Evaluation of facial expression feedback within self-report tools and an exploration of depression's symptomatology as facial cues

Hristo Ventzeslavov Valev, Dipl-Inf.
School of Computing and Communications
Lancaster University

A thesis submitted for the degree of
Doctor of Philosophy

November, 2021

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography.
Hristo Ventzeslavov Valev

Evaluation of facial expression feedback within self-report tools and an exploration of depression's symptomatology as facial cues

Hristo Ventzeslavov Valev, Dipl-Inf..
School of Computing and Communications, Lancaster University
A thesis submitted for the degree of *Doctor of Philosophy*. November, 2021.

Abstract

Self-reports are the most accurate form of assessing mood. They, can be administered frequently, and self-report tools are valuable for quantifying and monitoring one's mental state of well-being. Traditionally, self-reports are provided using numerical or graphical scales, however, those are known to be prone to systematic errors in their measurements. Alternatively, facial expressions are intrinsically connected to emotional experiences, are a tool for us to communicate our emotions. We are well-versed in enacting or recognizing facial expressions. Hence, those are suitable representations for mood. Tools relying on facial expressions can expand the space for mood self-report technologies.

Depression is an affective disorder, particularly pervasive in contemporary society. Its severity is typically measured on individual symptoms using screener questionnaires. However, when administered frequently, the assessment quality of those questionnaires is known to degrade significantly. Hence, by identifying salient features indicative of depression's symptomatology in the face, facial expression-based tools can capitalise on the strengths of self-reports and be used for assessing or monitoring depression's severity.

Herein, this thesis explores the design and implementation of four prototypes for mood self-reports iteratively. Three empirical studies evaluate the use of the method within three experimental contexts, by using text and images to elicit emotions in-situ and for monitoring mood in the wild. Therein, the method was evaluated quantitatively – by contrasting self-reports to those provided with the well-known visual analogue scale, and qualitatively – by identifying aspects of importance for facial expression-based tools and exploring user's preferences. Thereafter, an exploratory study was conducted identifying, and visualizing facial features indicative of symptoms of depression as a step towards creating disorder-specific self-report instruments. Finally, EmotionAlly, a prototype for contextualized assessment, tracking, and visualisation of mood using computer-generated facial expressions was developed, integrating findings from preceding quantitative and qualitative evaluations.

Publications

Two publications, shown below, have been created directly from the thesis, from which large portions of this published work is used within Chapters 4 and 5 respectively:

Hristo Valev, Tim Leufkens, Corina Sas, Joyce Westerink, and Ron Dotsch. "Evaluation of a Self-report System for Assessing Mood Using Facial Expressions." In: *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer International Publishing, 2019, pp. 231–241. DOI: 10.1007/978-3-030-25872-6_19

Hristo Valev, Alessio Gallucci, Tim Leufkens, Joyce Westerink, and Corina Sas. "Applying Delaunay Triangulation Augmentation for Deep Learning Facial Expression Generation and Recognition." In: *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, 2021, pp. 730–740. DOI: 10.1007/978-3-030-68796-0_53

The following contributions have been published as unindexed works. The former was published at *25th annual international CyberPsychology, CyberTherapy & Social Networking Conference* as an extended abstract and is based on content presented in Chapter 2, Section 2.5.2. The latter was published as a Poster at the *Society for Affective Sciences 2020 Annual Conference Symposium* and presents a sliver of the content presented in chapter 6. The contents are also available in the symposiums' collection of contributions on page 47.

Hristo Valev, Tim Leufkens, Corina Sas, and Joyce Westerink. "On the perception of facial expressions in affective disorders and potential technological uses." In: *25th annual international CyberPsychology, CyberTherapy & Social Networking Conference*. 2020

Hristo Valev, Tim Leufkens, Joyce Westerink, and Corina Sas. "An Interface with Computer-Generated Facial Expressions as an Alternative for Mood Self-Reports in an EMA Context." In: *Annual Conference Symposium*. Society for Affective Science, 2020, p. 47

The following patent has been created from work implied, but not explicitly discussed in detail, within this thesis. This patent is created in relation to the contribution described in further detail in [3].

Hristo Ventzeslavov Valev, Timmy Robertus Maria Leufkens, Joanne Henriëtte Desirée Monique Westerink, Dooren Marieke Van, Ee Raymond Van, Willem

Huijbers, Benito Maria Estrella Mena, and Adrianus Johannes Maria Denissen. "Apparatus for Determining and/or Assess Depression Severity of a Patient." Pat. WO2022101108A1. May 2022

Two further publications originating from the work described in this thesis are planned to be submitted. The first one will consist of the content described in Chapter 9. The second one will be an amalgam of multiple smaller contributions made throughout this thesis.

Contribution Statements

The work described in this Thesis is my own, however, the methods, analyses, experimental designs and concepts throughout this thesis benefited from the very valuable discussions with and suggestions by my supervisors Tim Leufkens, Joyce Westerink and Corina Sas.

Chapter 4

This Chapter was published as a conference paper at the Mindcare conference. Ron Dotsch ideated a reframe in the study design and is also a co-author in the respective paper. The Android application used in the experiment was of my own design and creation, except for the facial expression images contained within. Those were obtained from the D-VAMS scale website[1].

Chapter 5

The Radboud Faces Database (RafD) [8] was used to create the nuanced facial expressions. The dataset was obtained after contacting the responsible persons indicated at the RafD website[2]. The repository[3] contains code which applied a deconvolutional neural network and the RafD dataset. This project was used as a starting point for the work described in this Chapter. The content of this Chapter was published as a conference paper at the ICPR International Workshops and Challenges [2]. The facial expressions recognition experiments and interpretation of results were conducted by Alessio Gallucci, who is also the second author of the respective paper.

Chapter 6

A small portion of the results section in this Chapter was published as a poster at the Society for Affective Sciences conference in 2020 [4]. The conference, however, did not take place due to the, at the time, ongoing COVID-19 pandemic.

Chapter 7

The conceptualization of the multidimensional facial expressions scale as well as all technical implications were of my own design. Naturally, I have used existing frameworks which have been listed to aid me in the technical realization of the application. An inspiration for allocating facial expressions and their intensities over a polar coordinate system was Russel's circumplex model of affect [9].

[1] `http://dvams.com/dvams/menu_home_dvams.htm`

[2] `http://www.rafd.nl`

[3] `https://github.com/somewacko/deconvfaces`

Chapter 8

The experiment described in this Chapter employed the International Affective Picture Set (IAPS) [10] for eliciting emotions as well as the results from another work [11] which provided image ratings on categorical emotions.

Chapter 9

I was introduced to the Reverse Correlation - Classification Images method [12] and the rcicr R library [13] by Ron Dotsch. The image of the base face was created using the RADIATE dataset [14], which is an open-access facial expression image set.

Chapter 10

EmotionAlly was intended to be used in a clinical evaluation study as a collaboration within the AffecTech Consortium with my colleague Desirée Colombo. The content presented in this Chapter presents a subset of the complete functionality of EmotionAlly, relevant to my work. The technical realization of the application is of my own design, where my colleague contributed with ideas to improve some of its visual aspects and ideas relevant to parts of the application not presented in this thesis. Ultimately, no formal evaluation was conducted as the start of the study coincided with the first reported cases of COVID-19 in Europe.

Acknowledgements

First and foremost I would like to thank my direct supervisors Tim Leufkens and Joyce Westerink. The kindness, patience and support that I have received from them, summarized in their own words as their philosophy of the 'humane treatment of PhDs', has taught me a lot of valuable lessons. I know that I would have succeeded in passing it forward even if I'm able to apply those lessons only halfway through. I was lucky and I am grateful.

I also want to thank my academic supervisor and promoter Corina Sas. She has helped me a lot by teaching me to zoom out of the current fire and taking the time to look at the bigger picture.

I also want to thank my brother Krassimir, who has helped me in many tangible and intangible ways, which allowed me to embark on the PhD journey in the first place and helped me get to choose what my next one gets to be.

I also want to thank my parents Lubomira and Ventzeslav for bringing me into this world and for the lessons I have received or refused to remember.

Hereafter, I'd like to thank my colleagues at the Brain, Behaviour & Cognition department at Philips Research. I am lucky to have been given the opportunity to be a part of a group with so many smart and helpful people to learn from. Specifically, I want to thank Marieke van der Hoven for welcoming me into the BBC department and giving me the opportunity to pursue a PhD. On that note, I will take the opportunity to apologize for spending a chunk of my departments' budget on my mandated PhD travels. I also want to thank Inge for always being a very positive, kind and helpful person. I will always look back on the memories I have made during the years of my PhD and my time at the BBC department with fondness.

Additionally, I'm very grateful to have also been a part of the Philips PhD Community. The many lunch lectures, social events and fun conversations with Alessio, Hanne, Marko and many others was an excellent distraction from the work.

I want to thank my housemate Emanuele, the most patient impatient person, for putting up with my rants and complaints for years and for being a great (and now certified) friend. I also want to thank Vanina, Anna, Mark, Linda, and Pauline that always made time for me and made those 4 years much more enjoyable, despite the world seemingly slipping into (hopefully) temporary insanity at the time. And of course I also want to thank Aditya, Hyus, Samuil, Pavel, and Emi which have been with me for the full ride.

I would also like to thank my fellow AffecTech colleagues Desirée, Pavel, Shadi, Claudia, ChengCheng, Umair, Javier, Andrea, Dionne, Camille, Charles and many others for the many fun memories made on our training events all over Europe. I have, completely by accident, happened to see multiple sunrises on the many places that we went to.

I also want to thank Horizon 2020 research and innovation funding programme for awarding me the Marie Skłodowska-Curie Fellowship as part of the AffecTech Consortium. Being part of an interdisciplinary project has taught me that to truly

understand a problem you need to see it holistically through multiple lenses and perspectives and that knowing how to be a scientist and an engineer generalizes well in all matters of life. The PhD journey itself has taught me a lesson of humility, acceptance, and knowing my limits, which are very valuable skills that will unfortunately help me make better, but boring decisions for the future.

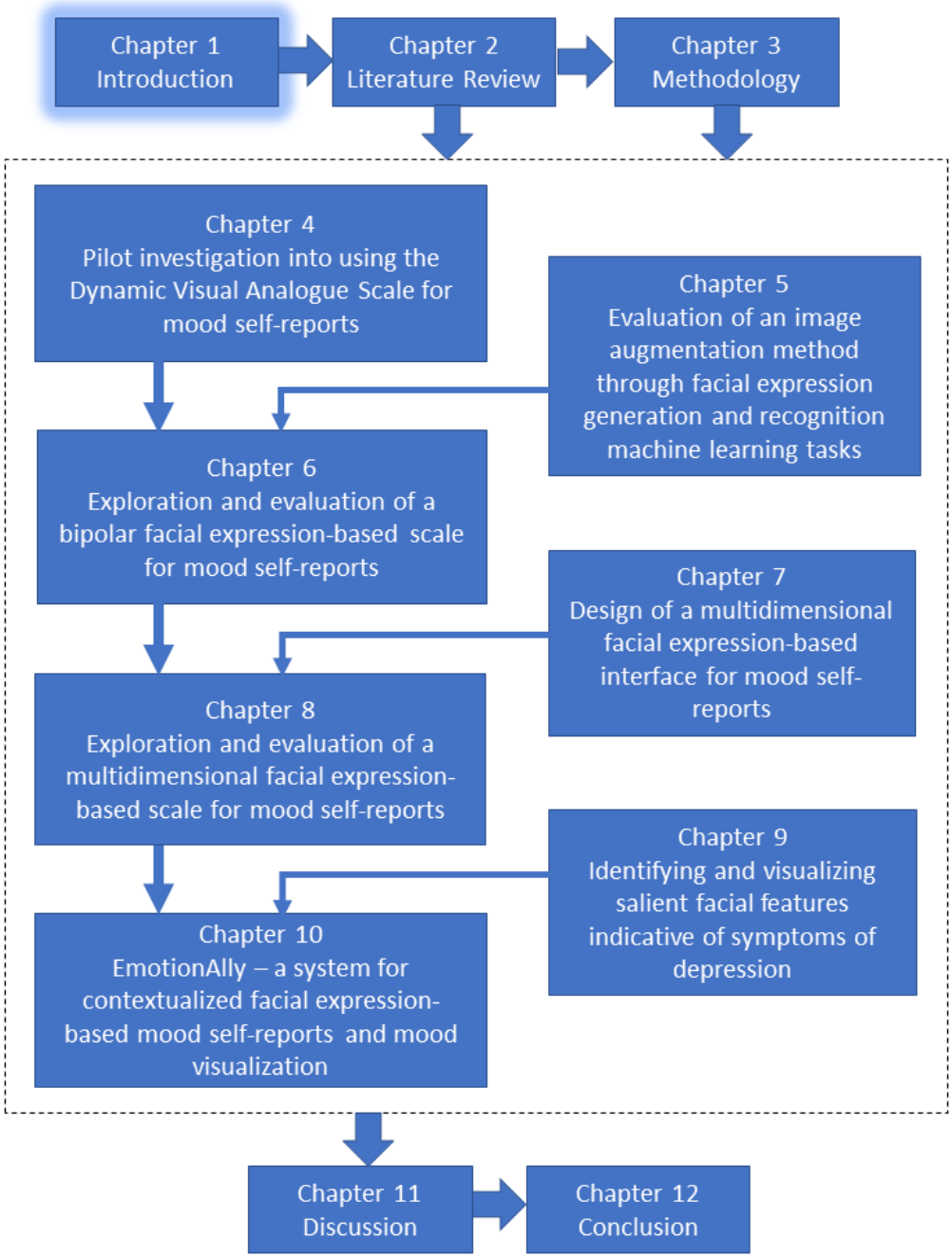
I also want to thank O'Sheas and the other reputable establishments for the many nights that I will neither forget nor remember, the people at Nieuwstraat and the TSH, TU/e Sports Institute, the participants in my studies and all the other people and places that have positively contributed to my life. And finally, I'd like to thank the Raspberry Pi Foundation for giving me a safe outlet for my pandemic frustrations and a great hobby that helped get on to my next vocation.

I am grateful to have had all those people in my life. My accomplishments wouldn't have been possible or worth it without any of them and I do apologize if I have missed to mention anyone.

Contents

List of Figures

List of Tables


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 1

Introduction

1.1 Problem Definition

Mood and its variations are an important aspect descriptive of the mental state of a person [15, 16]. Applications which rely on reporting mood can allow for tracking and visualising mood trends over time. Their use can allow persons to gain awareness of their own mood and factors which may influence it. Additionally, accurately assessing mood can allow tools to be developed which can recognize patterns of mood dysphoria and assist persons by suggesting suitable informative material, exercises or provide insights into factors affecting their mood.

The most accurate form of assessing mood is through self-reports [17]. Numerical or graphical scales such as the Likert and Visual Analogue Scales (VAS) [18, 19] are ubiquitous tools for providing self-reports and have been frequently used to measure mood. There are, however, criticisms regarding their use for mood assessments as they offer no particular inclination to capture mood [20, 21]. The reason being, assessing one's mood relies on translating a subjective mental state onto a scale, where the meaning of each numerical value or range of values is based on a persons subjective interpretation [20, 21]. Additionally, assessments provided with VAS scales are found to be susceptible to systematic patterns expressed as uneven utilization of their range, the coalescence of assessments near indicators denoting a scale's intervals, or an end-aversion bias in using the extreme ranges of the scale [22–24]. Subsequently, those appear to influence assessments negatively, thus reducing the assessment's quality. Consequently, measurements obtained using those methods possess multiple potential sources of measurement errors, which may hamper their use when precise self-assessments are required. Finally, numerical scales have been thoroughly investigated, explored, and validated, and subsequently leave little room for their further improvement. Finding suitable alternative means to represent and capture mood can improve the status quo or widen the space for mood self-report technologies.

Facial expressions are an excellent means of communicating mood in real-life. Additionally, a wealth of research has shown them to be both universal, and valid

indicators of mood [21, 25–27]. They are inherently linked to emotional experiences and are a visual tool for us to communicate emotions to the surrounding world [28]. As embodied representations of emotion they are intrinsically suitable for representing mood [21] and we are well-versed in enacting and recognizing those in others. Using facial expressions to represent mood within self-report tools, can prove to be a more accessible way to represent mood, as a facial expression's intensity is an indicator of the intensity of the portrayed emotion. Therefore, another benefit of using facial expressions for mood self-reports is that they can indicate emotions and their intensity non-verbally, thus increasing the accessibility of the method.

Assessing and tracking mood can be particularly valuable to monitor mood dysphoria, a symptom central to the symptomatology of affective disorders and depression in particular, a prevalent disorder in contemporary society [29]. Additionally, faces are the most informative and expressive feature of a person and encode a plethora of information such as age, gender, health, and disposition, among others, where we are known to make a variety of implicit social judgments based on those characteristics [30–32]. Hence, by exploring those implicit judgments, morphological characteristics in the face, descriptive of symptoms of depression, can be identified, which could, subsequently allow for depression-specific facial expression-based scales to be developed.

