

Multimodal Emotion Recognition using Mobile Devices

by

Kangning Yang

ORCID: [0000-0002-7106-0022](https://orcid.org/0000-0002-7106-0022)

A thesis submitted in total fulfillment for the
degree of Doctor of Philosophy

in the

School of Computing and Information Systems
Faculty of Engineering and Information Technology
THE UNIVERSITY OF MELBOURNE

November 2022

Abstract

Emotion recognition is an emerging interdisciplinary research topic that has received increasing attention over the past few years. It has the potential to play a vital role in mental health and well-being management and can aid in early diagnosis, ongoing monitoring, and timely interventions. With the rapid development of mobile and wearable devices, it has become easier to access users' affective data passively, continuously, and remotely. This creates new opportunities to help identify and interpret spontaneous human emotions in a less intrusive manner. This thesis first investigates the recent advances in mobile emotion recognition research, especially the state-of-the-art methodologies. Subsequently, we turn our attention to off-the-shelf commercial facial expression-based mobile emotion recognition systems. We identify and classify different distortions in real-world mobile device usage scenarios that can lead to poor image quality. On this basis, we present a replicable assessment protocol to analyse the performance of commercial emotion recognition systems that are based on facial features. Next, we presents an emotion recognition system with a hybrid multimodal architecture. It extracts high-level affective features from facial expression, speech, keystroke, and three types of physiological signals (blood volume pulse, electrodermal activity, and skin temperature) using attention-based deep learning architectures, and learns cross-modal correlations for fusion through several deep neural network layers. We demonstrate its robust performance by conducting a user study with 45 participants and explore different real-world usage scenarios where signals are partially available. Finally, we present an efficient hierarchical multimodal network that relies on self-attention and convolution mechanisms to model both local and global dependencies of multiple physiological signals. We evaluate the proposed system on two publicly available mobile multi-sensor datasets and achieve state-of-the-art performance. This thesis demonstrates the feasibility and effectiveness of using mobile devices to automatically and passively detect human emotions and provides novel approaches for designing robust multimodal mobile emotion recognition systems.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other materials used,
3. appropriate ethics procedure and guidelines have been followed to conduct this research,
4. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: Kangning Yang

Date: November 2022

Preface

This thesis is submitted in total fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Melbourne. The research presented here was primarily conducted at the School of Computing and Information Systems, The University of Melbourne under the supervision of Dr. Jorge Goncalves, Dr. Tilman Dingler, and Dr. Greg Wadley. The work was supported by the Melbourne Research Scholarship, awarded by The University of Melbourne.

The thesis comprises three peer-reviewed articles, referred to by roman numerals (Article I - III). All articles are included in full, prefaced with a brief introduction situating each article within the context of the thesis. While several collaborators have contributed to the articles, I declare that I am the primary author and have more than 50% contributions in each of the following publications:

Article I

Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer*, 37 (6):1447–1466, 2021. URL <https://doi.org/10.1007/s00371-020-01881-x>
Ethics ID: Not required.
Published by Springer, on June 2020.

Article II

Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing*, pages 1–17, 2021. URL <https://doi.org/10.1109/TAFFC.2021.3100868>
Ethics ID: 1955051.1, The University of Melbourne Human Ethics Advisory Group.
Published by IEEE, on July 2021.

Article III

Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, ICMR '22, page 562–570, New York, NY, USA, 2022. Association for Computing Machinery. URL <https://doi.org/10.1145/3512527.3531385>

Ethics ID: Not required.

Published by ACM, on June 2022.

I would like to acknowledge the contribution of the following sources of funding that supported my research:

- Melbourne Research Scholarship
- Australian Research Council Discovery Project (DP190102627)

Article I - III received minimal editorial editing in the regular publication process. No other third-party editorial assistance was provided in the preparation of the thesis. The study presented in Article II received ethics approval from The University of Melbourne Human Ethics Advisory Group under the ethics applications listed above.

Acknowledgements

First and foremost I am extremely grateful to my principal supervisor, Dr. Jorge Goncalves, for your invaluable advice, continuous support, and patience during my Ph.D. study. Your wise guidance helped me complete my academic research and writing of this thesis. At our regular meetings, you can always provide valuable feedback and guide me to see things holistically and dialectically. You are not only a trusted advisor and mentor, but a friend in my everyday life. Whenever I feel anxious, you are always there with a sympathetic ear. I could not have imagined my research life without your advice and support. You made me into a better person and a qualified researcher.

I am also thankful to the rest of my supervisors, Dr. Tilman Dingler and Dr. Greg Wadley, for their insightful comments and encouragement, as well as for the profound questions that prompted me to expand my research from all perspectives. It was your kind help and support that made my study and life in Australia a wonderful time. Besides my supervisors, I would like to extend my sincere thanks to my advisory committee chairs, Dr. Rui Zhang and Prof. Lars Kulik, who generously provided knowledge and expertise. Many thanks to Prof. Vassilis Kostakos, who leads the whole team to achieve a common goal.

My sincere thanks also go to Dr. Zhanna Sarsenbayeva and Dr. Benjamin Tag. Thank you for always being there when I needed support. Thank you for your selfless help with experimental design and paper revision. I would also like to thank Dr. Yue Gu, who brought me into the world of affective computing and deep learning. Thank you for the valuable discussions we had during my entire Ph.D.

I am very fortunate to work with the amazing team at The University of Melbourne. Many thanks to everyone in the HCI group: Weiwei Jiang, Difeng Yu, Qiushi Zhou, Brandon Syiem, Dr. Danula Hettiachchi, Dr. Senuri Wijenayake, Dr. Joshua Newn, Dr. Chu Luo, Sara Khorasani, Gabriele Marini, Henrietta Lyons, Nattapat Boonprakong, Ebrahim Babaei, Dr. Unni Krishnan, Dr. Andrew Irlitti, and Dr. Eduardo Velloso. It has been great fun working with you all! Chaofan Wang, my best friend, in particular thanks for your unconditional support and friendship. I really enjoyed our every-week food exploration. Mo Zhang, I have been so pleased to have you watch evening movies together. Jing Wei, thank you for your support and advice during my job hunting. I also thank Dr. Nicole Barbee and Allen Mari Pilaes. Thank you for your assistance during my Ph.D. study.

Besides, I would like to acknowledge The University of Melbourne for the financial support during my Ph.D. research.

Lastly, I would be remiss in not mentioning my family, especially my mom, my dad, and my fiancée (Yuxuan Huang). Without your tremendous understanding, unconditional love, and encouragement over the past few years, it would be impossible for me to complete my study. Yuxuan, thank you so much for all your sacrifices, love you forever!

Thank you!

Melbourne, August 2022

Kangning Yang

Contents

Abstract	i
Declaration	ii
Preface	iii
Acknowledgements	v
1 Introduction	1
1.1 Motivation	1
1.2 Contribution	3
1.2.1 Research Questions	3
1.2.2 Role of the Author	5
1.3 Thesis Outline	5
2 Background	7
2.1 Emotion Related Concepts	7
2.1.1 Emotion Generation	7
2.1.2 Discrete vs. Dimensional Emotion Model	8
2.2 Literature Selection	10
2.3 Overview of the Literature	11
2.3.1 Perception	11
2.3.2 Learning	25
2.3.3 Inference	43
3 Methodology	57
3.1 Research Approach	57
3.2 Data Analysis	59
3.3 Ethical Considerations	60
3.4 Limitations	61
3.5 Summary	63
4 Benchmarking Commercial Emotion Detection Systems	64
4.1 Introduction	64
4.2 Article I	65

5	Hybrid Multimodal Network	85
5.1	Introduction	85
5.2	Article II	86
6	Hierarchical Self-Attention Network	104
6.1	Introduction	104
6.2	Article III	105
7	Discussion	115
7.1	Mobile Emotion Sensing Systems	115
7.2	Benchmarking Commercial Mobile Emotion Recognition Systems	117
7.3	Developing Multimodal Mobile Emotion Recognition Systems	118
7.4	Open Challenges	120
7.4.1	Modality Choice	120
7.4.2	Ground Truth Establishment	121
7.4.3	Real Environment	123
7.4.4	Ethics, Privacy & Security	124
8	Conclusions and Future Directions	127
8.1	Summary of Contributions	127
8.2	Future Directions	128
8.3	Final Remarks	130
	Bibliography	131

Chapter 1

Introduction

1.1 Motivation

Emotions are the affective aspects of consciousness evoked by sensory stimuli, typically accompanied by physiological and behavioral changes in the body [4]. They play a crucial role in guiding people’s responses to events and situations, and impact decision-making, learning, communication, and situational awareness [5]. However, emotions can go awry and lead to negative repercussions on personal well-being. For example, long-term accumulation of negative emotions can not only lead to low mood or loss of interest in daily activities but can also induce mental health disorders such as depression and psychosis [6, 7].

Traditional psychological assessment methods usually depend on diaries, interviews, and questionnaires [8, 9]. However, these methods rely heavily on respondents’ retrospective recall and subjective assessment of affective variability or reactivity [10], the reliability of which cannot be guaranteed [11, 12]. Furthermore, these diagnostic measures can also be burdensome for respondents and may disrupt their daily activities [13]. Therefore, passive monitoring and automatic detection of affective states have become increasingly important.

Over the last few decades, researchers have attempted to empower computers with the ability to perceive human emotional states. A variety of data sources including facial expressions, voice, body gestures, and linguistics have been leveraged in research and

industry for emotion recognition [14]. Despite remarkable progress, there remain considerable challenges before current affect detection techniques can be effectively deployed in real-world settings. For instance, much of the existing work is limited to controlled scenarios, such as experiments conducted in laboratories or workplaces with sophisticated recording systems. Furthermore, the equipment used in these studies tends to be intrusive and expensive. For example, an electroencephalogram (EEG)-based system requires electrodes to be attached to an individual’s scalp, while an electrocardiogram (ECG)-based system requires sensors to be placed around an individual’s chest [15, 16]. These requirements not only reduce their applicability to real-world scenarios but also hinder participants from generating truly natural emotional responses. In addition, the lack of adequate realistic training data is an ongoing challenge, which further limits the development of emotion sensing techniques that can be used in the real world [17].

Fortunately, modern mobile devices are becoming increasingly low-cost and sensor-rich, providing new opportunities for solving the aforementioned challenges. Around the world, the number of smartphone users in 2022 is 6.648 Billion, accounting for 83.72% of the global population [18]. Mobile devices have truly become accepted by the general public as an indispensable part of daily life. This makes these devices an ideal platform to passively and continuously record users’ spontaneous responses and contexts without spatio-temporal limitations [19–21]. Compared with more traditional data acquisition devices (e.g., eeg-based or ecg-based systems), mobile devices are more ubiquitous and unobtrusive.

Moreover, data acquired by sensors of mobile devices are ecologically valid, have high-resolution, and span multiple modalities [22]. Apart from the frequently used behavioral and physiological signals mentioned above, mobile devices can collect a variety of contextual information such as location, audio environment, collocation with other devices, ambient lighting, or weather conditions [23]. This rich modality set can be efficiently exploited by multimodal emotion recognition systems delivering greater accuracy, robustness, and generalization, further promoting the development and deployment of emotion recognition systems in practical applications. Compared with unimodal approaches, multimodal systems can model human affective expression with higher fidelity. Importantly, the mapping between expressions and emotional states is generally a loose relationship [24, 25]. Hence, the multi-channel information is able to complement each other and thus forms a more complete and reliable flow for decision making [26]. This

also closely mimics the way humans perceive emotions, as we often receive information from multiple sources rather than a unimodal data source. For example, during conversations, we can better understand the intentions of the speaker by observing the expressions on their face while also focusing on what is being said [5].

These affordances and advantages of mobile devices open up a wide range of potential applications for mobile-based emotion recognition, covering areas such as healthcare [27], biometrics [28], education [29], and gaming [30]. For example, regarding education, emotion-based mobile applications can be used to recognize interest and boredom emotion states of students in online learning, further helping improve education quality, and strengthening the emotional interaction between teachers and students. With regard to mental health monitoring, mobile-based emotion detectors can allow clinicians to remotely monitor the emotional trajectories of patients, so as to gain a more comprehensive understanding of a patient’s ongoing psychological state and help improve the effectiveness of a treatment program. In this thesis, we define emotion recognition/sensing techniques that rely on mobile devices as mobile emotion recognition/sensing. We focus on robust mobile emotion recognition, adopting a multimodal structure to fully exploit the diversity and quantity of data collected from mobile devices, and exploring different approaches to build accurate recognition systems. We believe our work will facilitate the development of multimodal mobile emotion recognition frameworks and the deployment of emotion detectors into real-world contexts.

1.2 Contribution

1.2.1 Research Questions

Based on the research gaps identified above, the broader research question answered through this these is:

RQ: How can we build robust multimodal emotion recognition systems using mobile devices?

More specifically, I dissected this overarching research question into three main questions and sub-questions:

RQ1 What are the recent advances in emotion recognition using mobile devices?

- 1.1 What are the signals/modalities that can be collected from mobile devices?
- 1.2 What are the recent successful methods employed in this research area?
- 1.3 What can be further improved in mobile emotion recognition research?

RQ2 What are the limitations of current off-the-shelf commercial mobile emotion recognition systems that leverage facial features?

- 1.1 How can we objectively evaluate the performance of these commercial mobile emotion recognition systems?
- 1.2 What are the strengths and weaknesses of these emotion recognition systems when used in non-ideal scenarios?

RQ3 How can we take full advantage of the capabilities of mobile devices to develop robust and accurate emotion recognition systems?

- 1.1 How can we create suitable architectures that leverage the rich set of modalities that originate from mobile devices?
- 1.2 How can we efficiently extract informative features representing human affects from multiple inputs?

These research questions are investigated through a literature review and three research studies. First, we systematically investigated previous efforts for achieving effective emotion recognition using mobile and wearable devices. We divided mobile emotion recognition frameworks into three stages, namely perception, learning, and inference, and analyzed existing successful methods from these three perspectives (RQ1). Second, there has been an increase of available commercial platforms offering emotion recognition services in the market recently, particularly those that leverage facial features. As such, we selected five commonly used facial emotion recognition systems and established a replicable assessment method to evaluate and visualize their potential for use in real-world scenarios ([Article I](#)). We focused on facial expression-based systems in this study for two reasons: (1) most commercial services are based on facial emotion recognition; (2) based on our findings in the literature review, facial expression is still one of the widely used modalities when building multimodal systems, despite the reliability of facial expressions having come into dispute recently [31]. In addition, our findings

show that each system has its own strengths and limitations under different distortion conditions (RQ2). Regarding RQ3, we focused on multimodal learning, investigated different combinations of input modalities, and developed a hybrid deep learning framework to predict human emotions based on off-the-shelf mobile devices ([Article II](#)). We show that our system enables accurate and unobtrusive emotion detection, well above state-of-the-art performance reported in the literature. Furthermore, we introduced a novel end-to-end emotion recognition system ([Article III](#)), which focused on multiple physiological signals from low-cost wearable devices. The approach uses self-attention and convolution mechanisms to extract both global and local associations from multimodal physiological signals. The results show that the proposed framework outperforms state-of-the-art methods on published datasets and has good generalization ability.

1.2.2 Role of the Author

Three publications are included as main contributions in this thesis presented in Chapters [4](#), [5](#), and [6](#). They are published in leading, international, and peer-reviewed conferences/journals in the field of Affective Computing and Human-Computer Interaction.

I was the lead author for all the publications presented in this thesis and performed the majority of the work (over 50%). I provided the primary ideas behind the work, initiated the research, prepared the experimental and model designs, and performed programming tasks, technology deployment, participant recruitment, and data analysis. Additionally, I was the corresponding author for these publications. I managed the process of the article submissions, and revisions based on the received peer-reviews. I received valuable advice and feedback from my co-authors on designing and conducting the study and analyzing the data. Moreover, my co-authors contributed to the writing of these publications. Therefore, in the main chapters of this thesis, I use the scientific term “we” to reflect and thank my co-authors for their contributions.

1.3 Thesis Outline

The following sections of this thesis are organized as follows. Chapter [2](#) introduces the theoretical concepts of emotional models derived from psychology that are widely used

in mobile emotion recognition research. It then surveys previous work from three fundamental steps of a typical emotion recognition system (perception, learning, inference) and presenting current research methods, trends and shortcomings in this field.

Chapter 3 introduces the methodology adopted in the studies presented in the thesis (Article I–III). This chapter provides motivations behind methodological decisions for experimental design, model construction, data analysis, and ethical analysis, as well as a discussion of the limitations of these methods.

Chapter 4 presents a study establishing a replicable protocol for evaluating commercial facial emotion recognition systems and their appropriateness for use in real-world mobile contexts.

With the goal of developing robust mobile emotion recognition systems, Chapter 5 and 6 present two studies exploring the use of multimodal information and state-of-the-art deep neural networks as frameworks for emotion recognition. These two works leverage multimodal signals collected from off-the-shelf mobile and wearable devices, and demonstrate generalizability and state-of-the-art performance.

In Chapter 7, we reflect on the research contributions of the work presented in this thesis, and discuss the results and findings with regard to the research questions. Furthermore, we present the open challenges and future research directions of mobile emotion recognition. Finally, Chapter 8 concludes with a summary of the thesis.

Chapter 2

Background

In this chapter, we provide an overview of previous work on mobile emotion recognition research in the literature. We first introduce the terminology and concepts that are critical to understanding different emotion theories in Section 2.1 - including the discrete emotion model and the dimensional emotion model, and how to classify and process them from a technical perspective. In Section 2.2, we describe the search strategy and scope of this review, including the keywords in the search query and the criteria for selecting prior work relevant to this thesis. Section 2.3 details the three steps of a typical mobile emotion recognition framework and presents it in a way that readers can identify key perception, learning, and inference methods from the literature.

2.1 Emotion Related Concepts

2.1.1 Emotion Generation

According to the American Psychological Association (APA), an emotion is “a complex reaction pattern, involving experiential, behavioral, and physiological elements, by which individual attempts to deal with a personally significant matter or event”. The “modal model” of emotion (see Figure 2.1) explains that emotion generation begins with the occurrence of an eliciting stimulus, which after attention and appraisal of significance can give rise to an emotional response involving subjective experience, physiology, and behaviour [32, 33]. Human emotional responses are usually adaptive and guide people’s

responses to threats and opportunities in the environment [34]. For example, negative emotions may promote involuntary bodily responses that can minimize potential harm (e.g., feeling fear promotes increased blood flow in the extremities which can help people avoid or escape danger). Similarly, positive emotions can promote communication to strengthen social relationships or promote novel thought and action which serves to build personal resources [35].

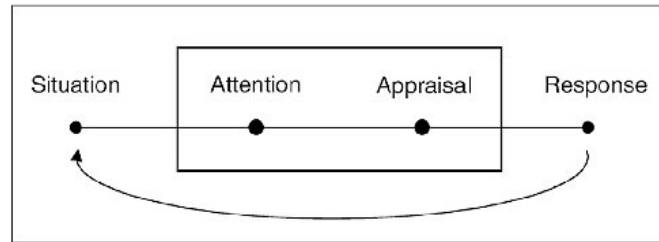


FIGURE 2.1: The modal model of emotion (from [36]).

While the terms “emotion” and “mood” are often used interchangeably by non-experts, psychologists define emotion as a usually “short and intensive” response to a particular event or person [37], while a mood is typically less intense and longer lasting [38]. Moods can continue for hours, days, or even weeks, and usually lack a definite stimulus [39]. For example, an angry emotion might be elicited by an insult, while an angry mood might arise slowly from a long-term dissatisfaction.

2.1.2 Discrete vs. Dimensional Emotion Model

The range of emotions is wide. It includes not only simple and easily recognizable categories such as fear, disgust, sadness, joy, anger, and surprise, and background affective states like enthusiasm or discouragement, but also more complex states such as embarrassment, shame, guilt, contempt, compassion, and admiration [40]. However, because of the subjective and complex nature of emotions, they are in fact difficult to define and describe accurately [41]. Psychologists have proposed two fundamental approaches to the classification of emotions: (1) the discrete emotion model and (2) the dimensional emotion model.

The discrete emotions model emerged from empirical studies by Ekman on cross-cultural emotion recognition, and evolutionary theory on the adaptive function of emotions [42], and beyond this from Darwin’s [43] empirical and theoretical work on emotions. This tradition holds that there are several distinct emotions that evolved to solve specific

problems of survival and reproduction and which are accompanied by particular expressive and physiological signals which can be measured and used to recognize emotions. In their cross-cultural work, Ekman and colleagues found that some facial expressions of emotion appear to be universally recognized, and identified six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [44]. After continued cross-cultural studies, Ekman clarified that each basic emotion is not a single affective state but a *family* (i.e., a group of things related by common characteristics) of related states [45]. In each emotion *family*, there are two components, a *theme* and *variations*, in which a *theme* refers to the core elements of the *family* or the shared characteristics of the *family* members, while *variations* reflect various influences including individual and contextual differences [32]. However, the existence of cross-cultural emotion prototypes has been challenged by other studies [31, 46–48]. Based on Ekman’s research, Robert Plutchik proposed the “wheel of emotions” that groups eight primary emotions into four pairs of polar opposites: joy versus sadness, trust versus disgust, fear versus anger, and surprise versus anticipation [49]. All other emotions are derived from these primary emotions. For example, optimism is a compound of anticipation and joy. In addition, each emotion can be expressed with varying degrees of intensity.

The dimensional emotion model emerged from critiques of discrete emotion theory which propose that emotions are constructed in the moment from underlying components, one of which is core affect, a subjective feeling composed of one, two or three underlying values [50–52]. Thus different emotions can be represented as different positions in a one, two or three dimensional space, with two being the most commonly used. The circumplex model of affect proposed by James Russell is one of the typical representatives [53]. It proposes that all affective states arise from two independent neurophysiological systems, in which one relates to valence describing an unpleasant–pleasant continuum while the other relates to arousal describing a deactivated–activated continuum [54]. Each emotion can be defined as a linear combination of the valence and arousal systems. Alternatively, the positive affect - negative affect (PANA) model [55] is defined by two primary unipolar dimensions: the horizontal axis scales low to high negative affect (NA), and the vertical axis scales low to high positive affect (PA). These two dimensions are characterized as “descriptively bipolar but affectively unipolar dimensions”, which means that although NA and PA may seem to be two poles of one dimension, they are in fact two independent and uncorrelated unipolar constructs [56]. The PANA model can be understood as a

45-degree rotation of the circumplex model, in which the NA and PA are predominantly defined by their high ends or activated ends. In turn, the low ends of NA and PA are typically considered to be neutral due to a lack of a specific type of activation [52, 57].

One three-dimensional model, the Pleasure-Arousal-Dominance (PAD) model, has been widely used [58]. Here, emotional states are defined using three independent and bipolar dimensions: pleasure–displeasure, arousal–nonarousal, and dominance–submissiveness. The extra dominance dimension measures the individual’s control over their emotions, which can more effectively distinguish some emotional states: for example, while fear and anger both have high displeasure and high activation, they are opposite in the dominance dimension since anger is a dominant emotion while fear is a submissive emotion.

From a neuroscience perspective, the different models of emotions result from different understandings of emotional mechanisms. The discrete emotion theory proposes that specific emotional processes arise from distinct neurobiological modules [59] so that each specific emotion can map to one discrete and independent neural system. Conversely, the dimensional emotion theory argues that there are common, underlying neurophysiological systems from which emotional states emerge [53, 54]. Despite ongoing debate, both discrete and dimensional emotion models have their respective strengths for research in affective computing. Discrete models are suited to classification problems, are easier for emotion labeling and more accordant with people’s description of emotions in daily life. Dimensional models can capture fine-grained emotional concepts with subtle differences and measure the similarity between emotional states; however, they are less intuitive for research participants to understand. In sum, both discrete and dimensional models have been widely applied in different tasks in the field of mobile emotion sensing: discrete models are mainly utilized in classification tasks with discrete emotion labels while dimensional models are mainly used in classification/regression tasks with continuous emotion labels [60]. We use both emotion models in the studies (Article I, II & III) presented in this thesis.

2.2 Literature Selection

We conducted a bibliographic search in three established digital libraries that contain the bulk of all mobile emotion sensing research: the *ACM Digital Library*, the *IEEE*

Xplore Digital Library, and *Springer Nature*. We used the same search queries, intended to capture related research on mobile emotion sensing: [(“mobile device” OR “wearable device” OR “mobile devices” OR “wearable devices” OR “smartphone” OR “wristband” OR “smartwatch” OR “smartphones” OR “wristbands” OR “smartwatches”) AND (“emotion detection” OR “emotion recognition” OR “emotion prediction” OR “affective state detection” OR “affective state recognition” OR “affective state prediction” OR “emotion monitoring”)].

We excluded papers that were not full-text research articles, such as tutorials, abstracts, workshops, posters, etc. We also applied a time filter in order to consider only publications between January 1, 2005 and April 30, 2022 (in order to put the focus on recent mobile devices). After this process, a total of 726 papers remained. We then analyzed each of the papers to ensure they were appropriate for this survey. We found that some papers were not relevant to our topic, since they were either not directly concerned with mobile emotion sensing research or did not feature original sensing technical research but instead utilized commercial services (e.g., Amazon Rekognition or iMotions) to conduct their studies (e.g., to enhance the user experience or design affective user interfaces). After excluding these papers, 150 papers remained for further analysis.

Furthermore, to help readers have a clear understanding, we describe and summarize research papers from three stages that form the overall architecture of a typical mobile emotion sensing framework: signal **perception**, feature representation **learning**, and emotion **inference** (shown in Figure 2.2, the vertical solid arrows in the figure indicate “includes”).

2.3 Overview of the Literature

2.3.1 Perception

Perception is the first stage in a mobile emotion sensing framework. The purpose of this stage is to collect emotion information through different types of sensors embedded in mobile devices. As stated in [61], a complete emotional experience is made up of three distinct components: a subjective experience, a physiological response, and a behavioral or expressive response. Each emotion includes an inner experience that can also be

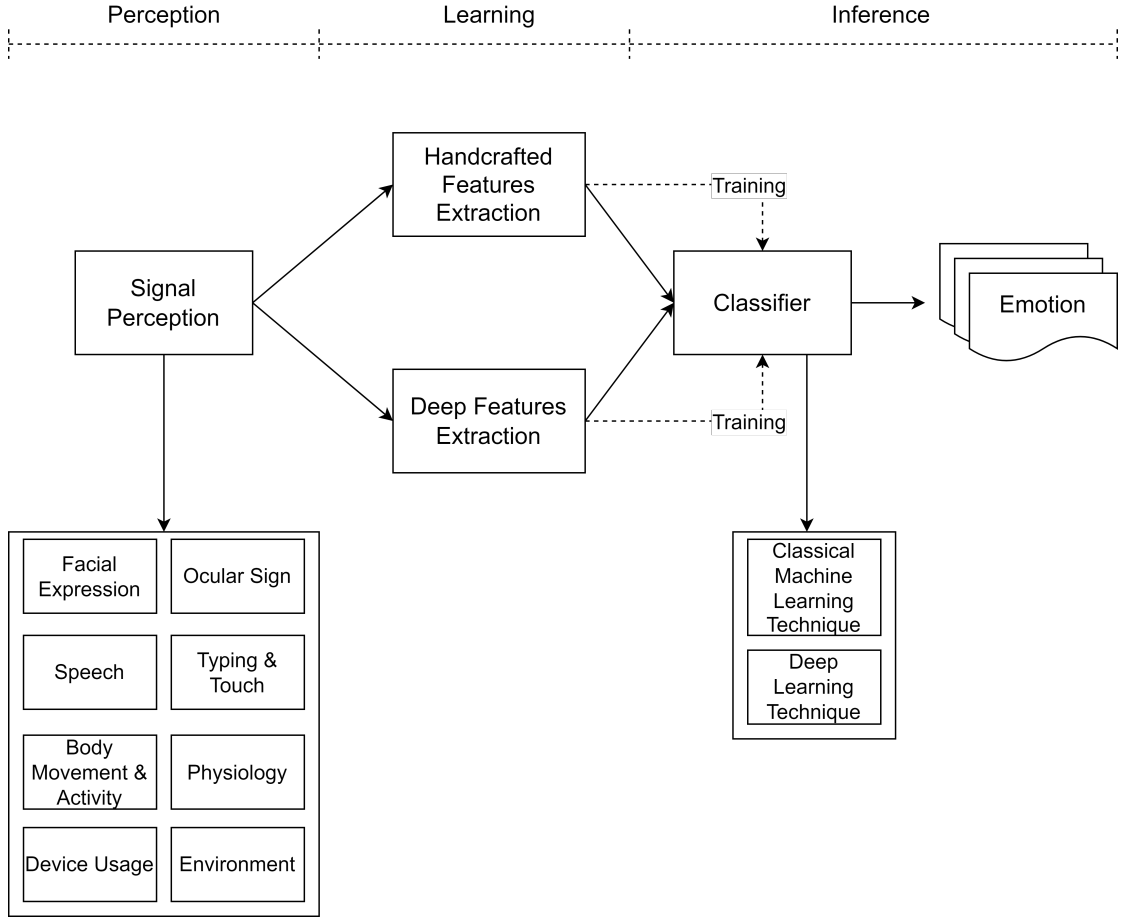


FIGURE 2.2: Mobile emotion sensing framework. In **perception**, we classify different modality data into eight categories; In **learning**, we classify current feature extractors into two categories; In **inference**, we discussed two categories of inference methods.

understood as a response and that is highly subjective. For example, when losing a loved one, some people may be full of remorse while others may be deeply saddened. While subjective experience cannot be directly sensed, physiological and behavioral responses (e.g., blushing, sweating, increased heart rate, or facial and body movements) generate data containing significant clues to the emotion being experienced. Collecting these signals and analyzing them to deduce the underlying emotional state is the basis of the emotion sensing framework.

With the development of mobile technology, most off-the-shelf mobile devices are equipped with a rich set of powerful sensors. Smartphones typically include an accelerometer, gyroscope, GPS, microphone, and camera, while today's smart wristbands offer photoplethysmograph (PPG), electrodermal activity (EDA) and temperature sensors. Due to the growing acceptance of these devices by the general public, it has become increasingly easier and more feasible to collect granular emotion data from individuals,

at larger scales than was previously possible. For example, the accelerometer allows the phone to sense the user’s activity states (e.g., walking, standing, or sitting); the GPS sensor allows the phone to determine the user’s location; the camera and microphone allow the phone to record the user’s facial expressions and vocal utterances; and the PPG and EDA sensors allow devices to detect the user’s physiological states. Advances in mobile device design are often accompanied by the introduction of new sensors: examples include the proximity sensor, used to detect the presence of nearby objects without physical contact, and the iris sensor, used for biometric recognition of users.

2.3.1.1 Emotion Sensing Modalities

Mobile devices can receive and log a breadth of information regarding users’ affective responses via on-board sensors, and use this as input for a given emotion sensing framework. In this section, we provide a brief overview of the different input data used in mobile emotion sensing. We classify this data into eight categories: facial expression, ocular sign, speech, typing & touch, body movement & activity, physiology, device usage, and environment (shown in Figure 2.2). Here, we use the term “modality” to include automatic input data (e.g., through sensing), manual/logged human input data (e.g., keystrokes), and the source type of data.

Facial Expression Previous work has highlighted that facial expression is a primary cue for understanding human emotions [62]. It is often possible to infer a person’s emotional states through reading their facial expression, especially during social interactions. For example, a smiling face may mean the person is happy, while a gloomy face may indicate they are experiencing negative feelings [45]. As a primary non-verbal channel for expressing emotions, facial cues have a long research history. Currently, still and video images are two principal data sources for analyzing facial expressions. These visual data reflect the movements of facial muscles or muscle groups over a short duration when an emotion is triggered. By calling different functions of the camera, mobile devices such as smartphones or tablets can capture their users’ facial expressions intermittently or continually. For example, Kosch et al. [63] used the front-facing smartphone camera as a tool for emotion detection based on facial expressions. In order to save computing power and reduce battery consumption, their application only recorded facial expressions when the user’s face was turned towards the smartphone screen.

Ocular Signs Different eye-related signals, such as pupil diameter, gaze distance, and eye blinking can be broadly categorized as ocular signs. They can be understood as special forms of facial expressions, but are relatively more difficult to observe and distinguish. This is because, compared to whole-face images, eye-region images cover a smaller area and contain less information on emotion-related facial changes and muscle movements [64]. Notwithstanding, as a form of natural interaction, ocular signs carry abundant information regarding cognitive activities [65, 66], and a number of studies have demonstrated that it is feasible to infer users’ emotional states from analyzing pupil signals [67, 68]. In the mobile emotion sensing domain, ocular data is typically measured using wearable eye trackers. For instance, Xing et al. [29] used Tobii glasses to collect users’ pupil diameter variation signals and applied the proposed emotion sensing method in MOOC education.

Speech One of the most natural means of human communication is speech. Similar to facial expression and ocular sign, speech can also transmit emotion information. During a conversation, speakers can easily integrate their emotions into their prosodic and acoustic characteristics. For example, when feeling sad, speech is often slow, low-pitched, and with little high-frequency energy; when feeling angry, produced speech is often fast, high-pitched, and with strong high-frequency energy [69]. Speech has been widely used in emotion sensing research, with many techniques and systems developed [70, 71]. In recent years, researchers began to consider mobile platforms, using the smartphone’s built-in microphone to capture human speech in diverse acoustic environments. Chang et al. [72] proposed AMMON (an emotion and mental health monitor), which is a speech analysis library for mobile phones. In another example, Lane et al. [73] presented Deep-Ear, a mobile audio framework built from Deep Neural Networks (DNNs), to support a variety of audio tasks (e.g., emotion sensing, speaker identification) in dynamic environments.

Typing & Touch Multiple studies have shown that users can manifest emotional signals during typing and touching behaviors when using interactive systems [74–77]. The ubiquity of touchscreen-based mobile devices and the increasing use of chat apps has made it more convenient and feasible to capture mobile-based interaction pattern data. For example, Lv et al. [76] used typing biometrics produced by a pressure sensor

keyboard to recognize six emotions. Similarly, Epp et al. [78] determined users' emotions by analyzing the rhythm of their typing patterns on a standard keyboard. Regarding touch interaction, Gao et al. [79] conducted a study to detect players' emotional states by using their touch behaviors during gameplay with an iPod Touch. Focusing on Android-based smartphones, Ghosh et al. [80] developed the TapSense application and recorded typing-related metadata to classify four emotional states (happy, sad, stressed, and relaxed). Specifically, they leveraged information such as the timestamp of when each tap event occurred, and the type of key input (e.g., alphanumeric keys, delete).

Body Movement & Activity Currently, research on sensing emotion through body movement and activity is focused on gesture, posture (body and head), body motions (e.g., gait patterns), and physical activities (e.g., walking, running, sitting, standing, and sleeping). A significant portion of this movement can happen unconsciously and unintentionally and is therefore not easy to disguise. Movement sensing also has the advantage that it involves only minimal or no disruption to users since motion features like acceleration, speed, and orientation can now be extracted from mobile device sensors without using sophisticated capture suits. For instance, Purabi et al. [81] utilized eSense, an in-ear multisensory stereo device equipped with an accelerometer and gyroscope designed to capture the underlying connection between head movements and corresponding traits and emotions. Using human gait signals, Hashmi et al. [82] proposed a method to identify emotions by means of body-mounted smartphones.

Physiology Another important source of signals that can reflect emotional states is physiology. These are objective signals involving electrical and hemodynamic activities of the nervous system [83]. The human nervous system consists of two main parts: the central nervous system (CNS - brain and spinal cord) and the peripheral nervous system (PNS - connecting the CNS with the limbs and organs). Affective neuroscience has long worked on exploring the latent links between emotional changes and activities of the nervous system, especially the activities of the CNS and autonomic nervous system (ANS - a division of the PNS conducting impulses from the CNS to cardiac muscles, smooth muscle, and glands, and which is thus in control of the fight-or-flight response) [84].

Commonly used signals include ECG, EEG, PPG, EDA, and skin temperature (SKT) [85]. These can be objectively measured via biosensors and are more difficult

than other types of signal to consciously conceal or manipulate as they are largely involuntarily activated [86]. In the mobile sensing field, Zhao et al. [87] presented an emotion sensing system based on a wearable wristband by leveraging Blood Volume Pulse (BVP), EDA, and SKT information. In another example, Jiang et al. [88] designed Memento, an emotion-driven lifelogging system on smart glasses, which senses the emotional changes of users through analysing EEG signals.

Device Usage Generally, device usage data can be divided into three classes: contact data, content data, and application data. Contact data reflects a user’s social connections, for which phone calls, text messages (SMS), and emails are the main data source. From these data, researchers can count the frequency of interactions that users have with their social contacts, such as the duration of each call or the number of SMS received and sent. Content data includes the text and emojis within messages such as SMS or email, as well as browser-related data like search history and bookmarks. Regarding application data, it is impossible to give a detailed account of each application since there are so many. Instead, common practice is to group applications into categories (e.g., Entertainment, Finance, Productivity, Social, Travel, Weather), and then analyse how users interact with these categories (e.g., time spent, launch frequency).

In an example that leveraged all three classes of device usage data, Sun et al. [89] proposed iSelf, a system that can automatically infer emotions from smartphone data primarily on SMS, calls, browser, and application usage data. Similarly, LiKamWa et al. [90] presented a smartphone system called MoodScope which recognizes users’ affective states based on usage data (SMS, calls, emails, web visits, application usage). Support for this approach comes from recent research which found a bidirectional relationship between user emotions and application use [91, 92]. Not only does the use of certain applications drive user emotions, but emotions tend to drive the use of particular applications [91].

Environment Environmental factors are known to influence the subjective feeling of users. However, unlike other modalities, environment-based sensor data is rarely used independently to infer users’ emotional states. In most cases, this type of data acts as an auxiliary signal to help in improving the performance of other emotion classifiers. In the mobile emotion sensing field, the environment modality is mainly related to contextual

information including monitoring sensor data from WiFi, GPS, Bluetooth, microphones and light sensors. WiFi data carries information on indoor position; GPS data carries information on outdoor position; the Bluetooth monitor can be used to detect other Bluetooth devices that are proximal; microphone audio reflects environmental noise levels (which plays a different role to speech), and the light sensor represents ambient illumination and can be used to infer how long an individual stays indoors or outdoors.

In an example of this approach, Lee et al. [93] classified users' emotions on mobile devices through their typing & touch characteristics and passive environment data (light sensor, temperature, weather, time, location). Similarly, Zhang et al. [94] leveraged environmental sensor data (microphone audio, light sensor, GPS, WiFi), smartphone usage patterns, and user activities to develop the MoodExplorer system for automatic emotion sensing.

Multimodality Multimodal emotion sensing means detecting affective expressions through the fusion of multi-sensor information [26, 95]. Compared to a unimodal approach, multimodality has several distinct advantages, such as providing higher fidelity models of human emotion and more continuous detection capability. This approach also benefits from people's tendency to express their emotions multimodally [5]. Multimodal systems do not ordinarily suffer from missing data problems, as if some signals are temporarily unavailable, the system can still rely on others.

Traditional multimodal emotion sensing tends to focus on collecting and processing audio, visual, and audio-visual data [95, 96]. In the mobile emotion sensing field, given the diversity of mobile and wearable devices, there is greater flexibility to combine different modalities such as physiology-body movement/activity, visual-physiology, visual-text, etc. Furthermore, some modalities (e.g., contextual information) that were rarely leveraged in traditional systems can be used more frequently. In an example study, Li and Sano [97] investigated the possibility of passive sensing and forecast of well-being based on the fusion of physiological and behavioral (human movement and sleep patterns) information collected by wrist-worn sensors. In another study, Chong et al. [98] proposed EmoChat, an online chatting application for mobile devices which combines facial expressions and text messages for emotion sensing.

2.3.1.2 Summary

Table 2.1 summarises research on the perception phase of mobile emotion sensing, described by the *modality* inputs, measurement *device* types, and *sensor* used (including the research work presented in this thesis). We only include papers that provide a clear description of the above items.

Overall, we find that more research has been conducted using a unimodal approach within the mobile emotion sensing field. This is because signals using the same modality usually have a similar data format and distribution, while also having lower complexity and computation power requirements, making one modality easier to work with. This is particularly true for earlier mobile devices, due to their limited processors and storage, as they were not well-suited to handle complex models. However, more recently researchers have recognised that this approach suffers from some notable problems. As D’mello and Kory [26] mentioned, the challenges of unimodal approaches are two-fold: missing data and reliability problems (e.g., human facial expression can be controlled consciously and disguised voluntarily as it belongs to the semi-voluntary response [99]). Additionally, signal noise seriously affects the performance of unimodal systems, particularly for speech-based sensing which involves inevitable noise from the ambient environment. In order to overcome these challenges, multimodal approaches have drawn increased attention from the mobile emotion sensing research community. Importantly, today’s mobile devices are more easily programmable and have much higher processing capabilities, which has accelerated the move towards multimodal approaches.

It is also worth pointing out that: (1) among the unimodal approaches, *Physiology*, *Facial Expression*, and *Typing & Touch* are the most commonly used modality inputs, while *Ocular Sign* and *Environment* are the most rarely used; while (2) among multimodal approaches, *Body Movement&Activity + Physiology*, *Facial Expression + Physiology*, *Facial Expression+Speech*, and *Body Movement & Activity+Device Usage+Environment* are the most commonly used modality combinations. We note that *Ocular Signs* have not been used often as the equipment needed is relatively more expensive and professional (and may even need to be made by the researcher) when compared to other modalities. Furthermore, unlike other modalities offering intuitive information, the data given by *Environment* are mostly latent and involve contextual information, and therefore, are more suitable for combining with other modalities rather than for

independent use. Overall, for both unimodal or multimodal approaches, *Physiology*, *Facial Expression*, and *Body Movement & Activity* have been widely adopted. This is likely due to their wide availability, and ability to work well both individually and in tandem with other modalities.

TABLE 2.1: Overview of mobile emotion sensing research conducted on the perception phase.

	Modality	Device	Sensor	Reference
Unimodality	Facial Expression	Smartphone/Tablet	Camera	[100–105]
		Eyewear	Photo-Reflective Sensor	[106]
		Kinect	3D Sensor	[107]
		Wearable Facial EMG	EMG	[108–111]
	Ocular Sign	Eye-tracking	Infrared Camera	[64]
	Speech	Smartphone	Microphone	[73, 112–114]
	Typing&Touch	Smartphone	Touchscreen	[80, 115–118]
			Touchscreen, ACC	[119]
			Touchscreen, ACC, GYRO	[120–122]
		Digital Pen, Tablet	Touchscreen	[123]
	Body Movement&Activity	Eearable	ACC, GYRO	[81]
		Smartphone	ACC	[124]
			ACC, GYRO	[82]
		Smartphone, Media Player	ACC, Touchscreen	[125]
	Physiology	Wireless	Radar Sensor	[126]
			RFID	[127]
		Smartphone	Camera	[128]
		VR Glass	EEG	[129, 130]

TABLE 2.1: Overview of mobile emotion sensing research conducted on the perception phase.

	Modality	Device	Sensor	Reference
Unimodality	Physiology	Wristband	ACC, EDA, PPG, SKT	[131]
			EDA, PPG	[132–134]
			EDA, Optical Sensor	[135]
			PPG	[136–138]
			EDA, PPG, SKT	[87, 139]
			Optical Sensor	[140, 141]
		Headset/Headband	EEG	[16, 88] [142–144]
		Chest-worn	ECG	[145]
		Wrist-worn	PPG	[146]
		Headset, Wristband	EEG, EDA, PPG	[147]
		Headband, Wrist-worn	EEG, EDA, PPG	[148]
		Wristband, Chest-worn	EDA, PPG, SKT, ECG	[149]
		Armband, Fingertip, Chest-worn	EDA, ECG, BR, SKT, HF, Pulse	[150]
	Device Usage	Smartphone	Software	[151–153]
			ACC, GYRO	[154]
	Environment	Smartphone	GPS	[155]

TABLE 2.1: Overview of mobile emotion sensing research conducted on the perception phase.

	Modality	Device	Sensor	Reference
Multimodality	Facial Expression + Speech	Smartphone/Tablet	Camera, Microphone	[156–159]
	Facial Expression + Body Movement&Activity	Smartphone	Camera, ACC	[160]
	Facial Expression + Physiology	Smartphone	Camera	[161, 162]
		Webcam, Wristband	Camera, EDA, PPG, SKT	[163]
	Facial Expression + Device Usage	Smartphone	Camera, Software	[63, 98]
	Speech + Body Movement&Activity	Smartwatch	Microphone, ACC, GYRO	[164]
	Speech + Device Usage	Smartphone	Microphone, Software, ACC, GYRO	[165]
	Typing&Touch + Environment	Smartphone	Touchscreen, ACC, Light Sensor, GPS, Web	[93]
	Body Movement&Activity + Physiology	Wristband	ACC, EDA, PPG	[166]
			ACC, EDA, PPG, SKT	[167]
			ACC, EDA, SKT	[97]
		Chest-worn	ACC, ECG, BR, SKT	[168, 169]
		Wristband, Chest-worn	EDA, PPG, SKT, ACC, ECG, EMG, RESP	[170]
		Armband, Chest-worn	ACC, EDA, HF, SKT, ECG, BR	[15]

TABLE 2.1: Overview of mobile emotion sensing research conducted on the perception phase.

	Modality	Device	Sensor	Reference
Multimodality	Body Movement&Activity + Device Usage	Smartphone	ACC	[171]
	Body Movement&Activity + Environment	Smartphone, Smartwatch	ACC, GPS	[172]
	Device Usage + Environment	Smartphone	Software, GPS	[90]
	Ocular Sign + Typing&Touch + Body Movement&Activity	Smartphone	Camera, Touchscreen ACC, GYRO	[173]
	Typing&Touch + Physiology + Body Movement&Activity	Smartphone, Kinect, Wristband	Touchscreen, Depth Camera, EDA, PPG	[30]
	Body Movement&Activity + Physiology + Environment	Smartphone, Smartclothing	ACC, Microphone, ECG	[174, 175]
		Smartwatch	ACC, Pedometer, HR Microphone, GPS	[176]
	Body Movement&Activity + Device Usage + Environment	Smartphone	ACC, Compass, GYRO, Screen, Software, Microphone, Light Sensor, GPS, WiFi	[94]
			ACC, Screen, Software, Microphone, Light Sensor, WiFi	[177]
			ACC, Software, GPS, Bluetooth, WiFi	[89]

TABLE 2.1: Overview of mobile emotion sensing research conducted on the perception phase.

	Modality	Device	Sensor	Reference
Multimodality	Body Movement&Activity + Device Usage + Environment	Smartphone	ACC, GYRO, GPS, Gravity, Pedometer, Light Sensor, Network	[178]
			ACC, GYRO, MAG, Software, GPS, Light Sensor, WiFi, Bluetooth, Screen, Camera	[179]
	Facial Expression + Speech + Body Movement&Activity + Physiology	Webcam, Kinect, Wristband	Camera, Microphone Depth Camera, EDA, PPG	[180]
	Facial Expression + Speech + Typing&Touch + Physiology	Smartphone, Wristband	Camera, Microphone, Touchscreen EDA, PPG, SKT	[2](Chapter 5)
	Body Movement&Activity + Physiology + Device Usage + Environment	Smartphone, Smartclothing	ACC, Software, GPS, ECG	[181]

ACC: Accelerometer; GYRO: Gyroscope; MAG: Magnetometer; RFID: Radio frequency identification; EMG: Electromyography; RESP: Respiration; BR: Breathing; HF: Heat Flux.

2.3.2 Learning

After the information is retrieved from sensors it is forwarded to the second, learning stage. In this stage, large-scale and ruleless raw data are preprocessed, and emotional features of different modalities are learned and represented. As shown in Figure 2.2, there are currently two categories of feature extractors, handcrafted feature-based extractors, and deep feature-based extractors. Handcrafted feature extraction is a two-phase analysis process that relies on traditional statistical analysis (e.g., mean value, standard deviation, Fast-Fourier Transform) to explore emotion-specific characteristics, and optimization analysis to reduce the dimension of calculated features and select the most effective features. By comparison, deep feature extraction uses deep neural network architectures (e.g., AlexNet [182], VGG16/19 [183]) as extractors to learn the inherent distribution of the raw data and further output the corresponding feature representations automatically [84]. Please note that here we classify feature extractors by the description in the original texts.

Overall, this phase aims to constitute a knowledge-based abstraction layer to retrieve the emotional cues or features in the superficial or latent space of collected sensor data. These features should convey significant information that characterizes a person’s emotional state.

2.3.2.1 Handcrafted Feature Extractors

In early studies, researchers mainly worked on handcrafted features in this learning phase [60]. A comprehensive handcrafted learning process can be divided into three steps: signal preprocessing, feature extraction, and feature reduction & selection.

Signal Preprocessing *Preprocessing* prepares the collected raw data before forwarding it to the formal feature extraction, and it includes operations like data filtering (e.g., incomplete and redundant data), artifact removal, and noise filtering [184–186]. As the first step, it focuses on segmenting, formatting, and restructuring raw data from different predefined or preconfigured sensors [187–189]. The objective is to provide high-quality reliable data, while also saving computation resources. There are different preprocessing approaches for different types of modality data. For instance, for video data (e.g.,

facial expression), frame sampler, target localization and alignment are the key contents of the preprocessing flow, which segments continuous video sequences into a series of representative frames [190, 191]; for acoustic data (e.g., speech), audio segmentation and silent region removal are often carried out to break audio data into frames, and there are different primary methods based on different scenarios (e.g., speaker segmentation, utterance-based segmentation, word-based segmentation) [192, 193]; while for sensor data (e.g., physiology), noise and artifact removal are usually used to enhance the reliability of the measurements, using different types of filters such as low-pass or band-pass, and different normalization methods such as Z-score or min-max normalization [194, 195].

Feature Extraction *Feature extraction* is the core step of handcrafted extractors. The aim is to find numerical attributes derived from the initial set of data that can describe its affective information. Handcrafted feature extraction requires manual feature engineering, which in many cases requires a good understanding of the background or domain of the given problem to create effective features. Over decades of research, engineers and scientists have developed various feature extraction methods for different data modalities.

