Kanbara, M. Kikumoto, Y. Fujiwara, R. Suzuki, and M. Osumi, "Using facial skin temperature to objectively evaluate sensations," *International Journal of Industrial Ergonomics*, vol. 19, no. 2, pp. 161–171, 1997.

[73] M. M. Khan, R. D. Ward, and M. Ingleby, "A survey of the recognition of emotions using thermal imaging," *Image and Vision Computing*, vol. 52, pp. 134–146, 2016.

[74] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, 2017.

[75] H. D. Critchley, "Electrodermal responses: what happens in the brain," *The Neuroscientist*, vol. 8, no. 2, pp. 132–142, 2002.

[76] M. Hamedi, S.-H. Salleh, T. S. Tan, and K. Ismail, "Emg-based facial gesture recognition through versatile elliptic basis function neural network," *Biomedical Engineering Online*, vol. 10, no. 1, pp. 1–22, 2011.

[77] T. Nguyen, K. Tran, and H. Nguyen, "Towards thermal region of interest for human emotion estimation," in *10th International Conference on Knowledge and Systems Engineering*, pp. 152–157, IEEE, 2018.

[78] D. Matsumoto, *The handbook of culture and psychology*. Oxford University Press, 2001.

[79] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[80] Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

[81] T. Nguyen *et al.*, "Towards thermal region of interest for human emotion estimation," *International Conference on Knowledge and Systems Engineering*, 2018.

[82] D. Leitão, P. Saleiro, M. A. Figueiredo, and P. Bizarro, "Human-ai collaboration in decision-making: beyond learning to defer," *arXiv preprint arXiv:2206.13202*, 2022.

[83] C. Punzi, R. Pellungrini, M. Setzu, F. Giannotti, and D. Pedreschi, "Ai, meet human: Learning paradigms for hybrid decision making systems," *arXiv preprint arXiv:2402.06287*, 2024.

[84] R. W. Picard, *Affective computing*. MIT press, 2000.

[85] S. Lemaignan, A. Andriella, L. Ferrini, L. Juricic, Y. Mohamed, and R. Ros, "Social embeddings: Concept and initial investigation," *Open Research Europe*, vol. 4, no. 63, p. 63, 2024.

[86] K. Gandhi *et al.*, "Understanding social reasoning in language models with language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[87] M. Sap, H. Rashkin, D. Chen, R. Le Bras, and Y. Choi, "Social iqa: Commonsense reasoning about social interactions," *arXiv preprint arXiv:1904.09728*, 2019.

[88] Y. Yuan, K. Tang, J. Shen, M. Zhang, and C. Wang, "Measuring social norms of large language models," *arXiv preprint arXiv:2404.02491*, 2024.

[89] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," vol. 26, 2013.

[90] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[91] S. Anwar, M. Beg, K. Saleem, Z. Ahmed, A. Javed, and U. Tariq, "Social relationship analysis using state-of-the-art embeddings," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1–21, 2023.

[92] Y. Zhang, D. Meng, and J. Han, "Context-aware affective graph reasoning for emotion recognition," *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2020.

[93] L. Devillers and R. Cowie, "Ethical considerations on affective computing: An overview," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1445–1457, 2023.

[94] A. Damasio, *Descartes' Error*. New York, NY, USA: Random House, 1994.

[95] L. Floridi, *The Ethics of Information*. Oxford, U.K.: Oxford University Press, 2013.

[96] R. Cowie, "The enduring basis of emotional episodes: Towards a capacious overview," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 98–104, Sept. 2015.

[97] E. Douglas-Cowie *et al.*, "Multimodal databases of everyday emotion: Facing up to complexity," in *Proceedings of Interspeech*, pp. 813–816, Sept. 2005.

[98] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York, NY, USA: Cambridge University Press/ICSLI, 1996.

[99] L. Floridi, "Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions," *Philosophical Transactions of the Royal Society A*, vol. 374, 2016. Art. no. 20160112.

[100] N. Lim, "Cultural differences in emotion: Differences in emotional arousal level between the east and the west," *Integrative Medicine Research*, vol. 5, pp. 105–109, June 2016.

[101] L. K. Barr, J. H. Kahn, and W. J. Schneider, "Individual differences in emotion expression: Hierarchical structure and relations with psychological distress," *Journal of Social and Clinical Psychology*, vol. 27, pp. 1045–1077, Dec. 2008.

[102] E. Portacolone *et al.*, "Ethical issues raised by the introduction of artificial companions to older adults with cognitive impairment: A call for interdisciplinary collaborations," *Journal of Alzheimer's Disease*, vol. 76, pp. 445–455, July 2020.

[103] R. Thaler and C. Sunstein, *Nudges*. Paris, France: Vuibert, 2017.

[104] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, p. e19, 2017.

[105] *Diagnostic and Statistical Manual of Mental Disorders*. Washington, DC, USA: American Psychiatric Association, 5 ed., 2013.

[106] G. N. Yannakakis and A. Paiva, "Emotion in games," in *Oxford Handbook on Affective Computing*, pp. 459–471, Oxford, U.K.: Oxford University Press, 2015.

[107] Y. Zhang, H. Liu, D. Zhang, X. Chen, T. Qin, and Q. Zheng, "Eeg-based emotion recognition with emotion localization via hierarchical self-attention," *IEEE transactions on affective computing*, vol. 14, no. 3, pp. 2458–2469, 2022.

[108] M. Khan, A. El Saddik, M. Deriche, and W. Gueaieb, "Stt-net: Simplified temporal transformer for emotion recognition," *IEEE Access*, 2024.

[109] X. Gong, C. P. Chen, and T. Zhang, "Cross-cultural emotion recognition with eeg and eye movement signals based on multiple stacked broad learning system," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 2, pp. 2014–2025, 2023.

[110] J. Tejada, R. M. K. Freitag, B. F. M. Pinheiro, P. B. Cardoso, V. R. A. Souza, and L. S. Silva, "Building and validation of a set of facial expression images to detect emotions: a transcultural study," *Psychological Research*, vol. 86, no. 6, pp. 1996–2006, 2022.

[111] R. V. Aranha, C. G. Corrêa, and F. L. S. Nunes, "Adapting software with affective computing: A systematic review," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 883–899, 2021.

[112] Y. Li, K. P. Tee, W. L. Chan, R. Yan, Y. Chua, and D. K. Limbu, "Continuous role adaptation for human–robot shared control," *IEEE Transactions on Robotics*, vol. 31, no. 3, pp. 672–681, 2015.

Part II

Included Publications