The concept of using facial expressions for self-reporting mood is not novel. Initially, two facial expression-based scales were developed, both relying on schematic representations, where The face scale uses a drawn androgynous character portraying a range of emotion intensities [21], while the Self-Assessment Manikin (SAM) depicts the principal components of valence, arousal, and dominance of the dimensional model of emotion [33]. Schematic representations of facial expressions, however, are of low-fidelity and cartoon-like, which cannot truly capture and convey the complexity of realistic facial expressions in all subtle variations. Consequently, such tools are inherently limited in the range of expressions or intensities they can portray.

Alternatively, in the field of Human-Computer Interaction (HCI), alternative representations for mood were explored by using emojis [34, 35], textual descriptions, colours [36, 37] or imagery [38–40]. Colours, or imagery, while expressive are rather abstract means to represent emotions and cannot reliably disambiguate emotions or their intensities. Textual descriptions, on the other hand, rely on a language proficiency, are susceptible to misinterpretations and consequently are not particularly suitable to be used for quick and frequent daily assessments. Additionally, the interpretation of emotion or emotion intensities in emojis is found to be inconsistent between persons [41].

Most recently, the Dynamic Visual Analogue Mood Scales (D-VAMS) [27], comprised of photographs of real persons enacting different expressions and expression's intensities was developed. While using realistic facial expressions, photographs offer little in terms of personalisation or customisation. Hence, leveraging machine learning methods for generating facial expressions can further

the development of facial expression-based technologies in their application for self-reporting mood. In addition, with the ubiquity of smartphones it is possible to provide self-reports in situ using smartphone prompts, enabling frequent assessments. The combination of machine learning methods for creating granular facial expressions used within a smartphone application could allow users to have better control over the facial expressions by matching the type and intensity of emotions they experience and self-report their mood quickly and unobtrusively to their day-to-day activities.

1.2 Research aim

The main aim of this thesis is to explore the use of facial expressions for representing and capturing mood or affective states. It was explored through the design and implementation of prototypes, each synthesizing a core idea, and evaluated within empirical studies with human participants. The evaluation consisted of a comparison to another established method and attempted to identify the strengths and weakness of the approach. Features, salient to users, were identified which steered the development of prototypes in subsequent studies and outlined future research directions. Furthermore, this thesis also aimed to identify technologies which can generate expressions portraying a range of intensities of emotion from arbitrary image-based representations of facial expressions. The developed prototypes were designed to be easily accessible by being used on commodity hardware, such as smartphones. Additionally, this thesis also includes an exploration of depressions' symptomatology, where symptoms were visualized as facial expressions as an initial step towards creating disorder-specific self-report instruments. Herein, throughout the thesis, the explored ideas and their embodiment within prototypes gradually increase in complexity. Formally, the aim of this thesis is addressed through the following research questions:

1.3 Research questions

A) How can facial expression-based methods be evaluated as a valid way to self-report mood? How would such tools compare to an established method, such as traditional numerical-based scales?

B) Which aspects of facial expression-based tools are valuable to users and which further capabilities are desired?

C) Which technologies can generate expressions depicting a range of emotion intensities using images of arbitrary expressions? Could those be used within applications on commodity hardware, such as smartphones?

D) Are there specific facial expression-based representations descriptive of affective states? If so, could those be used within self-report tools?

These research questions are broken down into the following objectives:

1.4 Research objectives

a) Develop a method for person-agnostic generation of facial expressions portraying various levels of intensity.

b) Empirically validate the use of facial expression-based interfaces for self-report of mood in naturalistic settings.

c) Identify through participant feedback qualitative requirements for facial expression-based technologies.

d) Continuously refine investigation prototypes through iterative design incorporating participant feedback at each step.

e) Explore whether and which symptoms of depressions' symptomatology possess distinct features identifiable in a facial expression.

1.5 Thesis' Main Contributions

Two types of contributions are derived from this thesis, described as technological and theoretical.

1.5.1 Technological Contributions

Data augmentation method for machine learning generation of person- and expression-agnostic facial expressions

This contribution consists of the application of Delaney Triangulation, a known technique that can morph arbitrary facial expressions or identities and create plausible intermediary blends thereof. Those synthetic representations were subsequently applied within machine learning generation and recognition tasks. Within the generative task, the application of the augmentation resulted in an increased quality in the generated faces, particularly so for those portraying intermediary intensities of emotion. For the recognition task, the augmentation allowed to encode richer facial feature variations for expressions of nuanced intensities, subsequently improving facial expression recognition results. In addition, the blending factor used for the creation of each augmented images created plausible soft labels for the facial expression category and intensity, that can be used in both generative and recognition tasks. The development and evaluation of this technique is described in Chapter 5.

Prototypes for bidirectional facial expression-based mood self-reports

This contribution consists of two smartphone application prototypes featuring a sad to happy facial expression scale using 1) the Dynamic Visual Analogue Mood Scales (D-VAMS) [27] and 2) computer-generated facial expressions using a generative model (FEAS). Those prototypes were used and evaluated within two empirical studies consisting of a mood-elicitation experiment and a mood-monitoring in the use of facial expressions for mood self-reports. The prototypes and their evaluation were described in Chapters 4 and 6 respectively.

Prototype for multidimensional facial expression-based mood self-reports (MFEAS)

This contribution consists of a smartphone application prototype featuring a multidimensional facial expression based scale consisting of the expressions for happiness, sadness, anger, disgust, fear, and the neutral expressions as created by a generative machine learning model. The interface is designed to accommodate an arbitrary number of facial expressions using a heuristic that organizes distinct facial expressions portraying varying intensities of emotion within a polar coordinate system inspired by the dimensional model of emotion [9]. The prototype was evaluated in a mood-elicitation laboratory experiment in Chapter 8.

Prototype for contextualized multidimensional facial expression-based mood self-reports and historical mood visualisation (EmotionAlly)

This contribution consists of a prototype (EmotionAlly) which allows users to provide mood self-reports using a multidimensional facial expression-based method, attach contextual information to their self-reports and visualize their mood through facial expressions using context or time as filters. The prototype encapsulates results from the iterative process of collecting user feedback from three empirical studies evaluating previously developed mood self-report prototypes in Chapters 4, 6 and 8. EmotionAlly itself is described in Chapter 10.

1.5.2 Theoretical Contributions

A heuristic for allocating arbitrary number facial expressions of varying intensities of emotion over a polar coordinate system

This contribution consists of a heuristic for the spatial allocation of an arbitrary number of facial expressions at varying intensities of emotion. The heuristic unambiguously maps the coordinates of a polar coordinate system to distinct facial expression and their intensities. It also positions emotions adjacently according to the likelihood of them being co-experienced using the dimensional model of emotion [9] as an emotion proximity metric. In addition, within this organization, blended expressions are created for each pair of adjacent expressions, bridging the

space between them. This heuristic was created as part of the development of a multidimensional facial expression-based prototype for mood self-reports, described in Chapter 7.

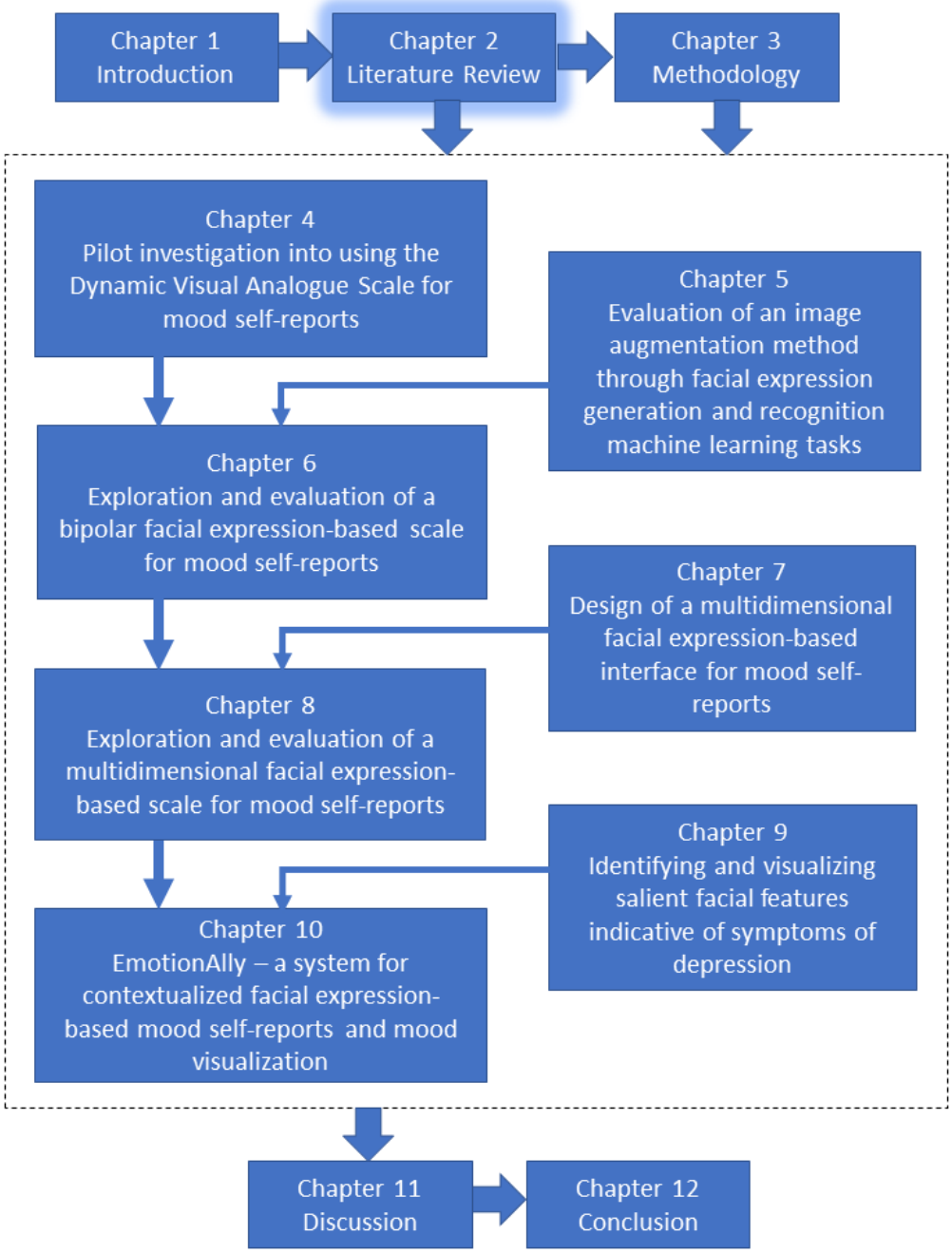
Visualisation of symptoms of depression as facial expressions

This contribution consists of 12 distinct models of symptoms of depression, descriptive of how laypersons perceive those to be reflected in the face. The symptoms were derived from the PHQ-9 (Patient Health Questionnaire) screener questionnaire [42] from 6 questions measuring unipolar symptoms and 3 – bipolar symptoms. The representations were created using a known reverse correlation classification images (CI) technique [43] using data collected in an online experiment with over 600 participants. Those models were described in Chapter 9.

1.6 Thesis Overview

Chapter 4 describes the use of an existing facial expression-based scale which was integrated within a smartphone application and evaluated quantitatively and qualitatively in a pilot study. The prototype used photographs of real persons' expressions of happiness and sadness and was evaluated using vignettes, i.e. short stories eliciting emotions such as amusement, awe, compassion and inspiration, to elicit positive and negative emotions. **Chapter 5** describes the use of a data augmentation method applied on the RafD facial expression dataset [44] and used in a machine learning generative and recognition tasks improving both the quality of generated facial expressions and facial expression recognition results. Thereafter, in **Chapter 6**, a bipolar happiness-sadness facial expression assessment scale (FEAS) was developed using the computer-generated facial expressions from the generative model described in Chapter 5 and the prototype design from Chapter 4 and was subsequently evaluated within a mood-monitoring experiment. **Chapter 7** describes the development of a prototype for expression-agnostic multidimensional facial expression-based assessment interface (MFEAS) using computer-generated facial expressions provided by the generative model described in Chapter 5. Then, in **Chapter 8**, MFEAS was evaluated in a lab experiment, using mood elicitation images from the International Affective Picture System (IAPS) [10]. The studies described in Chapters 6 and 8 were conducted in parallel and involved both quantitative and qualitative data collection and analysis. **Chapter 9** details an exploratory study, which identified facial features descriptive of symptoms of depression in an experiment conducted online. The visual representations for those symptoms were created using a reverse-correlation classification imaging (CI) technique [43]. Finally, **Chapter 10** describes EmotionAlly – a prototype system for self-reporting and visualising historical mood in a context-aware manner. In part, EmotionAlly was designed by integrating quantitative and qualitative findings from preceding chapters.

Chapter 3, Table 3.1 (p. 43) contains a thesis overview which links chapters, conducted studies, research objectives addressing the defined research questions, and employed research, data collection, and data analysis methods.


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 2

Literature Review

2.1 Introduction

This thesis aims to contribute to the space of self-assessment methods. Therein is an evaluation of the use of facial expressions to capture and represent mood as well an exploration of novel facial expression-based representations for affective states associated with depression's symptomatology. To achieve this aim, this thesis draws from the fields of Computer Science, Psychology, and Human-Computer Interaction (HCI), where those intersect within the domains of affective and digital health technologies. Specifically, this thesis contributes knowledge on facial expression-based mood self-assessment methods, design of facial expression-based technologies and furthers knowledge on depression's symptomatology.

Section 2.2 introduces prominent models of emotion, insights into the formation of emotions and their taxonomy and the interplay between emotions and mood in how they affect and modulate each other. Section 2.3 introduces tools from the fields of HCI and Psychology for self-reporting mood including facial expression-based ones. Section 2.4 introduces the relationship between facial expressions and particularities in how they are perceived. As this thesis investigates in part the utility of facial expressions indicative of symptoms of depression, Section 2.5 introduces affective disorders and depression in particular, its symptomatology, severity assessment methods and its influence on the perception of facial expressions. Finally, Section 2.6 bridges the topics introduced thus far and elaborates on their interconnectedness in the context of using facial expressions for mood self-reports.

2.2 Models of Emotion

Models of emotion are important for understanding facial expressions as the latter are an expression of distinct emotion states. In order to understand the function and origin of facial expressions and subsequently investigate their use in technological tools, an introduction of the two most widely accepted theories of emotion is required. Those are the Basic Emotion Theory (BET) and the dimensional theory

of emotion. Most commonly, those models are considered to be at odds with each other as the former posits that emotions are discrete entities, while the latter claims that emotions are mostly described by two independent dimensions of valence and arousal.

2.2.1 Basic Emotion Theory

BET posits that human beings have a limited number of basic emotions which are biologically hard-wired with distinct neural and physiological patterns [45, 46]. In this interpretation, BET conceptualizes emotions to be holistic, believing them to be distinct in their neural fingerprint. Contemporary understanding defines the basic emotions to be those for anger, disgust, fear, enjoyment, sadness and surprise [45]. They are thought to play an important role in addressing fundamental challenges for survival, where, for example, fear and anger are essential to the fight-or-flight strategies. Additionally, it is believed that basic emotions are integral in social functioning as well in communicating actionable information about the individual or the environment. Basic emotions are also thought to have persisted due to their biological and social utility in evolution and adaptation [47]. Furthermore, BET posits complex emotions such as amusement, awe, contentment, desire, embarrassment, pain, relief, sympathy, among others, to be constructed through the combination of multiple basic emotions and cultural conditioning [47]. Additionally, a common belief is that basic emotions are expressed unequivocally through distinct facial expressions of emotion. Those basic emotions are considered to be universally present and expressed in every culture[25]. However, such conceptualization has been criticized, disputing the composition of basic emotions and their cultural universality [48]. Most notably, a proposal of four basic emotions was made which posits the emotions of joy, sadness, fear and anger to be basic, while the ones for disgust and surprise are believed to have developed later for social functioning [49], a claim later reiterated [50, 51].

2.2.2 Dimensional Theory of Emotion

Earlier accounts on the dimensional theory of emotion posit that emotions are underpinned by three independent dimensions: pleasant-unpleasant, tension-relaxation, and excitation-calmness [52]. Subsequent research investigated this dimensional view and the contribution to variability of those principal components, where it was found that, in fact, two of those principal components are overlapping. This resulted in a new taxonomy of emotions defined by the dimensions of valence (pleasant-unpleasant) and arousal (active-passive) where emotions were arranged within a Circumplex Model of Affect. Valence, defines the positive or negative affect associated with an emotion, while arousal reflects the intensity of the experienced emotion, i.e. how agitating or stimulating it is. Additionally, a third dominance-submissiveness dimension was later identified, accounting for a small, but significant portion of variance in emotions [53]. The Circumplex Model

of Affect provides a visual representation of where and how emotions are located within a polar coordinate system according to their decomposition into the principal components for valence and arousal [9]. Additionally, biological accounts on the dimensional theory of emotion proposed that all affective states arise from cognitive interpretations of core neural sensations, which are, in turn, a product of two independent neurophysiological systems corresponding to its principal components [54]. Criticisms of the dimensional theory stem from the claim that its principal components are arbitrarily chosen [55, 56] as no neural structures that correspond and map the dimensions for valence and arousal have been consistently found [57].