Regarding facial expression, the most common features are based on Action Units (AUs) which define facial actions caused by the contraction of specific muscles, such that each emotion can be interpreted as a combination of AUs. For example, in the system proposed by Ekman and Friesen (Facial Action Coding System) [196], anger can be recognized as a combination of 4 AUs (Brow Lowerer, Upper Lid Raiser, Lid Tightener, and Lip Tightener). Masai et al. [106] leveraged skin deformations caused by the muscle movement of AUs to capture facial expressions, using sensors embedded in eyewear devices to capture skin deformations around the eyes. They could detect most muscle movements related to the target facial expressions including movements of the eyelids, eyebrows, nose, cheeks and mouth. In order to reduce the complexity of the movement encoder of facial muscles and speed up feature extraction, Zhong et al. [197] proposed adopting a relatively simple shape descriptor by using Pyramid Histogram of Oriented Gradients (PHOG) as feature descriptors to represent an image by its local shape and the corresponding spatial layout. An alternative to AUs, facial landmarks (i.e., the location of salient facial regions including mouth, eyebrows, eyes, nose, etc.) are also commonly

used features that can be tracked over time. Kosch et al. [63] and Pham and Wang [162] used OpenFace [198] and Affdex [199] respectively on Android smartphones to leverage the frontal camera to detect facial landmarks as indicators for emotions. Based on facial landmarks, Alshamsi et al. [157] also calculated the center of gravity (COG) of all the landmarks and extracted a characteristic vector to depict the spatial interrelationships between the COG and each landmark point. Furthermore, Suk and Prabhakaran [102] and Seanglidet et al. [103] applied Active Shape Models to obtain better positions of landmarks through iterative fitting. Besides these directly related expression features, some indirect elements related to the video or pictures themselves can also be used as facial features, such as color-related or saturation-related computations like Color Histograms [200] and color energy [201]. Kwon et al. [202] relied on a camera built into a glass-typed wearable device to capture facial expressions and used intensities in each pixel of the facial image as features. In another example, Hossain and Muhammad [101] first converted facial frames into gray scale images and then calculated Local Binary Patterns histograms to compose the feature vector.

In one of the few examples of using ocular sign for mobile emotion sensing, Xing et al. [29] first applied Discrete Wavelet Transform to obtain wavelet features from the denoised pupil diameter (PD) variation, then used statistical methods to further calculate statistical values (Max, Min, Mean, Range, Std, Median, etc.) as a supplement to create PD feature representation. Similarly, Fedotov et al. [203] confirmed the dependencies between eye gaze features and tourist satisfaction levels.

For speech, early psychological studies have found that some vocal parameters, especially pitch (F0), intensity, speaking rate, and voice quality, play an important role in the recognition of emotion and sentiment [204–206]. In the mobile emotion sensing domain, Yu [165] used a straightforward approach of calculating the pitch and volume change on a telephone conversation as acoustic features. Wu et al. [156] took this approach a step further by utilizing combined mel frequency cepstral coefficients (MFCCs) with its first and second order of derivatives as stable acoustic features. In a more robust approach, Lu et al. [113] not only used pitch related statistic values (e.g., standard deviation, the difference of max and min pitch) and MFCCs, but also adopted speaking rate, spectral centroid, high-frequency ratio, and TEO-CB-AutoEnv [207] as acoustic features. More recently, researchers have noted that the algorithms for extracting some of these features are too computationally-intensive for mobile devices. To solve this problem,

Deshpande et al. [208] designed a novel algorithm to extract the multi-dimensional Time Domain Difference (TDD) feature and achieved a reduction of 10% computational cost of extraction compared with MFCCs. Similarly, Provost and Narayanan [209] presented an emotion distillation framework to create emotion-specific features from original high-dimensional feature space in order to reduce computational complexity. Finally, as ambient noise is a major challenge in speech-based emotion sensing, Yang et al. [112] proposed a novel hybrid noise resilient algorithm to obtain the pitch feature in noisy environments and implemented it for Android devices.

Regarding typing & touch, feature extraction is mainly focused on users' touch or stroke behavior. There is a wide range of features that have been used to parametrically represent this physical behavior, such as touch frequency, touch pressure, finger contact area, typing speed, backspace, or special characters press frequency. Lee et al. [93] extracted typing speed, touch count, backspace key press frequency, special symbol press frequency, maximum text length, and erased text length as features to represent interactive communication between a user and a client. Dai et al. [117] focused on the length (distance of finger movement between the start position and end position), time, speed, and pressure of each stroke, and calculated their average, median, maximum, minimum, and variance values as an extracted typing & touch feature vector. Similarly, Ruensuk et al. [120] focused on touch area, pressure, touch count, hold-time, distance, speed, etc., and selected their descriptive statistics (mean, maximum, minimum, and standard deviation) as feature representation. Furthermore, Ghosh et al. [80, 210] noticed that overlapping typing events may occur in two different typing sessions (e.g., when switching one application to another), which are tagged with different emotional states, and thus, designed two different representations of typing speed. As well as interacting with mobile devices through fingers, digital pens or styli can also be used. For example, Han et al. [123] described the typing & touch feature set from a digital handwriting recording, which was calculated from each pen stroke event including the duration in time, stroke length in pixels, speed standard deviation, force average, average altitude angle of the stylus, etc.

For body movement & activity, accelerometer recordings during different kinematic motions have distinct statistical characteristics, enabling the extraction of representative features. For example, Rubin [168] computed the root sum squared over the three axial

accelerations ($\sqrt{Acc_x^2 + Acc_y^2 + Acc_z^2}$), and assumed different activity types (no/low/moderate/high activity) by defining different thresholds. Sun et al. [89] further extended it by taking into account external force, and calculated the acceleration magnitudes as follows: $\sqrt{Acc_x^2 + Acc_y^2 + Acc_z^2} - G(Gravity)$. Through two defined thresholds, they approximated the daily activity state of people. Except for specific daily activities, Adibuzzaman et al. [160] correlated accelerometer data with energy expenditure of a person, and achieved feature extraction by summing time integrals of accelerometer output over the three spatial axes, i.e. $\int_{t_0}^{t_0+T} |Acc_x| + |Acc_y| + |Acc_z| dt$. Di Lascio et al. [166] extracted 11 time-domain statistical features (including minimum, maximum, mean, standard deviation, dynamic change, slope, the absolute value of the slope, the mean and standard deviation of the first and second derivative) from the normalized accelerometer signal as a body-movement feature vector. In a more robust approach, Olsen and Torresen [124] considered both the time-domain and the frequency-domain and extracted features such as mean and standard deviation of acceleration, the standard deviation of mean peak acceleration, and power spectral density. Besides accelerometer output, data from other sensors such as the gyroscope also contain body movement & activity information. For example, Lee et al. [211] calculated time (average, standard deviation, average squared power), frequency (entropy level), and phase (percentage of the angle outside the defined control eclipse, weighting function of the angle outside the defined control eclipse) domain values from an accelerometer and a gyroscope as a head motion feature vector. In another example, Hashmi et al. [82] proposed a total of 29 unique time, frequency, and wavelet features as representations of gait patterns for each walking stride by means of accelerometer and gyroscope data.

BVP, EDA, ECG, and EEG are the commonly used physiological signals in the mobile emotion sensing literature [212]. The BVP signal is measured using a PPG sensor and indicates dynamic changes in blood volume in the peripheral blood flow by transmitting infrared light and measuring its absorption [213]. Feature extraction from BVP often concentrates on the time and frequency domains. For example, Di Lascio et al. [166] considered time-domain statistical features (slope, number of peaks, etc.) of BVP signals along with the mean and standard deviation of BVP pulses' amplitude and length. Meanwhile, as a kind of quasi-periodic signal, Wang et al. [167] and Zhao et al. [87] first utilized fast Fourier transform (FFT) to obtain the BVP signal in the frequency domain, and then partitioned the coefficients to obtain the low-, middle-, and high-frequency

band of the spectrum. Finally, they calculated the mean energy, the maximum energy, and the spectral entropy of each sub-band as features. Moreover, BVP signals can also be used to monitor heart rate (HR) and heart rate variability (HRV), thereby providing heart-related features. Specifically, they can be derived from the raw BVP signal by measuring the interbeat interval (IBI) (the time intervals between the peaks of the waveform). For example, Huynh et al. [30] focused on capturing the HRV features in both the time and the frequency domain. Firstly, they obtained the IBI measurements by detecting the systolic peak of the heartbeat waveform from the raw BVP data. On this basis, they calculated the mean and standard deviation of the intervals (SDNN), mean and standard deviation of the first and second derivative of the interval series, root mean square of successive interval differences (RMSSD), the standard deviation of successive interval differences (SDSD), and the number of successive interval pairs that differ by more than 50 ms and 20 ms (NN50 and NN20) as time-domain features. They also computed the powers of the low-frequency band and the high-frequency band in an HRV pattern's spectral as frequency-domain features. In another example, Di Lascio et al. [166] extracted HR and HRV related statistical features in the time domain. They computed HR as $60/\overline{IBI}$, where $\overline{IBI}$ is the mean of IBI; and quantified HRV by calculating the distance between adjacent peaks (NN), the SDNN, the SDSD, and the RMSSD.

The EDA signal “is a measure of neurally mediated effects on sweat gland permeability, observed as changes in the resistance of the skin to a small electrical current, or as differences in the electrical potential between different parts of the skin” [214]. The EDA signal reflects neurological control of the rate of sweat production in the glands of the extremities. EDA increases with excitement or nervousness, exemplified by the sweaty palms one may experience before giving a speech or being interviewed for a job. Overall, it consists of slow varying tonic sympathetic activity (named skin conductance level, SCL) and fast varying phasic sympathetic activity (named skin conductance response, SCR), in which SCR is considered to be more commonly used for short-term emotion sensing since it represents a transient response to external stimuli. For instance, Di Lascio et al. [166] first decomposed the EDA signal applying the cvxEDA algorithm [215], then extracted time-domain features from the isolated SCR signal, such as the mean and standard deviation of the first derivative, number of peaks, and peaks' amplitude.

Girardi et al. [147] took into consideration both SCL and SCR components. After using the cvxEDA algorithm, they extracted time-domain features from tonic and phasic parts, including mean tonic, phasic mean, and phasic AUC.

The ECG signal is the standard measurement used to capture electrical and functional activity of the heart. Rather than relying on distal pulse waves (e.g., from PPG) to indirectly trace heart activity, the electrodes of the ECG sensor are usually placed directly on the thorax to monitor cardiac activity. For the raw ECG signals, the QRS complex is the most visually obvious factor when tracing the cardiac cycle. It generally consists of three components (Q wave, R wave, and S wave), and represents the electrical impulse spreading through the ventricles. For ECG-based feature engineering, the analysis of the local morphology of the QRS waveform, and its time-varying and frequency-varying properties has been a standard method. Importantly, HRV time series can also be acquired from RR intervals (intervals between adjacent R waves) [216]. In one example, Hsu et al. [217] first determined the RR intervals by using the QRS detection algorithm, then calculated a total of 34 features from the time domain (e.g., SDNN, RMSSD, median value of RR intervals, mean absolute deviation of RR intervals) and the frequency domain (e.g., very-low-frequency range, low-frequency range, high-frequency range) as well as from nonlinear parameters (e.g., Poincaré plot analysis).

The EEG signal is directly related to the central organ of the human nervous system (i.e., the brain). EEG monitors the brain’s electrical activity by measuring voltage fluctuations resulting from ionic current within the neurons of the brain, and is recorded from multiple electrodes placed on the scalp. In one example, Barral et al. [218] explored frequency-based feature metrics by decomposing EEG data into several frequency-band components including theta (4-8Hz), alpha (8-12Hz), beta (12-30Hz), gamma1 (30-45Hz), and gamma2 (55-80Hz) bands. In another example, Jiang et al. [88] made use of Katz’s fractal dimension calculation [219], and applied it to the waveform to extract time-domain EEG feature metrics. Additionally, they noted that in mobile and dynamic scenarios, the EEG electrodes are normally attached loosely to users, which may cause electrodes to drift and lead to artificial changes to EEG signals. To solve this issue, they proposed a cross-correlation method to quantify electrode drift occurrence.

Device usage and Environment feature representations typically include social interaction records, along with application usage for the former, and noise, illumination,

location, and weather information for the latter. For example, LiKamWa et al. [90] utilized a background logger to capture users' social interactions via phone calls, SMS, and emails. Measures included a record of the number of exchanges the users had with their ten most frequently interacted contacts, the duration of phone calls, and the number of words used in text messages and emails to form the social interaction features. They also counted the launch instances and time spent on the different types of applications (e.g., categorized as {Built-in, Communication, Entertainment, Finance, Game, Office, Social, Travel, Utilities, Weather, Other, or "cannot categorize"}) as application usage features. For location information, they clustered the time series of location estimates using the DBSCAN algorithm [220]. In another example, Zhang et al. [94] used the longitude, altitude, and latitude of GPS data as the outdoor location features, and chose the frequency of the top 20 occurred SSIDs in the WiFi log as the indoor location features. Furthermore, they took the mean and variance of the volume of sounds logged by the microphone to indicate the volumes and displacements as well as defined a noise volume threshold range to calculate the noise ratio, silence ratio, and noise-silence ratio to represent the noise in the environment. For illumination features, the researchers used the mean and variance of the light sensor data values as well as calculated the dark ratio, bright ratio, and dark-bright ratio. Moreover, Lee et al. [93] considered 14 kinds of weather conditions defined by Google weather (e.g., fog, cloudy, drizzle, sunny) to describe the current environmental conditions around the user.

Feature Selection Generally, after feature engineering, the dimensionality of the obtained feature metric will be relatively high (from a few tens to hundreds). Despite efforts to obtain closely related features in the extraction process, not all features play an equal role in emotion detection. Inevitably, there will be some features that contribute less or carry overlapping information, which will increase the computational complexity and slow the classification evaluation process. Considering the limited battery, computation power, and storage of mobile devices, feature selection plays a pivotal role in the emotion sensing process. Additionally, using high dimensional feature vectors is prone to causing the overfitting problem if there is a limited amount of training data. Applying feature selection does not only reduce the number of features and the time cost of computing but also leads to more relevant feature subsets and enhanced generalization.

Collectively, feature selection strategies can be divided into three main categories:

- **Filter Methods:** Rely on scores calculated from various statistical tests to indicate the correlation between features and the outcome variable. The selection process is independent of any machine learning algorithms.
- **Wrapper Methods:** Aim to find the best feature combination by following a search approach (e.g., greedy search) to look through the space of possible feature subsets and evaluating their performance on a given machine learning model (usually choosing the model with better general effect, such as Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (kNN)).
- **Embedded Methods:** Combine the characteristics of filter and wrapper methods. It is implemented by embedding feature selection in the subsequent model training process, in other words, the feature selection is automatically carried out while training the model.

Each of these strategies has advantages and disadvantages. Filter methods have less time complexity and are less prone to overfitting, but have relatively strong subjectivity which may influence the final rate of classification/regression accuracy. Wrapper methods can find the optimal feature subset for the preset machine learning model, but usually require costly computation and come with a high chance of overfitting. Embedded methods have a time complexity between the above two methods and typically have penalization functions (e.g., L1 or L2 regularization) and built-in loss function to reduce overfitting. In summary, all of these methods have been applied in mobile emotion sensing. For example, Lee et al. [93] ranked all features by measuring information gain with respect to each emotional state and selected 10 features with the strongest correlation. Similarly, Barral et al. [218] first computed the Pearson correlation coefficient to reduce the possible redundant features and then used the Wilcoxon rank-sum test to remove non-informative features. For feature selection using wrapper methods, Olsen and Torresen [124] exploited Recursive Feature Elimination to recursively remove the feature having the lowest absolute weight until reaching the desired number of features. In another example, LiKamWa et al. [90] used Sequential Forward Selection to choose the more relevant features. This starts with an empty feature set and continually adds the feature that can maximize the performance of the model, in each iteration. Regarding feature extraction using embedded methods, Zhao et al. [126] adopted 1-norm SVM to optimize the selection of relevant feature subsets while training an SVM classifier.

Huynh et al. [30] utilized a Tree-based model (i.e., RF) with a mean decrease accuracy method to assess feature importance and determine a minimal set of features that can achieve the highest accuracy.

Apart from feature selection, another approach to reduce the dimension of the feature vector is dimensionality reduction. Unlike feature selection which selects and removes the features without changing them (i.e., the feature set after feature selection is actually a subset of the original set), dimensionality reduction aims to transform the original feature set into another lower-dimensional space, while retaining the meaningful properties and information. For instance, Tong et al. [221] used principal component analysis (PCA) to map the extracted EEG features into another set of the coordinate system with lower-dimensional space to achieve a reduction of feature dimensions. Xing et al. [29] also adopted the PCA method to reduce the original 199 features of the pupil diameter variation to 31 main features in order to reduce computing time and improve the final model performance.

2.3.2.2 Deep Feature Extractors

Although manual feature engineering has yielded good results, it is an empirical-based intervention, and the extracted handcrafted features are typically domain-specific, meaning that they do not generalize well to different application scenarios [222, 223]. Furthermore, as mentioned above, handcrafted-based approaches often rely on statistical variables as discriminating features, which neglects the function of the non-linear factors and lacks the abstraction ability of high-level feature associations. Moreover, while feature selection and dimensionality reduction are capable of filtering the less correlated features and avoiding the “curse of dimensionality”, studies have found that it leads to the discarding of large amounts of informational cues [224]. Specifically, given the complexity of human emotions, some filtered “noisy ” features may still contain important information like compound emotional cues.

In order to avoid these restrictions and build more flexible models, researchers have recently attempted to develop various high-level feature extractors based on deep learning approaches. Generally, there are three main categories of deep feature extractors. One is based on convolutional neural network (CNN) architectures. This method typically uses classical CNN networks (e.g., AlexNet [182] or VGG16/19 [183]) as the base model, and

aims to capture fine-grained feature patterns by local-perceiving convolutional kernels and weight-sharing translation equivariance [225]. Specifically, CNN is a hierarchical model that consists of multiple convolutional layers and pooling layers. A convolutional layer is composed of a stack of convolution operations, where multiple small arrays of numbers, called kernels, are applied to form an arbitrary number of feature maps, which represent different characteristics of the input. A pooling layer provides a downsampling operation that reduces the in-plane dimensionality of the feature maps in order to introduce a translation invariance to small shifts and distortions [226]. Sharma et al. [163] used a pre-trained VGG19 on facial data to extract the features, which provided a total number of 1000 features as output. Zhou et al. [181] adopted three convolution layers and three average pooling layers to extract features of ECG signals. Wu et al. [64] selected ResNet18 [227] as the base model and tailored it according to the application scenario. Specifically, they first considered the computation overhead and shrank the input size of the model from $224 * 224$ pixels to $64 * 64$ pixels. Second, in order to ensure the model can still learn fine-grained feature maps, they applied a $3 * 3$ kernel and removed the pooling layer at the beginning of the network. Then, they extended the network to 26 layers to maintain the model's extracting ability. Finally, their modified feature extractor could achieve 4 times faster speeds and only drop 5.3% accuracy compared to the original ResNet18, which makes it more suitable for mobile computing. Chong et al. [98] chose two light-weighted CNN networks, named mini-Xception [228] and TextCNN [229], as the basic models. They first used the fer2013 dataset [230] and NLP & CC 2013 dataset to respectively pre-train the mini-Xception model and the TextCNN model, and then finetuned them using their own collected mobile chatting data. On this basis, they extracted the output probabilities ($P(emotional\ states|facial\ expression)$ and $P(emotional\ states|text\ message)$) as the representations for facial expressions and text messages.

Another category of deep feature extractors is based on recurrent neural networks (RNNs). This type of extractors is usually used to extract features from time series-sensing data, with numerous variations (e.g., long short-term memory (LSTM) [231], or gated recurrent unit (GRU) [232]). They are distinguished by their "memory" as they can take information from prior inputs and remember or hold the important parts, thus influencing the current input and output. Unlike traditional deep neural

networks that assume the inputs and outputs are independent, the output of recurrent neural networks depends on the prior elements within the sequence, which enables them to form a much deeper understanding of a sequence and its context. Tizzano et al. [233] adopted an LSTM-based feature extraction module, which consists of two 35-unit LSTM layers. By feeding time-series 7-dimension (3 for accelerometer, 3 for gyroscope, 1 for heart rate) vectors, the module could output the learned feature representation. Similarly, Zhang et al. [177] constructed a two-stack LSTM network for feature extraction. Specifically, their developed network is composed of an input layer, two LSTM layers, and an output layer. The input of the input layer is a time-series vector $TS = \{TS_1, TS_2, \dots, TS_Q\}$. Different from the conventional operation that directly takes the last output of the last LSTM layer as the feature vector (i.e., many-to-one LSTM), they firstly set both LSTM layers to output the full sequence (i.e., many-to-many LSTM) and obtained another set of time-series prediction vector with the same length in the output layer $\hat{TS} = \{\hat{TS}_1, \hat{TS}_2, \dots, \hat{TS}_Q\}$, where $\hat{TS}$ is regarded as the predicted value of TS . Then by utilizing L-BFGS optimizer with the mean squared error (MSE) loss function $(\frac{1}{Q} \sum_{q=1}^Q (TS - \hat{TS})^2)$, they trained the two-stack feature extraction network preliminarily to capture the temporal dependency of the input time-series data. After the model converges, they finally took the hidden vector of the last LSTM cell in the second LSTM layer as the feature vector of the input time-series data.

Autoencoders (AE) are also popular neural networks for feature extraction. Unsupervised representation learning is the main advantage of this method. Generally, it consists of two parts, an encoder and a decoder. The encoder takes the original data sequence as input and compresses it to a lower-dimensional latent-space representation containing the essence of the input data, while the decoder tries to reconstruct the output from this reduced representation and make it as close as possible to its original input. Once the autoencoder has been well trained, the latent-space representation can be used to represent the feature of the input data. Ghosh et al. [116] used an LSTM-based encoder-decoder architecture for representation learning of raw keyboard interaction patterns. After minimizing the loss between the input sequence and the output sequence, they removed the decoder and used the output of the encoder as a representation vector. Beyond the basic encoder-decoder architecture, Wampfler et al. [115] and Li and Sano [97] respectively used three variations, variational autoencoder (VAE), denoising autoencoder (DAE) and recurrent autoencoder (RAE), to improve the robustness and richness

of the condensed information. Baghdadi et al. [144] leveraged the sparse autoencoder (SAE) that introduces the sparsity constraint on the hidden layers to find the most relative feature patterns matched with the input. After having the latent vector, they used it as the EEG feature representation.

In addition, there are other types of deep feature extractors. For example, Zhao et al. [234] combined CNN and RNN as a type of compact feature extractor for speech data. Specifically, they first used a binary convolution neural network to extract higher-level feature representations from log-mel spectrograms, and then fed them into a binary recurrent neural network to further learn contextual associations. Jiang et al. [235] proposed a hierarchical attention-based network as a feature extractor. Specifically, they first used bidirectional GRUs with an hourly-level attention mechanism to extract and aggregate important hidden states as daily representations. They then used the same bidirectional GRUs but with a daily-level attention mechanism to derive the overall representation of sensor data as the final feature vectors.

2.3.2.3 Summary

An overview of mobile emotion sensing methods used in the learning phase, described by the *modality* used, extractor *category*, and *feature* representations, is presented in Table 2.2 (including the research work presented in this thesis). We also provide a visual representation of the different modalities used for this phase (shown in 2.3).

Overall, we find that handcrafted extractors are still the mainstream of feature engineering. In the reviewed research, more than 80% of papers used handcrafted features as representations. In addition, although some researchers have begun to explore deep feature extractors in the mobile emotion sensing domain, the current approaches are still focused on elementary deep learning models compared with studies in a broader affective computing domain. This is likely due to the fact that modern deep models contain a large number of training parameters and involve complicated computations that require significant computational and memory resources. As a result, they need additional network compression and acceleration techniques to enable efficient deployment on mobile and embedded devices [236].

Furthermore, we note that: (1) for *Facial Expression*, handcrafted extractors are mainly focused on the tracking of feature points on facial landmarks, POIs, or expression contours; (2) for *Speech*, its handcrafted features can be grouped into four categories, continuous features, qualitative features, spectral features, and TEO-based features (we used these categories based on [237]); (3) for *Body Movement & Activity*, its related handcrafted extractors focus primarily on time-domain features, while the combinations of time-domain and frequency-domain features are more leveraged for *Physiology*; (4) for *Environment*, location information is the most commonly used as handcrafted feature representations; (5) for deep feature extractors, CNN- and RNN-based networks are used more frequently.

TABLE 2.2: Overview of mobile emotion sensing research conducted on the learning phase.

Modality	Category	Feature	Reference
Facial Expression	Handcrafted	AUs-based	[106, 108, 163, 180, 238]
		eigenvector-based	[160, 202, 239]
		landmarks/POIs/feature points-based	[63, 102, 104, 107, 108] [157, 161, 162, 180, 240, 241]
		histogram-based	[101, 107, 197, 200]
		facial EMG-based	[108–110]
	Deep	CNN-based	[98, 100, 105, 163, 242–244]
		RNN-based	[245]
		Attentional RNN-based	[2](Chapter 5)
Ocular Sign	Handcrafted	time-domain & frequency-domain	[29]
		time-domain	[173]
	Deep	CNN-based	[64, 246]
Speech	Handcrafted	spectral-based	[156, 157, 243, 247, 248]
		continuous & voice quality	[165]
		continuous & spectrum-based	[164]
		continuous & spectrum-based & TEO-based	[113]
	Deep	RBM-based	[73, 249]
		CNN-based & RNN-based	[234]

TABLE 2.2: Overview of mobile emotion sensing research conducted on the learning phase.

Modality	Category	Feature	Reference
Speech	Deep	Attentional RNN-based	[2](Chapter 5)
		Transformer-based & RNN-based	[250]
Typing&Touch	Handcrafted	keystroke event	[80, 210]
		touch event	[30, 117, 118, 121, 173]
		handwriting event	[123]
		keystroke & touch event	[93, 119, 120, 122]
	Deep	AE-based	[115, 116]
		Attentional RNN-based	[2](Chapter 5)
Body Movement&Activity	Handcrafted	time-domain	[15, 30, 81, 89, 94] [160, 166, 168, 170] [172, 173, 176, 178–181]
		time-domain & frequency-domain	[124, 164, 167]
		time-domain & frequency-domain & wavelet-domain	[82]
		time-domain & frequency-domain & phase-domain	[211]
	Deep	RNN-based	[177, 233]
		AE-based	[97]
		Attentional RNN-based	[235]

TABLE 2.2: Overview of mobile emotion sensing research conducted on the learning phase.

Modality	Category	Feature	Reference
Physiology	Handcrafted	time-domain	[88, 138, 139, 150] [161, 162, 166, 168] [176, 180, 202, 238, 251, 252]
		frequency-domain	[142, 143, 253, 254]
		time-domain & frequency-domain	[30, 128, 133, 140, 144] [147–149, 163, 167, 170] [211, 218, 239, 255–261]
		time-domain & nonlinear	[15]
		time-domain & frequency-domain & nonlinear	[87, 126, 127] [135, 136, 262, 263]
		time-domain & frequency-domain & space-domain	[129]
	Deep	CNN-based	[16, 131, 146, 181, 264–266]
		RNN-based	[233, 267]
		AE-based	[97, 144, 265]
		Attentional RNN-based	[2](Chapter 5)
		Transformer-based	[250] [3](Chapter 6)

TABLE 2.2: Overview of mobile emotion sensing research conducted on the learning phase.

Modality	Category	Feature	Reference
Device Usage	Handcrafted	logs	[63, 89, 90, 94] [151–153, 165] [171, 177, 179, 181]
		time-domain	[178]
		time-domain & frequency-domain	[154]
	Deep	CNN-based	[98]
Environment	Handcrafted	location-related	[90, 155, 172, 177, 181]
		communication-related & location-related	[89]
		illumination-related & communication-related & location-related	[178, 179]
		noise-related & illumination-related & location-related	[94]
		illumination-related & location-related & meteorology-related	[93]
		noise-related & location-related & meteorology-related	[176]
	Deep	RNN-based	[177]

POI: Points of Interest; LBP: Local Binary Pattern; TEO: Teager energy operator; RBM: Restricted Boltzmann Machine.

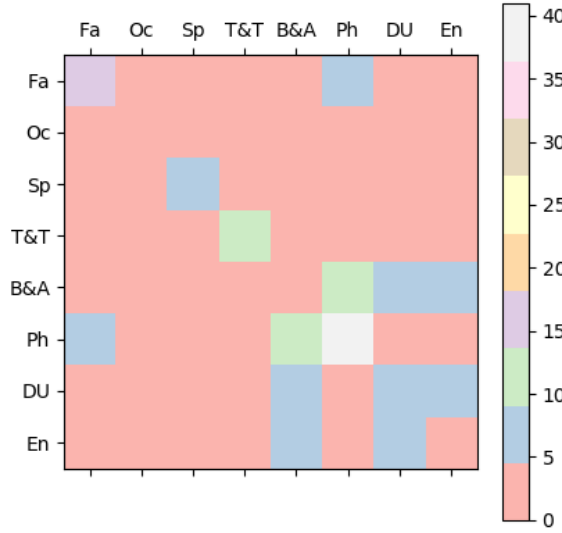


FIGURE 2.3: Visual representation of the different modalities used for mobile emotion recognition. Fa: Facial Expression; Oc: Ocular Sign; Sp: Speech; T&T: Typing&Touch; B&A: Body Movement & Activity; Ph: Physiology; DU: Device Usage; En: Environment.

2.3.3 Inference

Inference is the last stage in a mobile emotion sensing framework. In this stage, feature representations are finally mapped to emotional states; in other words, a connection is established between feature vectors and human emotions. Two major categories of inference methods are typically used [60]. One is classical machine learning methods, in which Naive Bayes (NB), Logistic Regression (LoR), Decision Tree (DT), kNN, and SVM are typical representatives. The other is deep learning methods, in which multilayer perceptron (MLP) is the frequently-used standard network. Coupled with different loss functions (e.g., cross entropy or mean squared error loss function), it can accomplish either classification or regression tasks. Here, we classify the fully connected or dense layers that usually connect below the deep feature extractors as the MLP inference network.

2.3.3.1 Classical Machine Learning Methods

Machine learning algorithms are the most commonly utilized method in deriving probable emotional states. Their usual learning paradigm is to build correlations between features and manually annotated emotion labels; also known as supervised learning. Some typical methods include SVM, kNN, LoR in emotion classification, and Support Vector Regression (SVR), Linear Regression (LiR) in emotion regression. For example, Shi et al. [179] built six classifiers including SVM, kNN, DT, AdaBoost, RF, and Gradient Tree Boosting (GTB) based on 116-dimensional handcrafted features. They carried out contrast experiments on both hybrid data and personal data of twelve participants to recognize emotional states based respectively on the discrete and dimensional emotion models. Their experimental results showed a recognition accuracy rate of 65.91% with hybrid data and 70.00% with personal data on five discrete emotions (happiness, sadness, fear, anger, and neutral), and 72.73% with hybrid data and 79.78% with personal data on a five-level valence scale (high displeasure, displeasure, neutral, pleasure, high pleasure), when using the RF classifier. In another example, LiKamWa et al. [90] considered emotional state as an underlying and slow-changing affect, and averaged all self-report ratings in each calendar day as the labels of the inference engine. They applied a least-squares multiple LiR on the feature table to perform the inference modeling, and evaluated the robustness of their system using leave-one-out cross-validation on two dimensions of emotion respectively. They achieved 93.1% estimation accuracy of daily pleasure averages and 92.7% estimation accuracy of daily activeness averages with a squared error under 0.25 using a personalized model. However, the estimation accuracy was significantly lower (66%, 67%) when using a general model. They further proposed a hybrid approach by incorporating personal data with prior knowledge from the instantiated general model, which performed with 72% accuracy with only 10 days of personalized training data.

In addition to these supervised methods, there is another group of machine learning methods that aim to discover the inherent distributions or hidden patterns in data without human intervention; this is known as unsupervised learning. Clustering is a major example, and groups data based on their similarities or differences. For example, Tizzano et al. [233] used the Gaussian Mixture Model (GMM)

to model the input data by estimating the parameters of a series of multidimensional Gaussian probability distributions to maximize the observing probability of those data. Specifically, they first built one GMM for each of the emotion classes to fit the input training data points, and then evaluated the classification of a new testing data point by checking the log-likelihood obtained from these trained models, that is, $output = \operatorname{argmax}\{GMM_{class_1}(x), GMM_{class_2}(x), \dots, GMM_{class_c}(x)\}$, where $GMM_{class_c}(x) = \ln p(X|\mu_c, \Sigma_c)$. Similarly, Lu et al. [113] also applied GMM as an inference framework. The framework created one GMM for each class (i.e., stress and neutral), and made decisions based on the likelihood function $p(X|\lambda)$ of each class with equal prior, where X is the acoustic feature vector and $\lambda(w, \mu, \Sigma)$ is the weight, mean, and covariance matrix parameters of the GMM model. They then investigated an adaptation model based on a non-iterative Maximum A Posteriori (MAP) scheme and achieved 81% and 76% accuracy for indoor and outdoor environments respectively. Their results demonstrated that the universal stress model can be customized well to different users and scenarios by using few new observation data.

2.3.3.2 Deep Learning Methods

Regarding deep learning-based inference networks, there are various structures such as MLP, CNN, RNN, etc. Among them, CNN and RNN along with their variations often serve as classifiers for inference. Alternatively, MLP (loosely referring to any feedforward Artificial Neural Network (ANN)) is a popular choice that can be implemented in a straightforward and simple way. It relies on multiple layers of perceptrons (or hidden layers) to learn the mapping function between input and output, $f(\cdot) : \mathbb{R}^{input} \rightarrow \mathbb{R}^{output}$. In each perceptron, there are a set of nonlinearly-activating neurons, which transform the values from the previous layer with a weighted linear summation, followed by a non-linear activation function to output the activated value: $\varphi(w^T x + b)$. For example, Gruebler and Suzuki [110] used manually extracted EMG features as affective vectors, and trained a two-layer feed-forward MLP to differentiate between smiling, frowning, and the absence of both. They set four neurons in the hidden layer, and used sigmoid as the activation function. By using back-propagation tuning, they achieve a high facial expression recognition rate on the side of the face.

Similarly, Schmidt et al. [131] built the four-layer MLP inference network for the CNN feature extractor. They used four fully-connected layers with the first three Relu and one last softmax as non-linear activation functions and achieved an average 45.5% mean F1 score over four different tasks (arousal, valence, stress, and state-trait anxiety inventory). In addition, they attempted to extend the single-task architecture to the multi-task architecture through sharing two fully-connected layers followed by multiple output branches, with the aim to predict labels of all tasks simultaneously. After training, their multi-task system also reached a comparable performance when compared to the single-task system. Golgouneh and Tarvirdizadeh [263] applied MLP and radial basis function (RBF) ANNs for continuous measurement of the stress index. By choosing 40 neurons for the hidden layer, their trained models could estimate the stress index with the correlation coefficient values of 0.86 and 0.74 and the average relative error of 0.35 and 0.42 on a subset of the combined features of PPG and GSR signals.

Moreover, there are other types of deep learning-based inference networks. Li and Sano [97] built a 2-layer stacked LSTM followed by a single dense layer as the generalized prediction model to forecast users' well-being states (scaled between 0 to 100). By feeding on daily-level temporal features auto-learned through a hierarchical recurrent autoencoder, the prediction mean absolute error (MAE) values of their general model reached 18.1 ± 0.3 for emotion (sadness and happiness), 19.3 ± 0.8 for health (sickness and health), and 19.9 ± 0.5 for stress (stress and calm). In another example, Ravindran et al. [258] leveraged a CNN architecture as the classification framework. They set three convolution-pooling blocks with Tanh non-linearity as an activation function followed by a global average pooling layer, to classify handcrafted HRV features on three emotional dimensions. After 5-fold cross-validation, they obtained a binary mean accuracy of over 60% on valence, arousal, and dominance. In addition, Baghdadi et al. [144] added a softmax layer on relevant EEG features to perform the classification task for anxious state detection. For self-assessment-based labels, they achieved 83.50% and 74.60% accuracy respectively for 2 and 4 anxiety levels detection.

2.3.3.3 Multimodal Fusion

With the shift toward increasingly multimodal emotion sensing systems, various fusion methods for data collected from multiple sources have been proposed by researchers.

Current multimodal fusion strategies for mobile emotion sensing can be summarized into two categories: feature-level fusion, and decision-level fusion.

Feature-level fusion is the most common method, in which features extracted from different modalities are directly concatenated to form the joint feature representation. Then, the combined feature representations are used as inputs to the final emotion inference network. Because the fusion process comes before the inference stage, this fusion method is also called *early fusion*. For example, Pham and Wang [162] learned a joint representation across PPG signals and facial expressions to recognize people’s emotional responses to mobile videos. Each modality was separately encoded with manual feature engineering and then merged to a joint representation by using a feature-level fusion approach. After that, a user-independent general model was evaluated on nine emotional response metrics by building an SVM binary classifier and utilizing the leave-one-subject-out method. Although feature-level fusion is easy and straightforward to implement, this method is often criticized for ignoring the synchronization among different modalities, especially when considering their differences in time formats and metric levels. Recently, Zhang et al. [177] considered synchronization among modalities, and made an improvement on feature-level fusion. They aligned data at the time-session level and adopted a hierarchical attention mechanism to incorporate feature sets from different sensors. Taking the extracted features from an individual sensor as input, they first leveraged an attention layer on the session-level to fuse them into a specific sensor feature representation. They then leveraged another attention layer on the sensor-level to further fuse the feature representation of each sensor to output a multi-sensor representation. After this step, they utilized MLP and a sigmoid function to obtain a prediction for emotion instability whose score was quantified by self-reported six basic emotion states (Ekman’s emotion model).

Decision-level fusion aims to combine the independent results from different single modality inference networks and make a final decision via some algebraic rules. Because this fusion process comes after the inference stage, this method is also called *late fusion*. For example, Alam and Riccardi [268] applied a majority voting strategy for decision-level fusion in their study. They trained five SVM classifiers on five feature sets respectively, and then combined the decisions from these classifiers by majority voting. Their experiments demonstrate that the decision combination provided significantly better results compared to the results of any single feature set in personality traits

prediction. In addition to these simple algebraic rules, some studies have also leveraged specific learning algorithms to fuse the decision results. Adibuzzaman et al. [160] used the Naïve Bayes algorithm to fuse the modalities of facial expression and energy expenditure of body movement at the decision level. Specifically, they used the conditional error distributions of each classifier to approximate the uncertainty of each classifier’s decision. The final combined decision of emotional state was the weighted sum of all individual outputs. Sharma et al. [163] used Ensemble Learning (EL) to combine decisions from the modalities of facial expressions and physiological signals to improve the overall performance of their approach. They designed seven different algorithms (linear, radial, and polynomial kernels of SVM and Gaussian processes, and a M5 model tree) for each modality, and obtained the fused results for cognitive assessment by using a weighted average from all the inference algorithms. Nevertheless, as different modalities are trained independently, decision-level fusion lacks the ability to learn the mutual correlations among modalities and the fusion process is relatively time-consuming.

2.3.3.4 Summary

An overview of mobile emotion sensing research conducted on the inference phase, described by the *task* category (classification or regression), *fusion* approach, used inference *method*, measured *emotions* and experimental *performance*, is presented in Table 2.3 (including the research work presented in this thesis).

Overall, we find that ML methods are the most extensively used inference structures. This is largely due to the fact that many of these ML methods are easier to implement, need less training data, and many software libraries such as scikit-learn, Weka, and OpenCV provide ready-to-use base ML structures. Furthermore, faster convergence speed and lower hardware requirements are also important benefits. Even for neural network-based feature extractors, ML methods can also be used as trade-off inference networks or baselines [98, 146, 163, 233]. Regarding multimodal fusion strategies, feature-level fusion has seen much wider use than decision-level fusion. In large part, this is because feature-level fusion enables the implementation of end-to-end architectures that have no need for extra pre-processing and feature extracting steps. This has practical significance for real-world deployment. On the contrary, decision-level fusion needs repetitive training for different classifiers, which is time-consuming and unfavourable

for real-time predictions. Among the papers we surveyed, over 72.2% used feature-level fusion (i.e., early fusion) in ML-based models, compared to 83.3% in DL-based models.

In most cases, emotion sensing is considered a classification problem, and the target of the proposed system is based on the discrete emotional theory. However, recently the dimensional emotional theory has increasingly drawn more attention due to its advantages [54], and researchers have attempted to explore the mobile emotion sensing problem from both a classification and a regression point of view. For example, by directly using continuous arousal/valence values or an average value over time as inference labels, taking a look at emotion sensing problems from a regression point of view [90, 97]; or by using thresholds to segment the continuous arousal or valence scales into different independent parts, realizing simple classification on the dimensional emotion model [87, 115]. Finally, regarding experimental performance, the accuracy of most studies has reached over 70 percent, while considering a plethora of different inputs and outputs. Nevertheless, it is important to note that these results are mostly based on existing public or self-collected laboratory data, which lacks thorough validation in realistic everyday scenarios. Therefore, there is still considerable ground to be covered before current emotion sensing technology can be integrated into everyday mobile and embedded devices.

TABLE 2.3: Overview of mobile emotion sensing research conducted on the inference phase.

	Task	Fusion	Method	Emotions	Performance	Reference
ML	cla	feature	SVM	laughter, non-laughter	Acc: 81%, Pre: 82%, Rec: 81%, F1: 80%	[166]
				relaxation, stress, fatigue	Acc: 99.52%	[211]
				six basic emotions, neutral	Acc: 97.26%	[157]
			Factor Graph	compound emotions	EM: 76.0%	[94]
			BN	six basic emotions, neutral	Acc: 67.52%	[93]
			LoR	anger, not anger	RMSE: 0.984	[165]
			RF	six basic emotions, neutral	F1: 79%	[63]
				three-level engagement	Acc: 85%, 77%	[30]
				anger, fear, happiness, sadness, neutral five-level valence	Acc_{dis_p} : 70.00%, Acc_{dis_g} : 65.91%, Acc_{dim_p} : 79.78%, Acc_{dim_g} : 72.73%	[179]
			GB, DT, RF, SVM	happiness, stress	Acc_{hap} : 88.42%, $F1_{hap}$: 86.18%, Acc_{str} : 73.94%, $F1_{str}$: 63.47%	[176]
		decision	NB	six basic emotions	Acc: 93%	[160]

TABLE 2.3: Overview of mobile emotion sensing research conducted on the inference phase.

	Task	Fusion	Method	Emotions	Performance	Reference
ML	cla	decision	SVM	anger, happiness, sadness, neutral	Acc: 87%	[156]
			HMM	six basic emotions, neutral	Acc: 51.64%	[98]
				positive, negative	Acc: 76.25%	
			RF	anger, fear, happiness, sadness, neutral arousal, valence	$F1_p$: 52.5%, 77.6%, 80% $F1_g$: 19.5%, 45.9%, 51.2%	[260]
		hybrid	SVM, DT, kNN	nine emotional states	Acc: 82.6%	[161]
		single	SVM	six basic emotions, neutral	Acc: 87%	[197]
				relaxation, stress	Acc: 79%, 75%	[146]
				depression	Acc: 90%	[152]
				normal, smile, anger	Acc: 80.6%	[107]
				six basic emotions, neutral, contempt	Acc: 92.8%	[106]
				anger, joy, pleasure, sadness	Acc_p : 87%, Acc_g : 72.3%	[126]
					Acc: 54.8%	[127]
					Acc: 54.55%, 81.82%	[259]
					Acc: 30.00%, 63.64%	
				arousal, valence	Acc: 73.64%, 74.41%	[256]
					Acc: 76%	[257]
					Acc: 70%	[123]

TABLE 2.3: Overview of mobile emotion sensing research conducted on the inference phase.

	Task	Fusion	Method	Emotions	Performance	Reference
ML	cla	single	SVM	anger, happiness, sadness, surprise	Acc: 89.85%	[200]
				affect, arousal, valence	peak F1: 86%, 88%, 86%	[120]
				arousal, valence	Acc: 75.56%	[87]
			NB	five-level frustration	Acc: 80.3%	[150]
				arousal, valence	Acc: 86.60%	[119]
			GMM	stress, neutral	Acc: 81%, 76%	[113]
			kNN	arousal, valence	Acc: 53.6% ~ 69.9%	[148]
				three-level stress	Acc: 85.3%	[263]
				mind wandering	Acc: 71.22%	[128]
				positive, neutral, negative	Acc: 97%	[252]
			XGBoost	arousal, valence	Acc: 68.4%, 80.1%	[262]
			RF	funny, not funny	AUC: 73%	[218]
				relaxation, stress	Acc: 84.13%	[149]
					Acc: 71%	[133]
				stress, anxiety, depression	Acc: 73.4%, 73.4%, 74.1%	[118]
				happiness, relaxation, sadness, stress	AUC: 73%	[80]
			DT	anger, happiness, pain, sadness	Acc: 65%	[139]
				enjoyment, frustration	Acc: 87.90%, 89.45%	[154]

TABLE 2.3: Overview of mobile emotion sensing research conducted on the inference phase.

	Task	Fusion	Method	Emotions	Performance	Reference
ML	cla	single	DT	arousal, valence	Acc: 72.54%, 68.21%	[143]
			SVM, HMM	six basic emotions	Acc: 70.6% (SVM) Acc: 65.2%(HMM)	[102]
			SVM, RF	six basic emotions	Acc: 86% ~ 95%	[82]
			SVM, AdaBoost	stress	Acc: 92.05%	[135]
			J48, kNN	sadness, dislike, joy, stress, normal no-idea, positive, negative	Acc_{six} : 92.64%, 94.39% Acc_{two} : 95.53%, 96.73%	[269]
			BN, kNN, C4.5	stress	Acc: 67.86%, 87.56%, 74.26%	[122]
			GBDT, RF, SVM	positive, negative	Acc: 81.3%	[129]
			NB, BN, SL, C4.5 BFTree, NBTree	six basic emotions, neutral	F1: 57% (overall) F1: 82% (DT-based)	[121]
	reg	feature	RiR	tenseness, excitement, happiness, relaxation, tiredness, boredom, upset, stress	MAE: 0.7631	[172]
			LiR	pleasure activeness	MSE_p : 0.075, MSE_g : 0.296 MSE_p : 0.085, MSE_g : 0.289	[90]
		single	SVM	arousal, valence	CC: 0.614, 0.610	[29]
			AdaBoost, RF	arousal, valence	Acc: 65.7%, 82.2%	[155]

TABLE 2.3: Overview of mobile emotion sensing research conducted on the inference phase.

	Task	Fusion	Method	Emotions	Performance	Reference
ML	reg	single	SVR, ANFIS	stress index	CC: 0.89, 0.93	[263]
DL	cla	feature	MLP	emotion instability	Acc: 78.33%, F1: 77.78%	[177]
				positive, negative	Acc: 89.2%, AUC: 86%	[2](Chapter 5)
			FFNN, LSTM	happiness, stress	Acc_{hap} : 64.74%, $F1_{hap}$: 64.74% Acc_{str} : 66.11%, $F1_{str}$: 66.11%	[176]
			Conformer	arousal, valence	$MacroF1$: 77.37%, 69.11% UAR : 79.42%, 80.6%	[3](Chapter 6)
		decision	LSTM	anger, fear, happiness, sadness, neutral arousal, valence	$F1_p$: 63.1%, 77.6%, 80.5% $F1_g$: 34.8%, 68.6%, 64.4%	[260]
		single	MLP	smile, neutral	–	[109]
				binary-level stress, five-level stress	Acc_b : 76.61%, Pre_b : 67.96% Rec_b : 90.91%, $F1_b$: 77.78% Acc_{five} : 47.57% $MicroF1_{five}$: 47.57% $MacroF1_{five}$: 47.06%	[235]
				binary-level anxiety, four-level anxiety	AUC: 83.50%, 74.60%	[144]
				happiness, relaxation, sadness, stress	AUC: 84%	[116]

TABLE 2.3: Overview of mobile emotion sensing research conducted on the inference phase.

	Task	Fusion	Method	Emotions	Performance	Reference
DL	cla	single	MLP	anger, fear, happiness, pleasure, neutral	Acc: 75.3% ~ 88.64%	[140]
				anger, frustration, happiness, sadness, neutral	Acc: 61.9%	[234]
				five-level valence	Acc: 86.9%	[100]
				arousal, valence, dominance	AUC: 82%, 84%, 82%	[115]
				arousal, valence, stress, state-trait anxiety inventory	mean F1: 45.5%	[131]
		feature	CNN	arousal, valence, dominance	Acc: 60.1%, 61.9%, 62.6%	[258]
				arousal, valence	Acc: 71.25%, 77.57%	[16]
	reg	feature	LSTM, LiR	emotion, health, stress	$MSE_{emo-p} : 14.1 \pm 0.2$ $MSE_{hea-p} : 12.4 \pm 0.3$ $MSE_{str-p} : 15.0 \pm 0.3$ $MSE_{emo-g} : 18.1 \pm 0.3$ $MSE_{hea-g} : 19.3 \pm 0.8$ $MSE_{str-g} : 19.9 \pm 0.5$	[97]
		single	MLP	arousal, valence	CC: 0.265, 0.209	[29]
			MLP, LSTM	arousal, valence	—	[247, 248]

TABLE 2.3: Overview of mobile emotion sensing research conducted on the inference phase.

	Task	Fusion	Method	Emotions	Performance	Reference
DL	reg	single	MLP-ANN, RBF-ANN	stress index	CC: 0.86, 0.74	[263]

hybrid: feature-level fusion + decision-level fusion; single: single modality; BN: Bayesian Network; HMM: Hidden Markov Model; GB: Gradient Boosting; GBDT: Gradient Boosting Decision Tree; SL: Simple Logistic; BFTree: Best-First Decision Tree; NBTree: Naïve Bayesian Decision Tree; RiR: Ridge Regression; AdaBoost: Adaptive Boosting; XGBoost: eXtreme Gradient Boosting; ANFIS: Adaptive Neuro-Fuzzy Inference System; FFNN: Feedforward Neural Network; Conformer: Convolution-augmented Transformer; Acc: Accuracy; Pre: Precision; Rec: Recall; Macro-F1: Average F1 score; UAR: Unweighted Average Recall; EM: Exact Match; RMSE: Root Mean Square Error; AUC: Area Under the ROC Curve; CC: Correlation Coefficient; Subscript p: personalized model; Subscript g: generalized model; Subscript dis: discrete emotions; Subscript dim: dimensional emotions.

Chapter 3

Methodology

This chapter describes the methodological decisions made in planning the research presented in this thesis. Specifically, we explain the motivations behind specific decisions adopted in experiment design, model development, and quantitative data analysis. Furthermore, we discuss ethical considerations associated with our experiments. This chapter focuses primarily on methodological factors that pertain throughout this thesis to ensure the continuity and rationality of our studies. More detailed descriptions of the methods used in each study can be found in the corresponding articles in subsequent chapters.

3.1 Research Approach

Through the literature review, we found that an increasing number of commercial platforms now offer cloud-based processing via Application Programming Interfaces (APIs) or Software Development Kits (SDKs), enabling developers and researchers to incorporate emotion recognition capabilities in their applications. Thus, we first focused on these off-the-shelf emotion recognition systems, and aimed to establish a replicable assessment method regarding their application potential in real-world scenarios. In [Article I](#), we selected five prominent commercial facial emotion recognition systems - Amazon Rekognition, Baidu Research, Face++, Microsoft Azure, and Affectiva. Each of them can be applied to mobile platforms and are capable of recognizing at least six basic emotion categories. We designed two experiments for evaluating these platforms: (1)

Experiment 1 leveraged standardized facial expression photos as the independent variable, and investigated and compared the five systems' performance; (2) Experiment 2 adjusted the input variables by adding common distortions, and examined how image distortions that often occur during phone use in the wild affect system performance. To simulate realistic distortions, we first identified five potential events that can lead to the poor quality of facial expressions taken by mobile phone cameras in real-world scenarios, i.e., *rotation*, *partial face*, *brightness*, *blur*, and *noise*. Then we manipulated the facial images of Experiment 1 to reflect these five events. On this basis, we conducted a quantitative study of the reduction in performance of each facial emotion recognition system using manipulated images. We found that there are some limitations for current commercial emotion recognition systems in terms of meeting robustness, accuracy, and overall performance requirements, which may limit the usefulness of such systems in real-world applications.

In order to develop more robust mobile emotion recognition systems, we applied a multi-modal structure, and extracted high-level informative features and associations through deep learning structures. In [Article II](#), we first introduced a hybrid deep multimodal system for extracting high-level features from both behavioral and physiological signals collected from off-the-shelf smartphones and wearable wristbands. It uses LSTM structure to capture temporal features from each input signal and constructs fine-grained modality representations. Considering that not all signals contribute equally to the final modality representation (e.g., not all frames have the same importance in the visual representation), we further introduced a feature attention mechanism to assign different importance weights and teach the model dynamic attention. We then concatenated all modality feature representations, used a multi-layer deep neural network to learn cross-modality associations, and classified emotions via a sigmoid classifier. Additionally, to evaluate the recognition capability regarding spontaneous emotions, we designed a replicable experimental protocol based on mobile devices, which utilizes visual stimuli to induce participants' emotions and mobile devices to collect spontaneous emotion responses. The experiment consists of four tasks: emotion induction, self-report, typing task, and speaking task. The following information is recorded: (1) facial features and three types of physiological data (BVP, EDA, and SKT) via smartphone front-facing camera and wristband respectively in emotion induction task; (2) manual labeling in

self-report task; (3) keystrokes data via Gboard¹ in typing task; and (4) audio data via smartphone microphone in speaking task. Based on this protocol, we conducted a user study with 45 participants, and collected a multimodal spontaneous mobile emotion database for further evaluation.

In [Article III](#), we proposed a hierarchical mobile emotion recognition system that augments convolutions with self-attention to select representative features and learn mutual dependencies. Compared with previous work, the proposed model has the following salient characteristics: (1) instead of using RNNs to model dependencies in time-series data as in previous research, we adopted the Convolution-augmented Transformer (Conformer) architecture [270] as an encoder to model sequences since it can effectively capture both position-wise local features and content-based global interactions in a more parameter-efficient way; (2) to extract more expressive features, we designed a convolutional front-end structure to capture fine-grained feature patterns and focus on information from different representation subspaces, through local-perceiving convolutional kernels and weight-sharing translation equivariance [225]. Furthermore, in this study, we adopted the dimensional emotion model and formalized the emotion recognition as five classification tasks: two-category and five-category classification on arousal and valence respectively, and a four-category classification on combined arousal-valence emotion space.

3.2 Data Analysis

The findings presented in this thesis are mainly derived through quantitative methods. We measured system performances ([Article I](#), [II](#), and [III](#)) through different evaluation metrics according to task requirements. In this section, we outline the quantitative methods used in the different studies presented in this thesis.

In [Article I](#), we first passed facial images through each system, received the corresponding emotion attributes, and selected the emotion with the highest confidence or probability value as the detection label. To evaluate performance, we used accuracy, precision, recall, and F1-score as evaluation metrics. Moreover, since sometimes the system may

¹<https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin&hl=en>

fail to detect a face in the image and thereby return a null response, we also added detection rate as a measure of the system’s ability.

In [Article II](#), we used deep learning techniques and proposed a novel attention-based LSTM emotion-sensing system. Compared with traditional machine learning techniques (e.g., DT or SVM), deep learning techniques can work on raw data and automate feature extraction and selection without the need to manually construct features [223]. To comprehensively evaluate the accuracy and robustness of our proposed system, we employed leave-one-participant-out cross-validation and demonstrated the performance for each participant. The cross-validation results were then averaged and combined with the area under the curve (AUC) score as overall performance metrics.

In [Article III](#), we used Conformer structure as the hierarchical encoder to better combine features extracted from multiple signals. In this study, we chose a publicly available multimodal affective dataset that uses off-the-shelf and low-cost mobile and wearable devices for data recordings and focuses on spontaneous emotions during naturalistic conversations. Since extreme-state emotions happen rarely for most people during usual communication, this results in imbalanced data distribution in this dataset. For a fair evaluation, we did not use accuracy as the evaluation metric like most other studies. Instead, we used the average F1 score (Macro-F1) and unweighted average recall (UAR) as our evaluation metrics, as they are very sensitive to the classification accuracy of the minority class. Moreover, we did two generalizability tests to further test the generalizability of our proposed model.

3.3 Ethical Considerations

We took multiple precautions to manage risks of harm to participants that might have arisen during the user study described in Chapter 5. The corresponding experimental procedures received ethical clearance from The University of Melbourne Human Ethics Advisory Group (Ethics ID: 1955051.1). Next, we detail the steps taken to ensure ethical conduct.

For this experiment, we aimed to use visual stimuli to elicit spontaneous emotions in participants. This is a common and effective approach used in Neurophysiology and Psychology research [271, 272]. We first selected affective pictures from the validated

subsets of the International Affective Picture System [273, 274] and the Nencki Affective Picture System [275, 276]. In order to exclude any risks that might be harmful to the physical and mental health of the participants, we avoided highly disgust-provoking and fear-provoking pictures, for example, we excluded pictures showing death (e.g. animal cadavers, bones), blood, or injection. We also explicitly stated that participants have the right to withdraw from the experiment at any point. In addition, student participants were informed that their decisions regarding participation in the project or withdrawing at any stage would not affect their grades.

Before the start of the experiment, we provided our participants with information through a Plain Language Statement explaining the objectives, procedures, data that will be collected, how the collected data will be used for the study, and the potential benefits of the research. Each participant's consent was established by signing and returning a Consent Form. Participants were given the opportunity to read the consent form before agreeing to participate in the study. We only commenced the experiment and data collection once the consent form was signed by the participant.

To ensure that the collected data would not be mapped to the identities of participants, we did not store and use participants' names, but used pseudonyms (e.g., P01, P02), in the experiments and publications. Moreover, regarding data storage and processing, we stored data on the University's secure servers, and encrypted it using SSL. Access to data is controlled through an authentication mechanism, and servers are protected with firewalls. Participants reserved the right to withdraw from the study at any point in time without consequences. In addition, all participants have the right to request the deletion of their data at any time.

3.4 Limitations

Although the studies presented in this thesis were carefully designed and executed, we acknowledge several limitations.

In [Article I](#), we conducted two experiments in order to assess the emotion recognition capabilities of five popular commercial facial emotion recognition systems. We acknowledge that most of these systems will regularly update their frameworks and corresponding network features and weights, which may improve their performance in

handling different distortions. However, our proposed assessment method is replicable, and can be easily used to reassess the effectiveness of these commercial systems or even automated for future evaluation of updated or new facial emotion recognition systems. Furthermore, we only assessed the impact of different distortion factors independently. It is possible that in a real-world scenario, multiple distortion types can occur simultaneously. We also acknowledge that there are currently commercial emotion recognition services that rely on other signals (e.g., speech). However, we chose to evaluate facial expression-based systems in part because most commercial services are based on facial expression recognition. In addition, according to our literature review, we found that facial expression is still one of the widely used modalities when building multimodal systems.

In [Article II](#), we ran experiments in laboratory settings to explore different usage scenarios. Nevertheless, we acknowledge that user behaviors under completely naturalistic conditions are likely to be more diverse and the collected signals would have more noise, which may pose a more severe challenge to the system’s disturbance rejection ability. Additionally, in the experiments, the demographics of the participants were skewed towards young people, which may have had an impact on our findings.

In [Article III](#), the data sample size in the final evaluation dataset is relatively small, which, to a certain extent, limits the learning ability of our proposed model and may create an overfitting problem in the training process. While we applied batch normalization and dropout function to overcome overfitting and internal covariate shift, a larger size of data samples could make it easier for the supervised training strategy to correct the generated biases and produce more robust systems [277]. Thus, further work is required to collect larger natural and spontaneous emotion datasets based on commercial mobile and wearable devices.

Furthermore, in [Article II](#) and [Article III](#), we did not deploy our systems into real-world contexts. Future work could develop mobile versions of the proposed systems and install them on users’ personal devices for real-world user studies.

3.5 Summary

This chapter provides an overview of the overall methodological approaches used in the studies presented in this thesis. We quantified the performance comparison regarding different commercial mobile emotion recognition services, and developed two novel mobile-based multimodal emotion recognition systems. The thesis aims to develop and improve models and systems for mobile emotion recognition, and advance the development of mobile emotion-sensing field. The following Chapters [4](#), [5](#), and [6](#) present original research articles on this topic, which have been published in peer-reviewed venues.

Chapter 4

Benchmarking Commercial Emotion Detection Systems

4.1 Introduction

In this chapter, we provide an investigation of the performance and limitations of off-the-shelf commercial facial expression-based mobile emotion recognition systems. Considering the diversity and uncontrollable nature of user behaviors and surrounding environments during in-the-wild smartphone usage, different facial image distortions are quite likely to occur. For example, in-the-wild facial photos from the front-facing camera may be only partially visible due to the limited camera view or the occlusion by the user's hairs, fingers, and clothing, or be distorted due to motion blur. In recent work [278], researchers explored how often users hold devices in a way that allows capturing their faces or eyes. By collecting 25726 in-the-wild photos taken from the front-facing camera of smartphones, their research shows that the full face is rarely visible in the front-facing camera's view, and the state-of-the-art face detection algorithm performs poorly against the non-ideal photos. Similarly, commodity facial emotion recognition services typically require full-face and high-quality images for good performance. However, few validation studies take into account the non-ideal conditions that can arise in real-world scenarios. Therefore, it is necessary to conduct a comprehensive validation study to explore how different distortions impact the detection accuracy of these emotion recognition systems.

To fill this gap, we designed two experiments. We manipulated standardized facial expression images to simulate different possible distortions in the real world and quantitatively evaluated each commercial system on these images. Our results show that each system has its own strengths and limitations under different distortion conditions. Rather than relying on one specific system, it would be a better choice to apply multiple systems and route facial images that need to be recognized to different systems based on the type of distortion. Importantly, our proposed evaluation method can be used for reliability assessment of future updated or new facial expression-based mobile emotion recognition systems (we have released our code on github¹). Further details of the related design, analysis, results, and discussion can be found in the attached publication, [Article I](#).

4.2 Article I

©Copyright is held by Springer-Verlag GmbH Germany, part of Springer Nature. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in:

Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer*, 37(6): 1447–1466, 2021. URL <https://doi.org/10.1007/s00371-020-01881-x>

¹<https://github.com/chickencaesar-yang/Benchmarking-commercial-emotion-detection-systems>

Benchmarking Commercial Emotion-Detection Systems using Realistic Distortions of Facial Image Datasets

Kangning Yang · Chaofan Wang · Zhanna Sarsenbayeva · Benjamin Tag · Tilman Dingler · Greg Wadley · Jorge Goncalves

Received: date / Accepted: date

Abstract Currently, there are several widely-used commercial cloud-based services that attempt to recognize an individual's emotions based on their facial expressions. Most research into facial emotion recognition has used high-resolution, front-oriented, full-face images. However, when images are collected in naturalistic settings (e.g., using smartphone's frontal camera), these images are likely to be far from ideal due to camera positioning, lighting conditions, and camera shake. The impact these conditions have on the accuracy of commercial emotion-recognition services has not been studied in full detail. To fill this gap, we selected five prominent commercial emotion recognition systems - Amazon Rekognition, Baidu Research, Face++, Microsoft Azure and Affectiva - and evaluated their performance via two experiments. In Experiment 1, we compared the systems' accuracy at classifying images drawn from three standardized facial expression databases. In Experiment 2, we first identified several common scenarios (e.g., partially visible face) that can lead to poor quality pictures during smartphone use, and manipulated the same set of images used in Experiment 1 to simulate these scenarios. We used the manipulated images to again compare the systems' classification performance, finding that the systems varied in how well they handled manipulated images that simulate realistic image distortion. Based on our findings we offer recommendations for developers and researchers who would like to use commercial facial emotion recognition technologies in their applications.

Keywords Affective computing · Facial emotion recognition · Commercial emotion recognition systems · Non-ideal conditions · Validation analysis

1 Introduction

A basic goal of most affective computing systems is to automatically recognize and react to human affective states during interactions with human beings [20]. To this end, emotion recognition is one of the most important sub-tasks of affective computing [82, 87], and has already been used in many real-world applications, including mental health systems [112], education systems [104, 56], and music recommendation systems [23]. Furthermore, emotions are known to play a crucial role in human daily lives, as they can aid decision-making, learning, communication, and situation awareness in human-centric environments [87]. Thus, reliable emotion recognition is seen as crucial for several other research fields including, but not limited to, Psychology, Social Sciences, and Computer Science.

Over the past few years, a variety of data sources including facial expressions, voice, body gestures, and linguistics have been leveraged for emotion recognition by researchers and industry [55]. Notwithstanding, given that facial expressions are the primary communication mode of non-verbal emotional expressions [33, 71], this particular data source is still one of the preferred choices in the emotion recognition domain. Recently, with the increasing availability of 3D sensors and corresponding devices, researchers are able to acquire 3D dynamic facial expressions [13, 14]. This methodology has the potential to provide more robust outcomes in practical applications [21, 107].

In our work, we focus on emotion recognition systems that use 2D facial expressions as their data source. Such systems have been widely applied with the advent of commercial platforms like Amazon Rekognition [6], Baidu Research [8], Face++ [101], Google Vision [51], Microsoft Azure [81], and Affectiva [2]. Most of these platforms offer cloud-based processing via an Application Programming Interface (API) or Software Development Kit (SDK) that enables developers

K. Yang, C. Wang, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, J. Goncalves

The University of Melbourne, Australia

K. Yang E-mail: kangning.yang@student.unimelb.edu.au

and researchers to include emotion recognition capabilities in their applications [59].

However, to the best of our knowledge, only limited research has been conducted on validating the reliability and accuracy of different facial emotion recognition systems [12, 19, 32, 98]. Furthermore, these studies have focused on model performance on selected databases rather than considering the imperfect conditions that might occur in real-world scenarios. For example, given the variation in individual habitual behaviours and the physical restrictions of smartphone cameras and surrounding environments, it is common that a user's face is only partially visible or that the light conditions are too dark or too bright in photos taken using front-facing smartphone cameras [64]. In such situations it is currently unclear which facial emotion recognition system will yield the best results.

To fill this gap, we conducted two experiments. First, we utilized three publicly available datasets containing emotion-labeled high-quality full-face photos to evaluate five commercial automatic facial emotion recognition systems. Then, we determined possible causes of poor-quality face images in naturalistic settings (i.e., a person using their smartphone), and classified them into the following categories: *rotation*, *partially-visible face*, *brightness*, *blur*, and *noise*. We then evaluated the five commercial automatic facial emotion recognition systems using manipulated images representing different examples of each category.

Thus, the contribution of this paper is three-fold:

- (1) We conducted an assessment of five commercial automatic facial expression recognition systems in a variety of conditions, such as high-quality full-face images and a variety of poor-quality images;
- (2) We identified different scenarios that can lead to poor-quality phone-camera images taken in real world scenarios, and created a set of images with different poor quality conditions (e.g., different degrees of blur or noise). This mechanism can be used in future work aimed at assessing the reliability of emotion recognition systems that rely on facial expressions;
- (3) We provide recommendations to the research community regarding the limitations and emotion recognition capabilities of each system.

2 Related Work

In this section, we describe research in topics related to our study, namely emotion recognition, facial emotional expression, and methods for facial emotion recognition systems.

2.1 Emotion Recognition

Emotional expression refers to how people communicate internal affective states to others through both verbal and non-verbal behaviour [3, 20, 108, 47, 52, 11]. During human communication, people can easily perceive the emotional states of others from multiple signals, for example, a happy emotional state from a smiley face and positive words, or a sad emotional state from a slight frown and a tearful tone. However, it is challenging to enable computers to deduce high-level affective states from low-level signal cues. This is mainly because of the gap between the extracted feature representation and human affective states, or the complexity of fusion patterns of different features [66, 54, 42].

Over the years, several approaches and algorithms have been proposed to detect people's emotional states. Some of them are based on the classification and analysis of facial expression including 2D and 3D representations [80, 94, 95, 13, 14], body language or posture [27, 79, 49], speech [72, 96], written text [4, 74] and physiological signals [57, 5]. Among the different available data sources, Ekman states that facial expression is "the most commanding" [33], and more than other behaviours can be "quite intense even when a person is alone" [33]. In other words, analysing facial expressions is one of the most direct ways to study human emotion.

2.2 Facial Emotional Expression

In Darwin's seminal work, *The Expression of the Emotions in Man and Animals* [28], he posited that "facial expressions are the residual actions of more complete behavioural responses, and occur in combination with other bodily responses – vocalizations, postures, gestures, skeletal muscle movements, and physiological responses" [71, 93]. More recently, researchers have theorised that all people express emotions via common facial expressions [37, 41, 61, 39, 77]. In other words, there exists a certain degree of universality in emotion representation through facial expressions, a position termed *the universality hypothesis*. Others question universality and propose that facial expressions of emotion vary with culture and language [83, 92, 62, 9]. Some researchers argue that people interpret facial expressions in terms of the social situations in which they occur [22]. Despite such controversy, both proponents and critics agree that human beings can express and obtain emotional information through facial movements [83]. Most of the commercial emotion recognition systems today exactly rely on facial expressions to achieve emotion recognition function and provide related services.

Furthermore, studies have shown that facial expressions are recognizable in terms of different discrete emotion categories [10, 20, 34, 38, 43, 60, 78]. In particular, works on emotion recognition using facial expressions have classified six

basic emotions that are typically recognizable: anger, disgust, fear, happiness, sadness, and surprise [35, 16, 103, 76]. Facial emotion analysis has been an active topic within the research community, and recently several commercial facial emotion recognition systems have become available that focus mostly on these six basic emotions. Our goal in this study is to evaluate how five commonly-used commercial emotion recognition systems perform when they are challenged with analyzing naturalistic facial photographs.

2.3 Methods for Facial Emotion Recognition Systems

A widely used approach for classifying facial expressions is based on the Facial Action Coding System (FACS) developed by Ekman and Friesen [40, 44]. FACS describes facial behaviours in terms of a set of specific Action Units (AUs) each of which is associated with an individual face muscle or muscle group [40, 44]. Using this approach, trained human coders can manually deconstruct and annotate nearly any possible facial activity as a combination of AUs. Follow-up work proposed a derivative of FACS called Emotion FACS (EMFACS), which focuses only on subsets of AUs that are likely to have universal emotional significance [45]. Performing analysis using FACS requires professional training and manual coding that is time- and labour-intensive, and until the 1980s most studies were completed by philosophers and psychologists [63, 111].

With advances in computer technologies, such as machine learning, researchers began to design software tools for automated AU annotation and emotion recognition in order to overcome the limitations of FACS [15, 70, 99, 105, 110]. Automatic facial emotion recognition systems work in three stages: face detection, feature extraction, and emotion recognition. Given a still image, the system first determines the regions where faces are present. Then, by detecting specific AUs or facial landmarks such as eyes, eyebrow, mouth, etc., the system extracts and forms feature groups to represent the facial expressions. Finally, by applying machine learning algorithms such as Naïve Bayes, Decision Tree, Linear Regression or Support Vector Machine, the system maps feature groups onto emotional states.

Recently, researchers have been further turning to 3D facial expressions, and propose to improve facial feature tracking and emotion recognition based on 3D face scans. Compared with 2D facial expressions, 3D facial expressions are more robust with respect to the uncertainties and some external disturbances [13]. Most of these works focused on building generic 3D models based on manually extracted facial landmarks and corresponding features [50, 75].

However, in terms of both 2D and 3D facial expressions, these extracted feature groups are low-level handcrafted features, and while having yielded promising classification ac-

curacy, they are domain-specific and do not perform well when generalized to varying real-world scenarios [88].

To address this issue, high-level visual features or representations were recently extracted using deep learning approaches such as Convolutional Neural Networks (CNNs) [24, 67, 88, 89] and Recurrent Neural Networks (RNNs), including Long Short-Term Memories (LSTMs) and Gated Recurrent Units (GRUs) [53, 91]. Compared with low-level handcrafted features, deep neural network feature extraction can learn hierarchical features that are more generalizable and representative [53]. Not only that, deep neural networks can also automate feature extraction and selection, and achieve a much better emotion recognition performance than traditional machine learning models. For this reason, most of the current commercial facial emotion recognition systems rely on deep learning approaches to build their models and provide relevant recognition services.

3 Study Overview

The aim of this study was to validate five popular commercial facial emotion recognition systems with respect to their performance in real-world usage scenarios, e.g., when people are using their smartphones. We designed two experiments: Experiment 1 investigated the five systems' performance on several facial emotion expression databases; Experiment 2 identified possible reasons for poor-quality photographs in real-world scenarios, and examined how such photographs impacted the systems' performance. Before detailing both experiments, we describe the emotion recognition systems that we used in both experiments.

3.1 Facial Emotion Recognition System

There are numerous facial analysis products on the market today, which provide multiple services typically including face detection, face verification, and emotion recognition. However, previous works tend to select a given product without carefully considering performance before embedding it into their technologies. The selected system, therefore, may sometimes fail to provide adequate performance given the system's characteristics. In this study, in order to evaluate the potential performance and limitation of recognition systems, we chose five widely used commercial products (see Table 1) fulfilling the following criteria:

- (1) has an existing API or SDK;
- (2) recognizes at least six basic emotional categories [36, 37, 35] (highlighted in Table 1);
- (3) returns a confidence or probability value for the detected emotion.

Table 1: Commercial Facial Emotion Recognition Services

Software Platform	Emotions Detected	Response	Approach
Amazon Rekognition	Angry, Disgusted, Fear, Happy, Sad, Surprised Calm, Confused, Unknown	Confidence Range: [0,100]	Deep Neural Network
Baidu Research	Anger, Disgust, Fear, Happy, Sad, Surprise Neutral	Probability Range: [0,1]	Deep Neural Network
Face++	Anger, Disgust, Fear, Happiness, Sadness, Surprise Neutral	Confidence Range: [0,100]	Deep Neural Network
Microsoft Azure	Anger, Disgust, Fear, Happiness, Sadness, Surprise Neutral, Contempt	Intensity Range: [0,1]	Deep Neural Network
Affectiva	Anger, Disgust, Fear, Joy, Sadness, Surprise Contempt	Probability Range: [0,100]	AFFDEX Algorithm

Table 2: Test Set (Static Pictures)

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
ADFES	25	22	23	23	23	21	22	159
RaFD	67	67	67	67	67	67	67	469
WSEFEP	30	30	30	30	30	30	30	210
Total	122	119	120	120	120	118	119	838

4 Experiment 1 - Performance Evaluation using Ideal Images

4.1 Databases

First, we introduce the three facial expression databases that we used in our experiments: the *Amsterdam Dynamic Facial Expressions Set* (ADFES¹) [106], the *Radboud Faces Database* (RaFD²) [68], and the *Warsaw Set of Emotional Facial Expression Pictures* (WSEFEP³) [84]. These contain pictures of various emotional expressions, validated by FACS coders and non-expert human judges [98]. In this study, we evaluated seven emotions that were present in all databases including the six basic emotions (angry, disgust, fear, happy, sad, surprise) and a neutral facial expression.

ADFES: This database contains both dynamic and static emotional expressions displayed by 22 models. The set includes 10 female and 12 male actors (who are either Northern-European or Mediterranean) showing ten emotions (anger, disgust, fear, joy, sadness, surprise, neutral, contempt, pride, and embarrassment). The used set consists of 159 static pictures.

RaFD: This database contains a picture set of emotional expressions displayed by 67 models. This facial database includes 25 females and 42 males (who are either Caucasian or Moroccan) expressing eight emotions (anger, disgust, fear,

happiness, sadness, surprise, neutral, and contempt). The used set consists of 469 static pictures.

WSEFEP: This database contains a set of 210 emotional pictures captured from 30 individuals. The full set includes 16 female and 14 male models covering seven emotions (anger, disgust, fear, enjoyment, sadness, surprise, neutral). The used set consists of 210 static pictures.

Similar to previous work [98], we combined the selected subsets from ADFES, RaFD, and WSEFEP to form a whole set as our validation database, which consists of 122 *angry*, 119 *disgust*, 120 *fear*, 120 *happy*, 120 *sad*, 118 *surprise*, and 119 *neutral* images (838 pictures in total, see Table 2).

4.2 Method

4.2.1 Overall Procedure

Due to the source codes of the commercial tools being encapsulated, i.e., the exact details of the inner workings of these tools are unknown and developers can only access the provided APIs or SDKs, we utilized a black-box testing approach [86].

The overall procedure to determine an emotion expressed by a person within an image is shown in Figure 1. The overall procedure can be divided into five stages as follows: (1) image selection: we chose every image from the aforementioned validation database, and inputted them to each of the five commercial systems; (2) image analysis: after each system completes the analysis, we received and saved the cor-

¹ <https://aice.uva.nl/research-tools/adfes-stimulus-set/adfes-stimulus-set.html>

² <http://www.rafd.nl/>

³ <http://www.emotional-face.org/>

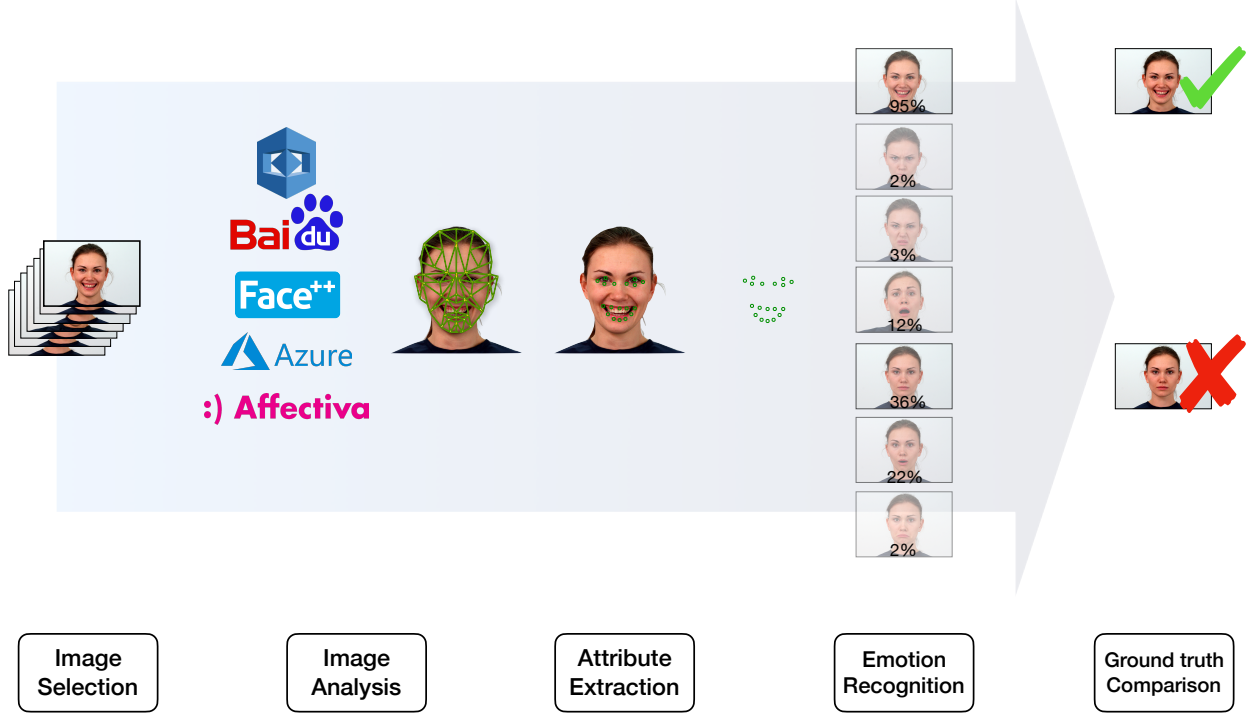


Fig. 1: Overall procedure

responding result responses; (3) attribute extraction: we extracted the emotional attributes along with their confidence or probability values from the results; (4) emotion recognition: we selected the highest confidence or probability value of all detected emotions, and marked this emotion as the detection label; (5) ground truth comparison: we compared the detection label with the ground truth label from the validation database to assess the recognition correctness.

If the detection label of a picture matched with the ground truth label for a given picture, that classification was labeled as “correct”, otherwise the classification was labeled as “incorrect”. If the system failed to detect the face in the picture, e.g., because it was only partially visible, there was no emotion recognition and an empty response: in such cases, the classification was labeled as “missed”.

4.2.2 Comparison Criteria

We used criteria similar to previous works to assess the emotion recognition systems [19, 70, 98], including **Matching Score**, **Precision**, and **F1 score**, which we describe next.

Matching score (MS) measures the system’s ability to correctly recognize an emotion category from a particular class (e.g., happiness, sadness, etc.) of emotional expression pictures, which is also called **true positive rate** or **recall**. A higher matching score means that the system has a higher accuracy on this type of emotional expression pictures. Specifically,

$$MS = \frac{\Sigma \text{true positive}}{\Sigma \text{true positive} + \Sigma \text{false negative} + \Sigma \text{missed}} \quad (1)$$

where “true positive” means the detection label matches the ground truth label; “false negative” means the detection label contradicts the ground truth label; “missed” means the system fails to detect an emotion (no detection label), among all images with the same annotated label.

Precision denotes the system’s reliability of recognizing results, which is also called **positive predictive value**. A higher precision value indicates a higher confidence level, when classifying an expression picture as a certain emotional category. It can be defined as:

$$\text{Precision} = \frac{\Sigma \text{true positive}}{\Sigma \text{true positive} + \Sigma \text{false positive}} \quad (2)$$

Table 3: Experiment 1 Results. Matching Score (%), Precision (%), F1 Score, Overall Accuracy (%), and Detection Rate (%).

		Amazon	Baidu	Face++	Microsoft	Affectiva
Surprise	MS	89.8	97.5	99.2	100.0	97.5
	Precision	73.1	92.0	68.8	74.7	49.8
	F1	0.806	0.947	0.812	0.855	0.659
Fear	MS	68.3	90.0	49.2	62.5	7.5
	Precision	82.8	97.3	95.2	100.0	50.0
	F1	0.749	0.935	0.649	0.769	0.130
Disgust	MS	83.2	91.6	95.0	93.3	89.9
	Precision	96.1	97.3	86.9	100.0	75.4
	F1	0.892	0.944	0.908	0.965	0.820
Happy	MS	100.0	100.0	100.0	100.0	94.2
	Precision	100.0	100.0	99.2	100.0	96.6
	F1	1.000	1.000	0.996	1.000	0.954
Sad	MS	90.8	98.3	82.5	90.0	63.3
	Precision	90.1	92.2	81.2	86.4	62.3
	F1	0.904	0.952	0.818	0.882	0.628
Angry	MS	77.1	96.7	48.4	48.4	50.8
	Precision	82.5	93.7	86.8	92.2	88.6
	F1	0.797	0.952	0.621	0.635	0.646
Neutral	MS	98.3	96.6	97.5	100.0	
	Precision	86.0	99.1	70.3	64.3	
	F1	0.917	0.978	0.817	0.783	
Overall Accuracy		86.8	95.8	81.5	84.7	67.0
Detection Rate		100.0	100.0	100.0	100.0	97.4

F1 Score is a balance between MS and Precision. We calculated an F1 score in which MS and Precision contribute equally by:

$$F1 = 2 \cdot \frac{Precision \cdot MS}{Precision + MS} \quad (3)$$

where the F1 score reaches its best value at 1, and worst value at 0.

4.3 Results

Table 3 depicts all detailed values of Experiment 1. Figure 2 shows the confusion matrix for each emotion recognition system.

We found that the emotion *happy* was the easiest to identify and classify. All emotion recognition systems achieved $\approx 100.0\%$ MS, Precision and F1 score, except for Affectiva (94.2% MS, 96.6% Precision, 0.954 F1 score). For *surprise* and *neutral*, Microsoft Azure had the highest MS (100.0% and 100.0%), but Baidu achieved the highest Precision value (92.0% and 99.1%). In other words, given a *surprise* (or *neutral*) annotated facial expression image, Microsoft Azure

was more likely to recognize the emotion correctly, however, if given a facial expression image without annotated ground truth, the *surprise* (or *neutral*) recognition result from Baidu had more credibility. In contrast, for *fear* Baidu had the highest MS (90.0%), but Microsoft Azure had the highest Precision value (100.0%). For *disgust*, Face++ achieved the highest MS (95.0%), yet Microsoft Azure performed the best on the Precision value (100.0%). Finally, for *sad* and *angry*, Baidu had the best performance on both MS (98.3% and 96.7%) and Precision value (92.2% and 93.7%). In terms of F1 score from each emotion category, Baidu performed the best on *surprise*, *fear*, *sad*, *angry* and *neutral* (0.947, 0.935, 0.952, 0.952, and 0.978); Microsoft Azure performed the best on *disgust* (0.965); while Amazon, Baidu, Face++, and Microsoft Azure performances were equally comparable to each other on *happy* (≈ 1.000).

We measured the **overall accuracy** of each emotion recognition system by globally counting the total number of true positives. Baidu had the highest overall accuracy value (95.8%), with Amazon second (86.8%); followed by Microsoft Azure (84.7%), Face++ (81.5%), and finally Affectiva (67.0%). Furthermore, while Affectiva failed to detect a face in 2.6% of

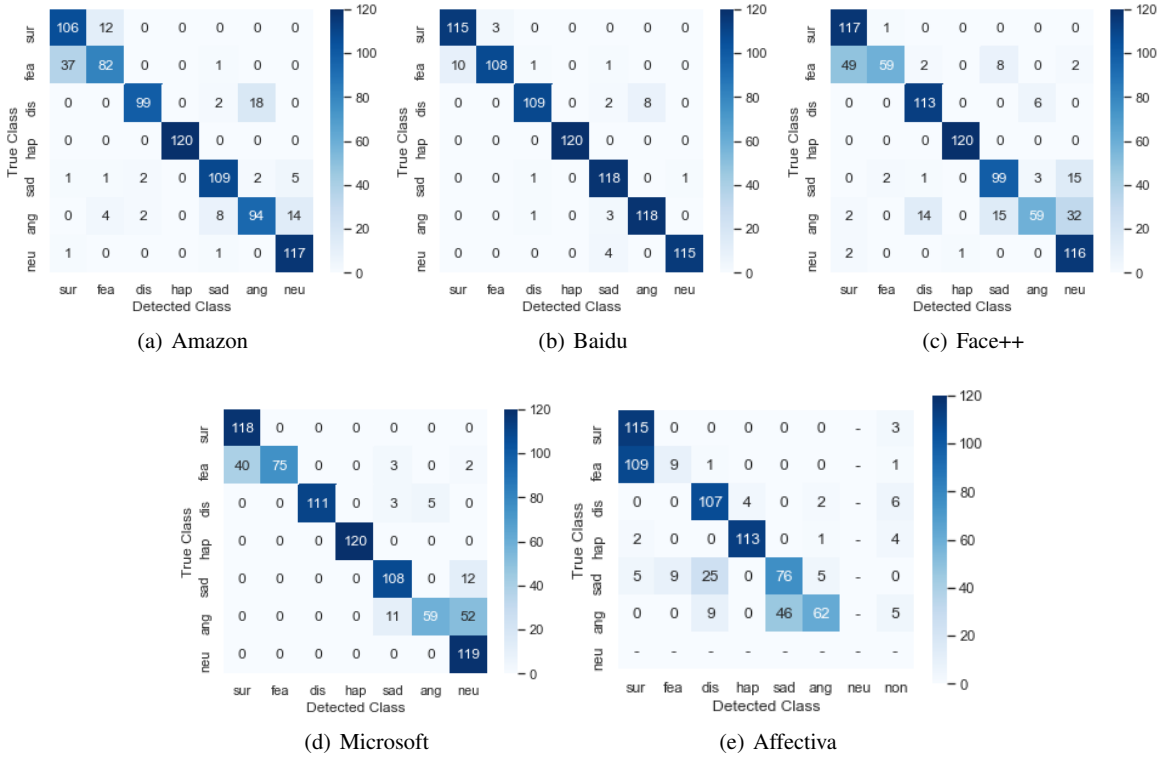


Fig. 2: Confusion Matrix for Experiment 1. (sur = *surprise*; fea = *fear*; dis = *disgust*; hap = *happy*; sad = *sad*; ang = *angry*; neu = *neutral*). Affectiva cannot detect *neutral* expression.

the pictures, other systems performed particularly well and did not produce any detection failures.

Figure 2 shows that *surprise*, *disgust*, *happy*, and *neutral* expressions were rarely confused with other emotions by the recognition systems. For other emotion categories we found the following:

- (1) Amazon, Face++, Microsoft Azure, and Affectiva usually underpredict the *fear* emotion but overpredict the *surprise* emotion;
- (2) Affectiva usually underpredicts the *sad* emotion but overpredicts the *disgust* emotion;
- (3) Face++ and Microsoft Azure usually underpredict the *angry* emotion but overpredict the *neutral* emotion, while

Affectiva usually underpredicts the *angry* emotion but overpredicts the *sad* emotion.

Generally, *fear* and *angry* expressions were the most challenging to identify and classify.

5 Experiment 2 - Performance Evaluation using Reduced-Quality Images

Facial emotion recognition systems typically perform best when inputting high-quality images that show a person's full face, as they are trained and tested on pristine image databases. However, in practice, such requirements are un-

Table 4: Augmented Test Set (Static Pictures)

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Rotation	732	714	720	720	720	708	714	5028
Partial Face	976	952	960	960	960	944	952	6704
Brightness	732	714	720	720	720	708	714	5028
Blur	366	357	360	360	360	354	357	2514
Noise	366	357	360	360	360	354	357	2514
Total	3172	3094	3120	3120	3120	3068	3094	21788

likely to be satisfied, as distortions are quite likely to be encountered during image acquisition, transmission, or storage [29]. For example, in recent work, where researchers continuously collected photos from the front-facing camera of 11 smartphone users for 2 weeks, only 29% of the time was the user’s full face visible, while most of the time the face was only partially visible [64]. Thus, during smartphone use it is not realistic to expect that images collected will always contain the users’ full faces. This is of particular importance as several previous studies have shown that the way people hold their smartphones can influence the performance of facial analysis algorithms [18, 64, 69].

This problem attracts much attention from researchers since it is impossible to avoid when implementing facial emotion recognition algorithms in real-life products. Especially for applications that need real-time facial emotion recognition capability (e.g., E-learning monitoring or mental health management systems), a more robust recognition framework that can (nearly) continually detect human emotion over time, regardless of image or video quality, is essential. Researchers have begun attempting to apply off-the-shelf recognition systems in different environments [97]. Thus, a comprehensive validation study of current commercial facial emotion

recognition systems on images of varying quality is warranted.

After a comprehensive consideration of in-the-wild user behaviours and image quality distortions, we identified five potential reasons for poor-quality natural photographs: *rotation*, *partial face*, *brightness*, *blur*, and *noise*. These are the most common distortions known to impact the performance of deep neural networks [65, 30, 90, 31, 26, 100]. We augmented the validation database of Experiment 1 using these five distortions (see Table 4), and then quantitatively investigated changes in performance for each of the facial recognition systems. To the best of our knowledge, we are the first to take into account both user behaviours and image quality distortions in a large-scale evaluation of commercial facial emotion recognition systems on multi-type and multi-granularity level.

5.1 Rotation

Rotation occurs when photos are taken from diverse angles by users. The rotation operation is achieved through a geometric transformation by the specified angle θ . We used OpenCV and Python to rotate each photo around its centre from -45° to 45° in steps of 15° , without scaling or

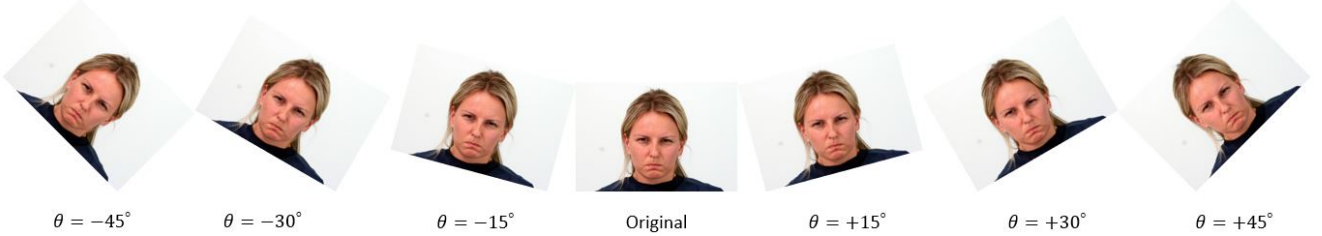


Fig. 3: Example of Rotation Manipulations

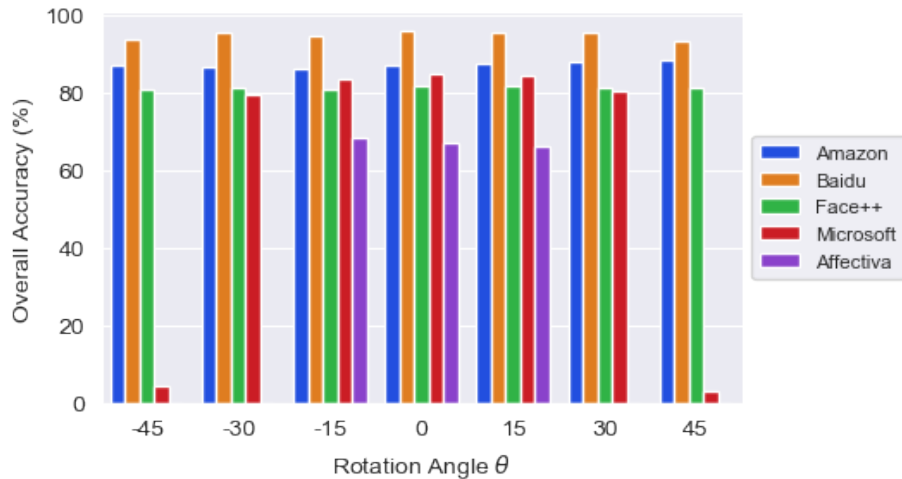


Fig. 4: Overall accuracy under different rotation angles

Table 5: Detailed results of Rotation Manipulations. OA = Overall Accuracy (%). DR = Detection Rate (%).

		Amazon	Baidu	Face++	Microsoft	Affectiva
$\theta = -45^\circ$	OA	86.8	93.4	80.8	4.2	0.0
	DR	100.0	100.0	100.0	5.1	0.0
$\theta = -30^\circ$	OA	86.4	95.6	81.0	79.5	0.0
	DR	100.0	100.0	100.0	99.5	0.0
$\theta = -15^\circ$	OA	86.0	94.6	80.8	83.3	68.3
	DR	100.0	100.0	100.0	100.0	96.9
Original	OA	86.8	95.8	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
$\theta = 15^\circ$	OA	87.4	95.6	81.6	84.3	66.2
	DR	100.0	100.0	100.0	100.0	94.0
$\theta = 30^\circ$	OA	87.7	95.2	81.2	80.2	0.0
	DR	100.0	100.0	100.0	99.3	0.0
$\theta = 45^\circ$	OA	88.1	93.2	81.0	3.1	0.0
	DR	100.0	100.0	100.0	5.0	0.0

cropping. We set 45° as the boundary value because pictures will be auto-rotated on some smartphone devices if they are tilted too far. For example, if taking photos in portrait mode on some devices, the resulting photos will be rotated by 90 degrees. Figure 3 shows an example of the image rotation. Here, a negative θ means clockwise rotation, and vice versa.

Table 5 and Figure 4 show the emotion recognition results for the rotated images. We found that Amazon, Baidu, and Face++ performed well and were stable to different θ . Their overall accuracies for the rotated images were almost as good as for the original images. However, Microsoft Azure and Affectiva were vulnerable to rotation manipulation. Their detection accuracies decreased significantly when the absolute value of θ increased. Overall, Baidu had the highest accuracy value, followed by Amazon and Face++.

There are several specific results worth mentioning: (1) Microsoft Azure’s performance was resilient when $|\theta| = 15^\circ, 30^\circ$; however, performance began to decrease when $|\theta| > 30^\circ$, and deteriorated further at $|\theta| = 45^\circ$ ($OA = 4.2\%$, $\theta = -45^\circ$; $OA = 3.1\%$, $\theta = 45^\circ$); (2) there was little change to Affectiva’s performance at $|\theta| = 15^\circ$; Affectiva presented a substantial performance decrease when $|\theta| > 15^\circ$, and it stopped functioning at $|\theta| = 30^\circ, 45^\circ$ ($OA = 0.0\%$); (3) compared with Microsoft Azure, Affectiva was more sensitive to rotation manipulation; (4) when $|\theta| = 15^\circ$, Microsoft Azure performed better than Face++.

In terms of detection rates, Amazon, Baidu, and Face++ had the best results with 100.0%. However, Microsoft Azure only detected a face in $\approx 5\%$ of the photos when $|\theta| = 45^\circ$; Affectiva performed worse as it failed to detect a face in all photos when $|\theta| = 30^\circ, 45^\circ$, further explaining why the overall accuracy of these two systems diminished with the change in rotation degrees.

5.2 Partial Face

Partially visible face images are common when devices are naturally held by users [64]. We considered both portrait and landscape orientations, and further classified partial face into eight sub-categories: left quarter (LQ), left half (LH), right quarter (RQ), right half (RH), up quarter (UQ), up half (UH), down quarter (DQ), and down half (DH). We implemented the CascadeClassifier in OpenCV [1], and then manipulated photos for the partial face experiment using two steps: first, we determined the rectangular region of the face using the CascadeClassifier; second, we shifted this region on the x-axis or y-axis through coordinate translation. Figure 5 shows an example of partial face manipulations in both x-axis and y-axis. We did not consider three quarters scale face translations (e.g., left three quarters or up three quarters), since the vast majority of facial landmarks (or AUs) are concentrated in brow, lid, eyes, nose, cheek, and lip [40]. When shifting the face in a photo by three quarters, it will leave only ears (left, right three quarters), chin (up three quarters), or forehead (down three quarters), that contribute little to facial emotion recognition.

An overview of the accuracy values and detection rates are shown in Table 6. For partial face manipulation on the x-axis, Figure 6(a) reveals that Amazon, Baidu, and Face++ were relatively resilient to partial face manipulation. Their performance began to worsen only when facing a high-level visible face change (LH or RH). Overall, Amazon showed the greatest resiliency and had the highest overall accuracy values at high-level visible face changes (LH and RH); Baidu performed best at low-level visible face changes (LQ and RQ). As for Microsoft Azure, it performed well at low-level visible face changes with the second-highest overall accu-

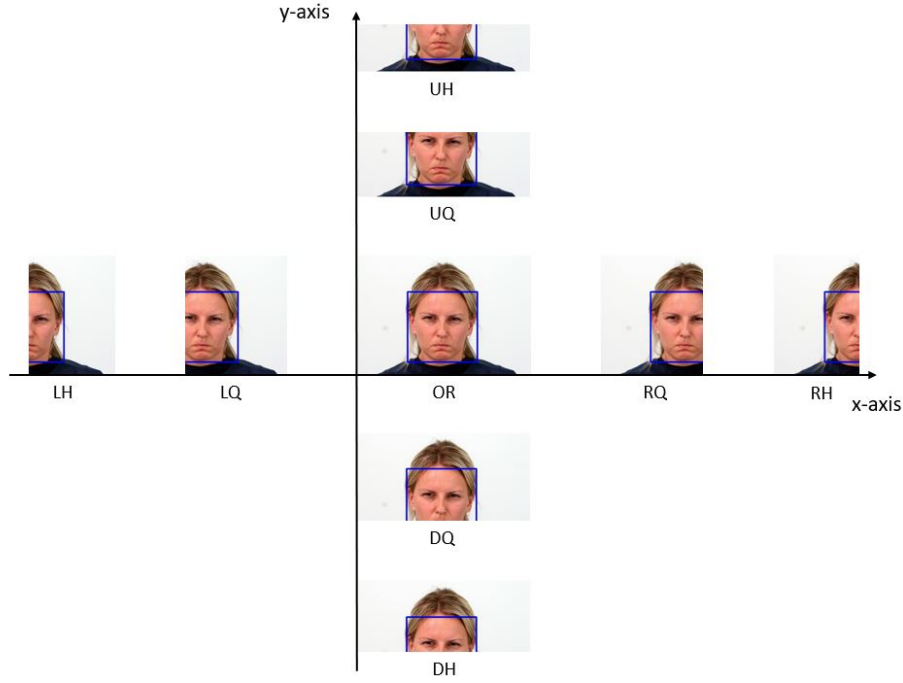


Fig. 5: Example of Partial Face Manipulations

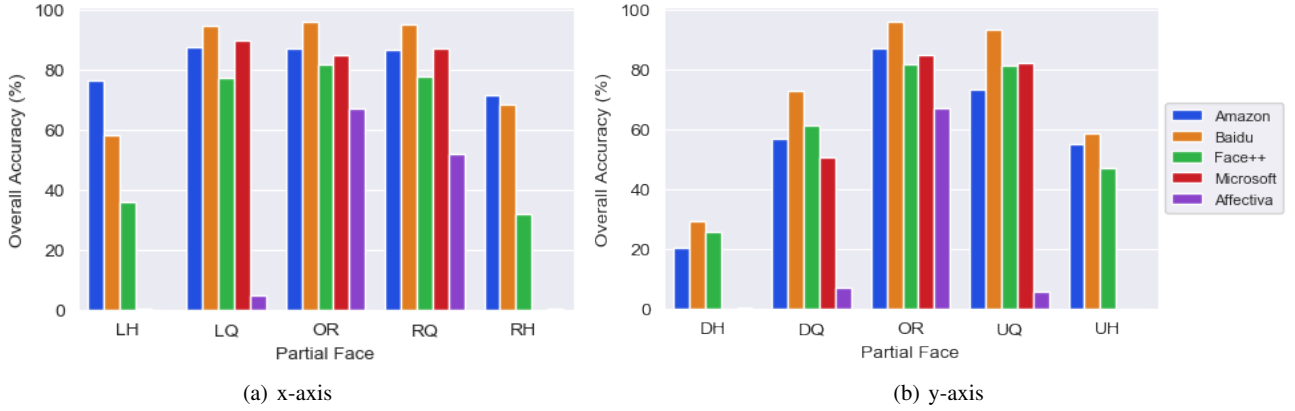


Fig. 6: Overall accuracy of partial face manipulations. OR = Original.

racy values (89.6% and 87.0%) for LQ and RQ. However, it was quite sensitive to high-level visible face changes, and struggled to detect any face for LH and RH. Affectiva had the lowest accuracy values and detection rates at all levels of partial face manipulation. It is also noteworthy that the overall accuracy of all systems was approximately symmetrical to the original image, except for Affectiva. This is probably due to similar characteristics (e.g., approximate number of facial landmarks) on both sides of the face.