2.2.3 Hybrid theory of emotion

Both BET and the dimensional model of emotion have been considered to be in competition with one another as they provide a different and somewhat conflicting interpretations of emotions. While both theories of emotion claim the genesis of emotions to be found in the brain and neural circuitry, they differ in the composition of their constituting elements. The former implies that each emotion has a distinct fingerprint in the brain, while the latter – that the principal components of the dimensional model are biologically-based structures. However, so far neither of them has yet produced compelling evidence proving itself or disproving the other. Both models however, provide functional value as a foundation for emotion research as there is sufficient evidence for regarding emotions both holistically and as constituting of a combination of principal components.

As proposed by multiple researchers over the years, these differences are not irreconcilable [57–60] and the two models could be, in fact, complementary to each other. Under the supposition that basic emotions also exhibit the same characteristics of intensity and pleasantness, there is no inherent contradiction between them [60]. Investigating the biological basis for both models, findings indicate that when observing facial expressions, both discrete emotion classification as well as a dimensional assessment of affect transpire simultaneously in the brain [59]. As such, their integration within a hybrid theory of emotion, which accounts for both categorical and dimensional processing of emotion, appears to be feasible. In the context of this thesis, this hybrid model of emotion offers the working framework for the approaches and solutions provided herein.

2.2.4 Emotions and Mood

Emotions and mood are often used interchangeably, particularly so in technical literature. However, those are fundamentally different concepts, distinctly defining internal emotion states. Contemporary understanding of mood and emotions categorizes moods as slow-moving diffusive states that can last for hours or days, while emotions, have a shorter duration, ranging from seconds to a few minutes [61]. Intense emotions or continuously self-reinforcing ones, however, can have a lasting impact on mood. For example, the experience of a negative emotion of

a sufficiently high intensity could dampen down or even negate the experience of an elated mood or alternatively heighten a negative one. Conversely, moods can also exert influence on how emotions are experienced and reinforce or negate them. This interdependent relationship between emotions and mood adds a layer of complexity to emotion research as empirically, neither mood nor emotions can be captured independently of one another. This is why both terms can be seen in scientific literature as used interchangeably, due to how they are related to one another, but fundamentally refer to completely different concepts.

Biologically, emotions are more clearly understood in a way that a triggered emotion coincides with measurable physiological and chemical changes in the body. Discrete emotions trigger the activation of specific neural circuitry and consequently neurochemicals are simultaneously released in the body [62], while moods are harder to their diffusive and longer-lasting nature. Changes in mood have a relatively low variance over time which makes them difficult to distinguish from background bodily or neurological processes. Current understanding of emotion theory suggests that some emotions are expressed through distinct facial expressions [25]. On the other hand, moods are not thought to be associated with distinct expressions, however facial expressions are considered to be valid indicators of mood [63, 64]

To measure emotions, empirical studies traditionally relied on emotion elicitation tools such as images or video segments. Those aim to evoke an emotional response, which is subsequently measured using various instruments. However, as moods are long-lasting and change rather slowly over time, it is difficult for mood to be reliably measured, or to account for variability in someone's mood. Although, the inseparability of both constructs makes it difficult to ensure whether assessments are indicative of mood or emotions, in cases where no discernible emotion-eliciting event has taken place, the assumption is that the assessment measures mood. Conversely, in the inverse case, the assessment would be indicative of the felt emotion.

2.2.5 Section Summary

This section presented the two prominent theories of emotion – the basic emotion theory and the dimensional theory of emotions. In sum, BET distinguishes between basic and complex emotions and views basic emotions holistically as distinct and inseparable entities, whereas complex emotions arise through the combination of basic emotions. Dimensional theory of emotion posits the existence of principal components which map out the space for all emotions. It arranges emotions based on the dimensions of valence and arousal where a spatial organization emerges allocating emotions within a two-dimensional Circumplex Model of Affect. Contemporary research proposes the fusion of both models into a hybrid model as they are perceived rather as complementing one another. This hybrid model neither disputes the existence of principal components nor the existence of basic emotions, but rather accommodates both within a single paradigm of hierarchical organization of emotion which accommodates BET and the taxonomy of basic and

non-basic emotions and both their decomposition into the principal components of the dimensional models of emotion. Finally, the difference between emotions and mood was introduced, how they mutually affect each other and the challenges they pose for emotion research in separately measuring either construct.

2.3 Methods and Tools for Assessing Mood: Analogue and Digital

This section introduces common methods developed in the literature and used in practice for assessing mood. It highlights their strengths, weaknesses, and their utility for representing or capturing mood. Additionally, the Ecological Momentary Assessments (EMA) framework is introduced as a means to provide quick assessments on easy-to-use instruments, frequently during the day, and the benefits of using EMAs.

2.3.1 Numerical or graphical scales

Numerical or graphical scales are well-known and used tools to measure attitudes. Likert-scales [18] typically feature 5, 7, or 9 numerically indexed discrete points indicating levels of agreement or disagreement with a question or a statement. They can be uni- or bipolar, where in the latter, the mid-point of the scale is used to indicate a neutral or impartial attitude [18]. Additionally, Likert-scales often feature anchors, typically as a single word or a short textual description, denoting the meaning at each extreme of the scale. Similar to Likert scales, Visual Analogue Scales (VAS) [19] are used to capture attitudes in an analogue manner. Therein, instead of discrete points, VAS are represented as a line, which allows participants to mark their response towards a particular subjects on any point over that line. The subsequent assessment is then quantified as a distance from either of the extremes in unipolar scales, or the centre in bipolar scales. VAS can also feature indicators over their range, denoting intervals or levels of agreement, designed to aid a person in providing their assessments. The difference being, that Likert-scales measure attitudes through a range of predefined discrete values, while VAS – as a continuous scale. Both scales are powerful instruments, as they allow capturing subjective opinions, attitude or feelings towards a particular subject, or an aspect of it as a numerical representation.

Despite their extensive use, however, there are criticisms regarding how those numerical scales are perceived by users and subsequently numerically quantify attitudes. For example, the presence of anchored labels or markings denoting intervals on a VAS scale is shown to prompt the coalescence of assessments near those indicators, while in-between ranges remained depressed [22]. VAS scales are found to be most precise when featuring a real-time feedback component, e.g. as a percentage, numerical value or through other means of feedback [22]. Additionally, distances between marked regions on a scale are not always perceived equally,

resulting in an offset in the provided assessments [20]. Additionally, the choice of a numerical range in Likert Scales (e.g. $[-3, 3]$ or $[0, 7]$) and their arrangement within a scale can, in turn, influence the interpretation of anchors [24]. Lastly, an end-aversion bias is identified in bipolar scales where regions lying towards the extremes of a scale remain underutilized [23].

In sum, assessments provided through numerical scales can be influenced by their textual labels (e.g. anchors), numerical indexing, or the number and frequency of indicators denoting increments on the scale, where each of those factors may impact the accuracy of provided measurements. With respect to measuring mood, scholarly works point out that mood is a highly subjective and often complex construct, difficult to be translated in a numerical equivalent [20, 21].

2.3.2 Abstract or Representational

The field of HCI has also investigated ways to assess mood as increasing one's awareness of their mood has a positive impact on their mental health and well-being [15]. Specifically, tools have been developed to capture and represent mood through modalities such as colours [36, 37] or abstract imagery [38–40]. For example, some such instruments are the Photographic Affect Meter (PAM) [38] which uses emotionally affective images to portray mood according to ratings of valence and arousal. It consists of 16 images, representing different points according to the circumplex model of affect, spatially allocated within the two-dimensional space of valence and arousal. Similarly, The moment [36] is an application that uses colours to span an emotion space, where colour shades represent the intensity of an emotion. While such approaches provide an easy and accessible way to represent mood, they are limited to the amount of emotional intensities they provide [65]. Additionally, while very expressive, colours cannot unambiguously represent distinct emotions or or abstract imagery are limited to the valence-arousal dimensions of the dimensional model of emotion. For those reasons, such tools can be useful for self-assessing one's mood, however at the expense of low resolution in their assessments.

2.3.3 Facial expression-based methods

The use of facial expressions to represent mood is not novel. Initially, the Face Scale was developed containing a variety of facial expressions varying in the intensity of portrayed emotion intended to be a non-verbal method for self-assessing mood, bridging the gap in literacy [21]. The Face Scale consists of 20 schematic faces spanning a range of emotions at incremental intensities from sad to happy and is aligned with BET. Subsequently, the Self-Assessment Manikin (SAM) was developed depicting the dimensions of valence, arousal and dominance [33]. SAM features a drawn mannequin using pictorial representations of those dimensions at different intensities, in concordance with the principal components of the dimensional model of emotion. Both instruments adhere to a particular model of emotion, and both have been successfully used for assessing mood [21, 33]. However, it is known

that schematic representations reduce the complexity inherent to realistic facial expressions [66]. Subsequently, schematic or drawn faces are of low-fidelity and consequently, are limited in the range of nuanced intensities they can portray.

More recently, the widespread use of smartphones and smartphone applications has allowed for the conceptualization and practical application of more modern facial expression-based ways to represent emotions, such as smileys [34, 35, 67]. Smileys are typically used to supplement emotional context to text-messages and while widely successful in achieving this task, they are, in essence, exaggerated portrayals of categorical emotions. That being said, smileys, similar to schematic or drawn faces are of relatively low fidelity and lack the ability to represent nuanced expressions, where in addition the interpretation of smileys between persons has also been found to be inconsistent [41].

Alternatively, scales using realistic faces have also been employed for self-reporting mood. One such tool is the Dynamic Visual Analogue Mood Scales (D-VAMS) [27] which features photographs of a man and a woman enacting various facial expressions arranged on a number of bipolar scales. While the scale is close to reality in using real facial expressions, the use of photographs limits the scale's range to a number of predefined static images portraying a fixed range of expressiveness, is not easily extensible to incorporate additional expressions and are difficult or nigh impossible to personalize beyond the predefined identities and expressions.

In sum, schematic faces are constrained to be low-fidelity cartoon-like representations of emotions and have limitations to the range of emotion intensities they can portray. On the other hand, photograph-based scales are fixed to a number of predetermined emotions and ranges of emotion intensities, while emojis do not provide sufficient nuance to be able to represent emotions beyond a categorical state. Hence, new methods for creating realistic-looking facial expressions using advancements in machine learning methods may allow to capitalize on creating realistic-looking facial expressions for mood self-reports, which are flexible in the range of intensities or expressions they can provide, while allowing a degree of customisation to users.

2.3.4 Ecological momentary assessments

Ecological momentary assessments (EMA) is a form of assessment, widely used typically in smartphone applications for patient monitoring and tracking. EMAs rely on notifying users periodically during the day with prompts to self-report on one or more modalities such as mood, behaviour patterns, symptoms, among others. They are extremely useful and have been used within a wealth of applications and scientific studies [6, 68, 69]. Additionally, EMAs are considered the most accurate form of self-report [70], as they allow a user to report on events close to their time of experience, thus mitigating inaccuracies in recalling past events, known as the retrospective recall bias [71]. Typically, they ensure that self-reports have ecological validity, i.e. that assessments generalize well in real-life settings [17]. EMAs rely on simple measurement tools such as the Likert or Visual Analogue Scales (VAS) [18,

19], which allow for assessments to be made quickly with little to no obstruction in day-to-day activities [70].

The research in this thesis will feature the development, design and evaluation of multiple smartphone application prototypes for self-reporting mood using facial expression feedback. As such, the use of EMA is pertinent to this thesis and will be applied throughout as a requirement in the development of various prototypes.

2.3.5 Section Summary

In this Section, tools from the fields of HCI and Psychology were introduced giving an overview of traditional methods to self-assess mood, their strengths, weaknesses and practical utility. Additionally, facial expression-based methods were introduced and possibilities of further developing such technologies were outlined leveraging novel advancements in machine learning methods for improving upon those approaches. Finally, the concept of ecological validity obtained through the use of EMAs was elaborated on, establishing its need as a prerequisite for improving the quality of assessments for, among others, mood self-report tools.

2.4 Facial expressions for mood self-reports

Facial expressions are inherently linked to emotional experiences and are powerful visual indicators to communicate our emotions to the surrounding world [28]. These are embodied representations and typically we are well-versed in enacting them or recognizing those in others. Facial expressions are externalized representations of affective states [72] and embody discrete emotions, making them a suitable medium to represent mood [21, 27, 63, 64]. Mood and emotions are different constructs, but while mood is not distinctly associated with a particular facial expression, a collection of emotions, however, are considered to be valid indicators of mood [21, 27, 63, 64]. A key benefit of using facial expressions for mood self-reports is that they are a non-verbal method. The advantage of using facial expression-based instruments is rooted in the biological basis of expressions, where they are considered to be a set of biologically rooted signals evolved for the purpose of communicating emotions [73]. An implication of this biological basis is that expressions are by large invariant over time and within-cultures. Therefore, using facial expressions as feedback for self-reports is generally more accessible, requiring no introspection, language proficiency, or familiarity with terminology.

2.4.1 Emotion Signaller-Decoder Framework

Facial expression processing and enactment have been traditionally regarded as separate paradigms, occurring independently [74]. However, in the context of communicating emotions, the enactment and interpretation of emotions are in fact interdependent [74]. Effectively communicating emotional information relies on the

enactment of an appropriate facial expression and its intensity. Similarly, effectively decoding that information from a facial expression relies on correctly inferring the categorical class and intensity of the portrayed emotion. In a typical social context, this relationship is burdened with multiple complexities. For example, social desirability bias and agency towards a particular social goal may prompt a modulation in one's own emotional response [75, 76]. However, within the context of an interaction with a facial expression-based interface, social factors are largely irrelevant.

Facial expressions have a particular signature comprising of different facial features, their spatial arrangement at the highest emotional intensity, and duration of the enactment of an expression. Therefore, an expressions' dynamics in terms of its facio-musculature activation and its unfolding over time is descriptive of emotion intensity [25, 26]. Facial expressions at their peak are clearly differentiated from one another, hence they are easy to be recognized categorically and cannot be interpreted as belonging to more than one emotion classes [25, 26]. Conversely, the experience of a weak, but sufficient in magnitude emotion will induce an expression of a minimal intensity in the signaller [77]. This is mediated by the perceived salience of the triggering stimuli and the signallers' own facial expressiveness. Factors which may impact facial expressiveness are, for example, congenital impairment, or other health conditions, among others [78]. For a decoder, the ability to detect a facial expression and its intensity are impacted by their facial expression literacy, such as individual differences, cultural or group familiarity, temporary or permanent impairment caused by a medical or congenital condition, individual learned experiences, cultural conditioning [48], among others.

Accounting for variations in producer's expressiveness to convey facial expressions and decoder's facial expression literacy to recognize and correctly interpret them creates an interpretation map between these actors [77]. Some subtle experienced emotions may lay beyond the ability of the signaller to express, or the ability of the decoder to accurately discern them; this is also the boundary where expressions may be misclassified. In short, for each subtle facial expression there is one at an intensity that is unequivocally produced by the signaller and recognized by the decoder as such, while those of higher intensities or at their peak are easily distinguished and recognized as they lay beyond this perceptual threshold. In the context of interfaces relying on facial expressions to convey emotions at varying intensities, technologies which are used to create those expressions need to be granular such that they can create a range of nuanced expressions which can be unequivocally recognized as such by users.

2.4.2 Individual and group variability in the interpretation of facial expressions

Faces encode a plethora of information and are the basis for making social judgments such as estimating age, gender and others [79–81]. The accuracy in perceiving

facial expressions and adequately estimating their emotion intensity is linked to ones' familiarity with a particular social or cultural group [82–84]. An analysis of the universality of facial expressions identifies that there are cross-cultural variations in the interpretation of facial expressions [48]. Additionally, works examining smiles in heterogeneous and homogeneous populations identify that persons belonging to more diverse societies adopt a facial expressiveness which is more accurately recognized across cultures [85]. It is known that cultural or social familiarity translates to more efficient and clear non-verbal communication through facial expressions. Furthermore, multiple studies identify increased accuracy in attribution of emotion to emotional faces favouring those within the same culture group [86–88].

Gender-differences also play a role in the attribution of intensity to emotional faces. In general, people tend to be better at recognizing the expressions of their own gender [89]. Beyond being more expressive, women appear to also have an advantage in identifying and estimating emotion from faces [89, 90]. Finally, there are also individual differences in the ability to discern emotions from faces [77].

In sum, individual- and group-differences are an important aspect to consider in building facial expression-based interfaces. Technologies used to create facial expressions need to be generic such that they can be customized and tailored to a users' social and cultural familiarity in allowing them to correctly infer emotion and its intensity from a face.