For partial face manipulation on the y-axis, Figure 6(b) shows that all systems were highly sensitive to visible face changes in portrait orientation. Even for low-level visible face changes, their performance decreased significantly. In

general, Baidu produced the best overall accuracy results; Amazon and Face++ performed just slightly worse. Microsoft Azure had a similar performance to Amazon and Face++ at low-level visible face changes, but failed completely to detect a face at high-level visible face changes (0.0% and 0.0% at both DH and UH). Affectiva had a large number of detection failures from low-level visible changes (12.9% detection rate at DQ and 11.8% detection rate at UQ). Furthermore, when examining the trend of the systems' performance, our results show that all systems, except for Affectiva, were always more sensitive to the downward direction manipulation (e.g., DQ and DH) than the upward direction manipulation (e.g., UQ and UH). This is likely due to these

Table 6: Detailed results of Partial Face Manipulations. OA = Overall Accuracy (%). DR = Detection Rate (%).

		Amazon	Baidu	Face++	Microsoft	Affectiva
LH	OA	76.5	57.9	35.7	0.1	0.0
	DR	100.0	88.8	69.5	0.5	0.0
LQ	OA	87.2	94.6	77.0	89.6	4.7
	DR	100.0	100.0	100.0	100.0	6.8
RQ	OA	86.3	94.8	77.8	87.0	51.7
	DR	100.0	100.0	100.0	99.6	78.4
RH	OA	71.6	68.3	31.9	0.0	0.1
	DR	100.0	96.2	66.2	0.0	0.1
Original	OA	86.8	95.8	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
DH	OA	20.4	29.4	25.8	0.0	0.1
	DR	99.8	58.2	99.5	0.0	0.1
DQ	OA	56.9	72.9	61.3	50.5	7.0
	DR	100.0	100.0	100.0	96.8	12.9
UQ	OA	73.0	93.1	81.2	82.1	5.6
	DR	100.0	100.0	100.0	100.0	11.8
UH	OA	54.9	58.5	47.0	0.0	0.0
	DR	100.0	90.3	95.2	0.0	0.3

systems having a bias towards features present in the lower part of a face.

5.3 Brightness

Brightness is a key factor that impacts the quality of images taken in real-world scenarios. A universally accepted approach is to divide brightness problems into two categories: overexposure and underexposure [58]. When taking a photo with strong lighting, the image sensor will capture too much light and make the photo lose some highlight details (called overexposure). On the contrary, when taking a photo with poor lighting, the image sensor will capture too little light and make the photo lose some shadow details (called underexposure). In this experiment, we varied the brightness of the images from 75% reduction (-75%) to 75% enhancement (+75%) in steps of 15% by decreasing/increasing the value of every pixel in each image. We note that an image will become completely black or white if the brightness is

reduced or increased by 100%. Figure 7 shows an example of the brightness levels we tested.

Results for all facial emotion recognition systems are shown in Table 7. Figure 8 depicts the overall accuracy of all systems under different brightness conditions. Our results show that the ranking of system in terms of overall accuracy remained consistent at low and medium degrees of brightness changes (25% and 50% reduction/enhancement). Baidu was the top-performing system; Amazon and Microsoft Azure ranked similarly (but Amazon performed better at a medium degree of enhancement; Microsoft Azure performed better at medium degree of reduction); followed by Face++ and Affectiva. However, for the highest degree of brightness changes (75% reduction/enhancement), the performance of all systems experienced a sharp decline: Baidu remained the relatively most accurate with dark images (36.5%); whereas Amazon achieved better performance on the overall accuracy for bright images (45.2%) with a high detection rate of 95.0%. Another interesting observation is that the detection rates of Amazon were relatively stable ($\geq 94\%$) across all

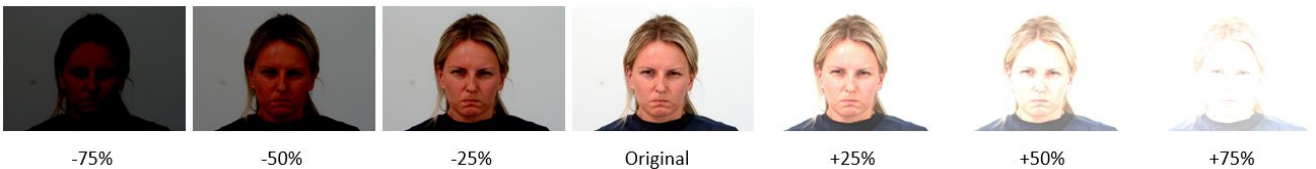


Fig. 7: Example of Brightness Manipulations

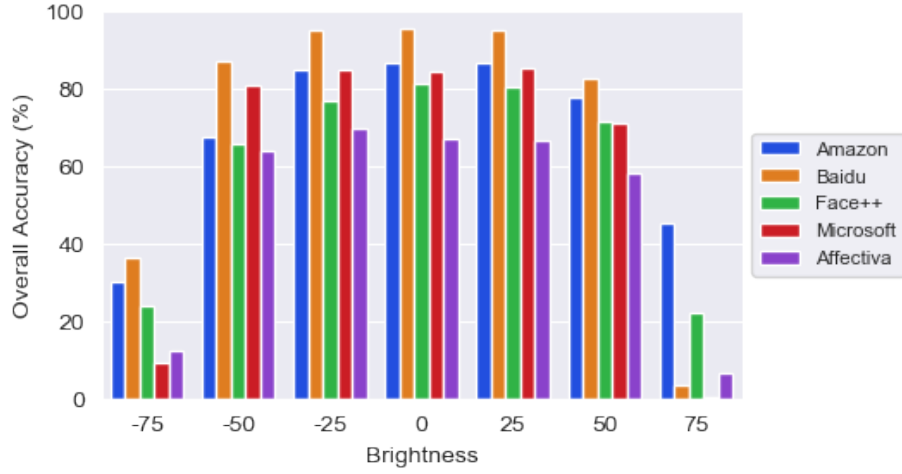


Fig. 8: Overall accuracy under different brightness conditions

Table 7: Detailed results of Brightness Manipulations. OA = Overall Accuracy (%). DR = Detection Rate (%).

		Amazon	Baidu	Face++	Microsoft	Affectiva
-75%	OA	30.6	36.5	24.2	9.7	12.8
	DR	94.9	69.2	77.0	14.8	29.5
-50%	OA	67.5	87.2	65.8	80.9	64.0
	DR	100.0	99.8	100.0	99.5	96.4
-25%	OA	85.1	95.0	77.1	84.8	70.0
	DR	100.0	100.0	100.0	100.0	98.2
Original	OA	86.8	95.8	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
25%	OA	86.8	95.2	80.6	85.4	66.6
	DR	100.0	100.0	100.0	100.0	96.8
50%	OA	78.0	82.6	71.8	71.1	58.1
	DR	100.0	90.7	99.6	92.4	90.8
75%	OA	45.2	3.8	22.4	0.5	7.0
	DR	95.0	5.6	53.0	0.8	12.2

degrees of brightness changes, which contributed towards relatively good performance in terms of overall accuracy of this system (Table 7).

5.4 Blur

Blur is a common factor in poor-quality images and can be caused by long exposure time, movement of the camera or inappropriate focus when using a smartphone or other hand-held cameras [109,29]. In this experiment, we used Gaus-



Fig. 9: Example of Blur Manipulations

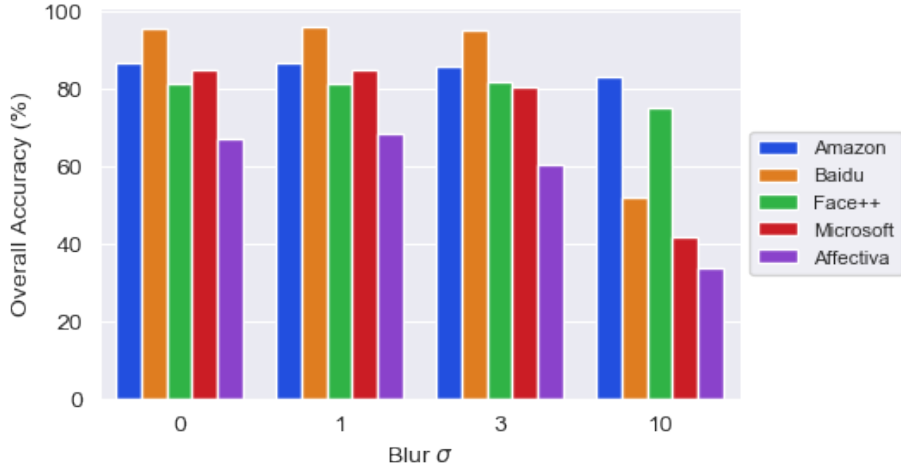


Fig. 10: Overall accuracy under different blur conditions

Table 8: Detailed results of Blur Manipulations. OA = Overall Accuracy (%). DR = Detection Rate (%).

		Amazon	Baidu	Face++	Microsoft	Affectiva
Original	OA	86.8	95.8	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
$\sigma = 1$	OA	86.8	96.1	81.5	85.0	68.6
	DR	100.0	100.0	100.0	100.0	96.9
$\sigma = 3$	OA	86.0	95.0	81.9	80.3	60.5
	DR	100.0	100.0	100.0	100.0	95.3
$\sigma = 10$	OA	83.2	52.2	75.1	41.8	33.5
	DR	100.0	100.0	100.0	100.0	82.8

sian Blur with different granularities implemented by the Python Imaging Library (PIL) to model and simulate image blur that can occur in a real-world scenario. As recommended by Gedraite and Hadad [46], we set the blur radius (the standard deviation of Gaussian) σ to three values: 1, 3, and 10, to represent low, medium, and high image blurring respectively. An example of this manipulation is shown in Figure 9.

We show the overall results in Table 8. Figure 10 reveals that all systems showed stable performance when it comes to low ($\sigma = 1$) and medium ($\sigma = 3$) levels of image blur. Specifically, Baidu achieved the best performance with 96.1% ($\sigma = 1$) and 95.0% ($\sigma = 3$) overall accuracy values; Amazon had the second-best performance with 86.8% and 86.0% accuracy values (on $\sigma = 1, 3$); Microsoft Azure (85.0%) performed better than Face++ (81.5%) at low-level blurring, but Face++ (81.9%) yielded a higher accuracy result than Microsoft Azure (80.3%) at medium-level blurring. Another finding was that Baidu, Microsoft Azure, and Affectiva showed a significant decrease at high-level blurring, in which their overall accuracies fell by 45.5% ($\frac{95.8-52.2}{95.8}$), 50.6%, and 50.0% respectively. However, there was little ef-

fect on Amazon and Face++. Therefore, at high-level blurring, Amazon worked best; Face++ was the second-best; followed by Baidu, Microsoft Azure, and Affectiva. In terms of stability, our results show that Amazon and Face++ performed better than the other systems.

5.5 Noise

Image noise is usually a form of electronic noise [17], and it is typically caused by the internal sensors of cameras [17, 109, 29]. In practice, noise can obscure image details, degrade its quality, or even make the image completely unusable. Previous works have commonly modeled image noise using Additive White Gaussian Noise (AWGN) [73, 85], since it can effectively mimic the noise that occurs in real world. Therefore, in this experiment, we also used AWGN to simulate natural noise in digital images. We set the standard deviation of the noise to be: 10, 40, and 70, to represent low, medium, and high noise levels respectively. Figure 11 shows an example of noise manipulations.



Fig. 11: Example of Noise Manipulations

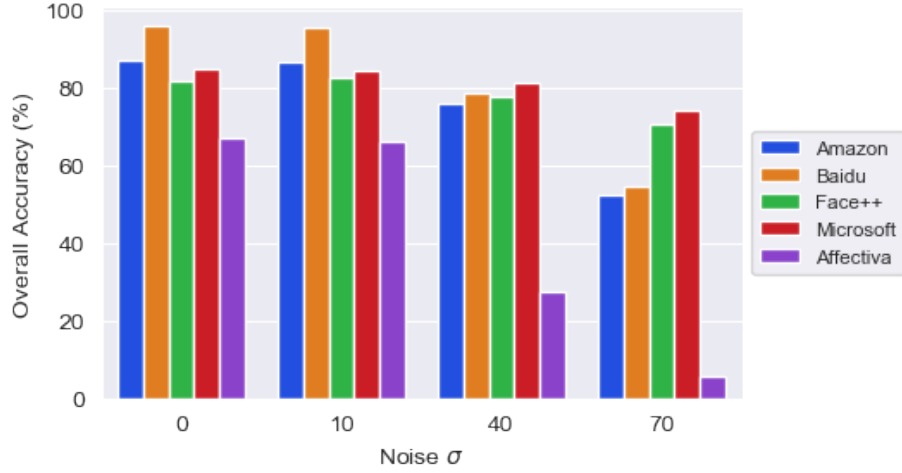


Fig. 12: Overall accuracy under different noise conditions

Table 9: Detailed results of Noise Manipulations. OA = Overall Accuracy (%). DR = Detection Rate (%).

		Amazon	Baidu	Face++	Microsoft	Affectiva
Original	OA	86.8	95.8	81.5	84.7	67.0
	DR	100.0	100.0	100.0	100.0	97.4
$\sigma = 10$	OA	86.3	95.2	82.6	84.4	66.1
	DR	100.0	100.0	100.0	100.0	98.6
$\sigma = 40$	OA	75.7	78.5	77.8	81.2	27.3
	DR	100.0	100.0	100.0	100.0	93.9
$\sigma = 70$	OA	52.4	54.7	70.3	74.2	5.6
	DR	100.0	95.9	99.2	99.8	30.5

Table 9 summarizes the overall results of the noise manipulations. Figure 12 shows that (1) at low noise level ($\sigma = 10$), all systems' performance remained basically unchanged, and kept the same ranking positions: Baidu, Amazon, Microsoft Azure, Face++, and Affectiva; (2) at medium noise level ($\sigma = 40$), all systems showed a decrease in accuracy values, but the performance of Microsoft Azure and Face++ deteriorated at a slower rate (by 4.1% and 4.5% respectively), whereas the performance of Baidu, Amazon and Affectiva deteriorated quicker (by 18.1%, 12.8%, and 59.3% respectively); (3) at high noise level ($\sigma = 70$), the performance of Baidu, Amazon, and Affectiva still fell much quicker compared to that of Microsoft Azure and Face++. In general,

Microsoft Azure and Face++ showed greater reliability in terms of processing images with medium- and high-level noise strength.

6 Discussion

In this study, we evaluated the performance of five commercial facial emotion recognition systems both with high-quality front facing images and images whose quality had been reduced by common real-world distortions (rotation, partial face, brightness, blur, noise).

6.1 Facial Emotion Recognition under Ideal Conditions

When identifying emotions from full-face high-quality photos (Experiment 1), the Baidu system had the highest overall accuracy value (95.8%), followed by Amazon (86.8%), Microsoft Azure (84.7%), Face++ (81.5%), and Affectiva (67.0%). Likewise, a recent validation analysis using similar facial expression databases [98] found that Affectiva had an overall accuracy value of 73%, which is comparable to our result of 67.0%. That analysis also reported that Affectiva usually confused *fear* with *surprise* and *anger* with *sadness*, which is in line with our findings. Furthermore, previous works using human judges, have shown that participants achieved an overall accuracy between 82% and 85% [70, 84, 106, 68]. Our results highlight that most of the popular commercial systems, particularly those using a deep neural network approach, can achieve accuracy rates comparable to human judges, and sometimes even outperform them.

Another noteworthy observation is that the tested systems were most accurate when identifying and classifying the *happy* emotion, and least accurate when processing the *fear* emotion. Similarly, previous works have shown that the *happy* emotion is also the easiest emotion to distinguish for human observers, and the *fear* emotion is the hardest [48, 68]. This similarity is not surprising given that automated facial emotion recognition systems rely on human-generated labels to improve their performance.

6.2 Facial Emotion Recognition with Reduced-Quality Images

From the results of Experiment 2, we find that the tested systems have different advantages and limitations depending on the particular reason behind the poor quality of the images. This is most likely because different systems rely on their own training datasets, with their own set of biases. This finding highlights that a *one system approach* for facial emotion recognition in real-world scenarios is far from ideal. Instead, a robust system that uses facial expressions for emotion recognition should first determine the quality of the image and the level of distortion, and then route the image to the facial emotion recognition system that is most likely to produce a reliable result (or discard the image if the quality is particularly poor).

Our recommendations are the following:

- For rotation, we recommend Baidu as the best system, and Amazon as the second-best;
- For partial face, on the x-axis, we recommend the use of Baidu for low-level visible face changes, and Amazon for high-level visible face changes; on the y-axis, we recommend Baidu as first choice;

- For brightness, we recommend using Baidu when encountering underexposure or low- and medium-level overexposure, and Amazon when encountering high-level overexposure. We do not recommend Baidu at high-level overexposure;
- For blur, we recommend using Baidu when processing low- and medium-level image blurring, and Amazon or Face++ system when processing high-level image blurring;
- For noise, we recommend the use of Baidu at low-level noise strength, and the use of Microsoft Azure or Face++ at medium- and high-level noise strength.

The reason why facial emotion recognition systems have poor overall accuracy rates in certain conditions is also noteworthy. There are two kinds of detection failures: face detection failure (“missed”) and emotion detection failure (“incorrect”). In detail,

- For Amazon Rekognition, emotion detection failure is the main issue, and present itself as an opportunity for further improvement;
- For Baidu and Face++, not only emotion detection but also face detection needs to be improved for high-level partial face and underexposure/overexposure manipulations;
- Microsoft Azure suffers from both emotion detection failure and face detection failure at high-level changes of rotation, partial face and brightness manipulations;
- Affectiva is very sensitive to most of the low- and high-level manipulations.

Finally, we note that the evaluated systems are likely to be improved in the future, and that other systems may emerge with a whole new set of advantages and limitations. Also, there are studies that show human beings have the capability to recognize very distorted images [30, 102, 25, 7, 31], which means existing recognition systems still have potential to overcome their limitations and achieve stronger robustness. Thus, another contribution of our work is the categorisation and operationalization of the distortions that can lead to poor-quality images and which are likely to occur in real-world scenarios, e.g., smartphone usage. These manipulations can be used as standard methods for evaluating other facial emotion recognition systems, or for further improving the five systems validated in this study. Not only that, researchers can also use our approach to build their own datasets for validation analysis according to different needs.

6.3 Limitations

There are limitations in our study. First, in Experiment 2, we analyzed and assessed the impacts of image manipulations using a relatively high-level of granularity (e.g., increments of 15 degrees for the rotation condition). Thus, a

finer-grained manipulation of the images may yield a more detailed validation study in the future. Second, we evaluated the impact of different distortion factors independently. Future work could investigate the potential for combination effects and even the prioritization of the individual factors. Third, it is likely that most of these systems will update their frameworks and corresponding network features and weights regularly, which may influence their performance under the different presented conditions. To tackle this issue, we recommend re-using or even automating our method to evaluate new commercial and non-commercial systems that may appear in the future.

7 Conclusion

In this study, we utilized three commonly-used facial expression databases and proposed a series of manipulations to simulate potential image quality reduction problems that are likely to occur during real-world smartphone usage. We conducted two experiments in order to assess the facial emotion recognition capabilities of five popular commercial emotion recognition systems: Amazon, Baidu, Face++, Microsoft Azure, and Affectiva. Through our experiments we found that each system has its own strengths and limitations under different distortion conditions (rotation, partial face, brightness, blur, noise). Based on our findings, we offer recommendations on how to achieve reliable facial emotion recognition results for applications in-the-wild, by selecting different systems depending on the nature of the captured image. Finally, we recommend the use of our image manipulation methods for future testing of facial emotion recognition system performance.

Compliance with Ethical Standards

Funding This work is supported by the Australian Research Council (DP190102627).

Conflict of Interest Authors, Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves, declare that they have no conflict of interest.

References

1. OpenCV: Cascade classifier training. URL https://docs.opencv.org/3.1.0/dc/d88/tutorial_traincascade.html#gsc.tab=0
2. Affectiva: Home - affectiva : Affectiva (2019). URL <https://www.affectiva.com/>
3. Albohn, D.N., Adams Jr, R.B.: Social vision: At the intersection of vision and person perception. In: Neuroimaging personality, social cognition, and character, pp. 159–186. Elsevier (2016)
4. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 579–586. Association for Computational Linguistics (2005)
5. AlZoubi, O., Calvo, R.A., Stevens, R.H.: Classification of eeg for affect recognition: an adaptive approach. In: Australasian Joint Conference on Artificial Intelligence, pp. 52–61. Springer (2009)
6. Amazon: Amazon rekognition – video and image - aws (2019). URL https://aws.amazon.com/rekognition/?nc1=h_ls
7. Bachmann, T.: Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? European Journal of Cognitive Psychology **3**(1), 87–103 (1991)
8. Baidu: Baidu ai (2019). URL <https://ai.baidu.com/docs#/Face-Detect-V3/top>
9. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D.: Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychological Science in the Public Interest **20**(1), 1–68 (2019)
10. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: Development and applications to human computer interaction. In: 2003 Conference on computer vision and pattern recognition workshop, vol. 5, pp. 53–53. IEEE (2003)
11. van Berkel, N., Goncalves, J., Koval, P., Hosio, S., Dingler, T., Ferreira, D., Kostakos, V.: Context-informed scheduling and analysis: Improving accuracy of mobile self-reports. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19 (2019)
12. Bernin, A., Müller, L., Ghose, S., von Luck, K., Grecos, C., Wang, Q., Vogt, F.: Towards more robust automatic facial expression recognition in smart environments. In: Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments, pp. 37–44. ACM (2017)
13. Berretti, S., Amor, B.B., Daoudi, M., Del Bimbo, A.: 3d facial expression recognition using sift descriptors of automatically detected keypoints. The Visual Computer **27**(11), 1021 (2011)
14. Berretti, S., Del Bimbo, A., Pala, P.: Automatic facial expression recognition in real-time from dynamic sequences of 3d face scans. The Visual Computer **29**(12), 1333–1350 (2013)
15. Bettadapura, V.: Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722 (2012)
16. Bourel, F., Chibelushi, C.C., Low, A.A.: Robust facial expression recognition using a state-based model of spatially-localised facial dynamics. In: Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 113–118. IEEE (2002)
17. Boyat, A.K., Joshi, B.K.: A review paper: noise models in digital image processing. arXiv preprint arXiv:1505.03489 (2015)
18. Bröhl, C., Mertens, A., Ziefle, M.: How do users interact with mobile devices? an analysis of handheld positions for different technology generations. In: International Conference on Human Aspects of IT for the Aged Population, pp. 3–16. Springer (2017)
19. Bryant, D., Howard, A.: A comparative analysis of emotion-detecting ai systems with respect to algorithm performance and dataset diversity. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 377–382. ACM (2019)
20. Calvo, R.A., D’Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on affective computing **1**(1), 18–37 (2010)
21. Carlotta Olivetti, E., Violante, M.G., Vezzetti, E., Marcolin, F., Eynard, B.: Engagement evaluation in a virtual learning environment via facial expression recognition and self-reports: A preliminary approach. Applied Sciences **10**(1), 314 (2020)
22. Carroll, J.M., Russell, J.A.: Do facial expressions signal specific emotions? judging emotion from the face in context. Journal of personality and social psychology **70**(2), 205 (1996)

23. Celma, O.: Music recommendation. In: Music recommendation and discovery, pp. 43–85. Springer (2010)
24. Chang, W.Y., Hsu, S.H., Chien, J.H.: Fatauva-net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 17–25 (2017)
25. Chen, Y., McBain, R., Norton, D.: Specific vulnerability of face perception to noise: A similar effect in schizophrenia patients and healthy individuals. *Psychiatry research* **225**(3), 619–624 (2015)
26. Cheng, Y., Jiang, B., Jia, K.: A deep structure for facial expression recognition under partial occlusion. In: 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 211–214. IEEE (2014)
27. Coulson, M.: Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior* **28**(2), 117–139 (2004)
28. Darwin, C.: The expression of the emotions in man and animals, anniversary ed., p. ekman, ed (1872)
29. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 eighth international conference on quality of multimedia experience (QoMEX), pp. 1–6. IEEE (2016)
30. Dodge, S., Karam, L.: Can the early human visual system compete with deep neural networks? In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2798–2804 (2017)
31. Dodge, S., Karam, L.: A study and comparison of human and deep learning recognition performance under visual distortions. In: 2017 26th international conference on computer communication and networks (ICCCN), pp. 1–7. IEEE (2017)
32. Dupré, D., Anđelić, N., Morrison, G., McKeown, G.: Accuracy of three commercial automatic emotion recognition systems across different individuals and their facial expressions. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 627–632. IEEE (2018)
33. Ekman, P.: Methods for measuring facial action. *Handbook of methods in nonverbal behavior research* pp. 45–90 (1982)
34. Ekman, P.: Expression and the nature of emotion. *Approaches to emotion* **3**, 19–344 (1984)
35. Ekman, P.: An argument for basic emotions. *Cognition & emotion* **6**(3–4), 169–200 (1992)
36. Ekman, P.: Basic emotions. *Handbook of cognition and emotion* **98**(45–60), 16 (1999)
37. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. *Journal of personality and social psychology* **17**(2), 124 (1971)
38. Ekman, P., Friesen, W.V.: Unmasking the face: A guide to recognizing emotions from facial clues. Ishk (2003)
39. Ekman, P., Friesen, W.V., Ellsworth, P.: Emotion in the Human Face: Guide-lines for Research and an Integration of Findings: Guidelines for Research and an Integration of Findings. Pergamon (1972)
40. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system: The manual on cd rom. A Human Face, Salt Lake City pp. 77–254 (2002)
41. Ekman, P., Friesen, W.V., O’sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., et al.: Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology* **53**(4), 712 (1987)
42. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition* **44**(3), 572–587 (2011)
43. Fridlund, A.J.: Evolution and facial action in reflex, social motive, and paralanguage. *Biological psychology* **32**(1), 3–100 (1991)
44. Friesen, E., Ekman, P.: Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* **3** (1978)
45. Friesen, W.V., Ekman, P., et al.: Emfacs-7: Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco **2**(36), 1 (1983)
46. Gedraite, E.S., Hadad, M.: Investigation on the effect of a gaussian blur in image filtering and segmentation. In: Proceedings ELMAR-2011, pp. 393–396. IEEE (2011)
47. Gellman, M.D.: Behavioral medicine. Springer (2013)
48. Goeleven, E., De Raedt, R., Leyman, L., Verschuere, B.: The karolinska directed emotional faces: a validation study. *Cognition and emotion* **22**(6), 1094–1118 (2008)
49. Goncalves, J., Pandab, P., Ferreira, D., Ghahramani, M., Zhao, G., Kostakos, V.: Projective testing of diurnal collective emotion. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp ’14, p. 487–497. New York, NY, USA (2014)
50. Gong, B., Wang, Y., Liu, J., Tang, X.: Automatic facial expression recognition on a single 3d face by exploring shape deformation. In: Proceedings of the 17th ACM international conference on Multimedia, pp. 569–572 (2009)
51. Google: Vision ai — derive image insights via ml — cloud vision api — google cloud (2019). URL <https://cloud.google.com/vision/>
52. Gross, J.J.: Emotion regulation: Past, present, future. *Cognition & emotion* **13**(5), 551–573 (1999)
53. Gu, Y., Li, X., Huang, K., Fu, S., Yang, K., Chen, S., Zhou, M., Marsic, I.: Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder. In: 2018 ACM Multimedia Conference on Multimedia Conference, pp. 537–545. ACM (2018)
54. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.: Hybrid attention based multimodal network for spoken language classification. In: Proceedings of the conference. Association for Computational Linguistics. Meeting, vol. 2018, pp. 2379–2390. NIH Public Access (2018)
55. Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., Marsic, I.: Multimodal affective analysis using hierarchical attention strategy with word-level alignment. *arXiv preprint arXiv:1805.08660* (2018)
56. van der Haar, D.T.: Student emotion recognition in computer science education: A blessing or curse? In: International Conference on Human-Computer Interaction, pp. 301–311. Springer (2019)
57. Heraz, A., Frasson, C.: Predicting the three major dimensions of the learner’s emotions from brainwaves. *International Journal of Computer Science* **2**(3), 187–193 (2007)
58. Hou, L., Ji, H., Shen, Z.: Recovering over-/underexposed regions in photographs. *SIAM Journal on Imaging Sciences* **6**(4), 2213–2235 (2013)
59. Howard, A., Zhang, C., Horvitz, E.: Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In: 2017 IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO), pp. 1–7. IEEE (2017)
60. Huang, D., De la Torre, F.: Bilinear kernel reduced rank regression for facial expression synthesis. In: European Conference on Computer Vision, pp. 364–377. Springer (2010)
61. Izard, C.E.: The face of emotion. (1971)
62. Jack, R.E., Garrod, O.G., Yu, H., Caldara, R., Schyns, P.G.: Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* **109**(19), 7241–7244 (2012)
63. Keltner, D., Ekman, P., Gonzaga, G., Beer, J.: Facial expression of emotion. Guilford Publications (2000)
64. Khamis, M., Baier, A., Henze, N., Alt, F., Bulling, A.: Understanding face and eye visibility in front-facing cameras of smartphones used in the wild. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, p. 280. ACM (2018)

65. Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M., Masquelier, T.: Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports* **6**, 32672 (2016)
66. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp. 3687–3691. IEEE (2013)
67. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, pp. 1097–1105 (2012)
68. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. *Cognition and emotion* **24**(8), 1377–1388 (2010)
69. Le, H.V., Mayer, S., Wolf, K., Henze, N.: Finger placement and hand grasp during smartphone interaction. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2576–2584. ACM (2016)
70. Lewinski, P., den Uyl, T.M., Butler, C.: Automated facial coding: Validation of basic emotions and faces aus in facereader. *Journal of Neuroscience, Psychology, and Economics* **7**(4), 227 (2014)
71. Lewis, M., Haviland-Jones, J.M., Barrett, L.F.: *Handbook of emotions*. Guilford Press (2010)
72. Litman, D.J., Forbes-Riley, K.: Predicting student emotions in computer-human tutoring dialogues. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 351. Association for Computational Linguistics (2004)
73. Liu, C., Freeman, W.T., Szeliski, R., Kang, S.B.: Noise estimation from a single image. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, pp. 901–908. IEEE (2006)
74. Ma, C., Osherenko, A., Prendinger, H., Ishizuka, M.: A chat system based on emotion estimation from text and embodied conversational messengers. In: *Proceedings of the 2005 International Conference on Active Media Technology, 2005.(AMT 2005).*, pp. 546–548. IEEE (2005)
75. Maalej, A., Amor, B.B., Daoudi, M., Srivastava, A., Berretti, S.: Local 3d shape analysis for facial expression recognition. In: 2010 20th international conference on pattern recognition, pp. 4129–4132. IEEE (2010)
76. Mao, X., Xue, Y., Li, Z., Huang, K., Lv, S.: Robust facial expression recognition based on rpca and adaboost. In: 2009 10th Workshop on Image Analysis for Multimedia Interactive Services, pp. 113–116. IEEE (2009)
77. Matsumoto, D., Ekman, P.: Facial expression analysis. *Scholarpedia* **3**(5), 4237 (2008)
78. Matsumoto, D., Keltner, D., Shiota, M.N., O'Sullivan, M., Frank, M.: Facial expressions of emotion. *Handbook of emotions* **3**, 211–234 (2008)
79. Matthews, O., Sarsenbayeva, Z., Jiang, W., Newn, J., Velloso, E., Clinch, S., Goncalves, J.: Inferring the mood of a community from their walking speed: A preliminary study. In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing, UbiComp '18*, p. 1144–1149 (2018)
80. McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., Graesser, A.: Facial features for affective state detection in learning environments. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29 (2007)
81. Microsoft: Face api – facial recognition software — microsoft azure (2019). URL <https://azure.microsoft.com/en-au/services/cognitive-services/face/>
82. Narwekar, A.A., Girju, R.: Uiuc at semeval-2018 task 1: Recognizing affect with ensemble models. In: *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 377–384 (2018)
83. Nelson, N.L., Russell, J.A.: Universality revisited. *Emotion Review* **5**(1), 8–15 (2013)
84. Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., Ohme, R.K.: Warsaw set of emotional facial expression pictures: a validation study of facial display photographs. *Frontiers in psychology* **5**, 1516 (2015)
85. Panigrahi, S.K., Gupta, S., Sahu, P.K.: Phases under gaussian additive noise. In: 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 1771–1776. IEEE (2016)
86. Patton, R.: *Software testing*. Pearson Education India (2006)
87. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* **37**, 98–125 (2017)
88. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2539–2544 (2015)
89. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional mkl based multimodal emotion recognition and sentiment analysis. In: 2016 IEEE 16th international conference on data mining (ICDM), pp. 439–448. IEEE (2016)
90. Rodner, E., Simon, M., Fisher, R.B., Denzler, J.: Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches. *arXiv preprint arXiv:1610.06756* (2016)
91. Rodriguez, P., Cucurull, G., González, J., Gonfaus, J.M., Nasrollahi, K., Moeslund, T.B., Roca, F.X.: Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics* (2017)
92. Russell, J.A.: Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin* **115**(1), 102 (1994)
93. Sander, D., Scherer, K.: *Oxford companion to emotion and the affective sciences*. OUP Oxford (2014)
94. Sarsenbayeva, Z., Ferreira, D., van Berkel, N., Luo, C., Vaisanen, M., Kostakos, V., Goncalves, J.: Vision-based happiness inference: A feasibility case-study. In: *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '17*, pp. 494–499. ACM, New York, NY, USA (2017)
95. Sarsenbayeva, Z., Marini, G., van Berkel, N., Luo, C., Jiang, W., Yang, K., Wadley, G., Dingler, T., Kostakos, V., Goncalves, J.: Does smartphone use drive our emotions or vice versa? a causal analysis. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, p. 1–15. New York, NY, USA (2020)
96. Schuller, B., Stadermann, J., Rigoll, G.: Affect-robust speech recognition by dynamic emotional adaptation. In: *Proc. Speech Prosody 2006, Dresden* (2006)
97. Sharma, P., Esengönül, M., Khanal, S.R., Khanal, T.T., Filipe, V., Reis, M.J.: Student concentration evaluation index in an e-learning context using facial emotion analysis. In: *International Conference on Technology and Innovation in Learning, Teaching and Education*, pp. 529–538. Springer (2018)
98. Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., Samson, A.C.: Facial expression analysis with affdex and facet: A validation study. *Behavior research methods* **50**(4), 1446–1460 (2018)
99. Swinton, R., El Kaliouby, R.: Measuring emotions through a mobile device across borders, ages, genders and more. In: *Proceedings of the ESOMAR Congress, Atlanta*, pp. 1–12 (2012)
100. Tan, X., Triggs, W.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE transactions on image processing* **19**(6), 1635–1650 (2010)
101. Technology, M.: Emotion recognition - face++ ai open platform (2019). URL <https://www.faceplusplus.com/emotion-recognition/>

102. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence* **30**(11), 1958–1970 (2008)
103. Towner, H., Slater, M.: Reconstruction and recognition of occluded facial expressions using pca. In: *International Conference on Affective Computing and Intelligent Interaction*, pp. 36–47. Springer (2007)
104. Useche, O., El-Sheikh, E.: An intelligent system framework for measuring attention levels of students in online course environments. In: *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, p. 452. The Steering Committee of The World Congress in Computer Science, Computer ... (2015)
105. Valstar, M.F., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: *Face and Gesture 2011*, pp. 921–926. IEEE (2011)
106. Van Der Schalk, J., Hawk, S.T., Fischer, A.H., Doosje, B.: Moving faces, looking places: validation of the amsterdam dynamic facial expression set (adfes). *Emotion* **11**(4), 907 (2011)
107. Violante, M.G., Marcolin, F., Vezzetti, E., Ulrich, L., Billia, G., Di Grazia, L.: 3d facial expression recognition for defining users' inner requirements—an emotional design case study. *Applied Sciences* **9**(11), 2218 (2019)
108. Visuri, A., Sarsenbayeva, Z., Goncalves, J., Karapanos, E., Jones, S.: Impact of mood changes on application selection. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 535–540. ACM (2016)
109. Yuan, L., Sun, J., Quan, L., Shum, H.Y.: Image deblurring with blurred/noisy image pairs. *ACM Transactions On Graphics (TOG)* **26**(3), 1 (2007)
110. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence* **31**(1), 39–58 (2008)
111. Zhang, L., Verma, B., Tjondronegoro, D., Chandran, V.: Facial expression analysis under partial occlusion: A survey. *ACM Computing Surveys (CSUR)* **51**(2), 25 (2018)
112. Zhang, Y., Chen, M., Huang, D., Wu, D., Li, Y.: idoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems* **66**, 30–35 (2017)

Chapter 5

Hybrid Multimodal Network

5.1 Introduction

The findings presented in Chapter 4 demonstrate that unimodal systems exhibit certain inherent limitations in meeting robustness, accuracy, and overall performance requirements [5, 26]. Meanwhile, from a physiological perspective, it is difficult to establish a high-fidelity emotional model and precisely analyze human emotional changes by using only one single signal [26, 84]. As human beings, we express our emotions through a series of actions, including facial expressions, voice, body movements, various gestures, and multiple physiological signals, such as heart rate and sweat, etc [17]. With the rapid development of signal processing and artificial intelligence, multimodal intelligent systems have obtained considerable development in the field of affective computing. Currently, most state-of-the-art frameworks rely on processing multimodal sources. However, with respect to the mobile emotion-sensing field, we find that unimodal approach is still the mainstream. This may be because a unimodal framework typically has lower computational complexity and battery consumption, making it more suitable for mobile devices. Fortunately, the processing power, battery capacity, and storage capacity of today's mobile and embedded devices are constantly growing. Multi-core processors, gigabytes of RAM, and all-day battery life are increasingly prevalent. Therefore, in this chapter, we proposed a hybrid multimodal framework that leverages the fusion of behavioral and physiological signals to detect and process underlying affective information. We collected a variety of behavioral and physiological signals from sensors embedded in off-the-shelf smartphones and commodity wearables, and evaluated our system in a user study with

45 participants. Experimental results show a high average accuracy of 89.2% (with 0.86 AUC) for binary emotion recognition, well above the historical classification accuracies using sophisticated models and multimodal data [279]. Furthermore, we explored real-world scenarios where data sources are partially available and provided suggestions on how to build robust mobile emotion recognition systems. Although in this paper we only focused on binary emotion recognition (limited by the complexity of experimental user study), our proposed model can be easily extended to multi-class emotion recognition by simply replacing sigmoid layer with softmax layer. More details can be found in the attached publication, [Article II](#).

5.2 Article II

©Copyright is held by IEEE. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in:

Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing*, pages 1–17, 2021. URL <https://doi.org/10.1109/TAFFC.2021.3100868>

Ethics ID: 1955051.1, The University of Melbourne Human Ethics Advisory Group.

Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition

Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves

Abstract—With the rapid development of mobile and wearable devices, it is increasingly possible to access users' affective data in a more unobtrusive manner. On this basis, researchers have proposed various systems to recognize user's emotional states. However, most of these studies rely on traditional machine learning techniques and a limited number of signals, leading to systems that either do not generalize well or would frequently lack sufficient information for emotion detection in realistic scenarios. In this paper, we propose a novel attention-based LSTM system that uses a combination of sensors from a smartphone (front camera, microphone, touch panel) and a wristband (photoplethysmography, electrodermal activity, and infrared thermopile sensor) to accurately determine user's emotional states. We evaluated the proposed system by conducting a user study with 45 participants. Using collected behavioral (facial expression, speech, keystroke) and physiological (blood volume, electrodermal activity, skin temperature) affective responses induced by visual stimuli, our system was able to achieve an average accuracy of 89.2% for binary positive and negative emotion classification under leave-one-participant-out cross-validation. Furthermore, we investigated the effectiveness of different combinations of data signals to cover different scenarios of signal availability.

Index Terms—Emotion recognition, mobile and wearable devices, behavioral signals, physiological signals, attention-based LSTM.

1 INTRODUCTION

EMOTIONS play a crucial role in our daily lives as they can aid decision-making, learning, communication, and situational awareness [1]. Emotions can go awry, however, and many of the most common and devastating mental illnesses involve emotional disorder [2]. In the last few decades, researchers have attempted to empower machines with human-like intelligence to automatically recognize and understand users' affective states, and further provide data-driven insights for well-being [3], [4]. Emotion recognition is an active and challenging research topic that directly contributes towards research efforts to create technologies that support emotional well-being [5], [6].

Research in psychology has shown that human emotions encompass complex combinations of subjective feelings, physiological and behavioral responses triggered by external stimuli [7], [8]. Recognizing these response signals can improve understanding of emotional expressiveness, and be crucial for clinical diagnostic and therapeutic procedures in mental health care [9]. Currently, two main approaches are used to identify human emotions. First, researchers have analysed a wide range of behavioral signals such as facial expressions [10], [11], speech [12], and gestures [13], all of which can be collected directly. Such signals, however,

often have reliability issues since people can disguise their emotions by controlling behaviors like facial muscles or intonation patterns in social communications [14]. Another challenge is that analysing these signals often requires an ideal technological setting. For example, analysing visual data (e.g., facial expression, body gestures and movement), requires pre-installed cameras with sufficient resolution and an appropriate viewing angle [15].

The second main approach to identify emotions analyses physiological data, such as electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), galvanic skin response (GSR), and heart rate (HR) to detect emotional changes in a person. These signals reflect activities of the human body's central nervous system (CNS) and autonomic nervous system (ANS) which have been shown to have intimate links with inner emotional changes [16], [17]. Compared to behavioral data, physiological data are considered to be more objective, as they typically reflect involuntary responses and are more difficult to consciously conceal or manipulate [18]. A significant amount of research in this area, however, relies on physiological sensing equipment of medical grade, which is often invasive, expensive, and needs technical guidance from professionals; this, therefore, limits their application in real-world scenarios. EEG, for example, requires electrodes to be attached to subjects' scalps, while GSR requires sensors to be placed on participants' hands or soles [19]. In addition, ambient environment noise and over-sensitive sensing devices pose barriers to physiological data acquisition [20].

In our work, we collect both behavioral and physiological signals from sensors embedded in off-the-shelf smartphones and commercial wearables in a way that can be

- K. Yang, C. Wang, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves are with the School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia, 3052.
E-mail: (kangning.yang, chaofanw)@student.unimelb.edu.au
E-mail: (zhanna.sarsenbayeva, benjamin.tag, tilman.dingler, greg.wadley, jorge.goncalves)@unimelb.edu.au
- Y. Gu was with the Department of Electrical and Computer Engineering, Rutgers University, New Jersey, USA, 08854.
E-mail: yg202@scarletmail.rutgers.edu

implemented for ordinary technology users in everyday life. In particular, we propose a deep, multimodal, mobile emotion-recognition system that leverages these signals in tandem to detect rapid and fine-grained fluctuation in emotion. On the one hand, deep learning methods have been shown to outperform previous state-of-the-art machine learning techniques in many tasks including noise reduction and high-level feature extraction [21], [22]. On the other hand, a multimodal approach has a few advantages over a unimodal approach: (1) it is able to deal with missing data problems which commonly arise in a unimodal signal [18]; (2) it is more capable of detecting emotions hidden behind social masks [23]. Moreover, with the development of mobile technology, modern smartphones and wearable devices are increasingly low-cost and sensor-rich, and allow us to unobtrusively detect user responses without spatio-temporal limitations, which contributes to efforts to deploy affect detection into real-world contexts. Potential applications of such a system are diverse. For example, traditional diagnosis of affective instability usually depends on interviews or questionnaires, which rely heavily on respondents' retrospective recall and subjective assessment of affective variability or reactivity [24]. Mobile-based emotion recognition could allow clinicians to monitor patients' emotional changes objectively and remotely, which could provide a more complete picture of patients' day-by-day mental states and help to improve diagnostic accuracy.

Moreover, to overcome the challenge of collecting real emotional responses [25], [26], [27], [28], [29], we utilize visual stimuli to elicit participants' emotions. This is a common and effective approach used in Neurophysiology and Psychology research [30], [31]. Then we used a smartphone to collect subjects' facial expressions, speech and keystrokes, and a wristband to collect their blood volume pulse (BVP), electrodermal activity (EDA) and skin temperature (ST). Next, we adopt a deep learning framework to identify human emotions from these raw data inputs. Compared with traditional statistical analysis methods and machine learning techniques (e.g., decision tree or support vector machine (SVM)) [32], deep learning techniques can work on raw data and automate the feature extraction and selection, which eliminates the need for data pre-processing and handcrafted feature construction [33]. Specifically, we develop a novel attention-based Long Short-Term Memory (LSTM) neural network classifier to train and examine the collected database. Lastly, by leveraging a modality-level fusion approach, we compare and analyze the relative recognition accuracy of different combinations of data signals, and provide recommendations to the research community regarding the recognition capabilities.

Thus, the contribution of our work is three-fold:

- 1) We proposed a novel attention-based LSTM sensing system that automatically detects human emotional states based on the fusion of behavioral and physiological data collected from an off-the-shelf smartphone and wearable wristband.
- 2) We designed a replicable experimental protocol based on mobile devices aimed at collecting spontaneous affective responses for future emotion-related research.
- 3) We conducted a study with 45 participants to evaluate our model and achieved a high average accuracy of 89.2% for binary (positive and negative) emotion recognition. We also compared the effectiveness of different combinations of behavioral and physiological signals in different scenarios.

2 RELATED WORK

2.1 Emotion Models

Emotions are often characterized by "short, intensive episodes", and triggered by "a particular event or person" [34]. These can manifest in relatively clearly recognizable states such as fear, disgust, sadness, joy, anger, and surprise, but also more complex states such as embarrassment, shame, guilt, contempt, compassion, and admiration [35], [36]. However, because of the subjective and personal nature of emotions, in most cases they are difficult to define accurately and describe quantitatively. Currently, a definitive description of emotions is still being debated within physiology and psychology research [37]. Throughout years of research, scholars have attempted to characterize and model human emotions from different aspects, and have derived two fundamental viewpoints: (1) the discrete emotion model and (2) the continuous multi-dimensional emotion model [38].

The discrete emotion model can be traced back to the mid-late 19th century with Charles Darwin's pioneering book, *The Expression of the Emotions in Man and Animals* [39]. In this seminal work, Darwin argued that human emotional expressions have evolved and (at least at some point in the past) were adaptive for survival [40], and that there exists a degree of universality in emotion expression [41]. Inspired by this, Ekman and his colleagues proposed that certain emotions appear to be universally recognized regardless of cultural background, and further identified six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [42]. Later, Plutchik proposed the "wheel of emotions" that groups eight primary emotions into positive and negative opposites, e.g., joy versus sadness [43]. This helps to illustrate that primary emotions can be expressed with varying intensity and mixed to form complex emotions.

The continuous multi-dimensional emotion models attempt to describe emotions through multiple dimensions rather than discrete labels [44]. The theory behind this model assumes that there are common, overlapping neurophysiological systems that all emotional states arise from, which challenges the discrete emotion theory that different emotions correspond to distinct expressions of the nervous system [45]. Among multi-dimensional emotion models, the valence-arousal model [45] is the most commonly used, in which the valence dimension describes the unpleasant-pleasant continuum and the arousal dimension represents the deactivated-activated continuum. Each emotion can be understood as an orthogonal combination of these two dimensions. For example, *stressed* can be conceptualized as an emotional state that is the product of unpleasant valence together with activated arousal.

In this paper, we aim to classify positive and negative emotions, which is of great importance for clinicians (e.g., the need to identify how often negative emotions take place

during some period of time) and researchers interested in well-being. In our study we use happy and sad scenes from validated visual stimuli databases to induce positive and negative emotions in our participants.

2.2 Emotion Recognition Technologies

To date, researchers have investigated a variety of different data sources to predict human affective states. Traditional emotion recognition technologies usually focus on visual or audio signals, i.e., facial expression and speech. Among a number of facial expression systems, the Facial Action Coding System (FACS) developed by Ekman and Friesen [46] was widely adopted by emotion researchers. FACS describes facial behaviours in terms of a set of specific Action Units (AUs), each of which is associated with an individual face muscle or muscle group. Based on this system, human coders can manually deconstruct any possible facial activity, mapping it to a combination of AUs. Subsequently, researchers developed software tools for automated facial coding [47], and employed machine learning techniques (e.g., nearest neighbor and bayesian networks) to map the extracted and selected features to different emotional categories [48], [49]. Speech is another data source that has been widely used for emotion recognition purposes, as it is one of the main channels used by human beings to transmit affective information. Early studies leveraged low-level acoustic descriptors and derivations (LLDs) with functional statistics as acoustic features for emotion analysis, especially some vocal parameters such as pitch, intensity, and speaking rate [50], [51]. Further studies found that pitch and energy-related features play a key role in the recognition of emotion, and proposed different methods, e.g., mel-frequency cepstral coefficients (MFCC), mel-frequency spectral coefficients (MFSC), log frequency power coefficients (LFPC) and linear prediction cepstral coefficients (LPCC), to derive these features from audio clips for emotional expression analysis [52], [53].

With the development of modern neuroscience, researchers have also begun to leverage physiological reactions in addition to behavioral responses, such as cardiovascular rhythms, the release of certain neuropeptides, or a change in gastrointestinal fluid [54], [55]. Inspired by these advances, affective computing studies began to investigate the link between emotions and numerous physiological signals from brain, heart, muscles and skin. Hsu *et al.* [56] adopted music emotion induction techniques to induce spontaneous emotions and recorded relevant ECG signals from 61 healthy participants. They designed a selection algorithm to extract the time-domain and frequency-domain ECG features and adopted generalized discriminant analysis (GDA) to reduce the dimensions of selected features. Using an LS-SVM classifier, they achieved an automatic ECG-based emotion recognition system. Zheng *et al.* [57] collected EEG signals and eye-tracking data from 5 subjects. They used film clips as stimuli to evoke emotions in subjects and built emotion recognition models based on two fusion strategies. By employing an SVM training method, their classifier was able to recognize three categories of emotional states: positive, neutral, and negative. Lee *et al.* [58] derived physiological signals from PPG, EMG, and inertial

measurement unit (IMU) sensors placed respectively on the earlobe, the upper trapezius muscle, and the back of the head. By training the SVM model, their proposed wearable system achieved a high accuracy rate of monitoring negative emotional states and demonstrated a high potential for implementation as a driver emotional-response monitoring system. Similarly, Perusquía-Hernández *et al.* [59] explored micro-expression detection by leveraging a wearable device that detects signals from distal facial EMG sensors. In another example, Zhao *et al.* [60] proposed EQ-Radio, a wireless system that can recognize emotions by using RF reflections off a person's body.

Later, the development and popularization of smartphones and many wearable devices such as smartwatches and wristbands has opened a new path toward emotion-aware computing. Compared with other techniques, mobile devices' increasingly lower cost, greater computational capacity, wider accessibility, and more unobtrusive sensing capability, can further facilitate affective sensing and computing research. For instance, Zhao *et al.* [14] developed an automatic emotion recognition system based on a wearable wristband. Specifically, using the embedded rich biosensors on the wristband, they collected three kinds of physiological signals, blood volume pulse (BVP), electrodermal activity (EDA), and skin temperature (ST) from 15 participants with emotion changes triggered by video stimuli. Then, they extracted a set of fine-grained features to represent physiological signals by adopting the sequence forward floating selection (SFFS) method, and classified different emotions through SVM method. In another example, Zhang *et al.* [61] proposed MoodExplorer, an automatic system for compound emotion detection based on smartphone sensing data. They first extracted different types of features from the environments, social contacts, APP usage and activities of individuals, then used a feature selection algorithm to choose the most significant features. By training a factor graph based machine learning model, they finally achieved 76.0% average recognition accuracy with 30 university students.

Besides using off-the-shelf wearable devices, researchers have developed several research prototypes equipped with different biosensors to explore mobile or wearable emotion sensing more flexibly. For example, Wu *et al.* [62] built a custom eye-tracking platform consisting of an infrared camera and a System-on-Chip (SoC) board to capture the single-eye-area images and further achieve accurate identification of a user's emotions. After conducting comprehensive experimental evaluations on 20 participants, their proposed system was shown to recognize seven-type emotions at 12.8 frames per second with a mean accuracy of 72.2%. In another example, Masai *et al.* [63] used 17 photo-reflective sensors embedded in the front frame of the eyewear to capture skin deformations caused by the movement of facial muscles. They then applied the SVM algorithm to the acquired data and achieved facial emotion recognition with around 75% accuracy for different usage scenarios. Although research wearable devices can potentially achieve higher emotion detection accuracy, we chose to explore commercially available devices in this study because of the following reasons: (i) they can be immediately used for further studies regarding the viability of emotion detection

in a mobile platform; (ii) our ultimate aim is to achieve mobile emotion sensing in real-world scenarios rather than only pursuing higher emotion recognition accuracy; from this perspective, commercial wearables are more appropriate since they are cheaper, easier to access, and have a more stable performance. Moreover, the off-the-shelf commercial wearable device we selected (i.e., the Empatica E4 wristband) has been widely used in other studies [64], [65], [66], [67], [68]. Lastly, we did not use commercial supporting software tools to process collected raw data, therefore, as for physiological signals we focused on (i.e., BVP, EDA, ST), it does not make a significant difference whether they are collected by commercial or research wearables.

More recently, deep learning has been widely used in emotion detection [69], [70], [71], [72], [73], due to its better performance than traditional machine learning methods. Compared with standard handcrafted features, deep neural networks can automate feature extraction and selection process, and learn high-level and non-linear robust features. For example, Shukla et al. [74] explored convolutional neural network (CNN) features from both audiovisual and EEG signals to recognize advertisement emotions. Rayatdoost et al. [75] presented a novel multimodal gated fusion method to learn the joint deep representation between EEG signals and facial behaviors. However, throughout the associated mobile affective computing literature, we note that while some attempts at detecting human emotion via mobile devices have begun to use deep learning [33], [76], [77], it is still a new and growing area of research that requires further work. This is partly due to the usage complexity of mobile devices as well as the rich variation in individual affective expression. Thus, in this study, we further explore deep learning frame-based mobile emotion recognition methods, and propose a novel attention-based LSTM structure that utilizes both a smartphone and a wristband to collect and fuse behavioral as well as physiological signals (i.e., facial expression, speech, keystroke, BVP, EDA, ST) to identify the users' emotional states. Because the focus of this paper is short-term mobile emotion recognition, we did not use the approach common in other mobile emotion recognition work of using indirect behavioral information from long-term sensor data (e.g., acceleration, gyroscope, Bluetooth, WiFi, GPS, etc.) as model inputs [61], [78], [79].

3 METHOD

3.1 Visual Stimuli

In recent years, there has been an increase in the availability of large-scale affect datasets. These datasets contain data originating from a single [80] or multiple sources [81], [82], and are validated using spontaneous [83], [84] or non-spontaneous [85], [86] emotion data. However, to the best of our knowledge, there is no publicly available labelled affect database that entirely relies on everyday mobile devices. This is of particular importance as users' spontaneous mobile behaviours are typically more complex and uncontrollable, which in most databases is not present or is treated as noise. Based on these considerations, we decided to collect a multimodal spontaneous emotion database exclusively using mobile devices.

Experimentally inducing emotions is the most rigorous means of analyzing the effects of emotions and has a rich history in psychology, neuroscience, and psychiatry [31], [87]. In our experiment, we chose static pictorial stimuli to evoke positive and negative emotions respectively. There are two main reasons why we chose this type of stimuli over other available methods: 1) visual stimuli (including static pictures and videos) are more efficient than other emotion induction techniques, such as music or autobiographical memory recall in most situations [31], and 2) unlike video stimuli, pictorial stimuli have several standard databases that have been repeatedly verified in both cross-user and cross-cultural conditions [88], [89]. Specifically, 10 affective pictures (5 for positive¹, 5 for negative²) were selected from two well-known affective picture databases, the International Affective Picture System (IAPS [90], [91]) and the Nencki Affective Picture System (NAPS [92], [93]). It is worth mentioning that there are some mobile or wearable emotion sensing works [94], [95], [96], [97] that also attempted to recognize emotions elicited by IAPS. But unlike their works, we focus on multimodality signals from both behavioral and physiological categories, given that their contained clues are complementary and attempting to mine comprehensive affective information, rather than biasing toward certain modalities.

3.2 Participants

We recruited 45 healthy participants via social media and posters placed around our university campus. We excluded data from 5 participants due to problems during the data collection process. All remaining participants were students or staff in our university. 52.5% (21) were males and 47.5% (19) were females. Their ages ranged from 18 to 36 years old (mean $\pm$ standard deviation, 23.9 ± 4.7 years). We ensured that none of the participants had a history of cardiovascular or neurological illness, and that all participants had normal or corrected-to-normal vision.

3.3 Procedure

Before each experiment started, we simultaneously connected the smartphone and wristband to a computer to synchronize their internal clocks with the computer's clock and assure that they had synchronized timestamps.

Upon arriving at our lab, participants were briefed on details of the experiment including the principal experimental tasks, the self-report questionnaires, and experimental devices. Participants were then asked to sign a consent form if they agreed with the experimental setup. Afterwards, each participant was asked to wear an Empatica E4 wristband³ on their non-dominant hand (to minimize motion [77]) with the assistance of a researcher to ensure the electrodes lined up on the bottom of the wrist and to avoid relocation during hand movements. Participants were also asked to hold a smartphone in their dominant hand. The experimental design was approved by the ethics committee of our university. Each participant was compensated with a \$20 gift voucher for their participation.

1. IAPS IDs: 1920; NAPS IDs: Animals_183_h, Faces_001_h, Faces_079_h, Faces_127_h.

2. IAPS IDs: 2205, 8010, 9421; NAPS IDs: Faces_155_h, People_143_h.

3. <https://www.empatica.com/research/e4/>

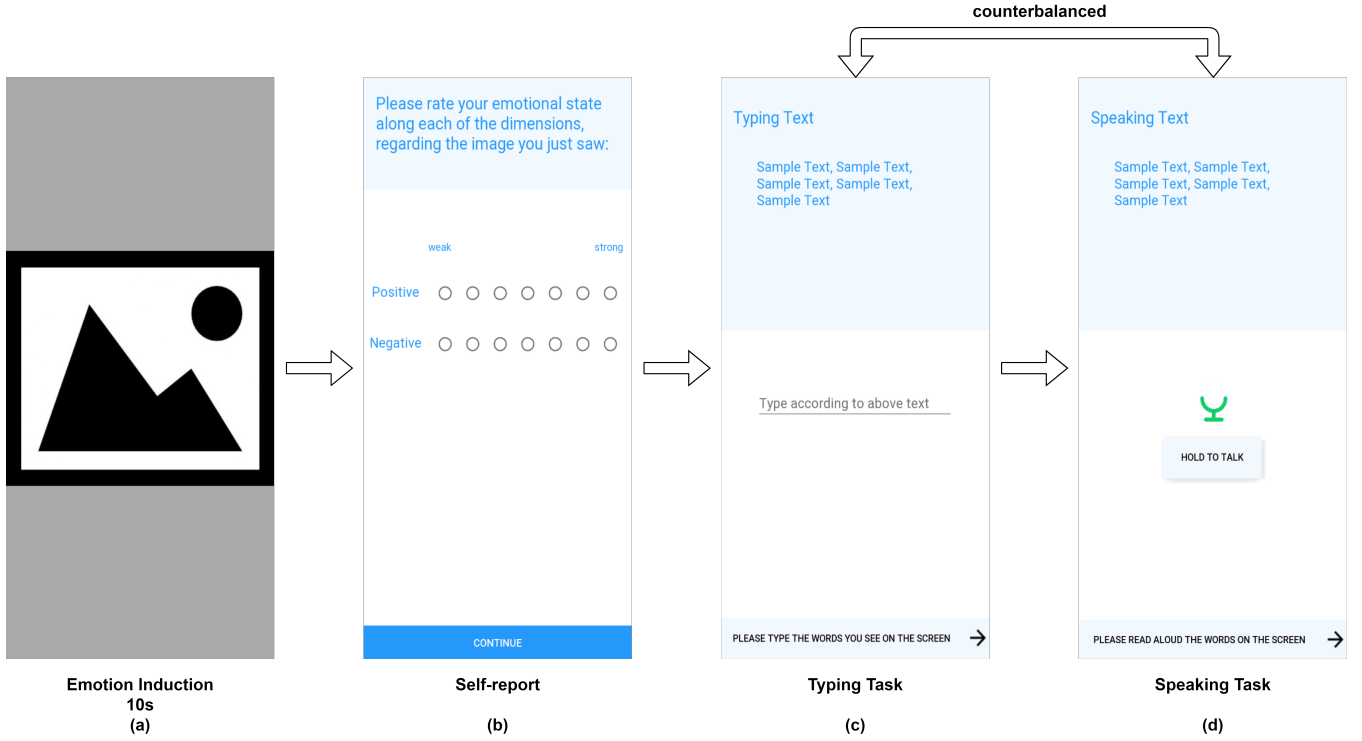


Fig. 1. The four stages of each experimental session, including *emotion induction*, *self-report*, *typing task*, and *speaking task*. In the *emotion induction* stage, one affective picture was presented for 10 seconds; in the *self-report* stage, participants completed a self-assessment reflecting on their current emotion after looking at the picture; in the *typing task*, participants had to type the text presented on the screen; and in the *speaking task*, participants had to read aloud the text presented on the screen.

Our experiment had a total of ten sessions. Each session consisted of four stages, including *emotion induction*, *self-report*, *typing task*, and *speaking task*. Fig. 1 shows the different stages of each experiment session. Firstly, one affective picture from our selected visual stimuli was presented in the center of the smartphone screen for 10 seconds (shown in Fig. 1a), in order to evoke one target emotion. This presentation duration has been shown to be more than sufficient for picture emotion elicitation [98], [99], [100], [101], [102]. Next, a questionnaire was triggered, asking participants to self-report their emotional state elicited by the displayed image (shown in Fig. 1b). Following recommendations from previous work [103], we used 7-point Likert scales to indicate the intensity of different emotions, with 1 indicating a *weak* feeling and 7 indicating a *strong* feeling. At this stage, we presented our participants with both positive and negative emotion choices and asked them to rate all emotions simultaneously in order to ensure that choices were independent (e.g., “5” for positive, and “1” for negative). We chose the emotion with the highest value as the final “ground truth”. After completing the questionnaire and pressing the “CONTINUE” button, the participants were required to type out in the dedicated text entry area some text that was displayed in the upper half of the smartphone screen (shown in Fig. 1c). Finally in stage 4, participants were asked to press and hold the “HOLD TO TALK” button, while reading out loud a set of sentences displayed on the top of the screen (shown in Fig. 1d). In short, we first induced an emotion in participants, then collected their self-reported emotion, and finally presented

them with typing and speaking tasks.

The ten affective pictures were presented to each participant in pseudo-random order where any two pictorial stimuli in the same category (happy or sad scene) would not appear in succession, so that all emotional scenes were spread evenly over the experiment. Furthermore, in order to avoid any possible sequence effects of *typing task* and *speaking task*, we also counter-balanced these two tasks, so that half of the participants encountered the *typing task* first, while the other half encountered the *speaking task* first. After each session, participants had a 10-second break to mitigate potential carryover effects of the previous emotional experience as suggested in the literature [19], [104]. In addition, after every two sessions, we added one extra session with a neutral picture to further strengthen this moderating impact. We also set a training session with a neutral picture to keep participants relax before the formal experiment. Thus, a complete experiment actually consisted of sixteen (10+5+1) sessions.

Moreover, to avoid the potential affective interference from the linguistic or semantic information of the sentences displayed in *typing task* and *speaking task*, we selected the “emotion-free” sentences from two phrases sets that are frequently used in text entry and speech synthesis studies: the MacKenzie and Soukoreff phrase set [105], and the CMU_ARCTIC database [106]. Then we used the IBM Watson natural language understanding API⁴ to ensure the selected sentences are indeed neutral expression. Further-

4. <https://www.ibm.com/au-en/cloud/watson-natural-language-understanding/resources>

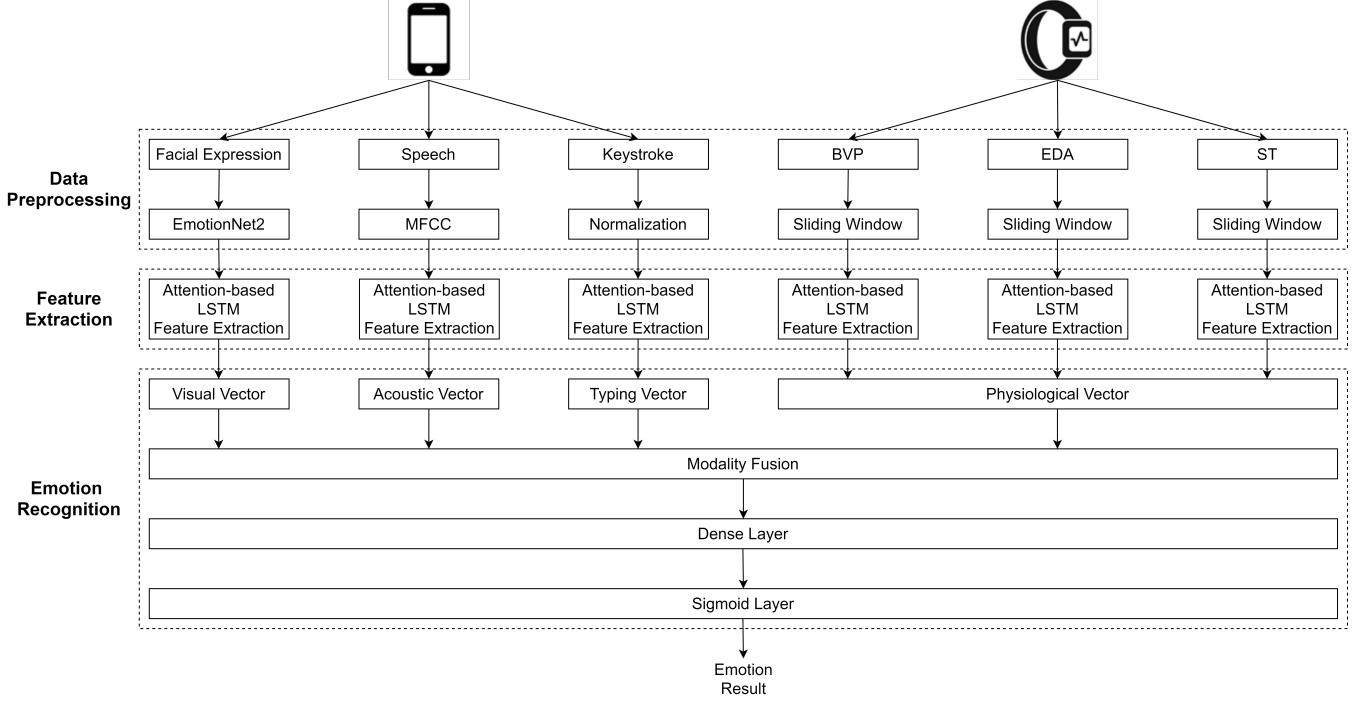


Fig. 2. Overall system structure. We collected facial expression and three types of physiological signals (BVP, EDA, ST) during the *emotion induction* stage, speech data during the *speaking task*, and keystroke data during the *typing task*. These data were passed to the preprocessing and attention-based LSTM feature extractor respectively, which outputted four modality-specific feature vectors. The concatenate and the dense layers fused these four vectors and learned the mutual association among modalities for final emotion recognition. The result was then compared with the ground truth collected in the *self-report* stage.

more, in order to avoid memorization, we did not use the same sentences for the *typing task* and *speaking task*. Instead, we used different sentences with similar Flesch-Kincaid scores that quantify the difficulty of the sentences used in mobile text entry [107], [108].

3.4 Data Collection

We developed a custom Android application as a data collection tool, which leverages the embedded front facing camera, microphone, and default software keyboard (Gboard⁵, with suggestion and correction features disabled, using the default Android widget named EditText) to monitor participants' behavioral data. Specifically, the application recorded the following data signals: (1) facial features when participants looked at the affective pictures; (2) participant's self-report questionnaire answers; (3) keystrokes (including each character typed and deleted, and the detailed time when each key-down event happened) according to the participant's typing behavior; and (4) raw audio data collected during the speaking task. We chose these specific behavioral modalities as they have been widely investigated in the emotion recognition field [109], [110], [111], [112] and have many high-fidelity feature models, rendering them more suitable for short-term emotion recognition research.

Regarding participants' physiological data, as suggested by [14], [65], we used the Empatica E4 wristband to capture blood volume pulse (from the photoplethysmograph (PPG)

sensor), electrodermal activity (from the electrodermal activity sensor), and skin temperature (from the temperature sensor) data with sample rate at 64 Hz, 4 Hz, and 4 Hz respectively⁶.

4 IMPLEMENTATION AND ANALYSIS

The overall structure of our proposed recognition system is illustrated in Fig. 2. The structure consists of three major parts: data preprocessing, feature extraction, and emotion recognition. We cover each part in more detail in the following sections.

4.1 Data Preprocessing

The system accepts raw data from facial expression, speech, keystroke, BVP signal, EDA signal, and ST signal as inputs. Through the data preprocessing module, the heterogeneous inputs can be formatted into specific representations, which can be effectively used in the following feature extraction network module.

Specifically, for the facial expression input, we first extracted the related frames with the speed of 30 fps by *FFmpeg*⁷, which converted the raw input video into a series of continuous image frames. Based on this, we built our facial feature extractor with EmotionNet2 [113] for each frame. EmotionNet2 is the extension of EmotionNet [114], which is a novel computer vision algorithm for emotion

5. <https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin&hl=en>

6. <https://support.empatica.com/hc/en-us/articles/201608896-Data-export-and-formatting-from-E4-connect->

7. <https://ffmpeg.org/>

recognition of human faces in photographs. It builds up on the ultra-deep ResNet architecture [115] and performs well especially in reducing the variance caused by background noise [116]. We used it as a trained feature-extractor for facial expression, and selected the output before the last layer as the feature representation of each frame. During training, we found that there were some instances where the extracted frames did not contain a face or the face proportion was too small. We removed these noisy frames by using the face detection algorithm embedded in EmotionNet2, and padded the frame stream at the end by duplicating the last frame from the video. Finally, each input is represented as a $N_f \times 512$ matrix, where N_f is the number of frames.

For the speech input, we used the *OpenSmile* toolkit [117] to directly extract mel-frequency cepstral coefficients (MFCC) from raw audio signals as acoustic representations. It computes MFCC features from 26 Mel-frequency bands computed from the fast fourier transform (FFT) power spectrum. In addition, delta (Δ) and acceleration ($\Delta\Delta$) coefficients are also appended to the MFCC. The extracted feature set thus contains 39 dimensions (12-MFCC, 12- Δ MFCC, 12- $\Delta\Delta$ MFCC, P, Δ P and $\Delta\Delta$ P, where P stands for raw energy of the input speech signal [118]). After normalization, the final acoustic representation is a 2-D array with $N_s \times 39$ dimensions, where N_s is number of extracted MFCC frames.

For the keystroke input, there are several different features that can be monitored when the user presses keys on a keyboard, such as pressure, finger movement, and typing speed [119]. We focused specifically on the duration features, i.e., the interval time between two consecutive key-down events on the smartphone virtual keyboard. Considering that most participants completed the *typing task* in a relatively short time slot, we regarded each raw input as a long window to observe the entire trends. After the normalization, we got a representation matrix with $(N_k - 1) \times 1$ dimensions, where N_k is the number of key-down events.

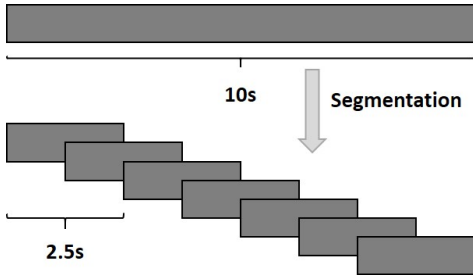


Fig. 3. Data segmentation. 10-seconds of physiological data was segmented into seven 2.5-second windows.

Regarding the physiological inputs, there are three types of data: (1) BVP measures the pulse-wave of the heart and the volume of the blood flowing through a vessel. It is obtained by the use of a PPG sensor embedded in the E4 wristband. This component illuminates the skin using a light-emitting diode and records the pulse waveform using a photo-diode. (2) EDA refers to the variation in the electrical characteristics of the skin. It is monitored by measuring the voltage between two electrodes by applying low-level current to the skin. It shows an electrodermal response of the human body to external stimuli. (3) ST

means the temperature of the human skin. It is measured by using the thermopile infrared sensor. It can describe the response of the vessels when stimulated by external factors. For example, the skin temperature will be warmer when the vessels dilate, while it will be colder when the vessels constrict. For each physiological input, we first standardized it through the z-score normalization. Then, as suggested in [15], we leveraged a 2.5-second sliding window with 50% overlap to achieve data segmentation. Considering that we only focused on the participant's physiological responses in *emotion induction* phase (10 seconds), a stream of input data was thus split into seven windows (shown in Fig. 3). Finally, the representation for each physiological input is a 2-D array with $N_p \times S_p$ dimensions, where N_p is the number of time windows, S_p is the size of data in one time window. Specifically, the value of S_p is 160, 10, and 10 for BVP, EDA, and ST respectively, based on their sample rates (64Hz, 4Hz, and 4Hz).

4.2 Attention-based LSTM Feature Extraction

In this step, we applied the LSTM structure with an attention mechanism as a high-level feature extractor. LSTM [120] is a variant of recurrent neural network (RNN) architecture, which has the ability to learn the long-term dependencies from time-series data. The key concepts of LSTM structure are the cell state and the gate structures, in which the cell state is similar to the “memory”, i.e., the information transferred in LSTM while the gate structures decide what information should be forgotten and updated. Specifically, there are three kinds of gate structures in LSTM, forget gate, input gate, and output gate. When passing through the forget gate, the unimportant information in the cell state from the prior steps will be forgotten; when passing through the input gate, the new information from current input will be used to update the cell state; and when passing through the output gate, the output information will be decided based on the new cell state. By using the LSTM structure, we can further extract the temporal associations from each input modality and form a more fine-grained representation. In addition, considering that not all subsamples in one input modality contribute equally to the final recognition (e.g., not all frames have the same importance in one input video), we also adopted an attention strategy [121] to denote the relative importance among subsamples and fuse them to a final informative feature vector. By adopting an attention mechanism, we can more effectively capture the temporal-spatial dependencies by assigning different importance weights to different subsamples. Furthermore, the fused informative vector usually has smaller dimensions, which can effectively reduce the training time [122]. The details are presented in Fig. 4.

Specifically, we first fed the representation obtained from data preprocessing phase into the LSTM in sequence:

$$h_i = LSTM(X_i), \quad i \in [1, N] \quad (1)$$

where X_i is the data of the i th time step of the input time-series representation, N refers to N_f , N_s , $N_k - 1$, and N_p for different inputs, LSTM means the LSTM cell, and h_i is the output of the hidden state, which can be viewed

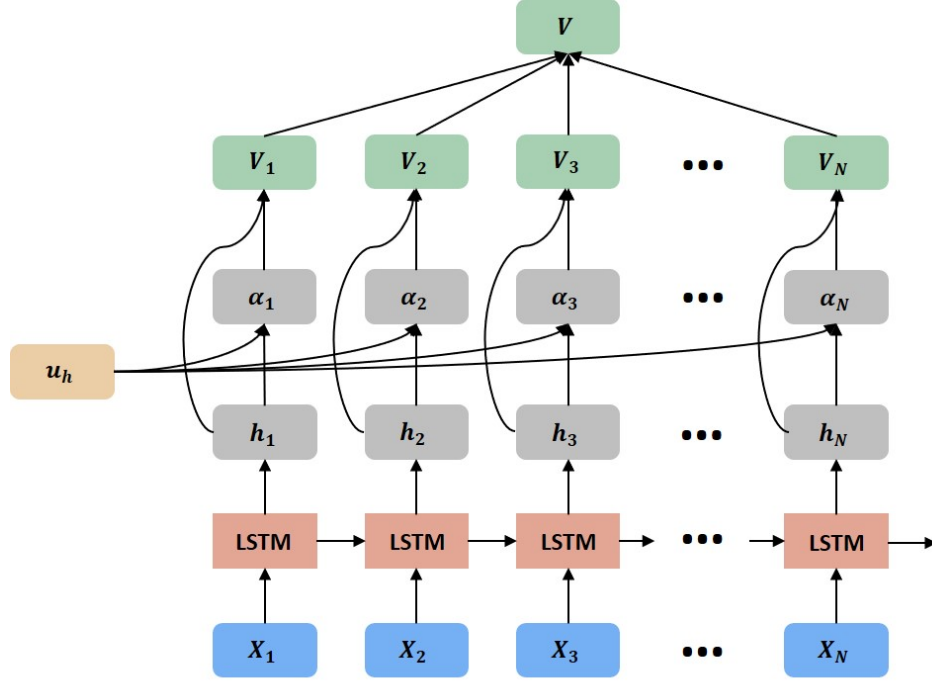


Fig. 4. Attention_based LSTM feature extractor. X_i stands for the data of the i th time step of the input time-series representation; LSTM stands for LSTM cell; h_i is the hidden representation of X_i ; u_h is the learnable context vector; α_i is the assigned importance weight for h_i ; V is the extracted informative feature vector.

as the latent information of X_i based on the previous time-stepping inputs $X_1, \dots, X_{i-1}$.

Then, similar to [123], we computed the attentive energy e_i of each h_i through a one-layer MLP, and the normalized importance weight α_i through a softmax function:

$$e_i = \tanh(W_h h_i + b_h) \quad (2)$$

$$\alpha_i = \frac{\exp(e_i^T u_h)}{\sum_i \exp(e_i^T u_h)} \quad (3)$$

where W_h and b_h are trainable parameters, and u_h is a trainable context vector with random initialization.

Finally, based on a weight sum of h_i and α_i , we calculated the high-level feature vector V by applying a dense layer. Concretely:

$$V = f_{Relu}(\sum_i \alpha_i h_i) \quad (4)$$

We considered the three physiological inputs (BVP, EDA, and ST) as a whole. On the one hand, all these three inputs are wristband-based signals, which can be obtained simultaneously. On the other hand, the emotional clues are hidden in multiple aspects of the individual's physiological signs, and it is difficult to identify the emotions based only on the incomplete clues extracted from one single physiological sign. Thus, we concatenated the feature vectors of BVP, EDA, and ST to form an entire physiological representation:

$$V_P = [V_{bvp}, V_{eda}, V_{st}] \quad (5)$$

4.3 Emotion Recognition

For the proposed recognition module, there are two major parts: modality fusion and decision making. We first used a concatenate layer to fuse the four modality-specific feature vectors (i.e., visual vector, acoustic vector, typing vector, and physiological vector). Then we passed the combined feature vector to a dense layer to further learn the associations across modalities by:

$$r = \tanh(W_r [V_V, V_A, V_T, V_P] + b_r) \quad (6)$$

where r is the final representation for all modalities, W_r and b_r are trainable parameters, V_V is a 128-dimension visual vector, V_A is a 128-dimension acoustic vector, V_T is a 64-dimension typing vector, and V_P is a 96-dimension (V_{bvp} : 64, V_{eda} : 16, V_{st} : 16) physiological vector. Finally, we made the classification by using r as input passed into a sigmoid layer which can output a final emotion classification result (positive or negative).

We implemented our proposed recognition system using the Keras framework with Tensorflow as the backend. We first pre-trained the feature extraction network for each modality respectively, and then tuned the entire network. We set LSTM in visual and acoustic modalities with 256 hidden states, in typing and BVP modalities with 128 hidden states, and in EDA and ST modalities with 64 hidden states. We set the learning rate to 0.01 and used the SGD optimizer. We used 16 as the batch size and trained the model for 100 epochs. Except for the attention layers, we used the ReLU activation function. To overcome the overfitting issue, we adopted dropout and batch normalization.

5 RESULTS

We collected a total of 2400 samples from all modalities, and 400 self-reports (40 participants $\times$ 10 affective sessions) on the spontaneous emotional state of participants. We removed 5 self-reports (1.25%) where the rating values were the same for both positive and negative emotions. Fig. 5 describes the distributions of the induced emotions (positive and negative), as collected for each affective picture and with pictures divided by their own category (happy and sad scenes). We first grouped the participants' self-reports based on the category of the used pictorial stimuli, then split the full range of them into seven bins according to different emotion feeling intensity (1-7 on the rating scales). Finally, the rating distribution for every induced emotion was plotted for each scenario category separately in each of the panels of Fig. 5.

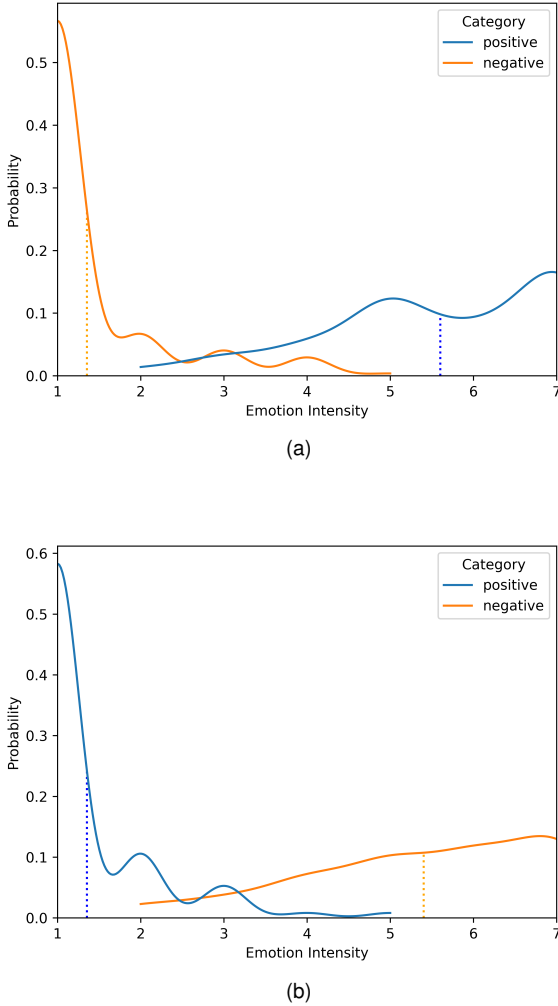


Fig. 5. Distributions of the self-report ratings of positive and negative emotions categories, together with the means of the respective distributions (dotted lines), for happy (a), and sad (b) scenes in selected stimuli database.

From the sub-figure (a) and (b), we can see that the distributions of both positive and negative emotional intensity ratings in all panels are clearly opposite. In the happy scene group, the mean value for positive emotion

is 5.60, while for negative emotion is only 1.35. Besides, 90.7% positive rating values were above the middle value of the scales(=4), and 99.5% negative rating values were below the middle value. Similarly, in the sad scene group, the mean values for negative and positive emotions were 5.40 and 1.35 respectively, with 88.2% negative rating values were above and 98.9% positive rating values were below the middle value. Although in the sad scene group, the negative distribution toward the largest value of the scale(=7) did not show a similar strong bias as the positive distribution toward the smallest value(=1) (the same situation also in the happy scene group), it still makes sense considering: (1) we did not select high-intensity pictures for ethical and experimental reasons; (2) high-intensity emotional changes happen rarely for most people during daily life. Moreover, it did not affect the results of our experiment, since we did not assume the class of affective stimuli as the “ground truth”, instead we focused on self-report and chose the emotion with the highest rating value in each self-report as the “ground truth”. It is worth mentioning that this practice is appropriate given that emotional responses vary from person to person.

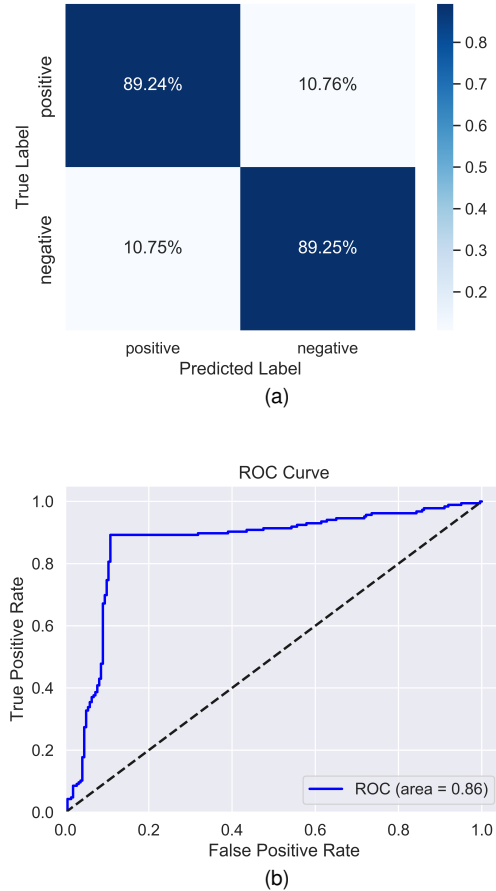


Fig. 6. Classification results of target emotions. (a) confusion matrix; (b) ROC curves.

We also calculated the time spent by participants on each experimental session. On average, each participant needed approximately 16 seconds to complete the emotion assessment and a total of 40 seconds (including induction stage)

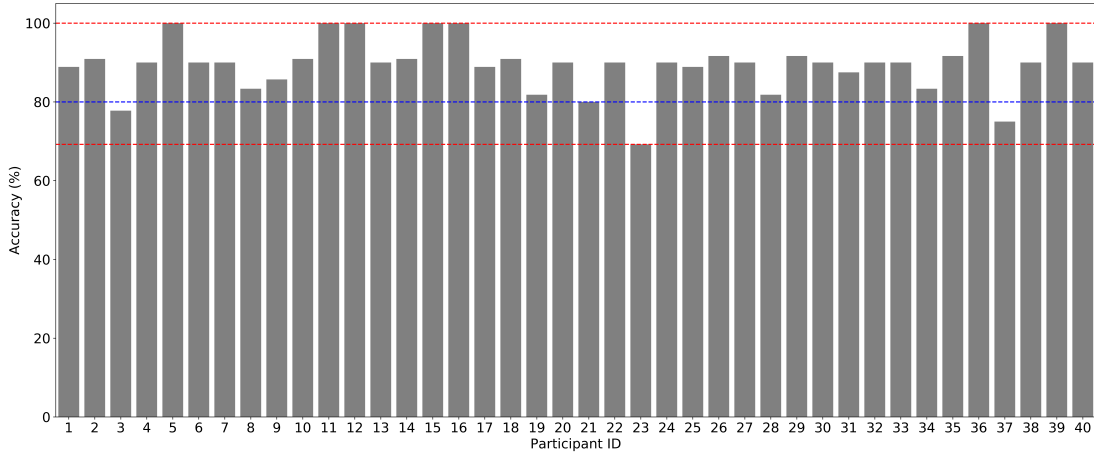


Fig. 7. Recognition accuracy for each participant.

to complete one entire session. Compared with other short-term emotion research with speech or typing tasks [19], [124], our experiment was less time consuming for each experimental session ensuring a more robust emotion induction for all stages.

5.1 Overall Performance

To investigate the robustness of our proposed system, we used leave-one-participant-out cross-validation to evaluate the classification performance. Fig. 7 describes the recognition results for each participant. We can see that our proposed system can recognize the emotional states of almost all participants (37/40, except participant ID 3, 23, and 37) with high accuracy greater than or equal to 80% (blue dotted line). The highest accuracy is 100% for seven participants (ID 5, 11, 12, 15, 16, 36, 39), while the lowest accuracy is 69% for participant ID 23 (two red dotted lines). Moreover, we also evaluated the system's performance on target emotions, and plotted their receiver operating characteristic (ROC) curves with the corresponding area under the curve (AUC) values (shown in Fig. 6). The results show that our model achieved a predictive performance of 89.24% for positive emotion and 89.25% for negative emotion (with 0.86 AUC). Considering that our model is a user-independent model, i.e., we trained the model from a set of participants and tested its performance from other unknown participants.

5.2 Impact of Different Scenarios

Our proposed system used four input modalities (visual, acoustic, typing, and physiological modality) collected from both a smartphone and a wearable wristband to analyze a person's emotional state. However, in real-world scenarios, it is likely that only a subset of modalities are accessible. For example, some people are not willing to wear a wristband in daily life. Or when using communication apps (e.g., Messenger, WhatsApp, Telegram), some users may prefer to chat with voice input, and some may tend to chat with keyboard typing. In such conditions, the user's habits and

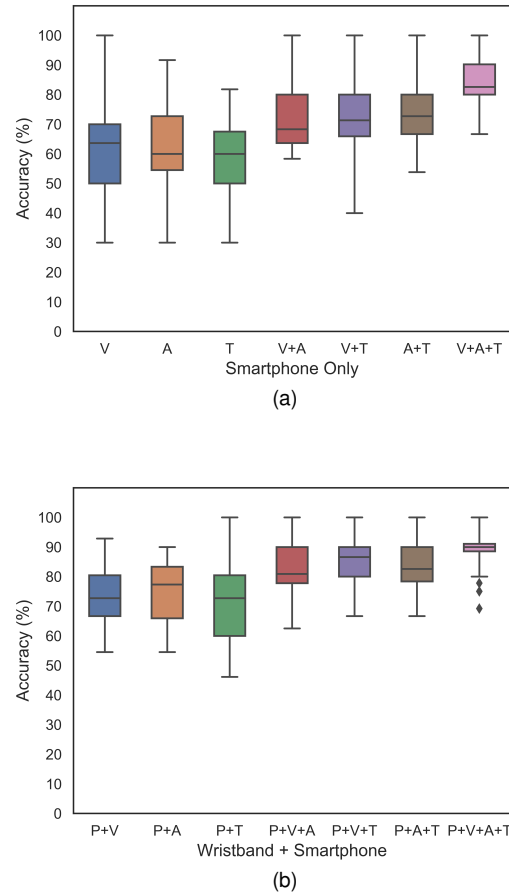


Fig. 8. Classification accuracy for different modality combinations in different scenarios. P: physiological, V: visual, A: acoustic, T: typing.

preferences determine the type of modality data that can be obtained. Thus, we need to explore whether our system can identify emotions using merely a subset of the data streams we investigated.

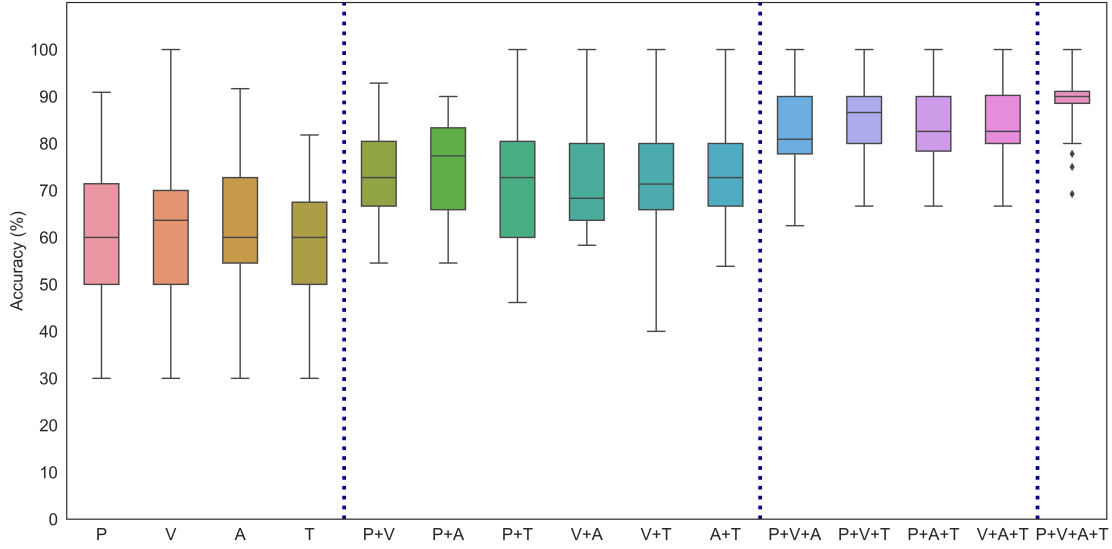


Fig. 9. Classification accuracy for all possible modality combinations. Unimodality (left), Two modalities (middle-left), Three modalities (middle-right), Four modalities (right). P: physiological, V: visual, A: acoustic, T: typing.

In this section, we first explore the performance of each kind of input combination in different real-world scenarios including smartphone-only scenario and wristband+smartphone scenario, as depicted in Fig. 8. We did not consider the wristband-only scenario here, since we regarded all three physiological inputs (BVP, EDA, and ST) as a whole (as explained in Section 4.2), thus, there was only one kind of input in this scenario.

From Fig. 8(a), we can see that in the smartphone-only scenario, three-modality combination (V+A+T) had the best performance, two-modality combinations (V+A, V+T, A+T) had a similar second-best performance, followed by one-modality inputs (V, A, T). Similarly, we see in Fig. 8(b) (i.e., wristband+smartphone scenario) that four-modality combination (P+V+A+T) had the best performance, followed by three-modality combinations (P+V+A, P+V+T, P+A+T) and two-modality combinations (P+V, P+A, P+T). These findings, on the one hand, corroborate the finding from previous work that multimodal accuracies were consistently better than unimodal accuracies [125], and on the other hand, extend its applicable scope to the mobile affective computing field.

Moreover, we put all possible inputs together for comparison, as depicted in Fig. 9. From the figure, we further found that the improvements of multimodal approaches were cross-scenarios. For example, three-modality combination in smartphone-only scenario (V+A+T) not only performed better than two-modality combinations (V+A, V+T, A+T) in the same scenario but also better than two-modality combinations (P+V, P+A, P+T) in wristband+smartphone scenario, i.e., three-modality combinations always had a better performance than two-modality combinations.

We also explored the impact of different combination strategies on classification accuracy of target emotions. As

TABLE 1
Classification accuracy of target emotions for all possible modality combinations. P: physiological, V: visual, A:acoustic, T: typing

	Positive (%)	Negative (%)
P	61.88	58.60
V	60.99	62.90
A	68.16	55.38
T	57.85	56.99
P+V	73.99	73.66
P+A	77.58	72.58
P+T	74.89	66.67
V+A	76.68	65.59
V+T	73.54	72.58
A+T	77.58	66.67
P+V+A	86.55	79.57
P+V+T	88.79	80.65
P+A+T	88.34	76.88
V+A+T	86.10	82.80
P+V+A+T	89.24	89.25

shown in Table 1, we found that for positive emotions, (1) A had the highest accuracy among unimodal inputs; (2) P+A, V+A, and A+T had a similar performance which was modestly better than P+V, P+T, and V+T among two-modality inputs; (3) P+V+T and P+A+T performed similarly with P+V+A+T, and were the best combinations among all inputs. Meanwhile, for negative emotions, (1) V was the best among unimodal inputs; (2) P+V, P+A, and V+T performed better than P+T, V+A, and A+T among two-modality inputs; (3) V+A+T had the best performance among three-modality inputs, followed by P+V+T and P+V+A; (4) P+V+A+T was the best among all inputs. Overall, each kind of input showed its strengths and limitations, and the performance for all combinations was equally able to recognize positive and negative emotions.

5.3 Comparison of Different Fusion Techniques

We also applied the decision-level fusion method to provide a more comprehensive analysis. Decision-level fusion is a kind of *late fusion* method, where different modalities are trained independently, and finally all their outputs are fused by using specific algebraic rules or learning algorithms. Here, we used three principles to achieve decision-level fusion: sum strategy, max strategy, and logistic regression algorithm. Sum strategy aims to sum the probabilities of the same emotions obtained from different classifiers and mark the emotion with the highest probability as the predicted label. Max strategy aims to select the higher probabilistic outputs of different classifiers as final results. Logistic regression processes the weighted combination of outputs from different classifiers by the logistic function.

TABLE 2
Classification accuracy for different fusion techniques. ML: modality-level, DL: decision-level, LR: logistic regression

	Accuracy (%)	Weighted F1-score
ML	89.2	0.89
DL-sum	65.5	0.65
DL-max	64.3	0.64
DL-LR	76.8	0.75

Table 2 shows the results of modality-level fusion, decision-level fusion based on sum strategy, decision-level fusion based on max strategy, and decision-level fusion based on logistic regression. We can see that, compared to modality-level fusion, the performance of decision-level fusion based on sum strategy dropped by 23.7% accuracy and 0.24 weighted F1-score, the performance of decision-level fusion based on max strategy dropped by 24.9% accuracy and 0.25 weighted F1-score, and the performance of decision-level fusion based on logistic regression dropped by 12.4% accuracy and 0.14 weighted F1-score. This is because decision-level fusion assumes all modalities are independent, and therefore cannot learn the mutual association among different modalities. By contrast, in our proposed system, we merged the hidden states of the feature extraction networks trained using every single modality, and allowed the system to learn the mutual correlation through several dense layers.

6 DISCUSSION AND FUTURE WORK

In this paper, we present a novel unobtrusive mobile emotion recognition system, which has the potential to contribute towards research efforts focused on emotional well-being. A large body of research has demonstrated the feasibility of leveraging digital technology to maintain and support a person's emotional well-being [126], [127], [128]. Due to the complexity and dynamic characteristics of human emotion, however, recognizing emotions accurately and in a timely manner is still an open challenge. In practice, achieving accurate and unobtrusive emotion detection would enable emotional health monitoring as well as triggering opportune emotional well-being interventions [129]. Our work tackles these challenges in a number of ways. The multimodal structure we use ensures the continuity of the data collection process and efficiently prevents a scenario

where there would be a lack of adequate realistic training data. The amalgamation of different modalities allows for a better integration of affective information from different channels, which leads to a more comprehensive understanding and judgment of users' emotional states.

Previous mobile emotion recognition work has mostly adopted traditional machine learning and feature engineering algorithms. Although these handcrafted approaches have yielded promising results, the low-level handcrafted features do not generalize well to different application scenarios. Besides, compared with current fast-growing deep learning networks, handcrafted-based approaches often need manual intervention to select the most discriminating features and the related thresholds from sensor data, which can be time-consuming, especially when the feature set is complex and non-linear. In addition, with the increase in size of the extracted feature set, traditional methods will need the support of dimensional reduction techniques to preprocess the features, which will lead to the loss of information of raw data [33]. To overcome these limitations, in this work, we adopted a deep learning approach, and designed a novel attention-based LSTM structure for our proposed mobile emotion-sensing system. Moreover, we did not impose strict restrictions on participants during the experiment, like holding the smartphone with a specific posture or typing with two hands to purposely get higher accuracy. Instead, we asked participants to imagine they were at home and behave normally. As expected, this led to some non-ideal data collection situations, e.g., the user's face being only partially visible or not visible at all. Nevertheless, our system can still detect the target emotions with high accuracy, which indicates strong robustness and disturbance rejection properties.

We further designed a replicable emotion elicitation protocol that leverages off-the-shelf mobile devices. This protocol uses pictorial stimuli to spontaneously evoke the different emotional states, as opposed to commonly used strategies in the emotion recognition databases where either participants are asked to imitate certain emotional facial expressions or speech, or professional actors are recruited to express certain emotions [85], [130], [131]. The latter approaches effectively enhance the distinctness and identifiability of different emotional responses and make it easier for sensors to pick up on them, but these posed expressions are not genuine emotional expressions. Thus, while such recognition systems may exhibit high accuracy, recent work has shown a dramatic drop in performance when actor-trained recognition systems process spontaneous facial expressions [132], [133].