2.4.3 Section Summary

In this section, understanding of emotional facial expressions and how they are perceived were presented, as well as factors pertaining to variations between cultures and among individuals within the same group. In particular, facial expression interpretation appears to be influenced by a variety of factors which contribute to the requirements in the design and implementation of facial expression-based tools. Specifically, this highlights the need for customizable interfaces, which are generic such that they would be able to accommodate to users' variations in their perception and familiarity with the enactment of facial expressions.

2.5 Affective disorders: Depression

Affective disorders are a group of conditions including depression, anxiety and bipolar disorder. Those are recurrent in nature, defined by periodic episodes of depression, mania or anxiety, during which certain behaviours or symptoms are exhibited. The main characterizing feature of depression is dysregulation in mood [91]. Mood dysphoria persists throughout the depressive phase of the disorder and is the most characteristic and pervasive symptom. Mood, however, also tends to naturally fluctuate throughout the day following a distinct diurnal pattern [92]. Such daily variations are an important and informative modality to capture due to

their centrality to the symptomatology of the disorder. Effectively capturing mood can be useful in identifying patterns preceding the development of a depressive episode [25].

2.5.1 Assessing depression through screener questionnaires

To diagnose depression a patient would usually undergo a psychiatric evaluation with a clinician. An important tool, aiding the diagnosis, are questionnaires which assess the presence and severity of symptoms in an objective and quantifiable way. Questionnaires are a standard part of clinicians' tool-set and have been reliably used in the therapeutic practice, i.e. Patient Health Questionnaire (PHQ) [42], Hamilton Depression Scale [93], Montgomery-Åsberg Depression Rating Scale [94]. Mostly, those questionnaires aim to assess a patient on the presence and perceived severity on a number of symptoms. The questionnaire scores are typically summed up and the final score indicates an absence or presence of mild, moderate, or severe depression. A patients' score is considered an objective measurement of their overall state of depression.

Nevertheless, questionnaires inherently possess a number of drawbacks. Their assessments are retrospective as they inquire about the presence and severity of symptoms within a period of two weeks to two months in the past. This type of retrospective assessment is susceptible to retrospective recall bias [95, 96], i.e. an incomplete or altered recollection of past events influenced by a persons' current mental state (e.g. recalling a generally positive memory with a negative connotation by a person living with depression). While such tools are effective in measuring the overall presence and severity of depression, they are unsuitable for frequent use. When administered more frequently than once every two weeks, the quality of provided assessments degrades significantly [97]. In addition, the practicality of frequently filling out a questionnaire requires moderate investment in time and effort, hampering its utility as a monitoring instrument.

Furthermore, questionnaires are known to obscure the symptom-composition of depression. Symptoms are scored on a numerical scale ranging from their absence to their consistent or frequent manifestation over a specified time period. The final score is calculated as the cumulative sum of individual symptom scores, where the demarcation between different levels of depression (e.g. mild, moderate, and severe) is based on how the final score measures up to predefined thresholds. In fact, recent studies established a baseline prevalence frequency for symptoms of depression in the distribution pattern of Patient Health Questionnaire (PHQ-9) [42] responses within the general population and found that some symptoms occur more frequently than others [98]. Moreover, symptom severity appears to diverge over the disorders' course-trajectory, indicating that depression may have multiple progression pathways [99]. Findings have also shown that symptoms have differential outcomes, whereas some are more likely to persist years after the formal diagnosis [99]. Therefore, questionnaire scoring may be flawed as it presupposes that symptoms are disjunct from one another and contribute equally

to the overall disorder severity. However, depression is a complex disorder with symptoms connected in a dynamic network of causality [99]. When comparing PHQ-9 to a semi-structured interview conducted by an experienced diagnostician, a meta-analysis concluded that PHQ-9 substantially overestimates depression prevalence [100]. Therefore, questionnaire-based assessments cannot be used as a single source of truth for diagnosing or assessing depression severity. While questionnaires are easy to access and use to assess one's depression severity, new tools may bridge the gap in being able to quantify depression severity, ideally in a way that accounts for individual symptoms and is suitable for frequent use.

In conclusion, using questionnaire based diagnostic instruments warrants caution. A reconsideration in how those instruments are scored in favour of a weighted-average approach considering the prevalence in the general population as well as symptom frequency may be more adequate. Questionnaires also rely on the correct interpretation of language in order to accurately score symptoms, where the accessibility to medical terminology may also lead to its misappropriation or misuse by the public. Furthermore, future work aiming to address this need should focus on a symptom-aware approach in building assessment tools. Assuming that symptoms of depression have a unique, distinct, and characteristic fingerprint expressed in the face, facial expression-based tools would be in a unique position to allow users assess their symptoms non-verbally where those assessments could also be administered frequently as EMAs provided suitable tools are developed.

2.5.2 Depression-induced biases in perceiving facial expressions

Depression appears to affect the way emotional faces are perceived and interpreted [101–105]. The literature distinguishes three types of biases characterizing the facial expression-processing disturbance: i) an **attention bias** in selectively attending to or dwelling on particular facial expressions, ii) a **processing bias** expressed as an increased latency in classifying facial expressions, and iii) a **perceptual bias** represented as misinterpreting emotion categories or intensities in specific facial expressions. A comparison between schematic faces, words and realistic facial expressions showed that those biases appear to manifest only when perceiving real human faces [106–108]. This implies that schematic faces may be perceived as a categorical representations of emotion, more similar to emotion words and as such, they are likely to possess exaggerated features making those categorical distinctions clearer at the expense of losing the complexity inherent to realistic faces [66].

Attention bias The attention bias can be measured as number of fixations drawn towards a particular facial expression, and dwelling, measured in the duration of those fixations. Literature indicates that depressed patients and those in remission tend to direct their gaze more often towards sad faces [101, 102]. Furthermore, non-clinical and clinically depressed patients tended to dwell significantly longer

on sad faces [109, 110] and to a lesser extent on negative faces [111]. Interestingly, also patients in remission were identified to selectively tend to sad faces [102]. Conversely, healthy participants were found to more often direct their gaze towards faces displaying happiness and avoid those of sadness [102], an effect which was not observed in either depressed patients or those in remission. The difference in tending to particular facial expressions between healthy participants and those with depression is further reinforced as depressed patients were found to dwell marginally less on happy faces [110].

Processing bias The processing bias is characterized by a latency in detecting, or reacting to a facial expression. Albeit limited, literature suggests a difference between depressive patients, those in remission and healthy controls. Depressed patients showed an increased reaction time to all facial expressions, an effect particularly emphasized in their response towards neutral faces [103, 104]. Additionally, they appear to be faster in recognizing happy expressions compared to sad and neutral ones, an effect also observed in healthy participants, but absent in sub-clinically depressed [104]. Patients in remission also appear to respond particularly slow to neutral faces, however they did not exhibit the same overall impairment for all expressions [104]. In contrast, healthy participants were quicker to respond to neutral expressions than sad ones [104]. In contrast, no significant difference in response times was found between a cohort of depressed and non-depressed women [112].

Perceptual bias The perceptual bias is expressed as an impairment in the ability to correctly identify emotion in a facial expression either categorically or by misinterpreting its intensity. A meta-analysis investigating emotion recognition acuity of depressed patients identified a general deficit in recognizing all basic facial expressions of emotion [25], except for that of sadness [105]. This analysis encompassed 22 papers examining responses to the six basic facial expressions [25] and is conclusive that there is an emotion recognition impairment in depressed patients [105]. However, some potential sub-group differences deserve further attention.

Consistent with previous findings, depression induces an overall impairment in the recognition of more subtle expressions of happiness and sadness [104], the neutral expression [113, 114] and that for anger and fear [112]. Furthermore, depressive patients require higher intensities of emotion compared to healthy participants to correctly identify the presence of emotion in all basic facial expressions of emotion [115, 116]. Sadness is an exception to the rule, where depressed patients appear to require less expressiveness in the face [115, 116]. As expected, recognition accuracy in both depressive and control groups increased when presented with more expressive faces, where the rate by which the recognition accuracy improved is inversely correlated to severity for those with depression [116]. Patients in remission and sub-clinical populations required more intensity of emotion only for the facial

expression of happiness, while displaying an improvement in their overall perception of affect for all other expressions [103, 104, 113]. In some cases, depressives also misclassified neutral faces as sad [104], which hints at a misattribution of affect in ambiguous expressions towards sadness.

2.5.3 Quantifying depression severity using biases in the perception of facial expressions

In sum, the perceptual bias is the most reliable and consistently replicated among the three ones. There seem to be distinct characteristics of perceptual biases descriptive of 1) severe depression 2) sub-clinical depression and remission and 3) absence of depression. Additionally, the extent to which perceptual biases are exhibited appears to correlate to depression severity. For example, severe depression impairs the recognition of all mild facial expressions, i.e. those portraying low intensity of emotion, except that for sadness. Additionally, some evidence even suggests a heightened recognition of sadness in facial expressions. A subset of the general impairment is attributed to sub-clinical depression and remission, where the perceptual bias affects only the recognition of mild expressions of happiness. This impairment is not observed in the absence of depression. This classification may be used as a diagnostic instrument where capturing and quantifying these biases could be an effective and unobtrusive way to assess depression severity. As those biases are only observed in the perception of realistic faces, facial expression-based tools which possess sufficient granularity to portray a wide range of nuanced expressions can capitalize on those biases. Assessing depression severity through identifying patterns or trends in the use of facial expression-based tools could be a way to track and monitor patients unobtrusively simply through the use of providing EMAs for self-reporting one's mood.

2.5.4 Section Summary

In this section, questionnaire-based methods for assessing depression severity were introduced. Additionally, their strengths, weaknesses were outlined as well as their applicability for monitoring patients. A research gap was identified that could allow facial expression-based representations of symptoms of depression to be used as assessment instruments instead. Finally, biases known to affect the perception of or interaction with real facial expressions were introduced and their diagnostic utility for assessing depression severity through facial expression-based tools.

2.6 Integrating facial expressions within contemporary theory of emotion as a valid construct for measuring mood

The integration of facial expressions as a medium to represent mood needs to adhere to principles posited by emotion theories. Both BET and the dimensional model can be seen as providing different levels of abstraction with BET being a holistic interpretation, while the dimensional theory – a reductionist approach based on its principle components. From the perspective of both models, the use of facial expressions for recording mood is not new. As elaborated on earlier, schematic and drawn faces have already been used for assessment purposes as happy-to-sad scale which was aligned with BET and SAM – with the dimension theory [21, 33]. However, there is limited research to date in using real facial expressions for mood self-report instruments. In addition, there are interesting properties inherent only to the perception of realistic human facial expressions. Tools relying on the use of facial expressions have the potential of capturing and quantifying facial expression biases to assess depression severity, or assuming that symptoms of depression possess distinct facial characteristics could allow to assessing individual symptoms of depressions.

Contemporary technologies for generating facial expressions such as deep learning or generative adversarial networks have achieved impressive results and have an even more promising future [117–121]. At present, those techniques are able to generate faces with increasing level of detail, without triggering the uncanny valley effect [122], expressed as a person's emotional response of repulsion towards an object that appears to resemble a human being, but is not an exact representation of one. There is potential that the future development of those technologies will pave the way for digital identities [123] virtual avatars [124, 125] and others. As such, the approach described within this thesis is not out of place or untimely. Studies, evaluating facial expression scales have established that users generally prefer human-like representations to abstract numerical ones [126, 127]. With further advancement of machine learning and 3D modelling techniques, it will be possible to create realistic and plausible enactments of various facial expressions. In turn, those could be fine-tuned and personalized to portray emotions at different levels of detail, intensity or control the enacting identity or type of expressions. Hence, further development in generating realistic facial expressions or using them for self-reports would advance research in this area.

In the context of depression, tools relying on realistic facial expressions for self-reports can be a very valuable source of information. A facial expression interface relying on EMAs for continuous monitoring could not only be used for mood-monitoring, but those assessments may also be used to quantify depression severity as depression appears to influence the perception of distinct facial expressions. Hence, it could be possible to fingerprint distinct states of depression over its course-trajectory according to the exhibited perceptual alterations. A feasible

way to do that would be to use a tool which relies on EMAs integrated with facial expression-based self-reports. Assessments provided via such a tool would be unobtrusive to patients as the underlying quantification could be computed based on provided self-reports at no additional cost in time and effort to a user. Additionally, depression severity could be measured within-subjects as those biases in the ability of a depressed person to detect emotion in a face are subjective and the patterns in their assessments can be compared to those provided during absence of symptoms. In turn, this approach is by default decoupled from variance introduced by population differences.

Group- and individual differences in the perception of facial expression pose a few challenges to such technologies. When considering the variance introduced by cultural factors such as group-belonging, it furthers the argument that facial expression-based tools need to be personalisable. The reason being that the accuracy in recognizing facial expressions improves when the observer belongs to the same culture-group or is familiar with the facial expressiveness of the enacting identity. While group-differences are a powerful instrument to derive and generalize on particular characteristics of a population, individual differences are also important as they describe the variance within that group. Tools relying on the representation of mood through facial expressions need to be user-centric and adhere to the particular perceptual acuity and ability to recognize facial expressions of the user. In sum, to capture deviations of small magnitude using such tools, it needs to be ensured that those deviations are not introduced by group or individual differences.

One of the most contentious topics in the literature revolves around the universality of basic facial expressions of emotion. Here, it is important to address this long-standing debate explicitly, as it put into question the fundamental supposition of the proposed approach and namely that facial expressions and their use within self-assessment tools would be unambiguously understood by their users. In particular, critics of the universality theory point out that the basic facial expressions of emotion should constitute only of those for anger, fear, happiness and sadness [49]. In addition, they argue against the veracity of the statement that facial expressions are perceived universally across cultures [128]. Research indicates that there are cross-cultural variations in the enactment of facial expressions as well as how accurately those are perceived by persons belonging to the same or different culture-groups. However, here it is important to delineate what this criticism does not dispute and namely the universal existence of basic emotions and their associated facial expressions in all cultures. It also does not imply categorical misattribution of emotion labels to basic facial expressions in culturally homologous experiments (e.g. emotion signaller and decoder belong to the same culture or the decoder is familiar with the expressions of the signaller). In sum, the existence of particular pan-cultural emotions, their respective facial expressions as well as the attributed core-meaning of emotion are not put into question. Also, it is not debated whether the emotional experience of one person differs to that of another regardless of cultural belonging. What is disputed, however, is whether some basic expressions are indeed 'basic' and that facial expressiveness is culture-invariant. Thus, tools

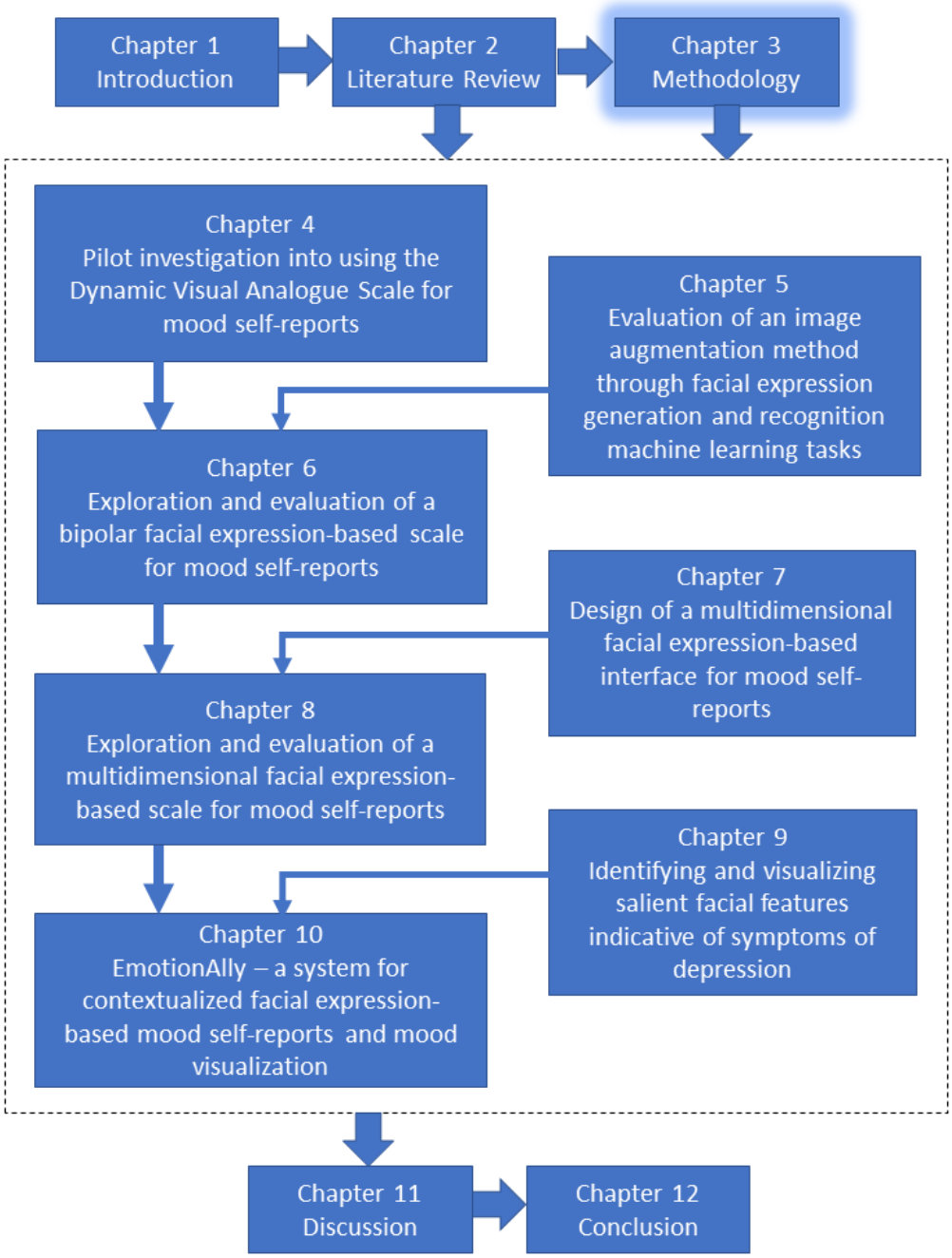
which utilize facial expressions should be able to accommodate for those differences such that they are generic and can create or visualize expressions for arbitrary identities (e.g. the portraying identity could be chosen to best suit the user) or expression-agnostic such that there would be no technological challenges in creating or visualising culture-specific facial expressions. Should those considerations be taken into account in the design and development of facial expression-based tools, the cultural variation aspect in the aforementioned debate would not be applicable.