Considering the spontaneous nature of our collected data, our approach still achieved high performance (89.2% average accuracy for 40 participants), well-above the historical classification accuracies using sophisticated models or multimodal data with performance ranged from 55 to 80 percent [76], [79], [127], [134]. This once again demonstrates that our model is robust. We did not further directly compare the performance of our model with previous work due to differences in the experimental aim (short-term emotion research) and experimental data (not collecting other smartphone usage data such as call/SMS logs or internet browser history).

We also explored the impact of different modality combinations. Overall, the best performance was achieved by the four-modality combination, followed by three-modality combinations, two-modality combinations, and finally unimodality. Whether in smartphone-only or wristband+smartphone scenarios, the more modality data collected, the higher accuracy the system achieved. Our results also showed that for a scenario with three modalities, the best-performance combination choice for positive emotion was P+V+T or P+A+T, and for negative emotion, V+A+T. For a scenario with two modalities, the best performance combination for positive emotion was P+A, V+A, or A+T, and for negative emotion, P+V, P+A, or V+T. For unimodality, the best performance for positive emotion was A, and for negative emotion was V. These findings suggest recommendations on how to achieve more reliable mobile emotion recognition. Generally, we recommend selecting different combinations depending on the available modality data.

6.1 Limitations

Our work has several limitations. First, the amount of data we collected was relatively small and the demographics of the participants was skewed towards young adults. Thus, in the future, we aim to recruit a larger sample of participants with a wider range of demographics to build a more robust validation database. Second, in this study, we focused on a narrow set of emotions (i.e., positive and negative emotions). Future work should consider a wider set of emotions to further investigate the robustness of our approach. Third, our experiment was conducted in a laboratory-based setting. In the future, we will focus on long-term user studies conducted in-the-wild, and further improve the proposed system by incorporating other types of smartphone usage data (including Call/SMS logs, location, application usage patterns, etc.).

7 CONCLUSION

In this study, we propose a novel mobile emotion recognition system that uses off-the-shelf mobile devices and attention-based deep multimodal architecture. Our system can predict a user's emotional states through comprehensively analyzing behavioral data (facial expression, speech, keystroke) recorded by a smartphone, with physiological signals recorded by a sensor-rich wristband. In order to evaluate the proposed system, we designed a replicable emotion experiment on mobile devices, including the induction of spontaneous emotion and the collection of multimodal affective responses. Through a leave-one-participant-out cross-validation, our model achieved an average accuracy of 89.2%. Finally, we explored emotion recognition in different scenarios where different data sources are available, and provided recommendations on how to achieve reliable mobile emotion recognition. Our work has the potential to inform the design of future in-the-wild applications that can help monitor real-time emotional well-being and provide emotion regulation recommendations.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council (DP190102627).

REFERENCES

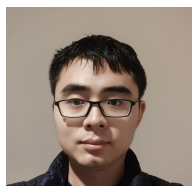
- [1] G. H. Bower, "Mood and memory," *American Psychologist*, vol. 36, no. 2, p. 129, 1981.
- [2] R. M. Nesse and P. C. Ellsworth, "Evolution, Emotions, and Emotional Disorders," *American Psychologist*, vol. 64, no. 2, p. 129, 2009.
- [3] R. W. Picard, "Affective computing: Challenges," *International Journal of Human Computer Studies*, vol. 59, no. 1–2, p. 55–64, Jul. 2003. [Online]. Available: [https://doi.org/10.1016/S1071-5819\(03\)00052-1](https://doi.org/10.1016/S1071-5819(03)00052-1)
- [4] J. J. Gross, "Emotion regulation," *Handbook of emotions*, vol. 3, no. 3, pp. 497–513, 2008.
- [5] J. Tao and T. Tan, "Affective computing: A review," in *Affective Computing and Intelligent Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 981–995.
- [6] Rui Guo, Shuangjiang Li, Li He, Wei Gao, Hairong Qi, and G. Owens, "Pervasive and unobtrusive emotion sensing for human mental health," in *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, 2013, pp. 436–439.
- [7] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in Psychology*, vol. 8, p. 1454, 2017. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpsyg.2017.01454>
- [8] J. E. LeDoux, "Emotion circuits in the brain," *Annual Review of Neuroscience*, vol. 23, no. 1, pp. 155–184, 2000, PMID: 10845062. [Online]. Available: <https://doi.org/10.1146/annurev.neuro.23.1.155>
- [9] K. Grabowski, A. Rynkiewicz, A. Lassalle, S. Baron-Cohen, B. Schuller, N. Cummins, A. Baird, J. Podgórska-Bednarsz, A. Pieniążek, and I. Lucka, "Emotional expression in psychiatric conditions: New technology for clinicians," *Psychiatry and Clinical Neurosciences*, vol. 73, no. 2, pp. 50–62, 2019.
- [10] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004, pp. 80–80.
- [11] Z. Sarsenbayeva, B. Tag, S. Yan, V. Kostakos, and J. Goncalves, "Using video games to regulate emotions," in *32nd Australian Conference on Human-Computer Interaction*, ser. OzCHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 755–759. [Online]. Available: <https://doi.org/10.1145/3441000.3441035>
- [12] H. Atassi and A. Esposito, "A speaker independent approach to the classification of emotional vocal expressions," in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, vol. 2, 2008, pp. 147–152.
- [13] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," *arXiv preprint arXiv:1402.5047*, 2014.
- [14] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "Emotionsense: Emotion recognition based on wearable wristband," in *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, 2018, pp. 346–355.
- [15] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong, "Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning," *Sensors*, vol. 19, no. 7, p. 1716, 2019. [Online]. Available: <https://www.mdpi.com/1424-8220/19/7/1716>
- [16] K. Vytal and S. Hamann, "Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis," *Journal of cognitive neuroscience*, vol. 22, no. 12, p. 2864–2885, December 2010. [Online]. Available: <https://doi.org/10.1162/jocn.2009.21366>
- [17] L. Li and J.-h. Chen, "Emotion recognition using physiological signals," in *Advances in Artificial Reality and Tele-Existence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 437–446.

- [18] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, Feb. 2015. [Online]. Available: <https://doi.org/10.1145/2682899>
- [19] E. L. Broek, "Ubiquitous emotion-aware computing," *Personal Ubiquitous Comput.*, vol. 17, no. 1, p. 53–67, Jan. 2013. [Online]. Available: <https://doi.org/10.1007/s00779-011-0479-9>
- [20] M. Mikhail, K. El-Ayat, R. El Kalioubi, J. Coan, and J. J. B. Allen, "Emotion detection using noisy eeg data," in *Proceedings of the 1st Augmented Human International Conference*, ser. AH '10. New York, NY, USA: Association for Computing Machinery, 2010. [Online]. Available: <https://doi.org/10.1145/1785455.1785462>
- [21] U. Agrawal, E. N. Brown, and L. D. Lewis, "Model-based physiological noise removal in fast fmri," *NeuroImage*, vol. 205, p. 116231, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1053811919308225>
- [22] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.
- [23] Y. Gu, X. Li, K. Huang, S. Fu, K. Yang, S. Chen, M. Zhou, and T. M. Piasecki, "Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder," in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 537–545. [Online]. Available: <https://doi.org/10.1145/3240508.3240714>
- [24] T. J. Trull, M. B. Solhan, S. L. Tragesser, S. Jahng, P. K. Wood, T. M. Piasecki, and D. Watson, "Affective Instability: Measuring a Core Feature of Borderline Personality Disorder With Ecological Momentary Assessment," *Journal of Abnormal Psychology*, vol. 117, no. 3, p. 647, 2008.
- [25] J. Goncalves, P. Pandab, D. Ferreira, M. Ghahramani, G. Zhao, and V. Kostakos, "Projective testing of diurnal collective emotion," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 487–497. [Online]. Available: <https://doi.org/10.1145/2632048.2636067>
- [26] M. Magdin and F. Prikler, "Are Instructed Emotional States Suitable for Classification? Demonstration of How They Can Significantly Influence the Classification Result in An Automated Recognition System," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 4, pp. 141–147, 2019.
- [27] H. Gunes and H. Hung, "Is automatic facial expression recognition of emotions coming to a dead end? the rise of the new kids on the block," *Image Vision Comput.*, vol. 55, no. P1, p. 6–8, Nov. 2016. [Online]. Available: <https://doi.org/10.1016/j.imavis.2016.03.013>
- [28] K. Yang, C. Wang, Z. Sarsenbayeva, B. Tag, T. Dingler, G. Wadley, and J. Goncalves, "Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets," *The Visual Computer*, vol. 37, no. 6, pp. 1447–1466, 2021.
- [29] A. Visuri, Z. Sarsenbayeva, J. Goncalves, E. Karapanos, and S. Jones, "Impact of mood changes on application selection," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 535–540. [Online]. Available: <https://doi.org/10.1145/2968219.2968317>
- [30] S. Koganemaru, K. Domen, H. Fukuyama, and T. Mima, "Negative emotion can enhance human motor cortical plasticity," *European Journal of Neuroscience*, vol. 35, no. 10, pp. 1637–1645, 2012.
- [31] E. Siedlecka and T. F. Denson, "Experimental methods for inducing basic emotions: A qualitative review," *Emotion Review*, vol. 11, no. 1, pp. 87–97, 2019. [Online]. Available: <https://doi.org/10.1177/1754073917749016>
- [32] E. Politou, E. Alepis, and C. Patsakis, "A survey on mobile affective computing," *Computer Science Review*, vol. 25, pp. 79–100, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1574013717300382>
- [33] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46–56, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253518300460>
- [34] D. Y. Dai and R. J. Sternberg, *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development*. Routledge, 2004.
- [35] A. Damasio, "Neural basis of emotions," *Scholarpedia*, vol. 6, no. 3, p. 1804, 2011.
- [36] L. Alba-Ferrara, M. Hausmann, R. L. Mitchell, and S. Weis, "The neural correlates of emotional prosody comprehension: Disentangling simple from complex emotion," *PLOS ONE*, vol. 6, no. 12, pp. 1–9, 12 2011. [Online]. Available: <https://doi.org/10.1371/journal.pone.0028701>
- [37] J. A. Russell, "Introduction to special section: On defining emotion," *Emotion Review*, vol. 4, no. 4, pp. 337–337, 2012. [Online]. Available: <https://doi.org/10.1177/1754073912445857>
- [38] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, 2018. [Online]. Available: <https://www.mdpi.com/1424-8220/18/7/2074>
- [39] C. Darwin, "The Expression of the Emotions in Man and Animals." *The Journal of the Anthropological Institute of Great Britain and Ireland*, vol. 2, 1873.
- [40] U. Hess and P. Thibault, "Darwin and Emotion Expression," *American Psychologist*, vol. 64, no. 2, p. 120, 2009.
- [41] M. Ghiselin, P. Ekman, and H. Gruber, "Darwin and Facial Expression: A Century of Research in Review." *Systematic Zoology*, vol. 23, no. 4, 1974.
- [42] P. Ekman, R. Levenson, and W. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983. [Online]. Available: <https://science.sciencemag.org/content/221/4616/1208>
- [43] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.
- [44] A. H. Pierce, W. Wundt, and C. H. Judd, "Outlines of Psychology." *The Philosophical Review*, vol. 17, no. 2, 1908.
- [45] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, p. 1161, 1980.
- [46] P. Ekman and W. V. Friesen, "Manual for the facial action coding system," *Consulting Psychologist*, vol. 104, 1978.
- [47] P. Lewinski, T. M. Den Uyl, and C. Butler, "Automated facial coding: Validation of basic emotions and FACS AUs in facereader," *Journal of Neuroscience, Psychology, and Economics*, vol. 7, no. 4, p. 227, 2014.
- [48] C. Juanjuan, Z. Zheng, S. Han, and Z. Gang, "Facial expression recognition based on pca reconstruction," in *2010 5th International Conference on Computer Science Education*, 2010, pp. 195–198.
- [49] S. Berretti, B. B. Amor, M. Daoudi, and A. Del Bimbo, "3D facial expression recognition using SIFT descriptors of automatically detected keypoints," *Visual Computer*, vol. 27, no. 11, pp. 1021–1036, 2011.
- [50] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.
- [51] P. N. Juslin and K. R. Scherer, "Vocal Expression of Affect," in *The New Handbook of Methods in Nonverbal Behavior Research*, 2008.
- [52] A. Milton, S. Sharmy Roy, and S. Tamil Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.
- [53] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.
- [54] K. A. Lindquist, "Emotions emerge from more basic psychological ingredients: A modern psychological constructionist model," *Emotion Review*, vol. 5, no. 4, pp. 356–368, 2013. [Online]. Available: <https://doi.org/10.1177/1754073913489750>
- [55] K. A. Lindquist and L. F. Barrett, "A functional architecture of the human brain: Emerging insights from the science of emotion," *Trends in Cognitive Sciences*, vol. 16, no. 11, pp. 533–540, 2012.
- [56] Y. Hsu, J. Wang, W. Chiang, and C. Hung, "Automatic eeg-based emotion recognition in music listening," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 85–99, 2020.
- [57] W. Zheng, B. Dong, and B. Lu, "Multimodal emotion recognition using eeg and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 5040–5043.
- [58] B. G. Lee, T. W. Chong, B. L. Lee, H. J. Park, Y. N. Kim, and B. Kim, "Wearable mobile-based emotional response-monitoring system for drivers," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 5, pp. 636–649, 2017.

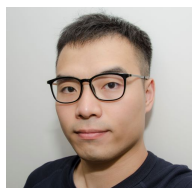
- [59] M. Perusquía-Hernández, M. Hirokawa, and K. Suzuki, "A wearable device for fast and subtle spontaneous smile recognition," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 522–533, 2017.
- [60] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 95–108. [Online]. Available: <https://doi.org/10.1145/2973750.2973762>
- [61] X. Zhang, W. Li, X. Chen, and S. Lu, "Moodexplorer: Towards compound emotion detection via smartphone sensing," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, Jan. 2018. [Online]. Available: <https://doi.org/10.1145/3161414>
- [62] H. Wu, J. Feng, X. Tian, E. Sun, Y. Liu, B. Dong, F. Xu, and S. Zhong, "Emo: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices," in *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 448–461. [Online]. Available: <https://doi.org/10.1145/3386901.3388917>
- [63] K. Masai, K. Kunze, Y. Sugiura, M. Ogata, M. Inami, and M. Sugimoto, "Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 4, Dec. 2017. [Online]. Available: <https://doi.org/10.1145/3012941>
- [64] E. Di Lascio, S. Gashi, and S. Santini, "Laughter recognition using non-invasive wearable devices," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 262–271. [Online]. Available: <https://doi.org/10.1145/3329189.3329216>
- [65] Z. Wang, Z. Yu, B. Zhao, B. Guo, C. Chen, and Z. Yu, "Emotionsense: An adaptive emotion recognition system based on wearable smart devices," *ACM Trans. Comput. Healthcare*, vol. 1, no. 4, Sep. 2020. [Online]. Available: <https://doi.org/10.1145/3384394>
- [66] K. Sharma, E. Niforatos, M. Giannakos, and V. Kostakos, "Assessing cognitive performance using physiological and facial features: Generalizing across contexts," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, Sep. 2020. [Online]. Available: <https://doi.org/10.1145/3411811>
- [67] Z. Sarsenbayeva, G. Marini, N. van Berkel, C. Luo, W. Jiang, K. Yang, G. Wadley, T. Dingler, V. Kostakos, and J. Goncalves, *Does Smartphone Use Drive Our Emotions or Vice Versa? A Causal Analysis*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–15. [Online]. Available: <https://doi.org/10.1145/3313831.3376163>
- [68] D. Girardi, N. Novielli, D. Fucci, and F. Lanubile, "Recognizing developers' emotions while programming," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ser. ICSE '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 666–677. [Online]. Available: <https://doi.org/10.1145/3377811.3380374>
- [69] Y. Gu, S. Chen, and I. Marsic, "Deep mul timodal learning for emotion recognition in spoken language," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5079–5083.
- [70] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 439–448.
- [71] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.
- [72] X. Zhou, J. Guo, and R. Bie, "Deep learning based affective model for speech emotion recognition," in *2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCoM/IoP/SmartWorld)*, 2016, pp. 841–846.
- [73] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jul. 2018, pp. 2225–2235. [Online]. Available: <https://aclanthology.org/P18-1207>
- [74] A. Shukla, S. S. Gullapuram, H. Katti, M. Kankanhalli, S. Winkler, and R. Subramanian, "Recognition of advertisement emotions with application to computational advertising," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.
- [75] S. Rayatdoost, D. Rudrauf, and M. Soleymani, "Multimodal gated information fusion for emotion recognition from eeg signals and facial behaviors," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, ser. ICMI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 655–659. [Online]. Available: <https://doi.org/10.1145/3382507.3418867>
- [76] S. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 200–213, 2020.
- [77] R. Wampfler, S. Klingler, B. Solenthaler, V. R. Schinazi, and M. Gross, "Affective state prediction based on semi-supervised learning from smartphone touch data," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–13. [Online]. Available: <https://doi.org/10.1145/3313831.3376504>
- [78] A. Exler, A. Schankin, C. Klebsattel, and M. Beigl, "A wearable system for mood assessment considering smartphone features and data from mobile ecgs," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1153–1161. [Online]. Available: <https://doi.org/10.1145/2968219.2968302>
- [79] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: Building a mood sensor from smartphone usage patterns," in *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 389–402. [Online]. Available: <https://doi.org/10.1145/2462456.2464449>
- [80] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, and X. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.
- [81] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.
- [82] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multi-modal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.
- [83] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.
- [84] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3438–3446.
- [85] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, 2008, pp. 1–8.
- [86] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 5 pp.–.
- [87] J. A. Coan and J. J. Allen, *Handbook of emotion elicitation and assessment*. Oxford university press, 2007.
- [88] B. Verschuere, G. Crombez, and E. Koster, "Cross cultural validation of the IAPS," *Ghent Belgium Ghent University*, pp. 14–17, 2007.
- [89] M. Riegel, A. Moslehi, J. M. Michałowski, Ł. Żurawski, M. Horvat, M. Wypych, K. Jednoróg, and A. Marchewka, "Nencki affective picture system: Cross-cultural study in europe and iran," *Frontiers in Psychology*, vol. 8, p. 274, 2017. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpsyg.2017.00274>

- [90] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," *NIMH Center for the Study of Emotion and Attention*, vol. 1, pp. 39–58, 1997.
- [91] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior Research Methods*, vol. 37, no. 4, pp. 626–630, 2005.
- [92] A. Marchewka, Ł. Żurawski, K. Jednoróg, and A. Grabowska, "The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database," *Behavior Research Methods*, vol. 46, no. 2, pp. 596–610, 2014.
- [93] J. M. Michałowski, D. Drożdżel, J. Matuszewski, W. Koziejowski, K. Jednoróg, and A. Marchewka, "The Set of Fear Inducing Pictures (SFIP): Development and validation in fearful and non-fearful individuals," *Behavior Research Methods*, vol. 49, no. 4, pp. 1407–1419, 2017.
- [94] D. Dai, Q. Liu, and H. Meng, "Can your smartphone detect your emotion?" in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2016, pp. 1704–1709.
- [95] M. Perusquía-Hernández, S. Ayabe-Kanamura, K. Suzuki, and S. Kumano, "The invisible potential of facial electromyography: A comparison of emg and computer vision when distinguishing posed from spontaneous smiles," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–9. [Online]. Available: <https://doi.org/10.1145/3290605.3300379>
- [96] R. Laureanti, M. Bilucaglia, M. Zito, R. Circi, A. Fici, F. Rivetti, R. Valesi, C. Oldrini, L. T. Mainardi, and V. Russo, "Emotion assessment using machine learning and low-cost wearable devices," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 576–579.
- [97] H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel, "Biocnn: A hardware inference engine for eeg-based emotion detection," *IEEE Access*, vol. 8, pp. 140 896–140 914, 2020.
- [98] S. R. Vrana, E. L. Spence, and P. J. Lang, "The Startle Probe Response: A New Measure of Emotion?" *Journal of Abnormal Psychology*, vol. 97, no. 4, p. 487, 1988.
- [99] E. Bernat, C. J. Patrick, S. D. Benning, and A. Tellegen, "Effects of picture content and intensity on affective physiological response," *Psychophysiology*, vol. 43, no. 1, pp. 93–103, 2006.
- [100] M. Codispoti, V. Ferrari, and M. M. Bradley, "Repetitive picture processing: Autonomic and cortical correlates," *Brain Research*, vol. 1068, no. 1, pp. 213–220, 2006. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S000689930501591X>
- [101] J. L. Rhudy, K. M. McCabe, and A. E. Williams, "Affective modulation of autonomic reactions to noxious stimulation," *International Journal of Psychophysiology*, vol. 63, no. 1, pp. 105–109, 2007.
- [102] M. K. Uhrig, N. Trautmann, U. Baumgärtner, R.-D. Treede, F. Henrich, W. Hiller, and S. Marschall, "Emotion elicitation: A comparison of pictures and films," *Frontiers in Psychology*, vol. 7, p. 180, 2016. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fpsyg.2016.00180>
- [103] M. Riegel, Ł. Żurawski, M. Wierzbza, A. Moslehi, Ł. Kłoczek, M. Horvat, A. Grabowska, J. Michałowski, K. Jednoróg, and A. Marchewka, "Characterization of the Nencki Affective Picture System by discrete emotional categories (NAPS BE)," *Behavior Research Methods*, vol. 48, no. 2, pp. 600–612, 2016.
- [104] A. Heraz and M. Clynes, "Recognition of emotions conveyed by touch through force-sensitive screens: Observational study of humans and machine learning techniques," *Journal of Medical Internet Research*, vol. 20, no. 8, 2018.
- [105] I. S. MacKenzie and R. W. Soukoreff, "Phrase sets for evaluating text entry techniques," in *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 754–755. [Online]. Available: <https://doi.org/10.1145/765891.765971>
- [106] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.
- [107] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.
- [108] Z. Sarsenbayeva, N. van Berkel, E. Velloso, V. Kostakos, and J. Goncalves, "Effect of distinct ambient noise types on mobile interaction," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 2, Jul. 2018. [Online]. Available: <https://doi.org/10.1145/3214285>
- [109] W. Y. Quack, D. Huang, W. Lin, H. Li, and M. Dong, "Mobile acoustic emotion recognition," in *2016 IEEE Region 10 Conference (TENCON)*, 2016, pp. 170–174.
- [110] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Hybrid attention based multimodal network for spoken language classification," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, aug 2018, pp. 2379–2390. [Online]. Available: <https://www.aclweb.org/anthology/C18-1201>
- [111] P. Pham and J. Wang, "Attentivevideo: A multimodal approach to quantify emotional responses to mobile advertisements," *ACM Trans. Interact. Intell. Syst.*, vol. 9, no. 2–3, Mar. 2019. [Online]. Available: <https://doi.org/10.1145/3232233>
- [112] S. Ghosh, K. Hiware, N. Ganguly, B. Mitra, and P. De, "Emotion detection from touch interactions during text entry on smartphones," *International Journal of Human-Computer Studies*, vol. 130, pp. 47–57, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1071581918304889>
- [113] M. Kennedy and A. Balint, "co60ca/emotionnet2," 2018. [Online]. Available: <https://github.com/co60ca/EmotionNet2>
- [114] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5562–5570.
- [115] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [116] I. Tautkute, T. Trzcinski, and A. Bielski, "I know how you feel: Emotion recognition with facial landmarks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1959–1959.
- [117] 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*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: <https://doi.org/10.1145/1873951.1874246>
- [118] F. Hassan, M. R. Alam Kotwal, M. M. Rahman, M. Nasiruddin, M. A. Latif, and M. Nurul Huda, "Local feature or mel frequency cepstral coefficients - which one is better for mfn-based bangla speech recognition?" in *Advances in Computing and Communications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 154–161.
- [119] H. Lee, Young Sang Choi, Sunjae Lee, and I. P. Park, "Towards unobtrusive emotion recognition for affective social communication," in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, 2012, pp. 260–264.
- [120] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [121] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [122] X. Zhang, F. Zhuang, W. Li, H. Ying, H. Xiong, and S. Lu, "Inferring mood instability via smartphone sensing: A multi-view learning approach," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1401–1409. [Online]. Available: <https://doi.org/10.1145/3343031.3350957>
- [123] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, jun 2016, pp. 1480–1489. [Online]. Available: <https://www.aclweb.org/anthology/N16-1174>

- [124] P. M. Lee, W. H. Tsui, and T. C. Hsiao, "The influence of emotion on keyboard typing: An experimental study using visual stimuli," *BioMedical Engineering Online*, vol. 13, no. 1, pp. 1–12, 2014.
- [125] S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 31–38. [Online]. Available: <https://doi.org/10.1145/2388676.2388686>
- [126] S. A. Cambo, D. Avrahami, and M. L. Lee, "Breaksense: Combining physiological and location sensing to promote mobility during work-breaks," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 3595–3607. [Online]. Available: <https://doi.org/10.1145/3025453.3026021>
- [127] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland, "Daily stress recognition from mobile phone data, weather conditions and individual traits," in *Proceedings of the 22nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 477–486. [Online]. Available: <https://doi.org/10.1145/2647868.2654933>
- [128] H. Yu, E. B. Klerman, R. W. Picard, and A. Sano, "Personalized wellbeing prediction using behavioral, physiological and weather data," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019, pp. 1–4.
- [129] L. Germine, R. W. Strong, S. Singh, and M. J. Sliwinski, "Toward dynamic phenotypes and the scalable measurement of human behavior," *Neuropsychopharmacology*, vol. 46, no. 1, pp. 209–216, 2021.
- [130] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0262885611000515>
- [131] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [132] 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.
- [133] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.
- [134] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Öhler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140–148, 2015.



Kangning Yang received his B. Eng. (2017) in automation from University of Electronic Science and Technology of China, China, and his M.S. (2019) in electrical and computer engineering from Rutgers, the State University of New Jersey, USA. He is current pursuing his Ph.D. at the University of Melbourne, Australia. His research interests include Emotion Recognition, Human Computer Interaction, and Deep Learning.



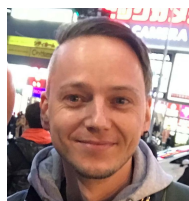
Chaofan Wang received his M.S. (2019) in information technology from the University of Melbourne, where he is currently pursuing the Ph.D. His research interests are Human Computer Interaction, Ubiquitous Computing, and Wearable Sensors.



Yue Gu received his Ph.D. (2020) in Electrical and Computer Engineering department from Rutgers, the State University of New Jersey, USA. Currently he works as an applied scientist in the Amazon Web Services & Amazon AI. His research interests include multimodal learning, affective computing, and speech recognition.



Zhanna Sarsenbayeva is currently a Postdoctoral Research Fellow in the School of Computing and Information Systems from the University of Melbourne. She received a Ph.D.(2020) in Computer Science and Engineering at the University of Melbourne. Her research interests include accessibility, ubiquitous computing, human-computer interaction, and affective computing.



Benjamin Tag is currently a Research Fellow at the School of Computing and Information Systems. He received his Ph.D. at the Graduate School of Media Design at KEIO University in Japan (2019), and he is actively involved in understanding human cognition by combining methods from the fields of cognitive psychology and pervasive computing. His recent research focuses on Digital Emotion Regulation, Cognitive Biases and the application of digital nudges to improve media literacy among technology users.



Tilman Dingler is currently a Lecturer in the School of Computing and Information Systems at the University of Melbourne. He received a PhD in Computer Science from the University of Stuttgart, Germany (2016). His research focuses on cognition-aware systems and technologies that support users' information processing capabilities.



Greg Wadley is currently a Senior Lecturer in the School of Computing and Information Systems at the University of Melbourne. He received his Ph.D. in Human-Computer Interaction from the University of Melbourne (2012). His research activities involves designing and evaluating technology interventions as well as studying the user experience and social impact of digital technologies.



Jorge Goncalves is currently a Senior Lecturer at the School of Computing and Information Systems at the University of Melbourne. He received his Ph.D. in Computer Science and Engineering from the University of Oulu (2015). His research interests include ubiquitous computing, human-computer interaction, crowdsourcing, affective computing, and social computing.

Chapter 6

Hierarchical Self-Attention Network

6.1 Introduction

Although RNNs have been firmly established for modeling and processing the temporal dependencies in sequential signals [232, 280], recently the Transformer and CNN based models have demonstrated better capabilities [281, 282]. Stemming from our work presented in Chapter 5, in this chapter, we further propose an efficient hierarchical mobile emotion recognition network that relies entirely on the convolution-augmented self-attention mechanism [270] to draw both content-based global and position-wise local dependencies of the input sequences and compute related feature map. Moreover, in order to study natural emotions arising in the wild and generalize the proposed framework to realistic scenarios, we used a multimodal naturalistic dialogue dataset as the validation dataset [283]. This dataset does not contain any intentional or fictitious emotional displays, instead relying on spontaneous emotional responses. By using this dataset we aimed to mitigate concerns regarding the authenticity of posed emotional expressions in affective computing research [284–286].

Our results show that the proposed architecture achieves state-of-the-art performance in all five classification tasks. The quantitative generalizability tests also highlight that

our system exhibits good convergence and generalization, and can handle different physiological signals. More details of this study can be found in the attached publication, [Article III](#).

6.2 Article III

©Copyright held by the authors. Publication rights licensed to ACM. This is the authors' version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in:

Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, ICMR '22, page 562–570, New York, NY, USA, 2022. Association for Computing Machinery. URL <https://doi.org/10.1145/3512527.3531385>

Mobile Emotion Recognition via Multiple Physiological Signals using Convolution-augmented Transformer

Kangning Yang
The University of Melbourne
Australia
kangning.yang@student.unimelb.edu.au

Benjamin Tag
The University of Melbourne
Australia
benjamin.tag@unimelb.edu.au

Yue Gu
Rutgers University
USA
yg202@scarletmail.rutgers.edu

Chaofan Wang
The University of Melbourne
Australia
chaofanw@student.unimelb.edu.au

Tilman Dingler
The University of Melbourne
Australia
tilman.dingler@unimelb.edu.au

Greg Wadley
The University of Melbourne
Australia
greg.wadley@unimelb.edu.au

Jorge Goncalves
The University of Melbourne
Australia
jorge.goncalves@unimelb.edu.au

ABSTRACT

Recognising and monitoring emotional states play a crucial role in mental health and well-being management. Importantly, with the widespread adoption of smart mobile and wearable devices, it has become easier to collect long-term and granular emotion-related physiological data passively, continuously, and remotely. This creates new opportunities to help individuals manage their emotions and well-being in a less intrusive manner using off-the-shelf low-cost devices. Pervasive emotion recognition based on physiological signals is, however, still challenging due to the difficulty to efficiently extract high-order correlations between physiological signals and users' emotional states. In this paper, we propose a novel end-to-end emotion recognition system based on a convolution-augmented transformer architecture. Specifically, it can recognise users' emotions on the dimensions of arousal and valence by learning both the global and local fine-grained associations and dependencies within and across multimodal physiological data (including blood volume pulse, electrodermal activity, heart rate, and skin temperature). We extensively evaluated the performance of our model using the K-EmoCon dataset, which is acquired in naturalistic conversations using off-the-shelf devices and contains spontaneous emotion data. Our results demonstrate that our approach outperforms the baselines and achieves state-of-the-art or competitive performance. We also demonstrate the effectiveness and generalizability of our system on another affective dataset which used affect inducement and commercial physiological sensors.

CCS CONCEPTS

• **Human-centered computing** → Ubiquitous and mobile computing; • **Computing methodologies** → Neural networks.

KEYWORDS

emotion recognition, physiological signals, off-the-shelf mobile devices, convolution-augmented transformer

ACM Reference Format:

Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2022. Mobile Emotion Recognition via Multiple Physiological Signals using Convolution-augmented Transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22)*, June 27–30, 2022, Newark, NJ, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3512527.3531385>

1 INTRODUCTION

The emotions we feel every day shape our behaviours and guide our decisions. Emotions can, however, go awry and lead to negative repercussions to an individual's well-being. For instance, long periods under stress and anxiety may not only impair mental health but also induce related diseases. Recent studies have demonstrated a high correlation between affective instability and psychosis [18]. Thus, automatic and accurate emotion recognition has increasingly become an important research topic as it can assist with early diagnosis, continuous monitoring, and can inform interventions for mental health and well-being.

In recent decades, researchers have explored different ways to empower machines with human-like perception of emotional states. Current automatic emotion recognition approaches can be categorized into two main types according to the signals used. One approach entails using human behavioral signals, such as facial expressions [4, 29, 43], voice [19], and gestures [28]. While these signals are typically easier to collect, the reliability of such methods cannot be guaranteed since these signals are semi-voluntary or voluntary responses [8, 26]. This means that people can easily disguise inner emotions by controlling their behaviors. For example, people

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMR '22, June 27–30, 2022, Newark, NJ, USA

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9238-9/22/06...\$15.00

<https://doi.org/10.1145/3512527.3531385>

can conceal their real emotions by showing a “poker face” in social communications [44]. The other approach entails using physiological signals, such as electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), galvanic skin response (GSR), and body temperature. Unlike behavioral signals, physiological signals originate from the activity of the central nervous system (CNS) and the autonomic nervous systems (ANS), which are involuntary responses and quite challenging to be controlled or hidden intentionally [31]. Therefore, approaches based on physiological signals are more reliable and universally applicable to discern the inner emotional feelings of human beings [14, 46].

However, emotion recognition based on physiological signals is still challenging. On the one hand, existing methods mainly rely on experience-based feature engineering to manually quantify numerical attributes that can characterize affective information contained in different physiological signals. Typically, the extracted features concentrate on the time and frequency domains, for example, the commonly used time-domain statistical features — minimum, maximum, mean, standard deviation [7], and frequency-domain features — energy distribution at different frequency bands [40]. Despite being widely adopted, these handcrafted features are essentially low-level feature representations, which do not generalize well in different scenarios and can not efficiently model the complex and non-linear spatio-temporal relationships among multiple physiological signals. Recent research has found that multiple physiological signals can more effectively reflect emotional changes than a single signal [31]. On the other hand, most affective databases (e.g., DEAP [14], MAHNOB-HCI [33]) used in physiological signals-based emotion studies are acquired by using artificial induction methods, i.e., by using specific music, pictures, or video clips as stimuli to induce the generation of certain emotional states from participants. Although such emotional portrayals could be considered as “spontaneous” emotional expressions, they do not represent natural emotional responses collected from in-the-wild social interactions and lack adequate contextual information. Moreover, the data acquisition in these databases typically relies on medical-level equipment because of its affordances, such as high sensitivity and high sampling rate. For example, the acquisition equipment used in the DEAP database supports a 512 Hz sampling rate for physiological responses, while the acquisition equipment used in the MAHNOB-HCI database supports a 1024 Hz sampling rate. Although high-quality physiological signals can be collected in this way, these facilities tend to be expensive and intrusive. For example, an EEG-based system requires a large number of electrodes to be attached to an individual’s scalp [10]. These hinder emotion recognition techniques from being integrated into everyday devices and being deployed into real-world contexts.

To address these issues, we design a deep multimodal architecture with a convolution-augmented transformer (conformer [11]) mechanism to classify emotions on different categories of arousal and valence. Specifically, it first aligns physiological signals by using the nearest-neighbor interpolation method to synchronize different sensor inputs. Then, it extracts high-level informative features from different signals through individual convolutional neural networks (CNNs). Lastly, by leveraging a conformer encoder, it can learn the latent local and global associations among different physiological signals. Moreover, with the development of mobile

technology, modern mobile and wearable devices are increasingly low-cost, sensor-rich, and lightweight. This makes them well-suited to detect multimodal physiological signals and emotional responses in an unobtrusive manner, overcoming the aforementioned inherent limitations from medical-grade acquisition equipment in traditional physiological emotion recognition research. Therefore, in this study, we evaluated our system using the K-EmoCon dataset [21], a publicly available multimodal sensor dataset acquired with off-the-shelf wearable devices in naturalistic conversations. We used four kinds of physiological signals collected from an Empatica E4 Wristband as inputs, blood volume pulse (BVP), electrodermal activity (EDA), heart rate (HR), and skin temperature (SKT), and conducted five classification tasks, two-category and five-category classification on arousal and valence respectively, and a four-category classification on combined arousal-valence emotion space. We also tested the generalizability of our system by conducting two additional experiments on the ASCERTAIN dataset [35].

Thus, the contribution of our work is two-fold:

- (1) We propose a novel conformer-based deep learning structure that detects human emotional states based on multimodal physiological signals from off-the-shelf devices. To the best of our knowledge, we are the first to apply a conformer mechanism in the field of affective computing.
- (2) We conduct extensive experiments to evaluate and test our proposed system. The results demonstrate that our model outperforms previous techniques, achieving state-of-the-art or comparable performance.

2 RELATED WORK

Despite a large number of studies on emotion recognition, most of them have focused on audio-visual signals [22] (e.g., facial expressions and speech), since they are seen as the most direct channels for human emotional expressions. However, these outward emotional signals can easily be modified or suppressed in social settings, reducing their reliability as signs of inner emotional feelings [31, 46]. In addition, recent work has argued that emotional facial expressions vary across races and cultures, and facial configurations can not act as a reliable signal for particular emotional states [1]. Furthermore, as faces and voices are directly linked to identity, these can raise significant privacy concerns, especially if they have to be uploaded to the cloud for processing [30].

For these reasons, the link between emotions and a range of physiological signals has recently received increased attention in the field of affective computing. For example, previous work has leveraged BVP, EDA, and SKT to recognise emotions [44]. They adopted several different types of film clips as stimuli to evoke participants’ emotions, and utilized an Empatica E4 wristband to collect the induced physiological responses. In another example, researchers proposed a method to recognise emotions relying on frontal EEG signals [41]. Specifically, they first selected four kinds of VR affective scenes containing happiness, fear, peace, and disgust. Then by using textile dry electrodes, they synchronously recorded three-channel frontal EEG signals when participants watched each VR stimulus. Unlike facial or acoustic data containing biometric information, physiological data as biosensed information may be

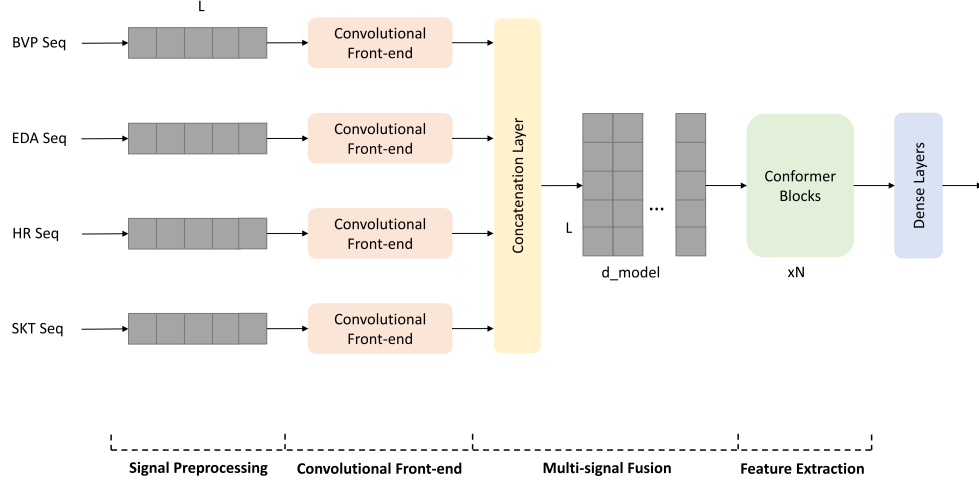


Figure 1: Overall structure of our proposed emotion recognition system.

used to describe our health status, behaviors, or emotions, but are not enough to identify us uniquely [30].

Researchers also proposed different approaches to analyze physiological data in order to detect an individual’s emotions. For example, Zhao et al. [44] first used adaptive band-pass filter and low-pass filter to eliminate the noise contained in the collected signals. Then, they extracted the distinct emotion-related features from the time domain, frequency domain, and nonlinear analysis. By adopting the *sequential forward floating selection method* to select the best feature set, they classified different emotions based on support vector machine (SVM). Similarly, Xu et al. [41] extracted three kinds of handcrafted features including the time domain, frequency domain, and space domain from collected frontal EEG signals. Then by using the *model stacking method*, they combined three models, Gradient Boosting Decision Tree, Random Forest, and SVM, to achieve EEG-based emotion recognition.

Researchers have, however, recently noted the limitations of handcrafted feature representations, and explored ways of designing deep learning-based architectures to extract fine-grained and higher-level features by taking full advantage of neural networks’ powerful feature abstraction ability. For instance, Wang et al. [39] presented a deep learning method to learn high-level feature representations from the raw EEG signals. They built a three-dimensional CNN and classified emotions on valence and arousal scales. Zitouni et al. [47] proposed a bidirectional long short-term memory (LSTM) neural network to extract informative features from four kinds of physiological signals, and predicted emotions into binary levels and quadrants of the arousal-valence space.

More recently, transformers were used to model long-range global context. For example, Wang et al. [38] designed a heartbeat-aware attention mechanism, and added it into transformer structure to enhance the alignment between encoded and decoded sequences. On this basis, they made arrhythmia classification from ECG signals. Similarly, Behinaein et al. [2] placed a convolutional front-end

before the transformer encoder to extract more informative representations, and achieved stress detection from ECG signals. Despite models with these structures being able to capture either local spatial dependencies, or temporal dependencies, or global information, they lack the capability to learn both local and global interactions simultaneously.

Our architecture is inspired by the conformer mechanism used in recent automatic speech recognition research [11] that combines the advantages of convolutions and self-attention mechanisms. In this paper we extend it to multimodal environments for emotion recognition, and design the conformer encoder to learn both position-wise local features and content-based global associations within and across different modalities.

3 METHOD

The overall structure of our proposed emotion recognition system is illustrated in Figure 1. There are four major parts of the system: signal preprocessing, convolutional front-end, multi-signal fusion, and feature extraction.

3.1 Signal Preprocessing

Our system accepts raw data retrieved from different sensors as sequence inputs. Through the data preprocessing module, the heterogeneous inputs will be formatted into specific representations, which can be effectively used in the following modules.

Specifically, in this study, we focused on four different kinds of commonly used physiological signals: (1) BVP (blood volume pulse), is related to the changes in blood volume in arteries and capillaries, and can be measured by a non-invasive optical sensor that detects changes in light absorption density of the localized tissue (e.g., skin) when illuminated [15]; (2) EDA (electrodermal activity), measures the variations in electric characteristics of the skin resulting from changes in sweat production and fluid concentration

in the sweat ducts, and can be monitored by the voltage changes between electrodes [15]; (3) HR (heart rate), is a measure of the functional activity of the heart, and can be estimated using the ECG or BVP signals; (4) SKT (skin temperature), can be measured by a thermopile infrared sensor.

Following the preprocessing operations stated in [47], we first normalized the raw signals of each subject separately to overcome individual differences which may vary due to age, gender, and personality. The normalization is performed based on the signals collected in the final 1.5 minutes of the relaxation period prior to each trial. Then, we leveraged the nearest-neighbor method and interpolated the lower frequency signals based on the highest sampling frequency to synchronize different physiological signals. Subsequently, each bio-signal can be represented as an array with the same length L , where $L = t \times f$, t is the time interval of emotional annotation, and f is the highest sampling frequency of four sensor inputs.

3.2 Convolutional Front-end

To extract more expressive features from physiological inputs, we applied the CNNs structure for each signal, which is able to capture fine-grained feature patterns by local-perceiving convolutional kernels and weight-sharing translation equivariance [3].

Inspired by [2, 38], the convolutional front-end module consists of two parts (shown in Figure 2). Each part starts with a 1D convolutional layer (1x3 padded convolution), then followed by a batch normalization and a rectified linear unit (ReLU).



Figure 2: Convolutional front-end structure.

We set the out-channels produced by the convolution to 8 and 16 respectively. After feeding the bio-signal outputted from the preprocessing phase, it will be converted into a feature matrix with $L \times 16$ dimensions.

3.3 Multi-signal Fusion

We considered the four physiological inputs (BVP, EDA, HR, and SKT) as a whole. It is crucial to leverage several physiological inputs as emotions are subjective feelings generated by a complex coordination of multiple neurophysiological systems [23], in other words, the emotional clues should be simultaneously reflected in multiple physiological signs.

Thus, in this module, we combined the four physiological signals together and treated them as an entire physiological embedding to represent the hidden affective information. Specifically,

$$P = [P_{bvp}, P_{eda}, P_{hr}, P_{skt}] \quad (1)$$

where P_{bvp} , P_{eda} , P_{hr} , and P_{skt} are feature representations of BVP, EDA, HR, and SKT retrieved from the convolutional front-end. We concatenated them over the time steps, and got the output with $L \times d_{model}$ dimensions (d_{model} is $64=16 \times 4$ in this case). Similar to natural language processing, here L can be regarded as the number

of words in one sentence, and d_{model} can be viewed as the size of the embedded word vector by *word2vec*.

3.4 Feature Extraction

We applied the conformer structure as feature extraction encoder to extract local and global associations within and across physiological signals.

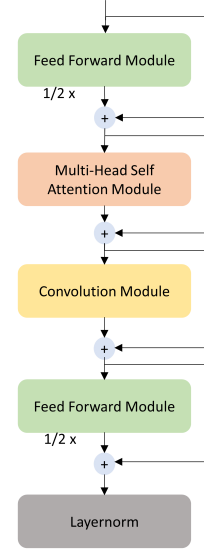


Figure 3: Conformer block structure [11].

The conformer mechanism presented by Gulati et al. [11] is a variant of the transformers. As shown in Figure 3, a conformer block consists of four modules, two feed forward modules, one multi-head self-attention module, and one convolution module. Firstly, the multi-head self-attention module uses a scaled dot-product attention mechanism to capture the dynamic global dependencies in the feature sequence [37]. It creates three vectors from each of the input feature vectors (the 64-dimensional embedding of each time step in this case): a query vector, a key vector, and a value vector. By computing the dot products of one query with all keys and applying a softmax function, it obtains the attention weights on all values, which determines how much focus to place on other parts of the input vectors as encoding one feature vector at a certain position. As defined by [38], given an input sequence $S = (S_1, S_2, \dots, S_L)$, the output sequence $\tilde{S}_h = (\tilde{S}_{h1}, \tilde{S}_{h2}, \dots, \tilde{S}_{hL})$ of the self-attention for a single head h can be computed by:

$$e_{ij} = \frac{(S_i W_q)(S_j W_k)^T}{\sqrt{d_k^h}}, j \in [1, L] \quad (2)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^L \exp(e_{ik})} \quad (3)$$

$$\tilde{S}_{hi} = \sum_{j=1}^L \alpha_{ij}(S_j W_v) \quad (4)$$

where $W_q, W_k \in \mathbb{R}^{d_{model} \times d_k^h}$ and $W_v \in \mathbb{R}^{d_{model} \times d_v^h}$ are learnable linear transformation matrices that generate the query, key, and value vectors, and $S_i \in \mathbb{R}^{d_{model}}$. The output of all heads are then concatenated together and condensed by another linear transformation:

$$\tilde{S} = \text{Concat}(\tilde{S}_1, \dots, \tilde{S}_n)W^o \quad (5)$$

where $W^o \in \mathbb{R}^{nd_v^h \times d_{model}}$ and n is the number of heads.

Secondly, the convolution module is responsible for enforcing locality, which contains a pointwise convolution with an expansion factor of 2, a gated linear unit (GLU) activation layer, a 1D depthwise convolution, a batch normalization, a swish activation layer, and another pointwise convolution projecting the number of channels back. By leveraging the channel-wise and spatial-wise learning capabilities, the convolution module is able to capture fine-grained local features, which largely complements the weakness of the self-attention mechanism. In this study, we concatenated features from four different physiological signals to form an entity. With the help of the convolution module, the system can better capture associations across signal-specific features.

For our feature encoder, we used 4 conformer blocks, and applied $n = 4$ parallel attention heads. For each head, we set $d_k^h = d_v^h = d_{model}/n = 16$. Results from feature extraction were flattened and passed into two dense layers to get the final representation for further decision making.

4 EVALUATION

4.1 Dataset

We first evaluated the performance of our proposed system on the K-EmoCon dataset [21]. We chose this particular dataset based on two criteria: (1) the dataset contains *spontaneous* emotional expressions in *naturalistic* environments or social interactions; (2) the data collection apparatus are *off-the-shelf*, *low-cost*, *mobile* and *wearable* devices. To the best of our knowledge, the K-EmoCon is the only publicly available multimodal affective dataset that is suitable for this experiment. Other widely used emotion datasets either consist of posed or induced emotions, or rely on expensive and intrusive data acquisition equipment that is not suitable for daily use. For example, the MAHNOB-HCI [33] and the DEAP [14] were considered to contain induced emotions and apply a number of intrusive electrodes [35]. Instead, the K-EmoCon dataset was collected via natural communication between individuals without professional training in acting. Specifically, the K-EmoCon contains two kinds of visual data (i.e., face and gesture), five kinds of physiological data (i.e., EEG, ECG, BVP, EDA, and SKT), speech audio, and accelerometer data recorded from 32 participants (20 males and 12 females). In our experiments, we only used four kinds of physiological signals captured by the Empatica E4 wristband, BVP, EDA, HR (which was derived from BVP), and SKT, since compared with chest-worn (used to detect ECG) and head-mounted (used to detect EEG) devices, a wristband is more ubiquitous and unobtrusive in everyday life and is without spatio-temporal limitations.

Moreover, the annotations for emotions in the K-EmoCon were measured on arousal and valence affective dimensions from the circumplex model of affect by James Russell [27]. Each affective

dimension was rated on five-point Likert scales. Accordingly, annotated emotion labels can be categorized into either five levels directly, or two levels based on the median value (i.e., 2.5) in which annotations with values ranging from 1 to 2 can be converted into *low* (L), and annotations with values ranging from 3 to 5 can be converted into *high* (H) [47]. To conduct a comprehensive assessment, we formalized the emotion recognition as five classification tasks: two-category arousal, two-category valence, five-category arousal, five-category valence, and four-category arousal-valence space (i.e., high arousal-high valence (HAHV), high arousal-low valence (HALV), low arousal-high valence (LAHV), and low arousal-low valence (LALV)).