Basic facial expressions of emotion are established to be biologically hard-wired and accompanying emotional experience as by evidence from congenitally blind people [129]. Thus, there is no argument to be made against using facial expressions being used as a construct for representing emotions. Taking that into account, irrespective of whether facial expressions are basic or complex, they could be used to represent emotions as long as they are meaning-invariant to the person or group of persons perceiving them. The benefits of extending facial expression tools to incorporate expressions beyond the basic ones would yield considerable advantage to the flexibility of the method as those would expand the assessment space. However, to date there is limited research and agreement on distinct expressions unambiguously conveying other emotions. However, should such expressions exist, a pertinent point for technologies used to create or visualize facial expressions is that they need to be expression-agnostic.

Finally, the enactment of facial expressions in social settings could be influenced by a multitude of social factors [76], however, in the personal interaction between user and technology, such considerations are a moot point, as the technology itself would not intentionally misrepresent emotions or enact judgments.

2.7 Chapter Summary

In this chapter, the core-concepts of emotion theory and emotion and mood were introduced as they shape our contemporary understanding on those subjects. Additionally, mood assessment methods used in practice in the fields of HCI and Psychology for self-reporting mood were presented. Subsequently, the use of facial expressions as a valid construct for representing emotions was introduced and argumented. Depression and its symptomatology were introduced highlighting possible applications of such tools for using facial representations of symptoms of depression for self-reports and in capitalizing on the manifestation of biases in the processing of faces as diagnostic tool to measure depression severity. Technologies that can create or visualize facial expressions within tools for self-reports were introduced highlighting their suitability for the task. Finally, theory and practice were bridged together in an exploration of what is known on facial expression- and non-facial expression-based assessment tools, social face perception outside and within the domain of affective disorders and how machine learning or 3D technologies can address issues spanning from the interplay of all those subjects.


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 3

Methodology

3.1 Introduction

The methodology adopted for this thesis integrates both quantitative and qualitative research methods with a stronger emphasis on quantitative analysis and interpretation of data obtained through multiple studies [130]. The main investigation principle is rooted within iterative development and follows the Rational Unified Process framework (RUP) [131], while borrowing some applicable principles from agile development [132]. The research conducted in this thesis investigates the basic concept of using facial expressions to represent emotions in multiple experimental studies. Following RUP, each chapter incorporates ideas and knowledge from previous ones.

3.2 Research Approaches

The research approach employed in this thesis is that of pragmatism [133–135]. In essence, the pragmatist approach postulates the role of action and change in a world of flux. '*The essence of society lies in an ongoing process of action – not in a posited structure of relations. Without action, any structure of relations between people is meaningless. To be understood, a society must be seen and grasped in terms of the action that comprises it*' [136, p. 71]. Actions are thus pivotal as meaningful conveyors of information and change. To achieve a purposeful goal, action needs to be informed by knowledge. Those characteristics are thus intertwined, such that knowledge informs action and gives it purpose. Pragmatism finds the practical consequences of an idea or a concept as foundational. This is formulated as a pragmatic principle as follows: '*Thus, we come down to what is tangible and practical as the root of every real distinction, no matter how subtle it might be; and there is no distinction of meaning so fine as to consist in anything but a possible difference of practice*' [137, p. 7].

Additionally, inquiry and constructive knowledge is central to applying pragmatism in research. This is best defined as follows: '*Inquiry is the controlled or*

directed transformation of an indeterminate situation into one that is so determinate in its constituents, distinctions and relations as to convert the elements of original situation into a unified whole' [138, p. 104]. Inquiry is '*as a natural part of life aimed at improving our condition by adaptation and accommodations in the world*' [139, p. 20]. Thus, inquiry is seen as an investigation into an aspect of reality with the purpose of exerting a controlled change in this part of reality and generating knowledge, where the interests of cognition and practical application overlap. A core idea of inquiry is to generate knowledge informing actionable change and improvement. This idea is further reinforced as follows: '*an empiricism which is content with repeating facts already past has no place for possibility and for liberty*' and outlines that pragmatism focuses on what is applicable and what might have a benefit in the future, rather than reporting on the current state of reality [140, p. 8]. As such, pragmatism views knowledge as an instrument to apply actionable and purposeful difference in practice.

Methodological pragmatism emphasizes the active role of the researcher in creating data and theories, where real-world experimentation is imperative. Through actions, inquiry or personal observations and informed by knowledge, a researcher examines the effects and success of different strategies. In action research, methodological pragmatism outlines the importance of continuous development, application and evaluation of acquired knowledge and strategies. Pragmatism also employs a pluralistic view on methods applied for the research purpose or empirical scenario as circumstances define their need [141].

3.2.1 Mixed Research Methodology

The combination of qualitative and quantitative methods at various stages in the research process are known and often employed under the umbrella of mixed methodology. This approach is commonly employed by the adherents of the pragmatism paradigm [142]. Contemporary understanding, considers quantitative analysis to be superior in order to inform generalizations about a population, while qualitative – in creating rich contextual data detailing particular aspects of importance to users [143]. The following sections describes the main quantitative and qualitative research methods employed within this thesis as well exploratory research which can either be quantitative or qualitative.

3.2.2 Quantitative Research Methodology

Quantitative methods enable collecting and evaluating measurable and verifiable data in order to identify and understand patterns and trends [144]. Technology, being ubiquitous in contemporary society, has eased the collection of data and enables the use of quantitative methods. Subsequently, modelling data using various techniques allows for its analysis and interpretation, which in turn informs research directions. For most of the thesis, a quantitative approach was selected as the

leading research method in particular for user studies described in Chapters 4, 6, 8 and 9.

3.2.3 Qualitative Research Methodology

Qualitative research is often associated with the philosophy of interpretivism [135]. However, an often overlooked fact is that qualitative research in information systems can also be achieved following the paradigm of pragmatism [133, 135]. As previously elaborated, pragmatism is associated with action, intervention and constructive knowledge, observed through the prism of its utility in action and its applicability. Knowledge of the multiple realities is therefore gained through an integration of a multitude of perspectives, encompassing both qualitative and quantitative research methods. This approach of mixed methods supports a more detailed understanding of research questions and results, leading to a balanced conclusion on the challenges and opportunities about a research problem. Thus, a qualitative researcher applying active and design research may also subscribe to a clear paradigmatic basis for their work. As such, pragmatism is considered an appropriate paradigm for both quantitative and qualitative investigations. Specifically, qualitative research methods were used in Chapters 4, 6, 8 and 10.

3.2.4 Exploratory Research Methods

Exploratory research aims to allow a researcher to be creative and gain insights on a subject. According to scholars, "*Social science exploration is a broad-ranging, purposive, systematic, prearranged undertaking designed to maximize the discovery of generalizations leading to description and understanding of an area of social or psychological life*" [145, p. 3]. It is argued that exploratory research should not use confirmatory mechanisms like hypotheses. Instead, it borrows from the principles of grounded theory, which postulates the construction of hypotheses and theories a-posteriori through the collecting and analysis of data [145]. As such, the topic of investigation needs to be clearly formulated as a question or a collection of questions. As data is collected and analysed, concepts, ideas and deductions 'emerge'. By adopting exploratory research to ones' tool kit, a researcher can create foundational knowledge on the matter in question. Exploratory research can add quality and insightful information and is vital to a study. Often an outside audience will be used, where internet-facilitated methods, which are interactive in nature are employed. Exploratory research can be used in a multitude of fields and can benefit greatly from a broad spectrum of research methods and analysis techniques. This type of research method adheres to the philosophy of pragmatism, as pragmatism employs a pluralistic view on research methods and analysis techniques whose outcome envisions a future of practical utility.

The exploratory research method was applied in Chapter 9 within a study that aimed to visualize symptoms of depression as facial expressions.

3.3 Principal Methods

3.3.1 Statistical modelling

Statistical modelling is a mathematically-formalized way to approximate reality [146]. A statistical modelling can make predictions about future observations and the accuracy of those predictions is a quantitative statement of how successful the model is. Statistical models traditionally attempt to generalize upon an aspect of reality from a constrained set of observations. Since it is impossible to observe everything at all times, the completeness of all use-cases in a given sample can enhance the success of the model. One of the most used models is linear regression analysis, which attempts to predict one variable by using input values from another on the merit of their historical correlation. Statistical modelling was used in Chapters 4, 6 and 8.

3.3.2 Data Visualization

Data visualisation is a generalization method for arranging data in a way that makes sense, such that it eases the process of finding patterns and aids in their interpretation [147]. Typically, it is used a first step preceding data analysis and modelling and as a final step to present findings about an identified trend. Various visualisation techniques exist, which emphasize on particular numerical characteristics of a dataset, relationship between variables or identified features. From an academic standpoint, data visualisation is a mapping of the underlying data to a visual arrangement. Various data visualisations have been used in all chapters of this thesis using the seaborn python library [148].

3.3.3 Data Mining

Data mining is a combination of statistical methods and data processing techniques, which aims to uncover patterns or correlations within a collection of observations [149]. It is exploratory in nature and is typically applied to large datasets, where those might be difficult to discern. Data mining can be used to increase the quantity of the data or in some cases its quality by finding patterns for implausible observations. Data mining was applied in Chapter 9 for the purpose of identifying disingenuous responses of respondents in an online-conducted experiment.

3.3.4 Machine Learning

Machine learning methods describe a family of models which make use of training data and are able to create representations of that data in order to provide predictions [150]. Within the field of machine learning, neural networks are a type of models best described as a series of algorithms that can identify underlying relationships through a process that mimics how the human brain operates. The

field distinguishes between two main types of neural network model categories – generative and classification networks. Generative networks attempt to create content based on learned criteria or constraints, while classification networks take input data and produce an estimate of categorical belonging.

Deep-learning or specifically deep neural networks are a type of neural network typologies which involve the use of multiple layers, where each layer specializes in identifying and aggregating features from the previous one [151]. Through the concatenation of multiple layers in sequence, complex organizations of features can be identified or generated. Neural network models are powerful tools to use when applied on complex or multi-modal data as they can use non-linearity to describe the classification space. That aspect makes them extremely versatile for a variety of tasks and consequently, they have been rapidly growing in popularity. Machine-learning methods were used in Chapter 5.

3.3.5 Reverse correlation

Reverse-correlation methods are a cluster of methods, which augment an existing signal with pseudo-random noise, where as a consequence, this modified signal may posses distinct characteristics delineating it from the original one, resulting in a change of meaning identifiable by a participant [12]. Those techniques are rooted in perceptual research and have been applied in a multitude of tasks. Specifically, reverse correlation was used in Chapter 9 for visualising symptoms of depression as facial expressions in noisy images by study participants.

3.4 Data collection methods

3.4.1 Dataset

A dataset is a collection of data, typically consisting of text, numbers, images or other mediums containing information [152]. A dataset can be heterogeneous, whereby it consists of multiple variables belonging to different modalities considered to be informative about a particular subject or homogeneous where a dataset consists of data of a single modality. It is a broad term to describe aggregated data [152]. In some instances, a dataset may bear relevance beyond the original purpose for its collection and may be used in secondary analyses. Secondary analyses is a viable research method to utilize previously collected information on a subject with the aim to generate new understanding or formulate new hypothesis.

All user-studies presented in this thesis involve the creation of a dataset resulting from a data collection phase, where data from multiple participants are combined and thereafter analyses are performed. In particular user-studies described in Chapters 4, 6, 8 and 9 rely on creating a dataset through data collection methods as outlined in Table 3.1 respectively. Additionally, Chapters 4 and 8 describe studies which rely on an existing mood-elicitation dataset in order to induce a

mood in a participant. Chapter 5 uses the Radboud Faces Database (RafD) [8], a facial expression dataset, used to create a data augmentation technique and train and evaluate a machine learning models applied to a generative and recognition tasks. Chapter 9 uses The racially diverse affective expression (RADIATE) face stimulus set [14], a facial expression dataset used to create a blended androgynous face.

3.4.2 Automatic logging

Automatic logging is a method for collecting data on digital applications or systems and consists of logic that automatically records values for system states, sensors or other modalities without requiring any user input [153]. It is a powerful instrument used in research, as it allows to either augment user-provided data with contextual information such as timestamps or location, among others, or can be used as a stand-alone data collection method.

Automated logging was used as part of all prototypes developed for user studies described in Chapters 4, 6, 8, 9. In particular, automatic logging in this thesis consisted mostly of collecting timestamps when a user interacted with a interface element in a prototype.

3.4.3 Self-reports

A self-report is a type of questionnaire, where respondents report on a characteristic or state of themselves without interruption or guidance [154]. A self-report may be submitted using a questionnaire or another measurement instrument that allows a person to rate a particular characteristic about themselves or the world.

A mood self-report in this thesis refers to reporting on ones' mood, where in the context of this thesis, various facial expression-based and visual analogue scales [19] were used in three different experiments described in Chapters 4, 6 and 8. Additionally, in Chapter 9, self-reports were provided by choosing one of two noise-augmented images of faces according to their likeness to a symptom of depression.

3.4.4 Questionnaires

A questionnaire is a research instrument consisting of multiple questions [155]. Those aim to collect information from participants regarding a subject matter. Questionnaires are an effective way to measure behaviour, attitude or preference in relation to particular characteristics of a process, application, device or others. Typically, a questionnaire may include open or close-ended questions or a combination of both. Collected data can be nominal, e.g. categorical such as gender or dichotomous such as 'yes' or 'no' responses, or ordinal, e.g. age, numerical measurements. Depending on the responses, data collected through questionnaires

can be analysed quantitatively, for ordinal or nominal data or qualitatively for open-ended questions [155]. Questionnaires are versatile as they can be administered in a pen and paper form, or digitally as part of a web-page or within a smartphone application, among others.

For investigating usability of prototypes used in Chapters 6 and 8, the Computer System Usability Questionnaire (CSUQ) [156] was used in a digital form integrated within the prototypes themselves. In Chapter 9, the Patient Health Questionnaire (PHQ-9) [42] was used in order to screen participants for presence and severity of symptoms of depression as well as to formulate concise descriptions for symptoms of depression used as experimental conditions. Further questionnaires have been used in user-studies described in Chapters 4, 6 and 8, tailored to assess characteristics specific to the prototypes used in each experiment.

Questionnaires featuring qualitative open-ended questions were used in user-studies described in Chapters 4, 6 and 8.

3.4.5 Interview

An interview is a qualitative research technique, which attempts to bring out informative elements on a particular idea, method or properties through dialogue [157]. It is the most direct and extensive method of capturing a users' perspective. Typically, interviews are either structured, semi-structured or unstructured. A structured interview presupposes a set of questions which will be posed to each participant, while an unstructured one does not and consists of a constructed conversation between researcher and participant. A semi-structured interview broadly implies the use of a number of prepared questions that provide a structure that allows a researcher to orient the exploration of topics of interest. Either type of interview comes with its own set of strengths and weaknesses, where structured interviews allow a researcher to draw more consistent and reliable interpretations. In some cases, particularly so with close-ended questions, responses can even be analysed quantitatively. Conversely, the unstructured interview is naturally more difficult to analyse and does not provide the same level of consistency, however it offers a wider breadth of information, which can be very valuable in evaluating existing aspects of an idea, tool, or design or identify future features of interest.

Within the scope of this thesis, a semi-structured interview was used in Chapter 8, consisting of a mixture of open- and close-ended questions.

3.5 Data transformation methods

3.5.1 Delaunay triangulation

Delaunay triangulation is a way to divide the surface or plane polygon into a set of triangles that can be computed for a given set of discrete points [158]. The method partitions the space in triangles under the condition that no point from the given

set of points lies within the circumference of any triangle. In that manner, the space is partitioned into a finite number of triangles, where the method maximizes the minimum angle of all the angles in all triangles contained within the Delaunay triangulation. This method has been applied for blending images of faces, where the input points are landmarks in an image descriptive of facial features. By controlling a blending factor, one can create a mixture between two or more distinct identities (e.g. images of different persons), or between different expressions (e.g. happiness and neutral), or both.

In Chapter 5, Delaunay triangulation was used to create an augmented version of the RafD dataset [8] that was used to improve the quality of generated facial expressions using a generative machine learning model and improve the recognition accuracy of images containing facial expressions in a facial expression classification task. Additionally, in Chapter 9, Delaunay triangulation was used to create an androgynous face from RADIATE [14], a dataset consisting of males and females of different ethnic backgrounds.

3.5.2 Neural Networks

Two categories of deep-learning neural network models have been evaluated in Chapter 5: a generative model used to create images of faces portraying varying intensities and expressions of emotion and two classification models for recognizing emotion from an image of a face. Images produced by the generative model have been used in prototypes for providing mood self-reports evaluated in Chapters 6 and 8 as well as EmotionAlly, a prototype system described in Chapter 10. The classification networks are referred to later in Chapter 9 regarding their ability to capture emotional content from a face.

3.5.3 Classification Images

Classification Images (CI) [43] is a reverse-correlation technique which is used to augment an image with pseudo-random noise, whereby the resulting image may resemble a psychological construct, known by a human observer. Typically, CI is employed within a classification task such as the two alternative force choice task (2AFC) [159], which implies that a participant is presented with two choices and is asked to select the one that fits best the experimental condition. Over multiple iterations where two such images are presented, a user reinforces salient features descriptive of the experimental conditions, while non-salient ones are smoothed out. Finally, by aggregating all choices made by a participant and collating their selections within a single image, a representation of the experimental condition emerges.

In Chapter 9, CI is used as a tool to modify an image of an androgynous face with patterned noise, where the resulting altered images were used within an 2AFC task. The task consisted of presenting 500 image-pairs to participants and required

the selection of the face that best resembled a person exhibiting a distinct symptom of depression.

3.5.4 User stories

User stories are usually a few sentences in simple language that outline a desired outcome [160]. A user story is an informal, general explanation of a software feature written from the perspective of the end user and are typically used in agile software development. User stories do not detail how a feature will be realized and should be seen as representing an end-goal from a user's perspective. The purpose of a user story is to articulate how a piece of work will deliver a particular value back. The use of user stories empowers users and allows them to highlight features considered of importance in contrast to informed decision-making or guesswork.

User stories were identified from user feedback collected throughout studies described in Chapters 4, 6 and 8. They were then used to define desired features in a system for self-reporting and visualising mood through facial expressions and steered the development of EmotionAlly, a prototype achieving those objectives described in Chapter 10.

3.6 Data analysis methods

3.6.1 Correlation analysis

In statistics, a correlation between two random variables or bivariate data is descriptive of their linear relationship [161]. A positive correlation would indicate that changes in either variable results in a change in the other in the same direction, where the strength of the correlation indicates how close that change is. Conversely, a negative correlation would still indicate a relationship between two variables, however, an increase in the value of one variable results in a decrease of the value in the other. Correlations are useful, because they can be used as an instrument to establish a relationship between two variables, where one variable can be used to predict the other. However, it is important to note that a correlation does not imply causation. That is, a correlation can merely speculate on one variable's expected value based on observed values of the other, but cannot be used to infer whether changes in one variable cause changes in the other. To make a claim about causality, alternative research methods or carefully designed experiments are needed.

Correlation analyses have been used as part of statistical modelling in user-studies described in Chapters 4, 6 and 8.

3.6.2 Regression analysis

In statistical modelling, regression analysis describes a set of processes for estimating a relationship between a dependent variable and one or multiple independent

variables [162]. Regression analysis integrate correlations such that as an outcome a model is created, where the model is trained on prior observations. The model has a predictive power based on that prior knowledge and can be used to predict the independent variable or variables given a dependent one. The most common regression analysis is a linear regression, where a line is found that approximates data of prior observations. Regression analysis is typically applied for predictions, where in some situations it can also be used to infer about the relationship between dependent and independent variables.

Linear regression models have been used in the analysis of self-reports provided with a facial expression-based and visual analogue scales [19] obtained in user-studies described in Chapters 4, 6 and 8.

3.6.3 Thematic analysis

Thematic analysis is a qualitative method for analysing data that involves reading into a dataset and identifying patterns of importance [163]. Thematic analysis is typically informed by information obtained from questionnaires, interviews or other data collections methods which allow participants to freely formulate their feedback. The process of thematic analysis involves clustering qualitative data into categories or codes, where codes refer to a particular construct or idea pertinent to the topic of investigation. Subsequently, codes can be added to themes which describe a particular aspect of a technology, tool or psychological construct, among others. Finally, themes allow the researcher to build narratives, which can tell a story from a user perspective regarding the topic of investigation. Thematic analysis presupposes the exemplification of claims or narratives by integrating vivid or evocative quotations from the collected dataset, substantiating the narrative.

Thematic analysis was used in Chapter 8 as a data analysis method applied on data collected through interviews.

3.7 Methodological Approach in this Thesis

There are four experimental studies conducted within this thesis described in Chapters 4, 6 and 8 that investigated the use of facial expressions for mood self-reports. Each of those three studies evaluated a prototype built for the specific needs of the respective study, where the prototypes used the principles of simple design. The study described in Chapter 9 focused on data collection and its interpretation. As a methodological approach, this thesis applied the Rational Unified Process (RUP) framework [131].

3.7.1 Experimental studies

In Computer Science, experimental design focuses on gathering measurements for testing hypotheses regarding the value of technological artefacts (often software

or hardware systems) [164]. The resulting combination of a functioning system, supported by empirical evidence is intended to be used as a basis to supplement evidence for theories. This is innately intertwined with the fundamental tenet of the scientific method which posits that any proposed theory has to be able to reliably explain and predict a part of reality. Providing evidence for a theory is achieved through reproducible experiments. Experimental software engineering emphasizes the use of empirical studies of all kinds to accumulate knowledge.

There are four experimental studies conducted within this thesis described in Chapters 4, 6, 8 and 9.

3.7.2 Simple design

Simple Design is a principle, where the rule is to keep things, as the name suggests, simple [165]. The acronym YAGNI descriptive of the concept, underlines the methodology of Simple Design [165]. YAGNI stands for 'You Aren't Gonna Need It' which suggests that if something is not required, it should be left out. By understanding and adopting the principles of Simple Design, costly detours and mistakes can be avoided by implementing only the most fundamental functionality. In turn, this ensures that the investigated concept remains undiluted with superfluous features and instead focuses on what is needed now.

In user-studies, created prototypes described in Chapters 4, 6 and 8 featured only the basic functionality needed for the study purposes.

3.7.3 Rational Unified Process

The rational unified process (RUP) framework [131] builds on the waterfall model. The waterfall model describes the end-to-end process of developing a system and itself is comprised of multiple fixed sequential phases. Each phase defines a particular thematic task and must be completed before the next one can begin, such that there is no overlap between phases in the development process.

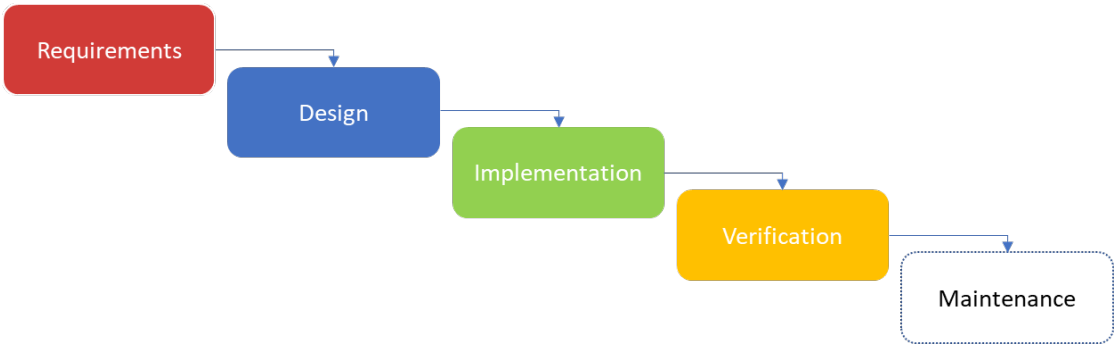


Figure 3.1: Typical process structure of the Waterfall model.

Typically, those phases are Requirements, Design, Implementation, Verification

and Maintenance as depicted in Figure 3.1. An explanation on how each phase was applied and documented within this thesis will be elaborated on below.

Similarly to the waterfall model, RUP is, in itself, also an iterative process. However, unlike the waterfall model, in RUP the building blocks are a full cycle of development as defined by the waterfall model. RUP allows to customize phases contained within each block and adapt them to suit a particular process. In the waterfall model, the maintenance phase is typically applied in situations where a product is already delivered to a consistent customer base and therein maintenance is defined as a continuous support of existing functionality over a changing technological landscape. Due to the fact that this thesis did not involve a fixed user base, this phase was omitted.

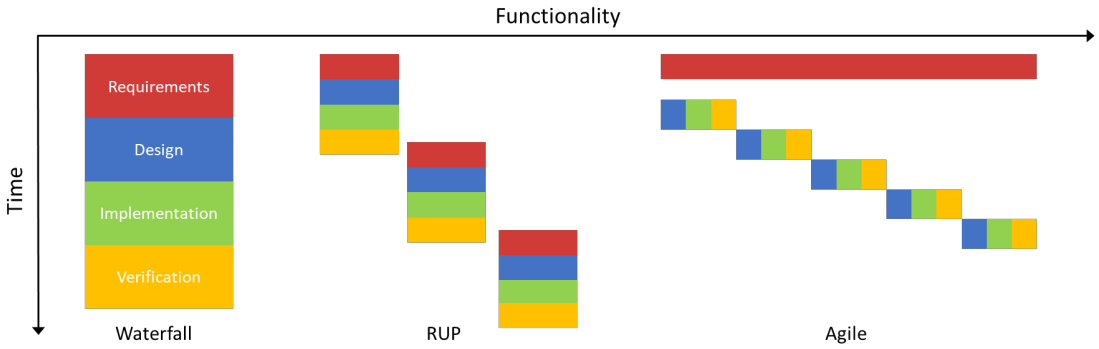


Figure 3.2: Structural differences between the Waterfall model, Rational Unified Process framework and Agile.

Figure 3.2 depicts and contrasts the waterfall model, RUP and agile methodological approaches. For the purpose of the investigation conducted in this thesis, RUP is ideal as it allows for iterative development and refinement of ideas through integrating qualitative and quantitative knowledge obtained sequentially in each user-study. As such, RUP has been employed as the working methodology for structuring the objectives in this work as well as the iterative development of prototypes.

Requirements The first phase within a block is to understand the goal and purpose of an objective. This phase consists of collecting requirements pertaining to the functionality, design and underlying technology which will be used throughout the project.

In Chapter 4, the development of the initial pilot prototype was informed by a literature review (Chapter 2) and uses an existing method for capturing mood using facial expressions. Subsequent user studies in Chapters 6 and 8 each build upon ideas collected from preceding quantitative and qualitative feedback, although the requirement phase was not made explicit in those chapters. In Chapter 10, requirements are formally defined through user stories, collected from feedback

received in user studies from preceding chapters and were explicitly formulated within that chapter.

System Design System design involves the conceptualization and definition of technological aspects. Traditionally, it is informed by the previous step, where stakeholders' requirements are formalized within achievable objectives. Additionally, this step also includes the definition of system requirements, hardware specifications and the underlying technology which will be used. In the context of this thesis, the system design and hardware constraints are predetermined for the complete scope of the project to utilize technological artefacts and methods usable on smartphone devices. The reason being is that complex systems relying on high-performance computing are undesirable as they are not ubiquitous and present a barrier-to-entry for potential users.

Chapter 7 describes the design and creation of a prototype subsequently used in a user study in Chapter 8. Additionally, Chapter 10 describes EmotionAlly – a system for providing and visualising contextualized mood self-reports using facial expressions. Although Chapters 4 and 6 also make use of distinct prototypes built for their respective study purposes, their functionality was simple and did not warrant an explicit elaboration beyond a functional description.

Implementation The implementation phase brings together the refined requirements from the system design phase into a working prototype. The generated artefact needs to adhere to the system and hardware constraints outlined in the system design phase and integrate functionalities informed by the requirements phase. The implementation phase of a project typically involves writing logic which accomplishes the required functionality. This phase was omitted in the description of each developed prototype as it does not contribute any novel knowledge with the exception of Chapter 10, where some implementation aspects had relevance to the design rationale and were explicitly elaborated upon.

Analysis In the analysis phase, the functionality of a prototype and its capabilities are evaluated in a deterministic manner such that each feature yields a consistent and predictable output. In the scope of this thesis, the analysis phase consists of conducting user studies using prototypes that investigate or focus on a particular aspect of the concept. This phase is contained within the results and discussion sections of each Chapter featuring a user study and contains elements of quantitative and qualitative data analysis methods. Studies involving an evaluation of a prototype were described in Chapters 4, 6 and 8. However, in Chapter 7 a system design for a prototype was evaluated in its ability to allow users to provide quick self-reports, wherein this aspect was highly relevant and was described in greater detail.

3.8 Study Ethics Procedures

Organizations that facilitate research with human subjects, including biomedical or behavioural experiments, require an approval from an institutional review board (IRB) for each experimental study. The work within this thesis investigates the use of facial expressions to represent various moods and affective states, where a substantial part is based on experimental studies with human participants. In order to protect participants of research studies and for the research to result in benefit and minimize the risk of harm, rigorous ethics procedures have been followed.

Each study dossier was examined and approved by the Internal Committee Biomedical Experiments (ICBE) at Philips Research. Thereby, this assessment encompasses not only the content and purpose of the study, but also a cost-to-benefit analysis establishing whether potential outcomes from the study exceed any risks posed to participants. Preceding the review of each study dossier by the ICBE, three assessments take place, (i) a *security assessment* ensuring that a prototype, device or the technical data transfer or retrieval method does not expose a participant's data, (ii) a *data and privacy assessment*, evaluating that only data strictly necessary for reaching the goal of the study is recorded and its storage location is accessible only by the responsible researcher(s), and (iii) a *risk-analysis* ranking by severity all feasible risks to participants that may arise through the use of the investigated device/software or study procedure and a mitigation plan for each of those risks. Provided each of those assessments has obtained an approval by a dedicated officer, specializing in either of those assessments, a study can be submitted for review by the ICBE. The ICBE Board comprises of domain-experts, regulatory, privacy and legal officers. Assessment of the study dossier is conducted on general quality, ethics, medical aspects, regulatory, privacy, legal, and brand image. Naturally, all studies described within this thesis in Chapters 4, 6, 8 and 9 have been approved prior to their execution.

The General Data Protection Regulation (GDPR) is a privacy and security act encompassing laws and regulations for collection and processing of personal data of EU-citizens. The study described in Chapter 4 took place prior to GDPR being in effect, however followed all previously outlined procedures. The studies described in Chapters 6, 8 and 9 were GDPR-compliant. Additionally, the study described in Chapter 9 was conducted online with no direct interaction between the researcher and participants, in-person or digital. As part of the *data and privacy assessment*, personal data was considered to be anonymous upon collection and following a de-anonymization assessment, it was established that individual participants could not be later identified.

For each study, consent was obtained explicitly in written form using a document created during the *data and privacy assessment*, apart from the study described in Chapter 9, where a click-through consent was deemed sufficient, since the data was considered anonymous upon collection.