Table 1: Sample distribution

	Two-class	Five-class	Four-class
Arousal		1: 104	
		2: 919	
	L: 1023	3: 1159	-
	H: 2157	4: 679	
		5: 319	
Valence		1: 69	
		2: 485	
	L: 554	3: 1815	-
	H: 2626	4: 728	
		5: 83	
Arousal-Valence			LALV: 172
			LAHV: 851
			HALV: 382
			HAHV: 1775

Specifically, we used the self-reported emotion labels as the ground truth. Since annotations in the K-EmoCon were labeled at every 5 seconds, the raw input size of BVP sequence, EDA sequence, HR sequence, and SKT sequence was respectively 1×320 , 1×20 , 1×5 , and 1×20 (BVP sampled at 64Hz, EDA at 4Hz, HR at 1Hz, and SKT at 4Hz), and $L = 320$ in this case. Additionally, the E4 data of 6 participants (Person IDs: 2, 3, 6, 7, 17, and 20) is missing due to the device malfunction or human error during data collection. Thus, the final evaluation dataset consists of 3180 data samples. Table 1 shows the sample distribution. We note that while there are imbalanced distributions of emotion labels in all five tasks, as stated in [21], *this imbalance is expected as emotion data is commonly imbalanced by its nature in the wild (i.e., people are more often neutral than angry or sad)*.

4.2 Baselines

To compare with the state-of-the-art physiological signals-based emotion recognition approaches, we chose the following methods as baselines: (1) Naive Bayes (NB) [47], (2) Support Vector Machine with linear kernel (SVM-LR) [35], (3) Support Vector Machine with radial basis function kernel (SVM-RBF) [35], (4) eXtreme Gradient Boosting (XGBoost) [47], (5) bidirectional LSTM (BiLSTM) [47]; and (6) transformer [2]. For the first four methods, we extracted

both the time-domain and frequency-domain features based on the proposed features in the literature [14].

4.3 Implementation Details

We implemented our proposed system using the PyTorch framework. We initialized the learning rate as 0.0001 and decayed the learning rate by $\gamma = 0.985$ every epoch. We used Adam optimizer, and applied batch normalization and dropout function to overcome overfitting and internal covariate shift [12].

To avoid data leakage, we did not use data re-sampling approaches [24] to solve the imbalanced data problem as they will change the original dataset itself. Instead, we chose focal loss [16] as the loss function of our model, which can automatically down-weight the contribution of easily classified samples and focus on hard misclassified samples by applying a modulating term to the cross-entropy loss. We used a 80–20 training and testing split, and further split 20 percent of the data from the training set as validation. We trained the model with 500 epochs and used early stopping with patience equal to 10. We also used 256 as the batch size and evaluated the model with 5-fold cross-validation. To make a fair comparison between the proposed system and baselines, we re-trained all models on the same training-testing set split.

4.4 Evaluation Metric

Considering that the classes of arousal and valence are heavily imbalanced, we did not use recognition accuracy as the evaluation metric like most studies. Instead, following [17, 24], we chose the average F1 score (Macro-F1) and unweighted average recall (UAR) as our validation metrics. These metrics give the same importance to each class, and are defined as the mean of class-wise F1 scores and recall scores respectively. Unlike accuracy or weighted F1 or weighted recall, their values will not be overwhelmed by the vast number of easily classified samples, which makes them suitable as the validation metrics of our experiment. Consider the prediction of two-class valence as an example. Its ground truth consists of 554 *low* and 2626 *high* samples. If there is a classifier that predicts *high* for all samples, it will have an accuracy of more than 82%, a weighted F1 score around 0.75, and a weighted recall score around 0.83. Despite having high values, we cannot say it is not a good classifier since it does not learn any informative features from the samples. Macro-F1 score and UAR can avoid this situation (only 0.45 and 0.50 in this example) because their values will be low if the model only performs well on the large number of common classes while performing poorly on the rare classes.

4.5 Results

The results of the emotion recognition from physiological signals on the K-EmoCon dataset are shown in Table 2. Our proposed system achieved better performance compared with all baseline approaches. For example, for five-class classification tasks, our model outperformed NB, SVM-LR, SVM-RBF, XGBoost, BiLSTM, and Transformer methods respectively by (1) 32.50%, 34.76%, 28.35%, 13.59%, 12.27%, 7.76% Macro-F1 increase, and 35.25%, 41.38%, 36.32%, 25.94%, 12.59%, 5.35% UAR increase on arousal; and (2) 26.17%, 30.57%, 27.42%, 7.76%, 6.52%, 3.78% Macro-F1 increase, and 43.91%,

Table 2: Comparison of emotion recognition performance using different approaches. A: arousal, V: valence, A-V: arousal-valence emotion space.

Approach	Task	Macro-F1(%)	UAR(%)
NB	Two-class A	36.16	53.78
	Two-class V	47.76	50.91
	Five-class A	23.43	30.28
	Five-class V	20.01	21.56
	Four-class A-V	23.54	31.96
SVM-LR	Two-class A	40.42	50.00
	Two-class V	46.48	50.56
	Five-class A	21.17	24.15
	Five-class V	15.61	20.53
	Four-class A-V	19.38	25.51
SVM-RBF	Two-class A	54.18	56.60
	Two-class V	45.94	50.30
	Five-class A	27.58	29.21
	Five-class V	18.76	22.01
	Four-class A-V	26.18	29.19
XGBoost	Two-class A	59.58	59.80
	Two-class V	54.88	54.79
	Five-class A	42.34	39.59
	Five-class V	38.42	35.74
	Four-class A-V	36.65	35.95
BiLSTM	Two-class A	74.50	76.53
	Two-class V	67.31	77.14
	Five-class A	43.66	52.94
	Five-class V	39.66	55.25
	Four-class A-V	47.33	56.34
Transformer	Two-class A	72.39	75.37
	Two-class V	60.36	73.11
	Five-class A	48.17	60.18
	Five-class V	42.40	62.54
	Four-class A-V	51.84	62.78
Our model	Two-class A	77.37	79.42
	Two-class V	69.11	80.60
	Five-class A	55.93	65.53
	Five-class V	46.18	65.47
	Four-class A-V	56.35	65.45

44.94%, 43.46%, 29.73%, 10.22%, 2.93% UAR increase on valence. Additionally, the recognition results were tested for statistical significance using a Wilcoxon signed-rank test, by comparing the Macro-F1 and UAR scores from different approaches over cross-validation folds. The test results show statistically significant improvements of our model over the baselines on all five tasks ($p < 0.05$) except two-class valence comparison on Macro-F1 with BiLSTM and five-class valence comparison on UAR with Transformer.

From the results, we can also observe that (1) BiLSTM performed slightly better than Transformer on 2-class tasks, while performing slightly worse than Transformer on multi-class tasks; (2) the major difference between our system and Transformer architecture is the conformer encoder, which further shows the importance of convolution modules that can effectively extract the local interactions

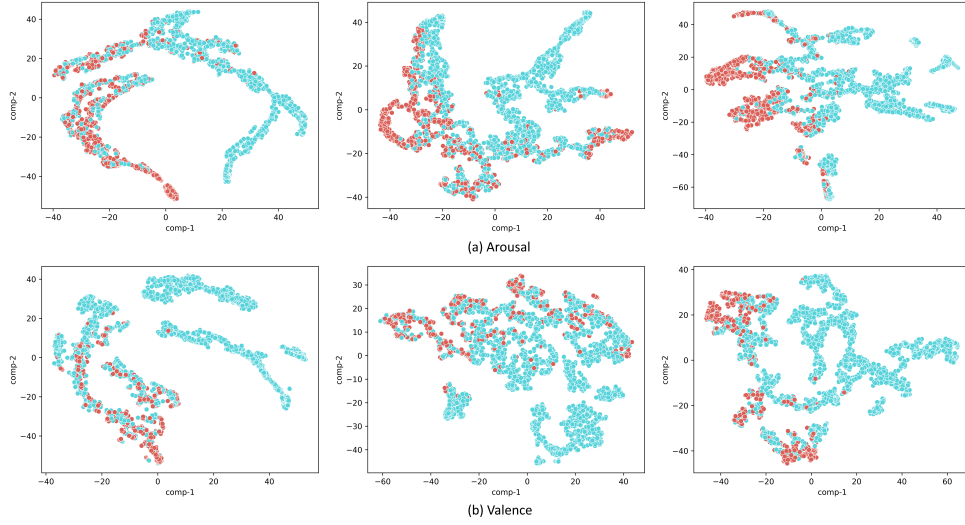


Figure 4: t-SNE visualisations of trained models for two-class tasks. Left is for the baseline BiLSTM approach; Middle is for baseline Transformer approach; Right is for our proposed model.

across different signal-specific features through the convolution mechanism. This also demonstrates the combination of convolutions with self-attention can improve the system’s performance.

We also presented a visualization of the extracted feature representation from the BiLSTM, the Transformer, and our model. For this, we used the t-distributed stochastic neighbor embedding (t-SNE) technique [36] which allows visualizing high-dimensional data in a two or three-dimensional map. Figure 4 shows the 2D t-SNE plots for two-class tasks.

As illustrated in the figure, the extracted representations of our proposed method are visibly more distinguishable by classes on both arousal and valence dimensions compared to the extracted representations from baseline methods. Moreover, this trend can also be found for other tasks, which further indicates the stronger ability of our model on learning the high-level and latent features and dependencies.

4.6 Generalizability Tests

To further test the generalizability of our proposed model, we performed additional experiments on another multimodal affective dataset, the ASCERTAIN dataset [35]. Despite emotional inducement being applied during the data collection (instead of relying on spontaneous affective responses like the K-EmoCon dataset), this dataset used wearable and commercial sensors for examining users’ physiological behavior, partially fulfilling the requirements of a more naturalistic and pervasive affective dataset. We designed two binary emotion recognition experiments to test our model’s generalizability.

As both the K-EmoCon and the ASCERTAIN datasets contain EDA sensor data, we first conducted a cross-datasets generalizability test. Specifically, we trained on the K-EmoCon dataset and then tested on the ASCERTAIN dataset. Considering that there

is only one input modality, we removed the BVP, HR, and SKT branches of our original structure and converted our model into a single modality emotion recognition system. Moreover, due to the different sampling frequencies (4Hz in the K-EmoCon versus 100 Hz in the ASCERTAIN), we used a downsampling method with an anti-aliasing filter to replace the interpolation in the preprocessing phase when testing the performance. On this basis, our model achieved 64.37% accuracy on arousal dimension and 61.88% accuracy on valence dimension, which is comparable to the unimodal performance of previous work that conducted both training and testing on the same ASCERTAIN dataset [45]. This result demonstrates that our proposed model is able to learn generalized and transferable physiological features rather than overfitting the data distribution of one particular dataset.

Table 3: Performance comparison on the ASCERTAIN dataset (accuracy in percentage)

Approach	Arousal	Valence
VM2HL-P [45]	72.54	68.53
Our model	71.79	69.11

We then followed-up with a second experiment to test the generalizability of our proposed model on other kinds of physiological signals. Considering the diversity of commercial physiological sensors and their recorded signals, our system should have good expansibility for different input modalities. Therefore, we further replaced the original input modalities of our model with the ECG, GSR, and EEG modalities from the ASCERTAIN dataset, and re-trained our model from scratch. Following the work by Zhao et al. [45], we re-trained our model under the same implementation details and applied the same evaluation metric. As can be seen in

Table 3, our model achieved 71.79% recognition accuracy on arousal and 69.11% accuracy on valence, which is equivalent to a 0.75% accuracy decrease and a 0.58% improvement respectively compared to the system proposed by Zhao et al. [45]. Despite leading to a 0.58% accuracy decrease on arousal dimension, it is important to note that we only used three modalities (we excluded facial landmarks because our work focused on physiological signals), while the system from the literature used all four modalities [45]. In other words, our system can achieve competitive performance with less data input, which is significant in real-world scenarios due to the common data missing problem [32].

5 DISCUSSION

With the rapid development of smart mobile and wearable devices, there has been a growing interest in ecological validity [13] and real-world application of emotion recognition techniques [25]. In this paper, we presented a novel emotion recognition system based on physiological signals from low-cost, off-the-shelf mobile and wearable devices, and evaluated its performance with spontaneous and natural emotion data collected in natural interactions.

It is worth noting that our proposed system has good generalizability in terms of different physiological signals. The system can efficiently extract high-level and generalized affective features from physiological signals, and perform well on spontaneous emotion recognition. Considering that smart mobile and wearable devices are increasingly sensor-rich, less intrusive, and becoming an essential and integral part of daily life, we can expect to see an increase in the need for robust mobile emotion recognition systems.

5.1 Towards Robust Multimodal Mobile Emotion Recognition

It is undisputed that human emotions are expressed through a multitude of signals. Thus, as Delplanque and Sander [5] argue, it is risky to rely on a single component to infer true emotional states. For example, while fear might be correlated to heart rate changes, and smiling can be an expression of happiness, neither signal can be used as a marker for one emotion, as the relation between emotions and such signals is "many-to-one" [5]. A smile can also be used to communicate non-emotional messages [1]. Consequently, as each of these signals can be influenced by numerous non-emotional phenomena, it is important that automated emotion detection systems rely on multiple signals directly related to expressions of emotion. However, it is a complicated task for computers to disentangle the multitude of signals that define emotion. Our proposed detection model was validated using a labeled dataset that contains multiple streams of physiological data recorded in naturalistic settings and outperformed different models.

While automated emotion detection and recognition promise to contribute to support human well-being and mental health, there are discussions about the potential risks and ethical concerns of this type of data, e.g., in medicine [34], where automated emotion detection has been used for mental health assessments. The criticism focuses mainly on the validity, the training, and ethical scaffolding necessary for a qualified assessment of the produced outputs [34]. Furthermore, automated pervasive sensing can raise serious privacy concerns, including the potential misuse of data, e.g., in the

workplace, where the detection of boredom by the employer could potentially lead to increased workload and stress for employees.

In comparison to vocal and facial features - while not impossible - it is more difficult to use physiological signals, such as the ones used in this study, to identify individuals, especially in naturalistic settings [9, 42]. Nevertheless, it is crucial to develop systems that process data in-situ without requiring the data to be sent through networks and cloud-services, as leaks of any kind of physiological or biometric data can put users' security and privacy at risk. Our system has the potential to run entirely on smartphones, whose processing power and memory are continuously increasing. Furthermore, a phone-based system can more easily give the user control over which specific data are being collected at any moment in time.

5.2 Limitations

Our work has several limitations. First, the amount of data samples in the final evaluation dataset is relatively small, which, to a certain extent, limited the learning ability of our proposed model. Compared with some well-known larger-scale datasets (e.g., ImageNet [6], LibriSpeech [20]) used in computer vision and speech recognition, our evaluation dataset only consists of 3180 samples. There is a need in emotion recognition research for larger natural and spontaneous emotion datasets based on commercial mobile and wearable devices. Second, we only considered the dimensional model (i.e., arousal and valence) to describe human emotions available in the used datasets. Further work is needed to test our approach using the discrete emotion model (i.e., anger, disgust, fear, happiness, sadness, and surprise). Third, we did not implement our system in a real-time field deployment as we first aimed to test its robustness and usefulness using publicly datasets. In the future, we aim to develop a mobile application version and install our system on users' personal devices to conduct a long-term user study.

6 CONCLUSION

In this paper, we propose a novel emotion recognition system based on multimodal physiological signals from off-the-shelf, low-cost mobile and wearable devices, which offers greater data privacy and the potential to be deployed into real-world contexts. Our system uses a convolutional front-end to embed each physiological signal, and employs the conformer structure to extract the local and global dependencies within and across different signals. We evaluated our system on a natural and spontaneous emotion dataset acquired with off-the-shelf devices. Our results show that our system outperformed the baselines, and achieved state-of-art or competitive performance. In addition, our generalizability experiment further established that our system has the ability to scale and handle different sensor signals. We discuss the potential of our system to be deployed into real-world scenarios for daily emotion monitoring and management. We also reflect on the potential risks and concerns related to pervasive emotion recognition approaches.

7 ACKNOWLEDGMENTS

This work was supported by an Australian Research Council Discovery Project (DP190102627).

REFERENCES

- [1] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.
- [2] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. 2021. A Transformer Architecture for Stress Detection from ECG. In *2021 International Symposium on Wearable Computers*. 132–134.
- [3] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. 2019. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3286–3295.
- [4] Ira Cohen, Ashutosh Garg, Thomas S Huang, et al. 2000. Emotion recognition from facial expressions using multilevel HMM. In *Neural information processing systems*, Vol. 2. Citeseer.
- [5] Sylvain Delplanque and David Sander. 2021. A fascinating but risky case of reverse inference: From measures to emotions! *Food Quality and Preference* November 2020 (2021), 104183.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.
- [7] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2019. Laughter recognition using non-invasive wearable devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 262–271.
- [8] Sidney K D mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *Comput. Surveys* 47, 3 (2015), 1–36.
- [9] Jorge Goncalves, Pratyush Pandab, Denzil Ferreira, Mohammad Ghahramani, Guoying Zhao, and Vassilis Kostakos. 2014. Projective Testing of Diurnal Collective Emotion. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. 487–497.
- [10] Hector A Gonzalez, Shahzad Muzaffar, Jerald Yoo, and Ibrahim M Elfadel. 2020. BioCNN: A hardware inference engine for EEG-based emotion detection. *IEEE Access* 8 (2020), 140896–140914.
- [11] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).
- [12] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.
- [13] John F Kihlstrom. 2021. Ecological validity and “ecological validity”. *Perspectives on Psychological Science* 16, 2 (2021), 466–471.
- [14] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.
- [15] Azadeh Kushki, Jillian Fairley, Satyam Merja, Gillian King, and Tom Chau. 2011. Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical sites. *Physiological measurement* 32, 10 (2011), 1529.
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.
- [17] Terrance Liu, Paul Pu Liang, Michal Muszynski, Ryo Ishii, David Brent, Randy Auerbach, Nicholas Allen, and Louis-Philippe Morency. 2020. Multimodal privacy-preserving mood prediction from mobile data: A preliminary study. *arXiv preprint arXiv:2012.02359* (2020).
- [18] Steven Marwaha, Matthew R Broome, Paul E Bebbington, Elizabeth Kuipers, and Daniel Freeman. 2014. Mood instability and psychosis: analyses of British national survey data. *Schizophrenia bulletin* 40, 2 (2014), 269–277.
- [19] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech communication* 41, 4 (2003), 603–623.
- [20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.
- [21] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 1 (2020), 1–16.
- [22] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.
- [23] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.
- [24] Jingyu Quan, Yoshihiro Miyake, and Takayuki Nozawa. 2021. Incorporating Interpersonal Synchronization Features for Automatic Emotion Recognition from Visual and Audio Data during Communication. *Sensors* 21, 16 (2021), 5317.
- [25] Mika Raento, Antti Oulasvirta, and Nathan Eagle. 2009. Smartphones: An emerging tool for social scientists. *Sociological methods & research* 37, 3 (2009), 426–454.
- [26] Erika L Rosenberg and Paul Ekman. 1994. Coherence between expressive and experiential systems in emotion. *Cognition & Emotion* 8, 3 (1994), 201–229.
- [27] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
- [28] Sriparna Saha, Shreyasi Datta, Amit Konar, and Ramadoss Janarthanan. 2014. A study on emotion recognition from body gestures using Kinect sensor. In *2014 International Conference on Communication and Signal Processing*. IEEE, 056–060.
- [29] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, and Jorge Goncalves. 2020. Does Smartphone Use Drive Our Emotions or Vice Versa? A Causal Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [30] Elaine Sedenberg and John Chuang. 2017. Smile for the camera: privacy and policy implications of emotion AI. *arXiv preprint arXiv:1709.00396* (2017).
- [31] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.
- [32] Yangyang Shu and Shangfei Wang. 2017. Emotion recognition through integrating EEG and peripheral signals. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2871–2875.
- [33] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (2011), 42–55.
- [34] Isabel Straw. 2021. Ethical implications of emotion mining in medicine. *Health Policy and Technology* 10, 1 (2021), 191–195.
- [35] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. 2016. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* 9, 2 (2016), 147–160.
- [36] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
- [38] Bin Wang, Chang Liu, Chuanyan Hu, Xudong Liu, and Jun Cao. 2021. Arrhythmia Classification with Heartbeat-Aware Transformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1025–1029.
- [39] Yi Wang, Zhiyi Huang, Brendan McCane, and Phoebe Neo. 2018. EmotioNet: A 3-D Convolutional Neural Network for EEG-based Emotion Recognition. In *2018 International Joint Conference on Neural Networks (IJCNN)*. 1–7. <https://doi.org/10.1109/IJCNN.2018.8489715>
- [40] Zhu Wang, Zhiwen Yu, Bobo Zhao, Bin Guo, Chao Chen, and Zhiyong Yu. 2020. EmotionSense: An Adaptive Emotion Recognition System Based on Wearable Smart Devices. *ACM Transactions on Computing for Healthcare* 1, 4 (2020), 1–17.
- [41] Tianyuan Xu, Ruixiang Yin, Lin Shu, and Xiangmin Xu. 2019. Emotion recognition using frontal eeg in vr affective scenes. In *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, Vol. 1. IEEE, 1–4.
- [42] Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Behavioral and Physiological Signals-Based Deep Multimodal Approach for Mobile Emotion Recognition. *IEEE Transactions on Affective Computing* (2021), 1–17.
- [43] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. 2021. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer* 37 (2021), 1447–1466.
- [44] Bobo Zhao, Zhu Wang, Zhiwen Yu, and Bin Guo. 2018. EmotionSense: Emotion recognition based on wearable wristband. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*. IEEE, 346–355.
- [45] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. 2018. Personality-Aware Personalized Emotion Recognition from Physiological Signals. In *IJCAI*. 1660–1667.
- [46] Junjie Zhu, Yuxuan Wei, Yifan Feng, Xibin Zhao, and Yue Gao. 2019. Physiological Signals-based Emotion Recognition via High-order Correlation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–18.
- [47] M Sami Zitouni, Cheul Young Park, Uichin Lee, Leontios Hadjileontiadis, and Ahsan Khandoker. 2021. Arousal-Valence Classification from Peripheral Physiological Signals Using Long Short-Term Memory Networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 686–689.

Chapter 7

Discussion

In this chapter, we reflect on the research contributions presented in Chapters 2, 4, 5, and 6. On this basis, we answer the proposed research questions defined in Chapter 1. In addition, we detail current open challenges and potential future directions in the thriving field of mobile emotion recognition.

7.1 Mobile Emotion Sensing Systems

RQ1: What are the recent advances on emotion recognition using mobile devices?

We aim to understand the recent successful methods adopted in mobile emotion sensing research and discuss trends in this field. To this end, we divide a typical affect detector for mobile and wearable devices into three stages: perception, learning, and inference. We categorize different studies and compare them with regards to these three aspects.

In Chapter 2, we show that in terms of the selection of input modalities, there are eight kinds of commonly used data sources in mobile emotion sensing: facial expression, ocular sign, speech, typing & touch, body movement & activity, physiological signals, device usage, and environment. Among them, facial expression, physiological signals, and body movement & activity are the most commonly used. Moreover, we also notice that despite the well-known limitations of unimodal systems [5, 26], the majority of recent state-of-the-art frameworks still rely on handling a single modality.

With regard to feature engineering, both manual extraction and automation extraction have been used. Early work normally used domain knowledge accompanied by functional statistics to manually extract low-level handcrafted features. However, handcrafted features are simply not expressive enough for complicated interrelationships [287]. Recently, with the rapid development of deep learning, various deep learning structures such as CNNs and RNNs were introduced in the mobile emotion sensing domain to automatically extract higher-level and finer-grained features. Overall, we find that handcrafted extractors are still the current mainstream of feature engineering, since deep learning feature extractors typically require significant computational power and memory resources. However, with the progress of wearable computing and wireless sensing technologies, this bottleneck is starting to dissipate.

Similarly, in order to infer final interpretations (i.e., human emotions) from abstract feature representations, there are two main approaches: machine learning approaches and deep learning approaches. Among the reviewed research, we find that machine learning approaches are the most widely used due to some benefits like less training time and lower hardware requirements. However, with the rapid development of cloud computing, the cloud can provide model training services at the back end, while mobile and wearable devices can be used as terminals to upload collected data sources and display final results. Recently, there is an emerging trend to leverage deep learning models in the field of mobile emotion sensing for inference.

In summary, through the literature review, we find that the recent research trend in the field of mobile emotion sensing is starting to lean toward multimodal learning. There has also been a growing interest in using deep learning approaches and multimedia architectures to incorporate into mobile and wearable devices. This also forms the basis of our effort to develop and improve multimodal mobile emotion recognition systems, which we describe in Chapter 5 and 6.

7.2 Benchmarking Commercial Mobile Emotion Recognition Systems

RQ2: What are the limitations of current off-the-shelf commercial mobile emotion recognition systems that leverage facial features?

While conducting the systematic review of emotion sensing research using mobile and wearable devices, we notice the advent of several commercial cloud-based services encapsulating and hiding the hard part of emotion recognition and offering simple-to-use APIs or SDKs [288]. For example, these include Amazon Rekognition, Face++, Microsoft Azure, etc. This makes it easier for developers and researchers to embed sophisticated emotion recognition functionality in applications.

With respect to these cloud-based commodity services, different data sources have been employed for training and establishing the relationship with human emotions, such as facial expression-based systems, speech-based systems, etc. Despite the wide variety of products, facial emotion recognition is well supported by most services and is easier to be integrated into everyday devices [288]. Thus, to answer the second research question, we evaluate the performance of commercial intelligent systems using five popular cloud-based facial emotion recognition services, in [Article I](#).

Prior validation and comparison research has only considered ideal conditions like full-face images [289–291], without taking into account the realistic non-ideal conditions like partially visible facial images or other distorted facial images. Thus, in [Chapter 4](#), we give full consideration to in-the-wild user behaviours and possible image distortions, and design an evaluation method to visualize the application potential of these systems in real-world scenarios.

We show that these intelligent systems all experienced varying degrees of performance degradation when processing simulated in-the-wild photos. Importantly, we find that these cloud-based emotion recognition systems have different performance and biases. For example, Baidu could tolerate lower-level image blurring, while Amazon and Face++ could be in favor of higher-level image blurring. This is most likely due to class imbalance in the datasets used to train these emotion recognition systems. Informed by our

findings, we recommend applying a multi-system approach and building a routing protocol to adaptively select a more appropriate system to deal with different real-world scenarios. On the other hand, these findings also demonstrate the limitations of unimodal affect detection, which, in contrast, multimodal approaches can easily overcome by flexibly basing decisions on more reliable channels. This also forms the motivation of our studies described in Chapter 5 and 6.

Meanwhile, we acknowledge that these commercial emotion recognition systems have regular updates and changes to their algorithms, which probably lead to better assessment outcomes. Therefore, our main contribution lays on the proposed critical evaluation method rather than the specific performance and limitations of these emotion recognition systems at the time of our study. In addition, researchers can also reuse our method to build their own validation datasets and better embed emotion recognition capabilities in their applications.

7.3 Developing Multimodal Mobile Emotion Recognition Systems

RQ3: How can we take full advantage of the capabilities of mobile devices to develop robust and accurate emotion recognition systems?

The superior performance of multimodal systems has been well established in the literature [5, 26]. Through RQ3, we aim to exploit deep learning approaches as well as multimodal architecture to develop more robust and accurate emotion recognition systems applied to mobile devices. We hope that our studies could further promote the development in the mobile affective computing area. In Chapter 5 and 6, we develop two deep multimodal mobile systems to predict human emotions based on large-scale multimedia data.

First, in Chapter 5, we investigate the feasibility of using off-the-shelf commodity smartphones and wristbands to detect human emotions. Our proposed system yields promising results with 89.2% average accuracy through analyzing behavioral and physiological

signals collected from a smartphone and a wristband. Particularly, we present and evaluate the impact of different modality combinations including one-modality input, two-modality input, three-modality input, and four-modality input, as well as smartphone-only scenario, wristband-only scenario, and wristband + smartphone scenario. We show that there is a positive correlation between performance and the number of input modalities. As expected, with more modality inputs, our system yielded higher accuracy. Our evidence strengthens the notion that multimedia data fusion can increase the accuracy and reliability of the estimates [292]. In addition, we also explain why we adopt deep learning structures, and how we use the attention mechanism to design the feature extractor on different signal branches and build modality-specific feature representation. Earlier attempts to use traditional statistical analysis methods and machine learning techniques for feature extraction and selection were unable to efficiently extract complex and nonlinear patterns in multimodal time-series signals. This makes it difficult to learn invariant characteristics in the data and establish abstract representations of the sensor data [223].

In Chapter 6, we make further improvements and introduce a more efficient hierarchical mobile emotion recognition system. We report significant performance improvements in predicting spontaneous human emotions when compared to other state-of-the-art emotion recognition approaches. Our experimental results demonstrate the constructed computing model is able to establish the complex mappings between naturalistic affective expressions and multimodal physiological signals. More importantly, in order to enhance the network generalization ability and improve training efficiency, we narrow our sights on wristbands, as they offer out-of-the-box support for unobtrusively collecting and monitoring multiple homogeneous body signals for long periods of time [293]. Given the involuntary activation of physiological signals, they cannot be changed or controlled by users consciously in most cases. As a result, the accuracy and reliability of this mode are more guaranteed [294, 295]. Additionally, to include a more robust feature extraction capability, we employ an advanced Conformer mechanism. Through generalizability tests and visualisation experiments, our analysis demonstrates that our system has the ability to efficiently extract high-level and distinguishable affective features. We envision a future where affect detectors can be integrated into everyday devices and make their way into our everyday lives for mental health and well-being management.

7.4 Open Challenges

7.4.1 Modality Choice

Currently, a variety of sensing modalities have been explored in mobile emotion sensing research. In Chapter 2, we summarize the different modalities that have been leveraged, finding that mobile emotion sensing research has evolved from traditional unimodal analyses to more complex multimodal analyses. The motivation behind this shift is that multimodal emotion sensing systems have significant accuracy and reliability advantages and offer wider application prospects. Not only are they more in line with the multimodal nature of expressions of emotions [54, 296], they are also more likely to lead to more reliable outcomes due to the presence of more (types of) data. The next frontier in emotion research clearly is the robust sensing of emotions in naturalistic settings. However, incomplete or missing data is an unavoidable problem when collecting emotion data, especially in uncontrolled settings. For example, visual data (e.g., facial expression, body movement & activity) is sometimes unavailable or untraceable due to inappropriate viewing angles, uncontrollable lighting, or camera occlusion [278]. In such scenarios, a purely vision-based system is virtually useless, whereas a multimodal system can mitigate such situations by temporarily basing its predictions on other available signals.

In a real-world deployment, however, the implementation of an extensive multimodal system may be restricted by the commercial availability of suitable mobile and wearable devices. The number of integrated sensors in a single device can limit multimodal approaches, forcing the use of multiple devices to collect a wider range of emotional response signals. Using several signal acquisition devices (e.g., a smartwatch in combination with a smartphone) is not always possible, although such scenarios are increasingly more commonplace in everyday life. Furthermore, the collection of more signals can also lead to higher computational costs, particularly if using multiple high-frequency sensor data. Ultimately, researchers and practitioners need to consider the trade-off between practical application and high-precision performance when choosing which and how many signals to collect.

Furthermore, there is no clear consensus regarding how to choose more appropriate modality combinations in different scenarios. In [Article II](#), we explore the impact of different combinations of four kinds of modalities (including facial expression, speech, typing & touch, and physiology), and report some preliminary conclusions. For example, for a scenario with three modalities, we show that the best performing combination for predicting positive emotion is *facial expression + typing & touch + physiology* or *speech + typing & touch + physiology*, and for predicting negative emotion is *facial expression + speech + typing & touch*. However, broad recommendations like these can only be followed when the necessary signals are available with sufficient quality. Furthermore, the diversity and complexity of real-world contexts (e.g., outdoor noisy environments with different types of noise and a wide range of signal-to-noise ratios in which it is challenging to collect clear speech signals) also means that modality choice needs to be dynamic. Further investigation is needed to provide a more comprehensive analysis of this problem.

7.4.2 Ground Truth Establishment

The establishment of ground truth also plays a crucial role in (mobile) emotion sensing research. In summary, previous work has highlighted three main types of ground truth [\[297\]](#):

1. Experienced emotion: felt and reported by the person himself/herself.
2. Perceived emotion: perceived by either receivers or third-person observers.
3. Expected emotion: expected due to the nature of the applied stimuli, or simulated by actors.

Experienced emotion seeks to assess subjective emotions by relying on the individual's own assessment. Survey-based instruments such as the Day Reconstruction Method (DRM) [\[298\]](#) and the Experience Sampling Method (ESM) [\[91, 299\]](#) are commonly used in mobile emotion research. These techniques typically entail asking participants to self-report their emotional experiences and states on smartphones periodically each day, or to recall how they felt at different times throughout the day [\[90, 91, 94\]](#). However,

self-report is prone to inaccuracy due to inter-individual differences, such as the inconsistent interpretation of the nomenclature, or the incongruity between the apparent and experienced emotion [297]. Furthermore, what and when to ask participants for their self-assessment is also a challenge, especially when conducting long-term mobile experiments outside the lab or exploring a wider range of emotions [13].

Perceived emotion determines ground truth based on annotations from a receiver who interacts with the subject and experiences the subject's emotions secondhand, or external observers who observe the subject's emotions and label the emotion by reaching an agreement. A significant difference between receivers and external observers is the contextual knowledge that receivers have but observers do not [283]. Annotation relies on manual labeling by specific people, which is restricted to lab-based studies in most cases, as observations are potentially obtrusive and infeasible in the wild. Standard arrangements include designing specific social interactions or dialogue scenes that engage participants; emotion annotations then are obtained from the receivers or external observers. Despite the fact that an annotation strategy is in accord with the expressive nature of human emotions, these indirect assessments do not necessarily represent the true emotions of a subject, particularly since the subject may sometimes disguise their emotions intentionally.

Expected emotion uses a predefined label determined by the nature of the applied stimuli or by the simulated emotional expressions from actors or non-actors. Although this strategy is simple to use without requiring additional manual annotation, it is also restricted to lab-based experiments and may not always produce reliable outcomes. First, human emotions are a highly subjective phenomenon and are sensitive to past experience [300]. This means different people are likely to generate different levels or even different classes of emotions in response to the same stimuli. Second, the emotional expressions deliberately performed by actors or non-actors are instances of posed emotion, which are far from spontaneous expressions of emotions. Previous work has shown a dramatic drop in performance when systems trained on posed expressions process spontaneous emotional expressions [95, 301]. Thus, we caution against assigning the same emotion labels to different people based on this approach.

Constantine and Hajj [297] state that using a combination of experienced and perceived

emotion is generally a better practice. However, perceived emotion annotations are challenging to obtain in naturalistic settings, considering the inherently limited control over conditions outside the lab. A logical consequence is to use survey-based instruments on mobile devices to measure experienced emotion (e.g., ESM), and to combine this with (remote) sensing. For instance, researchers can install software on participants' mobile phones to schedule ESM surveys in a pseudo-random manner based on the duration of the physical activity window and prompt them to complete the questionnaires through notifications; at the same time, mobile or wearable devices can capture a continuous stream of sensor data to build the prediction models (which can be pretrained on some public ground-truth datasets of spontaneous user emotions [283, 302]) and cross-referenced with the self-assessments. In addition, semi-supervised learning can also be a potential solution. For example, we can first collect and build a large-scale unlabeled emotion dataset. Then, through self-supervised contrastive learning [277, 303, 304], we can teach computers the necessary priors, rather than predefined inductive bias, to model the affective world. This way, only a small amount of manual annotation (which can be done by the users) is required to enable the system to handle downstream user-dependent emotion recognition tasks.

In [Article II](#) and [Article III](#), we adopted the experienced emotion as ground truth, since our aim is to develop personalized mobile emotion recognition system and hopefully benefit self mental health and well-being management.

7.4.3 Real Environment

Prior research has mostly been conducted in controlled lab environments sampling populations of young people (e.g., university students). Generalizations to real-world situations with a more diverse population, thus, require further research. Furthermore, when deploying mobile sensing solutions in the real world, there are some technical limitations (e.g., battery life) as well as other uncontrollable factors (e.g., environmental impacts), that require the deployed systems to be lightweight in order to provide reliable sensing results.

While controlled environments allow for relatively straightforward emotion elicitation, detection, and data analysis, in-the-wild studies pose a series of challenges [305]. On the one hand, mobile devices have to be computationally powerful enough to process

an ideally multimodal stream of data, in order to have alternative sources in case of interruptions in data collection. On the other hand, the use of processing power for research should not interfere with the regular usage of devices, e.g. by draining the battery too rapidly or slowing down the processor. Furthermore, in laboratory studies, it is relatively easy to identify sources and reasons for noisy data recording, whereas everyday settings come with a wide range of potential confounds. As mentioned before, emotions are expressed in a multitude of ways (e.g., through facial expressions, behavior, physiology, etc.), which makes it complicated to identify truly faulty or unnecessary data in real-world environments. Noisy data potentially contains vital information for accurately detecting emotions, thus, should not simply be deleted or ignored. Multimodal data streams potentially solve this issue, if systems are sufficiently “smart” to either filter the noise out of a signal, or substitute a noisy or faulty signal with an alternative that leads to a competitive result, while not significantly increasing the computational demand.

Another major challenge for real-world emotion sensing is the quest for ground truth. As we have seen above, there are three major approaches to collecting ground truth data when studying emotions. While ESMs and diary methods provide robust ground truth data, they require an interruption of participants’ activities, and rely on participants’ compliance. The burden on the user is relatively high and regular interruptions may distort their natural emotion experience. Moreover, research by van Berkel et al. [13] has shown that after three weeks of ESMs, compliance rates significantly decrease. Lack of compliance results in less accurate predictions, which consequently will require a longer, even more burdensome ground truth collection period. Emotions are highly subjective phenomena which not only involve responses to in-situ experiences but are influenced by past experiences.

7.4.4 Ethics, Privacy & Security

While the collection and processing of multimodal data streams using mobile devices promise to improve real-world emotion detection approaches and attenuate shortcomings of existing methods, it also poses a risk to user privacy and data security. The collection, (temporal) storage, and processing of emotion-related data directly via the users’ mobile devices offers a relatively secure setup, compared to external processing systems in which

sensitive data has to be transmitted through networks, shared with cloud services and other devices; there, the risk of an infected device, or malicious access to personal data is high [306]. Nevertheless, mobile devices are no fortresses protected from attacks; their limited processing power and energy supply render them incapable of continuously running powerful security software. It is crucial that systems are able to analyse data in situ, avoid potential leaks of biometric or highly sensitive data through connections with public networks, and minimize the storage times of such data, decreasing the risk that leaked data will provide a comprehensive image of the user.

The recent call for using ubiquitous technologies in the digital health domain [307], raises new questions regarding user privacy, security, and ethics. In the field of medicine, automated emotion detection has been used to assess users' mental health, which is highly sensitive information [308]. While a user may opt in for having their data collected to provide insights into their mental and physical health, potentially large amounts of rich data are shared with more or less serious scientific services. Simons et al. [309] have shown that many applications marketed as helpful training apps for improving brain functionality (e.g., memory) are in fact not effective, while collecting large amounts of data. Therefore, critical validation and scientific proof for the effectiveness of digital health applications, especially in the consumer market, have to be provided to protect users from potentially submitting their sensitive data to dubious services. Full automation of these systems comes with the potential risk of users losing control and knowledge of which data is collected and processed. While sensing technologies are specifically contested because of concerns about their scientific validity [31], Straw [308] also mentions that training and a better understanding of ethical implications are necessary for users to correctly estimate the potential impact of the data they provide. While a user may sign up for a smart speaker service, collecting speech samples from users to analyze their emotional state in-situ, can potentially lead to whole conversations being recorded and analyzed. This describes a clear invasion of the user's privacy beyond the desired service [310]. This has important implications for users' private and working lives. For example, if an automated system reports boredom in a workplace, employees might be overloaded with tasks and have recovery time shortened. Other domains under scrutiny include employee hiring, insurance assessments, and student evaluation¹. Additionally,

¹https://ainowinstitute.org/AI_Now_2019_Report.pdf, last accessed February 21, 2022

devices that are used to collect personal data, e.g., smart speakers, smartphones, or wearable sensors, have to be secured from external attacks, and data has to be encrypted in order to guarantee that it is not accessible by a third party. Moreover, regulations have to be put in place for times where data accidentally reaches a third party, e.g., what happens when a conversation about unfriendly work environments accidentally reaches an employer. Another example presented by Slavich et al. [310], describes a scenario where businesses have access to in-situ emotional data, e.g., from a stressed smartphone user. This information can be used to advertise specific medications, services, or products to vulnerable people. It is unclear when it is appropriate to use such information for commercial purposes and when it is not appropriate, as some users might opt in, specifically for these services.

Moreover, despite a large number of studies on emotion recognition, most of them have focused on audio-visual signals [5] (e.g., facial expressions and speech), since they are seen as the most direct channels for human emotional expressions. However, these outward emotional signals can easily be modified or suppressed in social settings, reducing their reliability as signs of inner emotional feelings [84, 295]. In addition, recent work has argued that emotional facial expressions vary across races and cultures, and facial configurations do not act as a reliable signal for particular emotional states [31]. Furthermore, as faces and voices are directly linked to identity, these can raise significant privacy concerns, especially if they have to be uploaded to the cloud for processing [311].

Emotion detection systems, however, cannot simply rely on a single modality in order to constrain the impact of potential security breaches, because, as discussed above, these do not provide sufficient information to detect emotion reliably, e.g, not every smile is necessarily an expression of joy [31]. Therefore, emotion detection systems have to be able to detect which modalities complement each other and enable reliable emotion detection, while limiting the collected data to a minimum. Delplanque and Sander [312] have shown that in order to clearly identify a person from emotional data alone, a multitude of such data would be necessary.

Chapter 8

Conclusions and Future Directions

8.1 Summary of Contributions

This thesis investigates the development and improvement of multimodal mobile emotion recognition systems. We survey the recent advances in technologies in the field of mobile emotion recognition, and highlight two important research trends - multimodal architecture and deep learning application. The thesis then discusses the capabilities and limitations of current off-the-shelf commercial mobile emotion recognition systems. To this end, we present a critical assessment framework that can evaluate and visualize the application potential of relevant systems in real-world scenarios. Our evaluation experiments show that a single-system and single-modal approach is not sufficient to tackle real-world distortions. All tested intelligent systems met with degrading performance when dealing with real-world scenarios.

Next, the thesis presents two mobile emotion recognition systems that leverage the advantages of multimodal learning and deep learning. The first system fuses multiple behavioral and physiological signals collected from off-the-shelf smartphones and wearable wristbands. By introducing an attention-based LSTM structure, the system can extract the informative feature from different data branches. We further improve the model by introducing a MLP network to learn the cross-modality correlations and better fuse different modalities. Results from our evaluation experiments demonstrate that the

proposed system achieves high accuracy in predicting positive and negative spontaneous emotional states.

The second proposed system utilizes multiple homogeneous body signals collected from a commercial wristband as inputs. We further improve the fusion by employing the Conformer mechanism to help the system simultaneously capture local and global dependencies among signals. We also design a convolutional front-end module with the aim to extract more comprehensive information from different subspaces. The results show that the proposed system outperforms state-of-the-art approaches and has good generalization performance and promotion of capacity. We argue that our work has the potential to be deployed into real-world scenarios for daily emotion monitoring and well-being management.

8.2 Future Directions

Firstly, most current mobile emotion sensing systems are networked and use cloud services. Specifically, there are mainly two modes, online mode and semi-offline mode. For online mode, mobile and wearable devices are used as terminals, while the cloud provides learning and inference services in the back-end. Multimodal signals are first collected by different mobile and wearable devices, and then uploaded to the cloud when the Internet or Wifi is available. After training in the cloud, the final results are received by the corresponding devices to display or store. Despite yielding promising performance [88, 313], the need for a high-speed and stable network link constrains development at the present stage. Moreover, outsourcing mental health-related tasks to remote servers may put users' sensitive data at risk. For semi-offline mode, the cloud services only involve partial computing or an initial training task. After that, the trained models and parameters will be downloaded and used to conduct inference tasks locally and continue updating on the local devices to ensure privacy [314]. Recently, with the rapid development of portable mobile devices, modern mobile devices with multi-core CPUs and gigabytes of memory are having increasingly processing capabilities. This creates new opportunities to deploy the affect detectors on local devices in a full-offline mode without the need for networks and cloud services, which could greatly reduce response time, communication costs, cloud burdens, and better protect users' privacy. Although there is some preliminary work in this area [73, 249], it is still challenging to directly translate sophisticated

emotion sensing models to mobile platforms, especially for current deep learning-based models.

Secondly, deep learning approaches have evolved rapidly and been applied to a variety of different application domains, including but not limited to computer vision, natural language processing, speech recognition, and machine translation. These approaches have also been increasingly used in the affective computing field, and a large number of new architectures have been put forward recently. However, most of these advanced models are proposed and developed for desktop computers or desktop GPUs and are not applicable to a variety of mobile and wearable devices or mobile GPUs. One of the main factors behind this issue is the diversity of computing environments, for example, some desktop-based CUDA libraries are not available for mobile GPUs [236]. There have been some recent attempts to design lightweight architectures for mobile devices, such as MobileRNN [315] or LiteTransformer [282], as well as some software frameworks intended to support ML/DL models on mobile and edge devices, such as TensorFlow Lite or PyTorch Mobile. However, such technologies are still in their early stages and far from sufficiently mature, particularly for applications with real-time requirements.

Last but not least, affective datasets are the basis for building an emotion-sensing model. Despite the establishment in recent years of an increasing number of multimodal emotion-labeled datasets (e.g., IEMOCAP [316], MAHNOB-HCI [302], DEAP [294], K-EmoCon [283]), the existing corpus is relatively small-scale. Compared with some popular datasets for computer vision (e.g., ImageNet [317]) which consist of millions of data samples, affective datasets usually only consist of hundreds or thousands of samples. The shortage of larger-scale affective datasets has become one of the major limits to further development and deployment of emotion sensing techniques. Moreover, data acquisition for most public datasets has relied on sophisticated devices and been conducted in lab environments using artificial emotion induction methods. Due to the lack of large-scale publicly available naturalistic affective interaction datasets that entirely rely on everyday mobile devices, most of the current mobile emotion sensing studies still rely on training data collected in bespoke user studies. Thus, it is difficult to directly reproduce and compare different techniques or apply the latest inference algorithms or architectures, as they may be trained on datasets containing different modality data acquired with different devices. Further research should explore the design and collection of large-scale multimodal spontaneous mobile affect datasets in the wild.

8.3 Final Remarks

The thesis outlines open challenges and future directions for mobile emotion recognition research. Better understanding of the modality choice, ground truth establishment, and corresponding ethics, privacy and security considerations enables further construction of robust mobile emotion recognition systems and faster application to our everyday lives. Furthermore, we foresee that future directions in this research area should focus on developing fully offline mode approaches, designing lightweight deep mobile architectures, and building large-scale naturalistic affective interaction datasets based on mobile and wearable devices. Ultimately, we believe that with the rapid development of mobile and wearable devices and increased interest in technology-enabled mental health interventions, there will be a growing need for accurate, robust, and ubiquitous emotion detectors that can be deployed and integrated into everyday interfaces and devices.

Bibliography

- [1] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The Visual Computer*, 37(6):1447–1466, 2021. URL <https://doi.org/10.1007/s00371-020-01881-x>.
- [2] Kangning Yang, Chaofan Wang, Yue Gu, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. *IEEE Transactions on Affective Computing*, pages 1–17, 2021. URL <https://doi.org/10.1109/TAFFC.2021.3100868>.
- [3] Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval, ICMR '22*, page 562–570, New York, NY, USA, 2022. Association for Computing Machinery. URL <https://doi.org/10.1145/3512527.3531385>.
- [4] James J Gross. Emotion regulation: Current status and future prospects. *Psychological inquiry*, 26(1):1–26, 2015.
- [5] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.
- [6] Steven Marwaha, Matthew R Broome, Paul E Bebbington, Elizabeth Kuipers, and Daniel Freeman. Mood instability and psychosis: analyses of british national survey data. *Schizophrenia bulletin*, 40(2):269–277, 2014.

- [7] Katherine S. Young, Christina F. Sandman, and Michelle G. Craske. Positive and negative emotion regulation in adolescence: Links to anxiety and depression. *Brain Sciences*, 9(4), 2019. ISSN 2076-3425.
- [8] Niall Bolger, Angelina Davis, and Eshkol Rafaeli. Diary methods: Capturing life as it is lived. *Annual review of psychology*, 54(1):579–616, 2003.
- [9] Daniel Kahneman, Alan B Krueger, David A Schkade, Norbert Schwarz, and Arthur A Stone. A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702):1776–1780, 2004.
- [10] Timothy J Trull, Marika B Solhan, Sarah L Tragesser, Seungmin Jahng, Phillip K Wood, Thomas M Piasecki, and David Watson. Affective instability: measuring a core feature of borderline personality disorder with ecological momentary assessment. *Journal of abnormal psychology*, 117(3):647, 2008.
- [11] Michael R Hufford, Saul Shiffman, Jean Paty, and Arthur A Stone. Ecological momentary assessment: Real-world, real-time measurement of patient experience. 2001.
- [12] Barbara L Fredrickson. Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion*, 14(4): 577–606, 2000.
- [13] Niels van Berkel, Jorge Goncalves, Lauri Lovén, Denzil Ferreira, Simo Hosio, and Vassilis Kostakos. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human Computer Studies*, 125(September 2018):118–128, 2019. ISSN 10959300.
- [14] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 2225. NIH Public Access, 2018.
- [15] Jin-Hyuk Hong, Julian Ramos, and Anind K. Dey. Understanding physiological responses to stressors during physical activity. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp '12*, page 270–279, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312240.