3.9 Datasets & Diversity and Inclusion

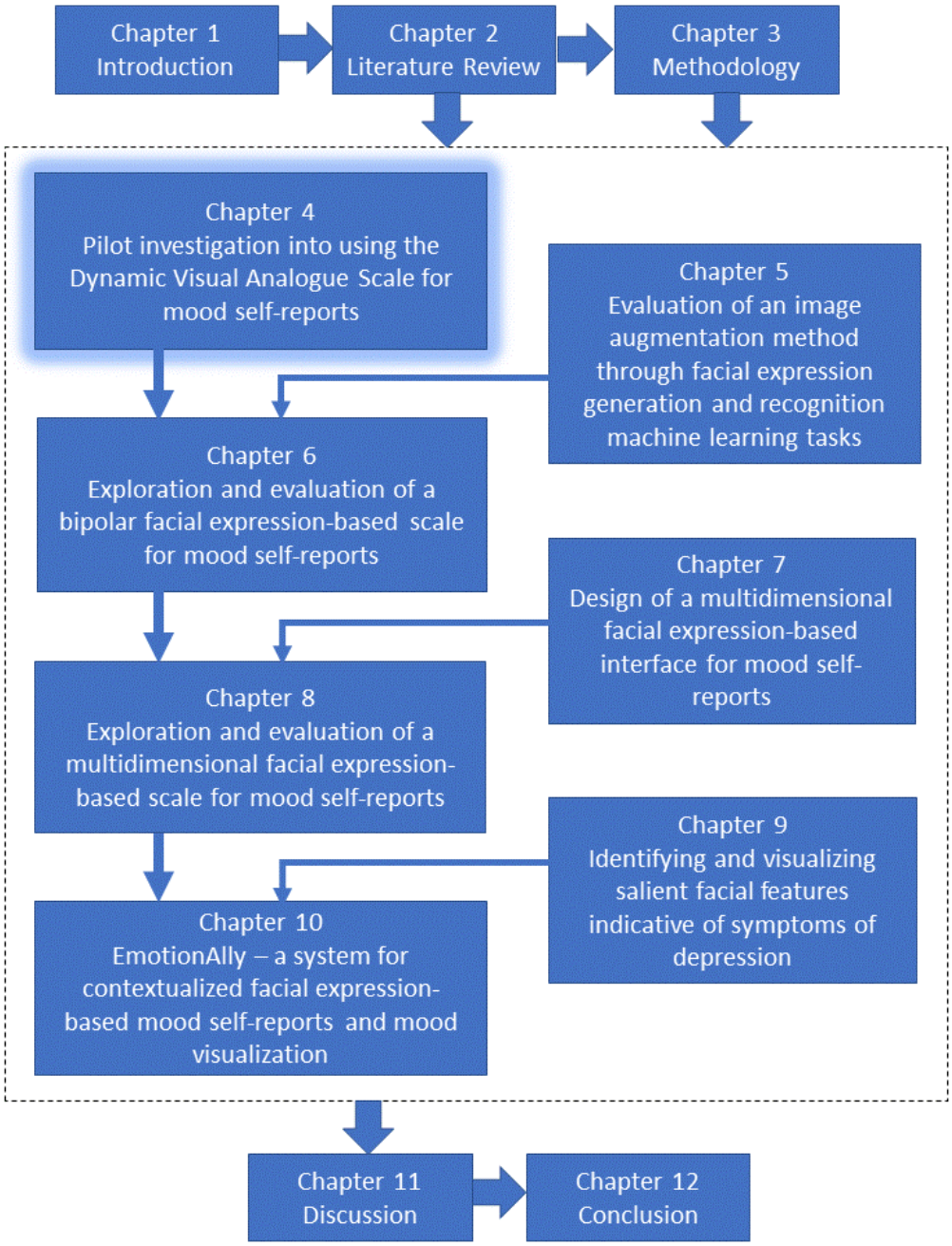
The datasets used within this thesis aimed to be representative of persons of various ethnic backgrounds. Two datasets constitute the majority of this work – the Radboud Faces Database (RafD) [8] and the Racially Diverse Affective Expression (RADIATE) Face Stimulus Set [14]. RafD portrays facial expressions of male and female persons of Dutch and Moroccan origin, where RADIATE encompasses a wider range of ethnic backgrounds featuring black, white, hispanic and asian adult male and female models. RafD has been used in Chapters 5, 6 and 7, where RADIATE – in Chapter 9.

3.10 Chapter Summary

This chapter outlined the research philosophy of pragmatism and how it was applied within the context of this thesis. Additionally, the quantitative, qualitative and experimental research methodologies used throughout the work described within this thesis were introduced. Therein, data collection and analysis methods were highlighted according to how and where they were used within the chapters to follow. Thereafter, the methodological approach in this thesis, and specifically RUP, was introduced as the working umbrella methodology throughout the work described in this thesis for steering the direction of investigation and the iterative development of prototypes and their experimental evaluation. Finally, the study ethics procedures followed prior to the execution of each study were outlined as well as the considerations made towards representing diversity in datasets used to facilitate this work.

Table 3.1: Overview of chapters and studies, employed data collection and analysis methods with links to the research questions they aim to address through research objectives.

	Study 1	Study 2	Study 3	System design	Study 4	Study 5	System design
Chapter	4	5	6	7	8	9	10
Research Question	A)	C)	A), B)	C)	A), B)	D)	C)
Research Objective	b), d)	a)	b), c), d)	d)	b), c), d)	e)	d)
Research Method	Quantitative Qualitative	Quantitative Qualitative	Quantitative Qualitative	–	Quantitative Qualitative	Exploratory	Qualitative
Data collection methods	Questionnaires Mood self-reports Automatic logging Dataset	Dataset	Questionnaires Mood self-reports Automatic logging Dataset	–	Questionnaires Mood self-reports Automatic logging Interview Dataset	Questionnaires Dataset	–
Data analysis methods	Statistical modelling Correlation analyses Regression analyses Data Visualization	Machine Learning Delaunay triangulation Data Visualization	Statistical modelling Correlation analyses Regression analyses Data Visualization	–	Statistical modelling Correlation analyses Regression analyses Data Visualization Thematic analysis	Reverse correlation – Classification Images Delaunay triangulation Data Visualization	User stories
Key Topic	Bipolar assessment scale	Data Augmentation Facial expression generation Facial expression recognition	Bipolar assessment scale EMA	System design	Multidimensional assessment scale	Depression symptoms Facial expressions	Multidimensional assessment scale Mood feedback Contextual mood System design


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 4

Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports

4.1 Introduction

Self-assessing mood is primarily done through numerical or graphical scales such as Likert or Visual Analogue Scales (VAS) (Chapter 2, Section 2.3.1). However, such scales are intrinsically not attuned to capture mood (Chapter 2, Section 2.3.1) and possess some drawbacks in how persons perceive and assign values to non-numerical mental constructs (Chapter 2, Section 2.3).

As a solution, alternative methods for providing self-reports have been explored. One such approach implies the use of facial expressions as a feedback mechanism through which a person can report their mood. Over the years, multiple facial expressions scales have been created and evaluated [21, 27, 126] (Chapter 2, Section 2.3.3). Some of those include using drawn or schematic faces to create scales portraying either discrete emotions, aligned with the basic theory of emotion or principal components, aligned with the dimensional model of emotion (Chapter 2, Section 2.2). It is known that schematic or drawn faces and smileys cannot represent a gradient emotion intensities. As a consequence, the limited continuity of the ranges of facial expressions they portray poses limitations on their ability to capture nuanced assessments. Additionally, with the development of mass-communication technologies, tools portraying emotions as smileys have also been investigated [34, 35, 67]. While using facial expressions as a construct to represent emotions and mood is intrinsically closer than a numerical representation (Chapter 2, Section 2.4), those types of scales are susceptible to drawbacks of their own (2, Section 2.4).

The work presented in this chapter aims to gather preliminary support for the conceptual idea of using realistic portrayals of facial expressions as means to portray and subsequently capture mood in a pilot experiment. In order to do that,

the Dynamic Visual Analogue Scale (D-VAMS) [27] which relies on photographs of a person enacting the facial expressions for happiness and sadness to capture emotional content was used. Herein, the main hypothesis will be investigated that facial expressions can be used to capture a variety of mood states. To achieve that, a pilot experiment was conducted with 11 participants, using vignettes as emotion elicitation material. Assessments were provided for each vignette on a facial expression-based scale as well as an equivalent visual analogue scale (VAS) [19]. The facial expression scale was implemented to feature the expressions for happiness and sadness as a continuous bipolar scale and similarly, VAS was also realized as a bipolar scale consisting of a horizontal slider with anchored labels denoting the emotion dimensions for happiness and sadness at both extremes. The efficacy of the facial expression-based mood scale approach was evaluated by comparing assessments from both scales. As VAS is a widely accepted and used method for providing self-assessments, a correlation analysis was conducted comparing both scales' performance on the stimulus set. Content of this Chapter was published as a conference article [1], reproduced with permission from Springer Nature.

4.2 Methods

4.2.1 Study Design

The pilot experiment used a within-subjects design with two independent variables – assessments provided with D-VAMS and VAS. The user experience was captured through an survey evaluating both the method of using facial expressions for assessments and the its practical realization within a developed prototype featuring a mixture of close- and open-ended questions.

4.2.2 Participants

Participants were recruited through fliers distributed across the High Tech Campus in Eindhoven, The Netherlands. They were required to be between 18 and 65 years of age with no history of mental illness and a good English proficiency. 11 participants took part in the pilot experiment (8F/3M, Age: M=29.8, SD=8.9). All participants were required to sign an informed consent form. Study participants were not remunerated for their effort.

4.2.3 Materials

A mobile application was developed for the Android operating system. It featured the happiness–sadness scale from the Dynamic Visual Analogue Mood Scales (D-VAMS) [27] and a happiness-sadness visual analogue scale (VAS) [19]. Both were designed to be bipolar and featured 101 discrete points. D-VAMS was represented

through 101 images consisting of 50 portraying a gradient of intensities representing happiness, 50 – of sadness, and an image depicting the neutral expression. The sadness dimension was allocated in the $[0-49]$ numerical range and the happiness – in the $[51-100]$ range. The lowest and highest points in those ranges portrayed images with the highest intensity of the respective emotions for sadness and happiness. The neutral expression occupied the centre-point in the scale. The images were obtained from the male sadness–happiness facial expressions set of D-VAMS [27], where those can be found at the D-VAMS project website [166]. Navigating through the scale was accomplished using gestures: sliding vertically upwards over the facial expression feedback image displayed increasingly happier expressions, while downwards – sadder ones.

VAS was designed as a bipolar pseudo-continuous horizontal slider with 101 discrete points. Anchored text denoted the emotion content at both poles of the scale (i.e. sadness and happiness). Figure 4.1 contains a screenshot of the VAS variant.

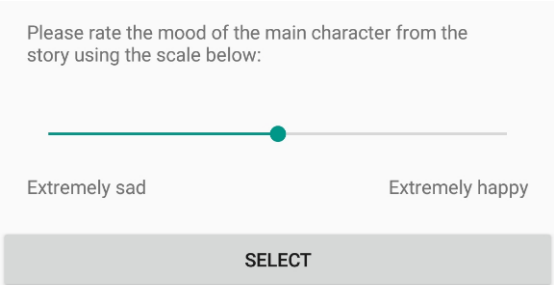


Figure 4.1: Screenshot of the VAS scale assessment as taken from the study application.

The stimuli used in the pilot experiment were 30 vignettes selected to convey a variety of positive and negative emotions. 15 conveyed positive emotions such as awe, pride, tranquillity, compassion, and others. Those were obtained from existing work and have been experimentally validated [167]. 15 vignettes were selected to convey negative emotions such as disappointment, grief, regret and were collected from various online resources. Those were not validated for their emotion content. 3 additional vignettes were selected and used for training to allow participants to familiarize themselves with the assessment application and study procedure. Those were not included in the subsequent analyses. The vignettes' content was paraphrased to portray a story from a third persons' perspective. Names common in the English language were sampled to replace the pronouns used in the original vignettes, which used a first-person narrative.

A user experience survey was given to participants investigating each assessment method and its functionality in the application. Thirteen questions aimed to investigate the methods' usefulness in terms of perceived usability for mood self-reports, accuracy of the provided assessments and user satisfaction. Ten questions allowed the participant to rate the user experience with the application, its responsiveness and intuitiveness. These close-ended questions in the user experience survey used a five-point Likert scale. Both scales were rated independently using separate questions rather than a relative comparison between both scales as a single question. Two questions were used for rating the preference between D-VAMS and VAS in terms of perceived speed in providing assessments. Two binary choice

questions prompted the participants whether they would be able to use either scale without instructions. Finally, four open-ended questions inquired about difficulties encountered during the pilot experiment and prompted for insights on how the scales could be improved. The questionnaire can be found in Appendix 4.A (p. 55).

4.2.4 Procedure

Participants that expressed interest in the study were sent an information letter and privacy notice detailing the content of the experiment, the task they are expected to accomplish and how their data will be used. Participants that agreed to participate were admitted to an on-site laboratory. On arrival, participants were handed a printed copy of the informed consent form and privacy notice and were given ample time to familiarize themselves with the contents. Additionally, they were informed that they could ask any questions pertaining to the experiment, their participation or the outcome of their data. After providing written consent, a mobile phone was provided with the study application pre-installed. Thereafter, participants were handed a copy of the training vignettes and the vignettes used in the experiment. The vignettes had a fixed predefined order, however, they were shuffled such that a mixture of one or few positive vignettes were followed by one or few negative ones.

Participants were instructed to start with the training vignettes and after reading each one, to use the smartphone application to rate the mood of the main character. Those ratings were provided twice, once with the D-VAMS scale and again with the VAS scale, where both scales were presented sequentially in a randomized order. Prior to each assessment, both scales were initialized in the neutral position, i.e. the slider was positioned in the middle of the scale and the facial expression feedback – to the neutral expression. For each assessment, the order in which D-VAMS and VAS appeared was randomized.

After completing all assessments, a user experience survey was administered with 26 open- and closed-ended questions. Each close-ended question was rated on a five-point Likert scale, where 1 was designated as the most negative score and 5 – the most positive one. Anchors were used to denote the extremities in each question.

4.2.5 Statistical analysis

Assessments provided with D-VAMS and VAS were compared using linear regression correlation models, where those were computed separately for each dimension (e.g. positive and negative for the respective vignette categories) and were examined for significance. The regression analysis aimed to establish to what degree assessments provided with each scale correlate to one another for each vignette dimension.

The application used automatic logging to record the time required to provide an assessment. Therein, four different timestamps were collected: when an assessment interface (e.g. D-VAMS or VAS) was initially displayed to a participant on the screen, when the assessment was completed by confirming through pressing the

'select' button (see Figure 4.1), when a user initiated an interaction with the assessment element by touching the facial expression feedback image or slider, and at the final interaction with the assessment element by releasing one's finger from the facial expression feedback image or slider. Two t-tests were conducted for both assessment scales for: the duration an interface was displayed on the screen, and the duration of the interaction with each assessment element (e.g. the image or slider). A few assessments' took a significant amount of time, caused by an interruption in the procedure due to a question asked by a participant or a self-enforced break. Those were filtered out using the three-sigma rule [168].

Within the administered user experience questionnaire, paired questions investigating aspects of the method or the application were compared using paired t-tests.

The numerical data were analysed using Python 3.6 and numpy and pandas libraries [169] and visualisations were created using the seaborn library [148]. For 13 assessments the application failed to record a value and those were discarded from subsequent analyses.

4.3 Results

Table 4.1: Table containing the sample size (n), mean (M) and standard deviation (mean(SD)) of mood assessments made with either D-VAMS or VAS grouped by vignettes' categorical dimension.

Dimension	Type	n	M (SD)
Happiness	D-VAMS	161	82.44 ($\pm$19.64)
Happiness	VAS	161	74.12 ($\pm$17.68)
Sadness	D-VAMS	156	19.80 ($\pm$16.46)
Sadness	VAS	156	22.03 ($\pm$16.08)

Table 4.1 presents characteristics of the mood assessments provided on positive and negative vignettes. Table 4.2 presents descriptive data of both linear regression (LR) models. The correlations between D-VAMS and VAS assessments were strong for both the happiness and sadness dimensions (Happiness ($r(161) = 0.85$, $p < .001$); Sadness ($r(156) = 0.85$, $p < .001$)). Figure 4.2 plots the fitted LR correlation models.