- [16] Hector A. Gonzalez, Shahzad Muzaffar, Jerald Yoo, and Ibrahim M. Elfadel. Biocnn: A hardware inference engine for eeg-based emotion detection. *IEEE Access*, 8:140896–140914, 2020.
- [17] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.
- [18] How many people have smartphones worldwide (apr 2022). URL <https://www.bankmycell.com/blog/how-many-phones-are-in-the-world>.
- [19] Geoffrey Miller. The smartphone psychology manifesto. *Perspectives on psychological science*, 7(3):221–237, 2012.
- [20] Tanzeem Choudhury and Alex Pentland. Sensing and modeling human networks using the sociometer. In *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, pages 216–222. IEEE, 2003.
- [21] Mika Raento, Antti Oulasvirta, and Nathan Eagle. Smartphones: An emerging tool for social scientists. *Sociological methods & research*, 37(3):426–454, 2009.
- [22] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. A survey on mobile affective computing. *Computer Science Review*, 25:79–100, 2017.
- [23] Abhinav Mehrotra, Veljko Pejovic, and Mirco Musolesi. Sensocial: a middleware for integrating online social networks and mobile sensing data streams. In *Proceedings of the 15th International Middleware Conference*, pages 205–216, 2014.
- [24] Lisa Feldman Barrett, Batja Mesquita, Kevin N Ochsner, and James J Gross. The experience of emotion. *Annual review of psychology*, 58:373, 2007.
- [25] James A Coan. Emergent ghosts of the emotion machine. *Emotion Review*, 2(3): 274–285, 2010.
- [26] Sidney K. D’mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.*, 47(3), February 2015.
- [27] Marwan Dhuheir, Abdullatif Albaseer, Emna Baccour, Aiman Erbad, Mohamed Abdallah, and Mounir Hamdi. Emotion recognition for healthcare surveillance systems using neural networks: A survey. In *2021 International Wireless Communications and Mobile Computing (IWCMC)*, pages 681–687. IEEE, 2021.

- [28] Andrew McStay. Emotional ai, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society*, 7(1):2053951720904386, 2020.
- [29] Baixi Xing, Lekai Zhang, Junying Gao, Ritai Yu, and Ruimin Lyu. Barrier-free affective communication in mooc study by analyzing pupil diameter variation. In *SIGGRAPH ASIA 2016 Symposium on Education*, SA '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450345446.
- [30] Sinh Huynh, Seungmin Kim, JeongGil Ko, Rajesh Krishna Balan, and Youngki Lee. Engagemon: Multi-modal engagement sensing for mobile games. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1), March 2018.
- [31] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. ISSN 21600031.
- [32] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169–200, 1992.
- [33] Iris B. Mauss, Loren McCarter, Robert W. Levenson, Frank H. Wilhelm, and James J. Gross. The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2):175, 2005. ISSN 15283542.
- [34] Randolph M Nesse and Phoebe C Ellsworth. Evolution, emotions, and emotional disorders. *American Psychologist*, 64(2):129, 2009.
- [35] Cara A Palmer and Candice A Alfano. Sleep and emotion regulation: an organizing, integrative review. *Sleep medicine reviews*, 31:6–16, 2017.
- [36] James J Gross and R Thompson. Emotion regulation: Conceptual foundations. In *J. J. Gross (Eds.) Handbook of Emotion Regulation*. 2007.
- [37] David Yun Dai and Robert J. Sternberg. *Motivation, Emotion, and Cognition: Integrative Perspectives on Intellectual Functioning and Development*. Educational Psychology Series. Taylor & Francis, 2004.
- [38] Christopher Beedie, Peter Terry, and Andrew Lane. Distinctions between emotion and mood. *Cognition & Emotion*, 19(6):847–878, 2005.

- [39] William N Morris. A functional analysis of the role of mood in affective systems. In *Emotion*. Sage Publications, Inc, 1992.
- [40] Antonio Damasio. Neural basis of emotions. *Scholarpedia*, 6(3):1804, 2011.
- [41] Susan S. Brierley. *An Introduction to Psychology*. Routledge Revivals. Taylor & Francis, 2019. ISBN 9780429620324.
- [42] John Tooby and Leda Cosmides. The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In *Handbook of emotions*, 3rd ed., pages 114–137. The Guilford Press, New York, NY, US, 2008.
- [43] Charles Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, 1872.
- [44] Paul Ekman, Robert W. Levenson, and Wallace V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210, 1983. ISSN 00368075.
- [45] Ekman Paul and V Friesen Wallace. Unmasking the face: a guide to recognizing emotions from facial clues, 1975.
- [46] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.
- [47] James A Russell. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102, 1994.
- [48] Nicole L Nelson and James A Russell. Universality revisited. *Emotion Review*, 5(1):8–15, 2013.
- [49] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.
- [50] Wilhelm Max Wundt and Charles Hubbard Judd. *Outlines of psychology*. W. Engelmann, 1902.
- [51] Harold Schlosberg. The description of facial expressions in terms of two dimensions. *Journal of experimental psychology*, 44(4):229, 1952.

- [52] David Watson, David Wiese, Jatin Vaidya, and Auke Tellegen. The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of personality and social psychology*, 76(5):820, 1999.
- [53] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.
- [54] Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.
- [55] David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219, 1985.
- [56] Michael A Zevon and Auke Tellegen. The structure of mood change: An idiographic/nomothetic analysis. *Journal of personality and social psychology*, 43(1):111, 1982.
- [57] David C Rubin and Jennifer M Talarico. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, 17(8):802–808, 2009.
- [58] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.
- [59] Jaak Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Affective Science. Oxford University Press, USA, 2004. ISBN 9780195178050.
- [60] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Trans. Multimedia Comput. Commun. Appl.*, 15(3s), December 2019. ISSN 1551-6857.
- [61] Don H Hockenbury and Sandra E Hockenbury. *Discovering Psychology*. Worth Publishers, 2007. ISBN 9780716774778.
- [62] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

- [63] Thomas Kosch, Mariam Hassib, Robin Reutter, and Florian Alt. Emotions on the go: Mobile emotion assessment in real-time using facial expressions. In *Proceedings of the International Conference on Advanced Visual Interfaces, AVI '20*, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375351.
- [64] Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. Emo: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services, MobiSys '20*, page 448–461, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379540.
- [65] Chin-An Wang and Douglas P Munoz. A circuit for pupil orienting responses: implications for cognitive modulation of pupil size. *Current opinion in neurobiology*, 33:134–140, 2015.
- [66] Benjamin Tag, Andrew W Vargo, Aman Gupta, George Chernyshov, Kai Kunze, and Tilman Dingler. Continuous Alertness Assessments: Using EOG Glasses to Unobtrusively Monitor Fatigue Levels In-The-Wild. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*, pages 464:1—464:12, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2.
- [67] Dania Al-Omar, Areej Al-Wabil, and Manar Fawzi. Using pupil size variation during visual emotional stimulation in measuring affective states of non communicative individuals. In *International Conference on Universal Access in Human-Computer Interaction*, pages 253–258. Springer, 2013.
- [68] Francesco Onorati, Riccardo Barbieri, Maurizio Mauri, Vincenzo Russo, and Luca Mainardi. Reconstruction and analysis of the pupil dilation signal: Application to a psychophysiological affective protocol. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5–8, 2013.
- [69] Carl E Williams and Kenneth N Stevens. Vocal correlates of emotional states. *Speech evaluation in psychiatry*, pages 221–240, 1981.
- [70] Chul Min Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, 2005.

- [71] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.
- [72] Keng-hao Chang, Drew Fisher, and John Canny. Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of PhoneSense*, 2011, 2011.
- [73] Nicholas D. Lane, Petko Georgiev, and Lorena Qendro. Deeppear: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 283–294, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335744.
- [74] Marc Exposito, Javier Hernandez, and Rosalind W. Picard. Affective keys: Towards unobtrusive stress sensing of smartphone users. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '18, page 139–145, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359412.
- [75] A. Kolakowska. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *2013 6th International Conference on Human System Interactions (HSI)*, pages 548–555, 2013.
- [76] Hai-Rong Lv, Zhong-Lin Lin, Wen-Jun Yin, and Jin Dong. Emotion recognition based on pressure sensor keyboards. In *2008 IEEE International Conference on Multimedia and Expo*, pages 1089–1092, 2008.
- [77] Philippe Zimmermann, Sissel Guttormsen, Brigitta Danuser, and Patrick Gomez. Affective computing—a rationale for measuring mood with mouse and keyboard. *International journal of occupational safety and ergonomics*, 9(4):539–551, 2003.
- [78] Clayton Epp, Michael Lippold, and Regan L. Mandryk. *Identifying Emotional States Using Keystroke Dynamics*, page 715–724. Association for Computing Machinery, New York, NY, USA, 2011. ISBN 9781450302289.
- [79] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Trans. Comput.-Hum. Interact.*, 19(4), December 2012. ISSN 1073-0516.

- [80] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. Evaluating effectiveness of smartphone typing as an indicator of user emotion. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 146–151, 2017.
- [81] Sharmin Akther Purabi, Rayhan Rashed, Mirajul Islam, Nahiyen Uddin, Mahmuda Naznin, and A. B. M. Alim Al Islam. As you are, so shall you move your head: A system-level analysis between head movements and corresponding traits and emotions. In *Proceedings of the 6th International Conference on Networking, Systems and Security, NSysS '19*, page 3–11, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450376990.
- [82] Muhammad Arslan Hashmi, Qaiser Riaz, Muhammad Zeeshan, Muhammad Shahzad, and Muhammad Moazam Fraz. Motion reveal emotions: Identifying emotions from human walk using chest mounted smartphone. *IEEE Sensors Journal*, 20(22):13511–13522, 2020.
- [83] Katherine Vytal and Stephan Hamann. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of cognitive neuroscience*, 22(12):2864–2885, 2010.
- [84] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals. *Sensors*, 18(7):2074, 2018.
- [85] Lan Li and Ji-hua Chen. Emotion recognition using physiological signals. In *International conference on artificial reality and telepresence*, pages 437–446. Springer, 2006.
- [86] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 8:1454, 2017.
- [87] Bobo Zhao, Zhu Wang, Zhiwen Yu, and Bin Guo. Emotionsense: Emotion recognition based on wearable wristband. In *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCCom/IOP/SCI)*, pages 346–355, 2018.

- [88] Shiqi Jiang, Zhenjiang Li, Pengfei Zhou, and Mo Li. Memento: An emotion-driven lifelogging system with wearables. *ACM Trans. Sen. Netw.*, 15(1), January 2019.
- [89] Boyuan Sun, Qiang Ma, Shanfeng Zhang, Kebin Liu, and Yunhao Liu. Isself: Towards cold-start emotion labeling using transfer learning with smartphones. *ACM Trans. Sen. Netw.*, 13(4), September 2017.
- [90] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, page 389–402, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450316729.
- [91] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, and Jorge Goncalves. Does smartphone use drive our emotions or vice versa? a causal analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080.
- [92] Abhinav Mehrotra, Fani Tsapeli, Robert Hendley, and Mirco Musolesi. Mytraces: Investigating correlation and causation between users' emotional states and mobile phone interaction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3), September 2017.
- [93] Hosub Lee, Young Sang Choi, Sunjae Lee, and I. P. Park. Towards unobtrusive emotion recognition for affective social communication. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pages 260–264, 2012.
- [94] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4), January 2018.
- [95] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.
- [96] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Hybrid attention based multimodal network for spoken language classification.

- In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2379–2390, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
- [97] Boning Li and Akane Sano. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(2), June 2020.
- [98] Luyao Chong, Meng Jin, and Yuan He. Emochat: Bringing multimodal emotion detection to mobile conversation. In *2019 5th International Conference on Big Data Computing and Communications (BIGCOM)*, pages 213–221, 2019.
- [99] Erika L Rosenberg and Paul Ekman. Coherence between expressive and experiential systems in emotion. *Cognition & Emotion*, 8(3):201–229, 1994.
- [100] Natalia Efremova, Mikhail Patkin, and Denis Sokolov. Face and emotion recognition with neural networks on mobile devices: Practical implementation on different platforms. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–5, 2019.
- [101] M. Shamim Hossain and Ghulam Muhammad. An emotion recognition system for mobile applications. *IEEE Access*, 5:2281–2287, 2017.
- [102] Myunghoon Suk and Balakrishnan Prabhakaran. Real-time facial expression recognition on smartphones. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 1054–1059, 2015.
- [103] Yean Seanglidet, Bu Sung Lee, and Chai Kiat Yeo. Mood prediction from facial video with music “therapy” on a smartphone. In *2016 Wireless Telecommunications Symposium (WTS)*, pages 1–5, 2016.
- [104] Yara Rizk, Maya Safieddine, David Matchoulian, and Mariette Awad. Face2mus: A facial emotion based internet radio tuner application. In *MELECON 2014 - 2014 17th IEEE Mediterranean Electrotechnical Conference*, pages 257–261, 2014.
- [105] Haopeng Wang, Diana P. Tobón V, M. Shamim Hossain, and Abdulmotaleb El Saddik. Deep learning (dl)-enabled system for emotional big data. *IEEE Access*, pages 1–1, 2021.

- [106] Katsutoshi Masai, Kai Kunze, Yuta Sugiura, Masa Ogata, Masahiko Inami, and Maki Sugimoto. Evaluation of facial expression recognition by a smart eyewear for facial direction changes, repeatability, and positional drift. *ACM Trans. Interact. Intell. Syst.*, 7(4), December 2017. ISSN 2160-6455.
- [107] Filip Malawski, Bogdan Kwolek, and Shinji Sako. Using kinect for facial expression recognition under varying poses and illumination. In *International Conference on Active Media Technology*, pages 395–406. Springer, 2014.
- [108] Monica Perusquía-Hernández, Saho Ayabe-Kanamura, Kenji Suzuki, and Shiro Kumano. The invisible potential of facial electromyography: A comparison of emg and computer vision when distinguishing posed from spontaneous smiles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–9, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702.
- [109] Monica Perusquía-Hernández, Mazakasu Hirokawa, and Kenji Suzuki. A wearable device for fast and subtle spontaneous smile recognition. *IEEE Transactions on Affective Computing*, 8(4):522–533, 2017.
- [110] Anna Gruebler and Kenji Suzuki. Design of a wearable device for reading positive expressions from facial emg signals. *IEEE Transactions on Affective Computing*, 5(3):227–237, 2014.
- [111] Monica Perusquía-Hernández, Felix Dollack, Chun Kwang Tan, Shushi Namba, Saho Ayabe-Kanamura, and Kenji Suzuki. Smile action unit detection from distal wearable electromyography and computer vision. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021.
- [112] Na Yang, He Ba, Weiyang Cai, Ilker Demirkol, and Wendi Heinzelman. Bana: A noise resilient fundamental frequency detection algorithm for speech and music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12): 1833–1848, 2014.
- [113] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. Stresssense: Detecting stress in unconstrained acoustic environments using

- smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 351–360, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312240.
- [114] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: A mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, page 281–290, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605588438.
- [115] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, and Markus Gross. Affective state prediction based on semi-supervised learning from smartphone touch data. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080.
- [116] Surjya Ghosh, Shivam Goenka, Niloy Ganguly, Bivas Mitra, and Pradipta De. Representation learning for emotion recognition from smartphone keyboard interactions. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 704–710, 2019.
- [117] Daxiang Dai, Qun Liu, and Hongying Meng. Can your smartphone detect your emotion? In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1704–1709, 2016.
- [118] Fabrizio Balducci, Donato Impedovo, Nicola Macchiarulo, and Giuseppe Pirlo. Affective states recognition through touch dynamics. *Multimedia Tools and Applications*, 79(47):35909–35926, 2020.
- [119] Subrata Tikadar and Samit Bhattacharya. A novel method to build and validate an affective state prediction model from touch-typing. In *IFIP Conference on Human-Computer Interaction*, pages 99–119. Springer, 2019.

- [120] Mintra Ruensuk, Hyunmi Oh, Eunyong Cheon, Ian Oakley, and Hwajung Hong. Detecting negative emotions during social media use on smartphones. In *Proceedings of Asian CHI Symposium 2019: Emerging HCI Research Collection*, Asian-HCI '19, page 73–79, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366793.
- [121] Hyun-Jun Kim and Young Sang Choi. Exploring emotional preference for smartphone applications. In *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pages 245–249, 2012.
- [122] Ensar Arif Sağbaş, Serdar Korukoglu, and Serkan Balli. Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques. *Journal of medical systems*, 44(4):1–12, 2020.
- [123] Jiawen Han, George Chernyshov, Dingding Zheng, Peizhong Gao, Takuji Narumi, Katrin Wolf, and Kai Kunze. Sentiment pen: Recognizing emotional context based on handwriting features. In *Proceedings of the 10th Augmented Human International Conference 2019*, AH2019, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450365475.
- [124] Andreas Fsrøvig Olsen and Jim Torresen. Smartphone accelerometer data used for detecting human emotions. In *2016 3rd International Conference on Systems and Informatics (ICSAI)*, pages 410–415, 2016.
- [125] Céline Coutrix and Nadine Mandran. Identifying emotions expressed by mobile users through 2d surface and 3d motion gestures. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 311–320, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312240.
- [126] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, MobiCom '16, page 95–108. Association for Computing Machinery, 2016. ISBN 9781450342261.
- [127] Qian Xu, Xuan Liu, Juan Luo, and Zhenzhong Tang. Emotion monitoring with rfid: an experimental study. *CCF Transactions on Pervasive Computing and Interaction*, 2(4):299–313, 2020.

- [128] Phuong Pham and Jingtao Wang. Attentivelearner: improving mobile mooc learning via implicit heart rate tracking. In *International conference on artificial intelligence in education*, pages 367–376. Springer, 2015.
- [129] Tianyuan Xu, Ruixiang Yin, Lin Shu, and Xiangmin Xu. Emotion recognition using frontal eeg in vr affective scenes. In *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, volume 1, pages 1–4, 2019.
- [130] Feng Kuang, Lin Shu, Haoqiang Hua, Shibin Wu, Lulu Zhang, Xiangmin Xu, Yunhe Liu, and Man Jiang. Cross-subject and cross-device wearable eeg emotion recognition using frontal eeg under virtual reality scenes. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3630–3637, 2021.
- [131] Philip Schmidt, Robert Dürichen, Attila Reiss, Kristof Van Laerhoven, and Thomas Plötz. Multi-target affect detection in the wild: An exploratory study. In *Proceedings of the 23rd International Symposium on Wearable Computers, ISWC '19*, page 211–219, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368704.
- [132] Marco Maier, Chadly Marouane, and Daniel Elsner. Deepflow: Detecting optimal user experience from physiological data using deep neural networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19*, page 2108–2110, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.
- [133] Philip Gouverneur, Joanna Jaworek-Korjakowska, Lukas Köping, Kimiaki Shira-hama, Pawel Kleczek, and Marcin Grzegorzek. Classification of physiological data for emotion recognition. In *International Conference on Artificial Intelligence and Soft Computing*, pages 619–627. Springer, 2017.
- [134] Jaime Andres Rincon, Ângelo Costa, Paulo Novais, Vicente Julian, and Carlos Carrascosa. Using non-invasive wearables for detecting emotions with intelligent agents. In *International Joint Conference SOCO'16-CISIS'16-ICEUTE'16*, pages 73–84. Springer, 2016.

- [135] Sunder Ali Khowaja, Aria Ghora Prabono, Feri Setiawan, Bernardo Nugroho Yahya, and Seok-Lyong Lee. Toward soft real-time stress detection using wrist-worn devices for human workspaces. *Soft Computing*, 25(4):2793–2820, 2021.
- [136] Gloria Cosoli, Angelica Poli, Lorenzo Scalise, and Susanna Spinsante. Heart rate variability analysis with wearable devices: Influence of artifact correction method on classification accuracy for emotion recognition. In *2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6, 2021.
- [137] Preethi M., Nagaraj S., and Madhan Mohan P. Emotion based media playback system using ppg signal. In *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 426–430, 2021.
- [138] Aaron Frederick Bulagang, James Mountstephens, and Jason Teo. Multiclass emotion prediction using heart rate and virtual reality stimuli. *Journal of Big Data*, 8(1):1–12, 2021.
- [139] David Pollreisz and Nima TaheriNejad. A simple algorithm for emotion recognition, using physiological signals of a smart watch. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2353–2356, 2017.
- [140] Chun-Chieh Hsiao, Wen-Dian Zheng, Ren-Guey Lee, and Robert Lin. Emotion inference of game users with heart rate wristbands and artificial neural networks. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pages 326–329, 2018.
- [141] Nam T. Nguyen, Nhan V. Nguyen, My Huynh T. Tran, and Binh T. Nguyen. A potential approach for emotion prediction using heart rate signals. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pages 221–226, 2017.
- [142] Krishna Bairavi and K. B. Kumar Sundhara. Eeg based emotion recognition system for special children. In *Proceedings of the 2018 International Conference on Communication Engineering and Technology, ICCET '18*, page 1–4, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450364546.

- [143] Beatriz García-Martínez, Antonio Fernández-Caballero, Arturo Martínez-Rodrigo, and Paulo Novais. Analysis of electroencephalographic signals from a brain-computer interface for emotions detection. In *International Work-Conference on Artificial Neural Networks*, pages 219–229. Springer, 2021.
- [144] Asma Baghdadi, Yassine Aribi, Rahma Fourati, Najla Halouani, Patrick Siarry, and Adel Alimi. Psychological stimulation for anxious states detection based on eeg-related features. *Journal of Ambient Intelligence and Humanized Computing*, 12(8):8519–8533, 2021.
- [145] Roberto Covello, Giancarlo Fortino, Raffaele Gravina, Antonio Aguilar, and John G. Breslin. Novel method and real-time system for detecting the cardiac defense response based on the ecg. In *2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 53–57, 2013.
- [146] Francesca Gasparini, Alessandra Grossi, and Stefania Bandini. A deep learning approach to recognize cognitive load using ppg signals. In *The 14th Pervasive Technologies Related to Assistive Environments Conference, PETRA 2021*, page 489–495, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450387927.
- [147] Daniela Girardi, Nicole Novielli, Davide Fucci, and Filippo Lanubile. Recognizing developers’ emotions while programming. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, ICSE ’20*, page 666–677, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371216.
- [148] Rita Laureanti, Marco Bilucaglia, Margherita Zito, Riccardo Circi, Alessandro Fici, Fiamma Rivetti, Riccardo Valesi, C Oldrini, Luca T Mainardi, and Vincenzo Russo. Emotion assessment using machine learning and low-cost wearable devices. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 576–579, 2020.
- [149] Victoriano Montesinos, Fabio Dell’Agnola, Adriana Arza, Amir Aminifar, and David Atienza. Multi-modal acute stress recognition using off-the-shelf wearable devices. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2196–2201, 2019.

- [150] Brandon Taylor, Anind Dey, Daniel Siewiorek, and Asim Smailagic. Using physiological sensors to detect levels of user frustration induced by system delays. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 517–528, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335744.
- [151] Shehan R. Gunarathne, Jenomi De Silva, E. M. C. P Ekanayake, I. Samaradiwakara, P. S. Haddela, and P. A Fernando. Intellemo: A mobile instant messaging application with intelligent emotion identification. In *2013 IEEE 8th International Conference on Industrial and Information Systems*, pages 627–632, 2013.
- [152] Shiqi Yang, Ping Zhou, Kui Duan, M Shamim Hossain, and Mohammed F Alhamid. emhealth: towards emotion health through depression prediction and intelligent health recommender system. *Mobile Networks and Applications*, 23(2): 216–226, 2018.
- [153] Martin G Salido Ortega, Luis-Felipe Rodríguez, and J Octavio Gutierrez-Garcia. Towards emotion recognition from contextual information using machine learning. *Journal of Ambient Intelligence and Humanized Computing*, 11(8):3187–3207, 2020.
- [154] Orestis Piskioulis, Katerina Tzaflkou, and Anastasios Economides. Emotion detection through smartphone’s accelerometer and gyroscope sensors. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 130–137, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383660.
- [155] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. Emma: An emotion-aware wellbeing chatbot. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, 2019.
- [156] Yu-Hao Wu, Shu-Jing Lin, and Don-Lin Yang. A mobile emotion recognition system based on speech signals and facial images. In *2013 International Computer Science and Engineering Conference (ICSEC)*, pages 212–217, 2013.
- [157] Humaid Alshamsi, Veton Kepuska, Hazza Alshamsi, and Hongying Meng. Automated facial expression and speech emotion recognition app development on smart phones using cloud computing. In *2018 IEEE 9th Annual Information Technology*,

- Electronics and Mobile Communication Conference (IEMCON)*, pages 730–738, 2018.
- [158] Rita Francese and Pasquale Attanasio. Supporting depression screening with multimodal emotion detection. In *CHIItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, CHIItaly '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450389778.
- [159] Tae-Hoon Lee, Hyeong-Joon Kwon, Dong-Ju Kim, and Kwang-Seok Hong. Design and implementation of mobile self-care system using voice and facial images. In *International Conference on Smart Homes and Health Telematics*, pages 249–252. Springer, 2009.
- [160] Mohammad Adibuzzaman, Niharika Jain, Nicholas Steinhafel, Munir Haque, Ferdous Ahmed, Shiekh Iqbal Ahamed, and Richard Love. Towards in situ affect detection in mobile devices: A multimodal approach. In *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, RACS '13, page 454–460, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450323482.
- [161] Phuong Pham and Jingtao Wang. Attentivevideo: A multimodal approach to quantify emotional responses to mobile advertisements. *ACM Trans. Interact. Intell. Syst.*, 9(2–3), March 2019.
- [162] Phuong Pham and Jingtao Wang. Understanding emotional responses to mobile video advertisements via physiological signal sensing and facial expression analysis. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, page 67–78, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450343480.
- [163] Kshitij Sharma, Evangelos Niforatos, Michail Giannakos, and Vassilis Kostakos. Assessing cognitive performance using physiological and facial features: Generalizing across contexts. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3), September 2020.
- [164] Jun Gu, Bin Gao, Yuanpeng Chen, Long Jiang, Zhao Gao, Xiaole Ma, Yong Ma, Wai Lok Woo, and Jikun Jin. Wearable social sensing: Content-based processing methodology and implementation. *IEEE Sensors Journal*, 17(21):7167–7176, 2017.

- [165] Yuan-Chih Yu. A cloud-based mobile anger prediction model. In *18th International Conference on Network-Based Information Systems*, pages 199–205, 2015.
- [166] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. Laughter recognition using non-invasive wearable devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, PervasiveHealth'19, page 262–271, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361262.
- [167] Zhu Wang, Zhiwen Yu, Bobo Zhao, Bin Guo, Chao Chen, and Zhiyong Yu. Emotionsense: An adaptive emotion recognition system based on wearable smart devices. *ACM Trans. Comput. Healthcare*, 1(4), September 2020. ISSN 2691-1957.
- [168] Jonathan Rubin, Hoda Eldardiry, Rui Abreu, Shane Ahern, Honglu Du, Ashish Pattekar, and Daniel G. Bobrow. Towards a mobile and wearable system for predicting panic attacks. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 529–533, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335744.
- [169] Luis Cruz, Jonathan Rubin, Rui Abreu, Shane Ahern, Hoda Eldardiry, and Daniel G. Bobrow. A wearable and mobile intervention delivery system for individuals with panic disorder. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, MUM '15, page 175–182, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336055.
- [170] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356923.
- [171] Rasam Bin Hossain, Mefta Sadat, and Hasan Mahmud. Recognition of human affection in smartphone perspective based on accelerometer and user's sitting position. In *2014 17th International Conference on Computer and Information Technology (ICCIT)*, pages 87–91, 2014.

- [172] Zack Zhu, Héctor F. Satizábal, Ulf Blanke, Andres Perez-Urbe, and Gerhard Tröster. Naturalistic recognition of activities and mood using wearable electronics. *IEEE Transactions on Affective Computing*, 7(3):272–285, 2016.
- [173] Mintra Ruensuk, Eunyoung Cheon, Hwajung Hong, and Ian Oakley. How do you feel online: Exploiting smartphone sensors to detect transitory emotions during social media use. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4), December 2020.
- [174] Pedro Carmona, David Nunes, Duarte Raposo, David Silva, Jorge Sa Silva, and Carlos Herrera. Happy hour - improving mood with an emotionally aware application. In *2015 15th International Conference on Innovations for Community Services (I4CS)*, pages 1–7, 2015.
- [175] David Nunes, Jorge Sá Silva, Carlos Herrera, and Fernando Boavida. Human-in-the-loop connectivity management in smartphones. In *International Conference on Wired/Wireless Internet Communication*, pages 159–170. Springer, 2016.
- [176] Jannik Roessler and Peter A Gloor. Measuring happiness increases happiness. *Journal of Computational Social Science*, 4(1):123–146, 2021.
- [177] Xiao Zhang, Fuzhen Zhuang, Wenzhong Li, Haochao Ying, Hui Xiong, and Sanglu Lu. Inferring mood instability via smartphone sensing: A multi-view learning approach. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1401–1409, 2019. ISBN 9781450368896.
- [178] Luobing Dong, Yueshen Xu, Ping Wang, and Shijun He. Classifier fusion method based emotion recognition for mobile phone users. In *International Conference on Broadband Communications, Networks and Systems*, pages 216–226. Springer, 2019.
- [179] Dianxi Shi, Xi Chen, Jing Wei, and Ruosong Yang. User emotion recognition based on multi-class sensors of smartphone. In *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pages 478–485, 2015.
- [180] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems. In *Proceedings of the 2020 International Conference on Multimodal*

- Interaction*, ICMI '20, page 315–323, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375818.
- [181] Ping Zhou, Yixue Hao, Jun Yang, Wei Li, Lu Wang, Yiming Miao, and Jeungeun Song. Cloud-assisted hugtive robot for affective interaction. *Multimedia Tools and Applications*, 76(8):10839–10854, 2017.
- [182] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [183] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [184] Boualem Boashash. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic press, 2015.
- [185] Hossein Bidgoli. *Encyclopedia of Information Systems: EJ*, volume 2. Academic Press, 2003.
- [186] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, 2020.
- [187] Tohari Ahmad and Mohammad Nasrul Aziz. Data preprocessing and feature selection for machine learning intrusion detection systems. *ICIC Express Lett*, 13(2):93–101, 2019.
- [188] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.
- [189] Kah Phooi Seng, Li-Minn Ang, and Chien Shing Ooi. A combined rule-based & machine learning audio-visual emotion recognition approach. *IEEE Transactions on Affective Computing*, 9(1):3–13, 2016.
- [190] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1-2):160–187, 2003.

- [191] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.
- [192] Min Xu, L-T Chia, and Jesse Jin. Affective content analysis in comedy and horror videos by audio emotional event detection. In *2005 IEEE International Conference on Multimedia and Expo*, pages 4–pp. IEEE, 2005.
- [193] Tomas Pfister and Peter Robinson. Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis. *IEEE Transactions on Affective Computing*, 2(2):66–78, 2011.
- [194] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427, 2004.
- [195] Vipin Gupta, Mayur Dahyabhai Chopda, and Ram Bilas Pachori. Cross-subject emotion recognition using flexible analytic wavelet transform from eeg signals. *IEEE Sensors Journal*, 19(6):2266–2274, 2019.
- [196] Ekman Paul and Wallace V Friesen. Facial action coding system: a technique for the measurement of facial movement. *Consulting Psychologists Press*, 1978.
- [197] Zhao Zhong, Gang Shen, and Wenhui Chen. Facial emotion recognition using phog and a hierarchical expression model. In *2013 5th International Conference on Intelligent Networking and Collaborative Systems*, pages 741–746, 2013.
- [198] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.
- [199] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 3723–3726, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340823.

- [200] Hyun-Jun Kim and Young Sang Choi. A peak detection method for understanding user states for empathetic intelligent agents. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 261–265, 2013.
- [201] Jianwei Niu, Shihao Wang, Yiming Su, and Song Guo. Temporal factor-aware video affective analysis and recommendation for cyber-based social media. *IEEE Transactions on Emerging Topics in Computing*, 5(3):412–424, 2017.
- [202] Jangho Kwon, Da-Hye Kim, Wanjoo Park, and Laehyun Kim. A wearable device for emotional recognition using facial expression and physiological response. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5765–5768, 2016.
- [203] Dmitrii Fedotov, Yuki Matsuda, Yuta Takahashi, Yutaka Arakawa, Keiichi Yasumoto, and Wolfgang Minker. Towards estimating emotions and satisfaction level of tourist based on eye gaze and head movement. In *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 399–404, 2018.
- [204] Iain R Murray and John L Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.
- [205] Christer Gobl and Ailbhe Ní Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1-2):189–212, 2003.
- [206] Patrik N. Juslin and Klaus R. Scherer. Vocal Expression of Affect. In *The New Handbook of Methods in Nonverbal Behavior Research*. 2008.
- [207] John HL Hansen, Wooil Kim, Mandar Raurkar, Evan Ruzanski, and James Meyerhoff. Robust emotional stressed speech detection using weighted frequency sub-bands. *EURASIP Journal on Advances in Signal Processing*, 2011:1–10, 2011.
- [208] Gauri Deshpande, Venkata Subramanian Viraraghavan, Mayuri Duggirala, and Sachin Patel. Detecting emotional valence using time-domain analysis of speech signals. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3605–3608, 2019.

- [209] Emily Mower Provost and Shrikanth Narayanan. Simplifying emotion classification through emotion distillation. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4, 2012.
- [210] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. Tapsense: Combining self-report patterns and typing characteristics for smartphone based emotion detection. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350754.
- [211] Boon Giin Lee, Teak Wei Chong, Boon Leng Lee, Hee Joon Park, Yoon Nyun Kim, and Beomjoon Kim. Wearable mobile-based emotional response-monitoring system for drivers. *IEEE Transactions on Human-Machine Systems*, 47(5):636–649, 2017.
- [212] Jiayu Shu, Mangtik Chiu, and Pan Hui. Emotion sensing for mobile computing. *IEEE Communications Magazine*, 57(11):84–90, 2019.
- [213] Erik Peper, Fred Shaffer, and I-Mei Lin. Garbage in; garbage out—identify blood volume pulse (bvp) artifacts before analyzing and interpreting bvp, blood volume pulse amplitude, and heart rate/respiratory sinus arrhythmia data. *Biofeedback*, 38(1):19–23, 2010.
- [214] Hugo Critchley and Yoko Nagai. *Electrodermal Activity (EDA)*, pages 666–669. Springer New York, New York, NY, 2013.
- [215] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2016.
- [216] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.
- [217] Yu-Liang Hsu, Jeen-Shing Wang, Wei-Chun Chiang, and Chien-Han Hung. Automatic ecg-based emotion recognition in music listening. *IEEE Transactions on Affective Computing*, 11(1):85–99, 2020.

- [218] Oswald Barral, Ilkka Kosunen, and Giulio Jacucci. No need to laugh out loud: Predicting humor appraisal of comic strips based on physiological signals in a realistic environment. *ACM Trans. Comput.-Hum. Interact.*, 24(6), December 2017. ISSN 1073-0516.
- [219] Michael J Katz. Fractals and the analysis of waveforms. *Computers in biology and medicine*, 18(3):145–156, 1988.
- [220] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD’96, page 226–231. AAAI Press, 1996.
- [221] Laiyuan Tong, Jinchuang Zhao, and Wenli Fu. Emotion recognition and channel selection based on eeg signal. In *2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA)*, pages 101–105, 2018.
- [222] Yue Gu, Xinyu Li, Kaixiang Huang, Shiyu Fu, Kangning Yang, Shuhong Chen, Moliang Zhou, and Ivan Marsic. Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM ’18, page 537–545, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657.
- [223] Eiman Kanjo, Eman M.G. Younis, and Chee Siang Ang. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, 49:46–56, 2019.
- [224] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- [225] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019.

- [226] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629, 2018.
- [227] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [228] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*, 2017.
- [229] Yoon Kim. Convolutional neural networks for sentence classification, 2014.
- [230] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013.
- [231] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [232] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [233] Giuseppe Romano Tizzano, Matteo Spezialetti, and Silvia Rossi. A deep learning approach for mood recognition from wearable data. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5, 2020.
- [234] Huan Zhao, Yufeng Xiao, Jing Han, and Zixing Zhang. Compact convolutional recurrent neural networks via binarization for speech emotion recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6690–6694, 2019.

- [235] Jyun-Yu Jiang, Zehan Chao, Andrea L. Bertozzi, Wei Wang, Sean D. Young, and Deanna Needell. Learning to predict human stress level with incomplete sensor data from wearable devices. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 2773–2781, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763.
- [236] Yanjiao Chen, Baolin Zheng, Zihan Zhang, Qian Wang, Chao Shen, and Qian Zhang. Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions. *ACM Comput. Surv.*, 53(4), August 2020.
- [237] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.
- [238] Youssef Mohamed, Giulia Ballardini, Maria Teresa Parreira, Séverin Lemaignan, and Iolanda Leite. Automatic frustration detection using thermal imaging. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, page 451–460. IEEE Press, 2022.
- [239] Jangho Kwon, Jihyeon Ha, Da-Hye Kim, Jun Won Choi, and Laehyun Kim. Emotion recognition using a glasses-type wearable device via multi-channel facial responses. *IEEE Access*, 9:146392–146403, 2021.
- [240] Leandro Y. Mano, Gabriel T. Giancristofaro, Bruno S. Façal, Giampaolo L. Libralon, Gustavo Pessin, Pedro H. Gomes, and Jó Ueyama. Exploiting the use of ensemble classifiers to enhance the precision of user's emotion classification. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, EANN '15, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335805.
- [241] Yingjun Dong and Hiroki Sayama. Mutual-information-based feature selection for facial emotion recognition on light-weight devices. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2455–2461, 2019.
- [242] Ghulam Muhammad and M. Shamim Hossain. Emotion recognition for cognitive edge computing using deep learning. *IEEE Internet of Things Journal*, pages 1–1, 2021.

- [243] Peisong Liu and Xiaoping Wang. A lightweight multi-modal emotion recognition network based on multi-task learning. In *2021 International Conference on Neuromorphic Computing (ICNC)*, pages 368–372, 2021.
- [244] Romulus-Cristian Moraru and Angel Caçaron. Automated recognition of facial expressions and gender in humans implemented on mobile devices. In *2021 International Aegean Conference on Electrical Machines and Power Electronics (ACEMP) & 2021 International Conference on Optimization of Electrical and Electronic Equipment (OPTIM)*, pages 331–337, 2021.
- [245] Fwa Hua Leong. Deep learning of facial embeddings and facial landmark points for the detection of academic emotions. In *Proceedings of the 5th International Conference on Information and Education Innovations*, pages 111–116, 2020.
- [246] Yingying Zhao, Yuhu Chang, Yutian Lu, Yujiang Wang, Mingzhi Dong, Qin Lv, Robert P. Dick, Fan Yang, Tun Lu, Ning Gu, and Li Shang. Do smart glasses dream of sentimental visions? deep emotionship analysis for eyewear devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(1), mar 2022.
- [247] Anderson R. Avila, João Monteiro, Douglas O’Shaughneussy, and Tiago H. Falk. Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks. In *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 360–365, 2017.
- [248] Anderson R. Avila, Zahid Akhtar, João F. Santos, Douglas O’Shaughnessy, and Tiago H. Falk. Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. *IEEE Transactions on Affective Computing*, 12(1):177–188, 2021.
- [249] Petko Georgiev, Sourav Bhattacharya, Nicholas D. Lane, and Cecilia Mascolo. Low-resource multi-task audio sensing for mobile and embedded devices via shared deep neural network representations. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3), September 2017.

- [250] Woan-Shiuan Chien, Huang-Cheng Chou, and Chi-Chun Lee. Self-assessed emotion classification from acoustic and physiological features within small-group conversation. In *Companion Publication of the 2021 International Conference on Multimodal Interaction, ICMI '21 Companion*, page 230–239, New York, NY, USA, 2021. Association for Computing Machinery.
- [251] Nikolaos S. Katertsidis, Christos D. Katsis, and Dimitrios I. Fotiadis. Intrepid, a biosignal-based system for the monitoring of patients with anxiety disorders. In *2009 9th International Conference on Information Technology and Applications in Biomedicine*, pages 1–6, 2009.
- [252] Bohdan Myroniv, Cheng-Wei Wu, Yi Ren, and Yu-Chee Tseng. Analysis of users' emotions through physiology. In *International Conference on Genetic and Evolutionary Computing*, pages 136–143. Springer, 2017.
- [253] Cristina Gena, Claudio Mattutino, Simone Pirani, and Berardina De Carolis. Do beis detect user's engagement? the results of an empirical experiment with emotional artworks. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP'19 Adjunct*, page 387–391, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367110.
- [254] Berken Utku Demirel, Ivan Skelin, Haoxin Zhang, Jack J. Lin, and Mohammad Abdullah Al Faruque. Single-channel eeg based arousal level estimation using multitaper spectrum estimation at low-power wearable devices. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 542–545, 2021.
- [255] Fadi Al Machot, Mouhannad Ali, Suneth Ranasinghe, Ahmad Haj Mosa, and Kya-makya Kyandoghere. Improving subject-independent human emotion recognition using electrodermal activity sensors for active and assisted living. In *Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference, PETRA '18*, page 222–228, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450363907.
- [256] Zhong-Min Wang, Shu-Yuan Hu, and Hui Song. Channel selection method for eeg emotion recognition using normalized mutual information. *IEEE Access*, 7: 143303–143311, 2019.

- [257] N Beckmann, R Viga, A Dogangün, and A Grabmaier. Measurement and analysis of local pulse transit time for emotion recognition. *IEEE Sensors Journal*, 19(17): 7683–7692, 2019.
- [258] Akshay Sujatha Ravindran, Sho Nakagome, Dilranjan S. Wickramasuriya, Jose L. Contreras-Vidal, and Rose T. Faghih. Emotion recognition by point process characterization of heartbeat dynamics. In *2019 IEEE Healthcare Innovations and Point of Care Technologies, (HI-POCT)*, pages 13–16, 2019.
- [259] Cheng He, Yun-jin Yao, and Xue-song Ye. An emotion recognition system based on physiological signals obtained by wearable sensors. In *Wearable sensors and robots*, pages 15–25. Springer, 2017.
- [260] Nastaran Saffaryazdi, Yenushka Goonesekera, Nafiseh Saffaryazdi, Nebiyu Daniel Hailemariam, Ebaso Girma Temesgen, Suranga Nanayakkara, Elizabeth Broadbent, and Mark Billingham. Emotion recognition in conversations using brain and physiological signals. In *27th International Conference on Intelligent User Interfaces, IUI '22*, page 229–242, New York, NY, USA, 2022. Association for Computing Machinery.
- [261] Reika Takeshita, Aya Shoji, Tahera Hossain, Anna Yokokubo, and Guillaume Lopez. Emotion recognition from heart rate variability data of smartwatch while watching a video. In *2021 Thirteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, pages 1–6, 2021.
- [262] Sheng-Hui Wang, Huai-Ting Li, En-Jui Chang, and An-Yeu Andy Wu. Entropy-assisted emotion recognition of valence and arousal using xgboost classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 249–260. Springer, 2018.
- [263] Alireza Golgouneh and Bahram Tarvirdizadeh. Fabrication of a portable device for stress monitoring using wearable sensors and soft computing algorithms. *Neural Computing and Applications*, 32(11):7515–7537, 2020.
- [264] Zhongke Gao, Yanli Li, Yuxuan Yang, Na Dong, Xiong Yang, and Celso Grebogi. A coincidence-filtering-based approach for cnns in eeg-based recognition. *IEEE Transactions on Industrial Informatics*, 16(11):7159–7167, 2020.

- [265] Soumya Jyoti Raychaudhuri, Soumya Manjunath, Chithra Priya Srinivasan, N Swathi, S Sushma, KN Nitin Bhushan, and C Narendra Babu. Prescriptive analytics for impulsive behaviour prevention using real-time biometrics. *Progress in Artificial Intelligence*, 10(2):99–112, 2021.
- [266] Lam Huynh, Tri Nguyen, Thu Nguyen, Susanna Pirttikangas, and Pekka Siirtola. Stressnas: Affect state and stress detection using neural architecture search. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, UbiComp '21, page 121–125, New York, NY, USA, 2021. Association for Computing Machinery.
- [267] M. Sami Zitouni, Cheul Young Park, Uichin Lee, Leontios Hadjileontiadis, and Ahsan Khandoker. Arousal-valence classification from peripheral physiological signals using long short-term memory networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 686–689, 2021.
- [268] Firoj Alam and Giuseppe Riccardi. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, page 15–18, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450331296.
- [269] Ali Mehmood Khan and Michael Lawo. Recognizing emotion from blood volume pulse and skin conductance sensor using machine learning algorithms. In *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, pages 1297–1303. Springer, 2016.
- [270] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Ji-ahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.
- [271] Satoko Koganemaru, Kazuhisa Domen, Hidenao Fukuyama, and Tatsuya Mima. Negative emotion can enhance human motor cortical plasticity. *European Journal of Neuroscience*, 35(10):1637–1645, 2012.

- [272] Ewa Siedlecka and Thomas F Denson. Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, 11(1):87–97, 2019.
- [273] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3, 1997.
- [274] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4): 626–630, 2005.
- [275] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior research methods*, 46(2):596–610, 2014.
- [276] Jarosław M Michałowski, Dawid Drożdziel, Jacek Matuszewski, Wojtek Koziejowski, Katarzyna Jednoróg, and Artur Marchewka. The set of fear inducing pictures (sfip): Development and validation in fearful and nonfearful individuals. *Behavior Research Methods*, 49(4):1407–1419, 2017.
- [277] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.
- [278] Mohamed Khamis, Anita Baier, Niels Henze, Florian Alt, and Andreas Bulling. *Understanding Face and Eye Visibility in Front-Facing Cameras of Smartphones Used in the Wild*, page 1–12. Association for Computing Machinery, New York, NY, USA, 2018. ISBN 9781450356206.
- [279] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized multitask learning for predicting tomorrow’s mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213, 2017.
- [280] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

- [281] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [282] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*, 2020.
- [283] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7(1):1–16, 2020.
- [284] Michael T Motley and Carl T Camden. Facial expression of emotion: A comparison of posed expressions versus spontaneous expressions in an interpersonal communication setting. *Western Journal of Communication (includes Communication Reports)*, 52(1):1–22, 1988.
- [285] Rebecca Jürgens, Annika Grass, Matthias Drolet, and Julia Fischer. Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *Journal of nonverbal behavior*, 39(3):195–214, 2015.
- [286] Patrik N Juslin, Petri Laukka, and Tanja Bänziger. The mirror to our soul? comparisons of spontaneous and posed vocal expression of emotion. *Journal of nonverbal behavior*, 42(1):1–40, 2018.
- [287] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544, 2015.
- [288] Albrecht Schmidt. Cloud-based ai for pervasive applications. *IEEE Pervasive Computing*, 15(01):14–18, 2016.
- [289] Arne Bernin, Larissa Müller, Sobin Ghose, Kai von Luck, Christos Grecos, Qi Wang, and Florian Vogt. Towards more robust automatic facial expression recognition in smart environments. In *Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments*, pages 37–44, 2017.

- [290] De'Aira Bryant and Ayanna Howard. A comparative analysis of emotion-detecting ai systems with respect to algorithm performance and dataset diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 377–382, 2019.
- [291] Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea C Samson. Facial expression analysis with affdex and facet: A validation study. *Behavior research methods*, 50(4):1446–1460, 2018.
- [292] Hairong Qi, Xiaoling Wang, S Sitharama Iyengar, and Krishnendu Chakrabarty. Multisensor data fusion in distributed sensor networks using mobile agents. In *Proceedings of 5th International Conference on Information Fusion*, pages 11–16, 2001.
- [293] Akane Sano, Sara Taylor, Andrew W McHill, Andrew JK Phillips, Laura K Barger, Elizabeth Klerman, Rosalind Picard, et al. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. *Journal of medical Internet research*, 20(6):e9410, 2018.
- [294] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.
- [295] Junjie Zhu, Yuxuan Wei, Yifan Feng, Xibin Zhao, and Yue Gao. Physiological signals-based emotion recognition via high-order correlation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–18, 2019.
- [296] Shinsuke Shimojo and Ladan Shams. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509, 2001.
- [297] Layale Constantine and Hazem Hajj. A survey of ground-truth in emotion data annotation. In *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 697–702, 2012.

- [298] Jorge Goncalves, Pratyush Pandab, Denzil Ferreira, Mohammad Ghahramani, Guoying Zhao, and Vassilis Kostakos. Projective testing of diurnal collective emotion. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, page 487–497, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329682.
- [299] Niels van Berkel, Jorge Goncalves, Simo Hosio, Zhanna Sarsenbayeva, Eduardo Velloso, and Vassilis Kostakos. Overcoming compliance bias in self-report studies: A cross-study analysis. *International Journal of Human-Computer Studies*, 134: 1–12, 2020.
- [300] Rob Boddice. The history of emotions: Past, present, future. *Revista de Estudios Sociales*, (62):10–15, 2017.
- [301] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133, 2014.
- [302] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.
- [303] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020.
- [304] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.
- [305] Benjamin Tag, Jorge Goncalves, Sarah Webber, Peter Koval, and Vassilis Kostakos. A retrospective and a look forward: Lessons learned from researching emotions in-the-wild. *IEEE Pervasive Computing*, 21(1):28–36, 2022.
- [306] Hadjer Fellah, Chaker Mezioud, and Mohamed Chaouki Batouche. Mobile cloud computing: Architecture, advantages and security issues. In *Proceedings of the*

- 3rd International Conference on Networking, Information Systems & Security*, NISS2020, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450376341.
- [307] Homero Rivas and Katarzyna Wac. *Digital Health Scaling Healthcare to the World*. Springer, 2020. ISBN 9783319614465.
- [308] Isabel Straw. Ethical implications of emotion mining in medicine. *Health Policy and Technology*, 10(1):191–195, 2021. ISSN 22118845.
- [309] Daniel J. Simons, Walter R. Boot, Neil Charness, Susan E. Gathercole, Christopher F. Chabris, David Z. Hambrick, and Elizabeth A.L. Stine-Morrow. Do “Brain-Training” Programs Work? *Psychological Science in the Public Interest, Supplement*, 17(3):103–186, 2016. ISSN 15396053.
- [310] George M. Slavich, Sara Taylor, and Rosalind W. Picard. Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress*, 22(4):408–413, jul 2019. ISSN 1025-3890.
- [311] Elaine Sedenberg and John Chuang. Smile for the camera: privacy and policy implications of emotion ai. *arXiv preprint arXiv:1709.00396*, 2017.
- [312] Sylvain Delplanque and David Sander. A fascinating but risky case of reverse inference: From measures to emotions! *Food Quality and Preference*, (November 2020):104183, 2021. ISSN 09503293.
- [313] Ping Wang, Luobing Dong, yueshen xu, Wei Liu, and Ningning Jing. Clustering-based emotion recognition micro-service cloud framework for mobile computing. *IEEE Access*, 8:49695–49704, 2020.
- [314] H. Brendan McMahan and Daniel Ramage. Google ai blog: Federated learning: Collaborative machine learning without centralized training data. 2017. URL <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>.
- [315] Qingqing Cao, Niranjan Balasubramanian, and Aruna Balasubramanian. Mobirnn: Efficient recurrent neural network execution on mobile gpu. In *Proceedings of the 1st International Workshop on Deep Learning for Mobile Systems and Applications*, pages 1–6, 2017.

-
- [316] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemo-cap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.
- [317] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.