Figure 4.3 visualizes the assessments for each vignette with both D-VAMS and VAS as box plots. What becomes evident is that most assessments on vignettes' have overlapping interquartile ranges (IQR) for the two scales. The exceptions were vignettes [1,7,11,29] for positive vignettes. Furthermore, most IQR lie within their expected category dimensions on both scales. That is, positive vignettes' IQRs are above 50 with the exception of assessments provided with D-VAMS on

Table 4.2: Linear regression model parameters describing the fit of VAS scores as predicted by D-VAMS scores, for mood assessments made in the emotion categories for happiness and sadness separately. The columns indicate the linear regression slope (s), intercept (i), correlation coefficient (r), significance value (p), standard error (SE), intercept standard error (iSE)

Dimension	s	i	r	p	SE	iSE
Happiness	0.77	11.04	0.85	.000	0.04	3.18
Sadness	0.83	5.62	0.85	.000	0.04	1.07

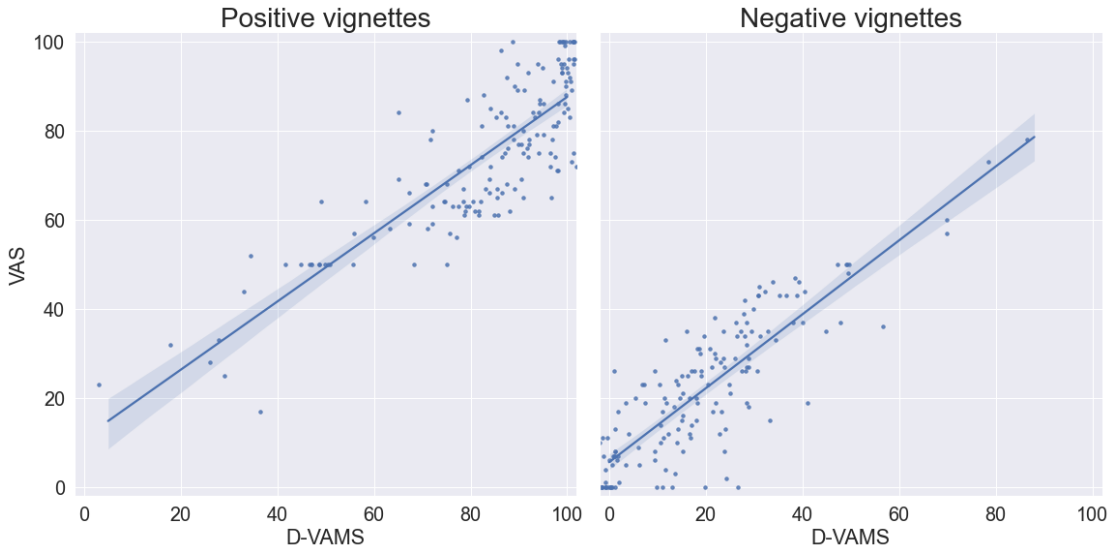


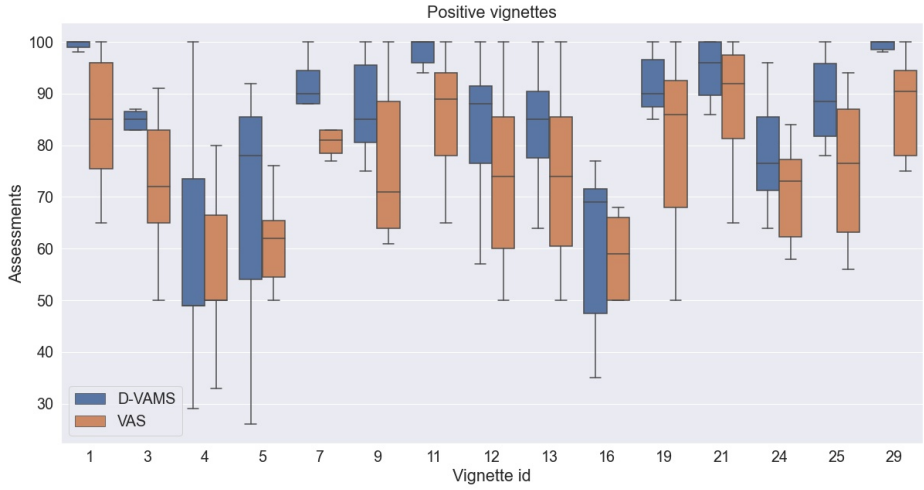
Figure 4.2: Regression model fit of VAS assessments regressed on D-VAMS assessments split by vignette dimension. The visualisations uses jitter on the x-axis to improve the legibility of overlapping assessment data-points.

vignettes 4 and 16, which encompass a small portion of the sadness dimension. For the negative vignettes, the exception was vignette 14 with VAS, where its IQR included that of the neutral expression.

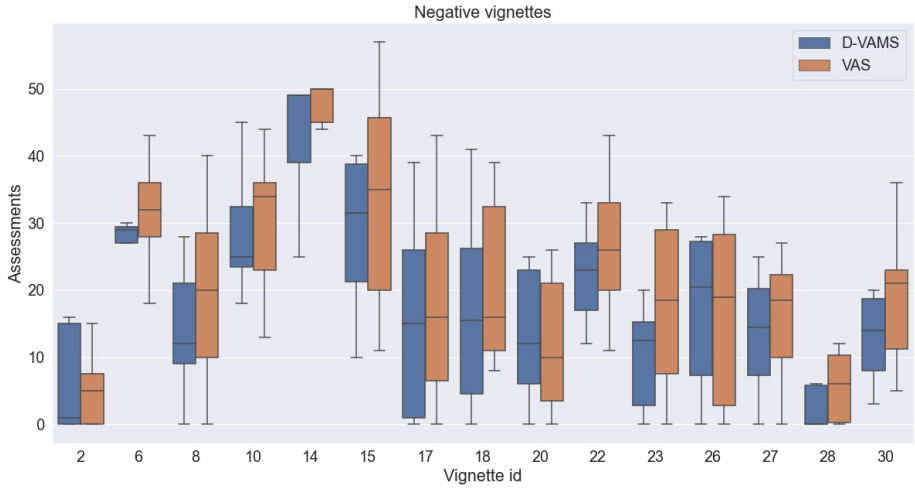
The mean values for completing an assessment (e.g. the time difference between the interface being displayed to a user and a user submitting an assessment) with D-VAMS was 5.10 seconds and for VAS – 3.71 seconds. Those were statistically significant as well ($t = 6.26, p < .000$). The time spent solely on interacting with the interface was 2.93 seconds for D-VAMS and 1.06 seconds for VAS. Those were also statistically significant ($t = 12.40, p < .000$).

Table 4.3 features the results obtained from the user experience survey (see Appendix 4.A). None of the quantitatively comparable results reached statistical significance.

Nine participants provided additional feedback on the open-ended questions. It revealed that the dimensions for happiness and sadness were insufficient to capture



(a) Box plot containing positive vignettes' Interquartile ranges (IQR), their minimum and maximum assessment values on D-VAMS and VAS scales as enumerated by their order of appearance in the vignette list.



(b) Box plot containing negative vignettes' Interquartile ranges (IQR), their minimum and maximum assessment values on D-VAMS and VAS scales as enumerated by their order of appearance in the vignette list.

Figure 4.3: Box plots of mood assessments for positive and negative vignettes split by emotion dimension and assessment scale.

the different mood states described in the vignettes: *"I think there is more to the emotional spectrum than just happiness or sadness. Other emotions might be relevant to depression as well. Such as fear, disgust, anger, disappointment, frustration, satisfied, grateful, relaxed, nervous, challenged."* Interestingly, another

Table 4.3: Answers on the user experience survey investigating aspects of the assessment method and prototype implementation containing mean (M), standard deviation (SD), t- and p-value t-test scores (see Appendix 4.A). Each question was rated on a five point Likert scale. Higher score is better.

Method	D-VAMS M(SD)	VAS M(SD)	t(df=10)	p
Ease of use	4.09 (±1.0)	3.73 (±0.96)	-0.83	0.42
Suitability for mood	3.73 (±1.29)	3.64 (±0.88)	-0.18	0.86
Accuracy	3.73 (±0.96)	3.73 (±0.62)	0.0	1
Satisfaction	4.0 (±1.28)	3.18 (±0.72)	-1.77	0.1
Application	**D-VAMS M(SD)**	**VAS M(SD)**	**t(df=10)**	**p**
User experience	4.45 (±0.89)	3.73 (±1.19)	1.85	0.08
Ease of use	4.18 (±1.19)	3.73 (±0.75)	1.02	0.32
Responsiveness	4.36 (±0.88)	3.6 (±0.92)	1.84	0.08
Intuitiveness	4.0 (±1.04)	4.09 (±0.67)	-0.23	0.82
Preference	3.91 (±1.44)	3.73 (±0.86)	0.34	0.73

participant pointed out that they liked the fact that D-VAMS featured a real face instead of a more conventional cartoon-like character: *"I like the use of a real person and not a cartoon or smiley-type of representation."* One participant for D-VAMS and no-one for VAS noted that they would need instructions before using the respective scale.

For the two questions which directly assess the preference for either D-VAMS or VAS, participants favoured D-VAMS with an average score of 2, where 1 corresponded to every participant strongly favouring D-VAMS and 5 – VAS. In the direct comparison between D-VAMS and VAS for their speed of assessment, participants rated on average both scales to be equally fast.

4.4 Discussion

First, it needs to be acknowledged that the present study was conducted as a pilot and aimed to obtain insights into the applicability of the method of using facial expressions for mood self-reports. Therein, it was investigated whether a facial expression-based scale would yield results comparable to the more conventional VAS. Hence, the assessment interfaces were designed to be as simple as possible, featuring the dimensions for happiness and sadness only. Those aimed to capture positive and negative emotions conveyed through 30 vignettes.

The results indicate that the vignettes used as emotion elicitation material were captured through the happiness and sadness emotion classes. That is, the interquartile range (IQR) of assessments provided with both D-VAMS and VAS for positive and negative vignettes were contained within their expected sub-scale

ranges (e.g. positive vignette were rated consistently on the happiness dimension and negative ones – on the sadness dimension). This is evidence that the stimulus material was cohesive and representative of what it was supposed to elicit, despite the fact that negative vignettes were not validated.

The LR models resulted in a strong correlation on both happiness and sadness dimensions consisting of $r = 0.85$ between both assessment methods. This is an indication that both scales are similar, as they allowed participants to measure the stimulus material through assessments which were near-similar to one-another.

When considering time required to provide an assessment, doing so with VAS is clearly faster than with D-VAMS. This is not surprising, due to the fact that for VAS, the complete range of the scale was accessible to participants. Conversely, providing an assessment with D-VAMS required a participant to use a gesture that incrementally modified the facial expression feedback until it reached the one desired for the assessment.

While the results from the user experience questionnaire did not reach significance results, the quantitative analysis revealed that both scales perform similarly to each other. Questions investigating the D-VAMS and VAS method of providing mood self-reports and scale implementation separately were also not significant. However, in the direct preference comparison between D-VAMS and VAS, participants indicated an overall preference for D-VAMS. Additionally, comparing the perceived speed for providing an assessments, participants thought both scales to be equally fast. This indicates that there is promise in the method for assessing mood through facial expressions. For example, qualitative feedback by participants indicated the desire to see more facial expressions that can be used for self-reports. This warrants a more thorough evaluation of the method by, for example, using a larger participant sample, utilizing methods for creating more nuanced and detailed facial expressions, or different experimental contexts such as a mood-monitoring experiment.

4.5 Conclusion

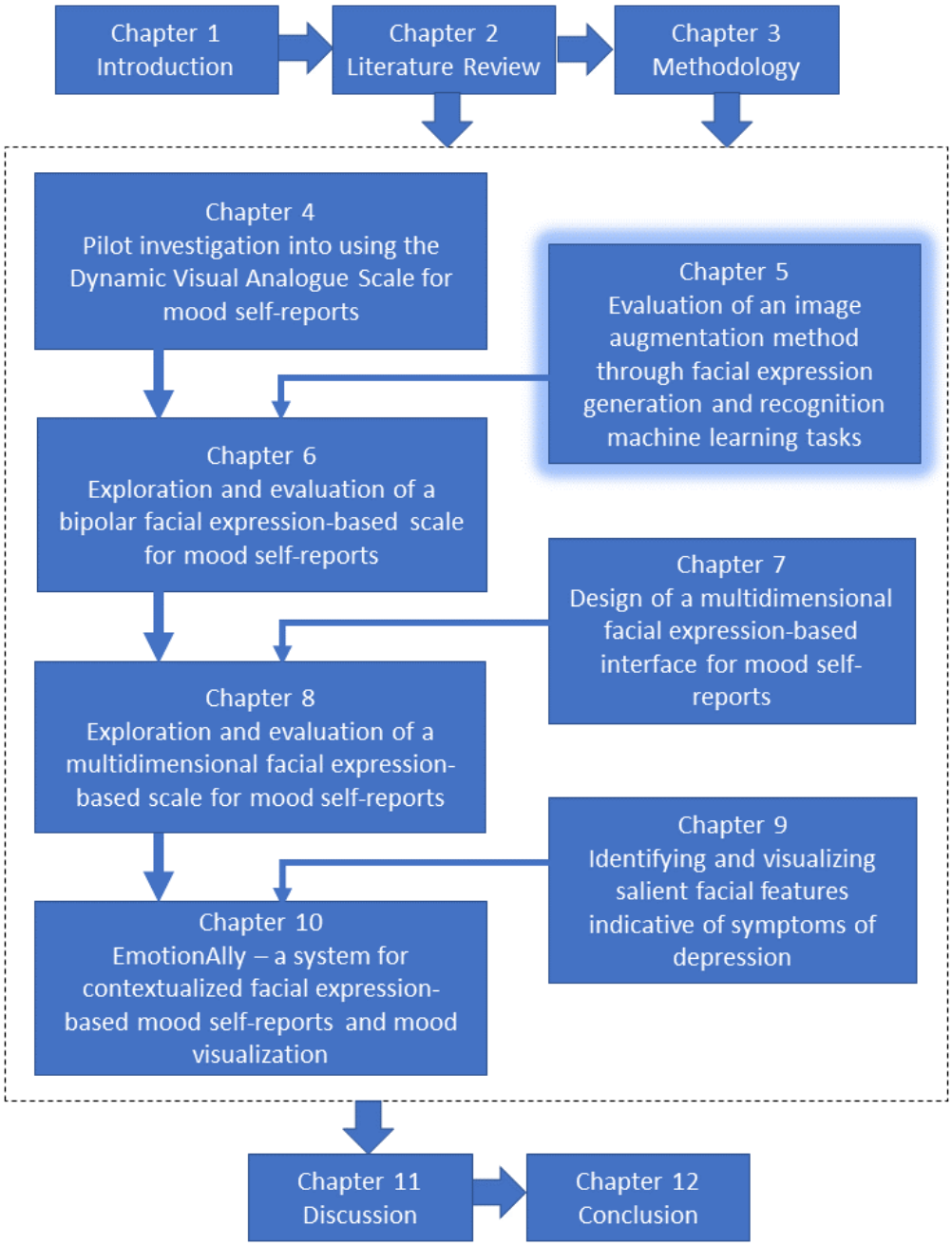
This pilot experiment explored whether facial expressions for happiness and sadness can adequately be used to capture a variety of mood states. To this end a pilot experiment was set up in which 11 participants rated 30 vignettes using a happiness-sadness facial expression-based scale (D-VAMS) and a visual analogue scale (VAS). Results indicated that assessments provided with D-VAMS and VAS are comparable to one another, indicating that both scales allow participants to rate emotional content similarly. Furthermore, it appeared that both scales can effectively capture complex mood states on the happiness and sadness emotion dimensions presented as a bipolar scale. In conclusion, while a VAS scale leaves little room for improvement, a facial expression-based one could be further improved.

4.6 Chapter Summary

The results of this pilot experiment indicated that the facial expressions for happiness and sadness are indeed able to capture an overall positive and negative emotional experience through the expression of happiness and sadness. The results from the user experience survey highlighted a desire for further facial expressions to be included in such tools such that they could represent a broader variety of mood states.

4.A User experience questionnaire

Participant information							
What is your age?							
What is your gender?							
Method-related questions							
How would you rate the ease of use of the method working with the slider?	(difficult)	☐	☐	☐	☐	☐	(easy)
How would you rate the ease of use of the method working with the image?	(difficult)	☐	☐	☐	☐	☐	(easy)
How suitable was the slider for capturing mood?	(not suitable)	☐	☐	☐	☐	☐	(suitable)
How suitable was the image for capturing mood?	(not suitable)	☐	☐	☐	☐	☐	(suitable)
How accurate do you consider the slider is in capturing the mood?	(not accurate)	☐	☐	☐	☐	☐	(accurate)
How accurate do you consider the image is in capturing the mood?	(not accurate)	☐	☐	☐	☐	☐	(accurate)
How satisfying was the slider to work with?	(very dissatisfying)	☐	☐	☐	☐	☐	(very satisfying)
How satisfying was the image to work with?	(very dissatisfying)	☐	☐	☐	☐	☐	(very satisfying)
Which assessment method would you personally prefer?	(image)	☐	☐	☐	☐	☐	(slider)
Which method, according to you, was faster to use for mood assessment?	(image)	☐	☐	☐	☐	☐	(slider)
Do you have any comments regarding either of the mood capturing methods?							
Application-related questions							
What was your experience using the image within the application?	(negative)	☐	☐	☐	☐	☐	(positive)
What was your experience using the slider within the application?	(negative)	☐	☐	☐	☐	☐	(positive)
How easy to use did you find working with the image within the application?	(difficult)	☐	☐	☐	☐	☐	(easy)
How easy to use did you find working with the slider within the application?	(difficult)	☐	☐	☐	☐	☐	(easy)
How responsive was the image assessment within the application?	(unresponsive)	☐	☐	☐	☐	☐	(responsive)
How responsive was the slider assessment within the application?	(unresponsive)	☐	☐	☐	☐	☐	(responsive)
How intuitive do you consider working with the image within the application is?	(not intuitive)	☐	☐	☐	☐	☐	(very intuitive)
How intuitive do you consider working with the slider within the application is?	(not intuitive)	☐	☐	☐	☐	☐	(very intuitive)
Do you think you would use the image for mood assessment?	(unlikely)	☐	☐	☐	☐	☐	(likely)
Do you think you would use the slider for mood assessment?	(unlikely)	☐	☐	☐	☐	☐	(likely)
Do you think you can use the image approach within the application for assessing mood without a clarification from a technical person?	(no)	☐				☐	(yes)
Do you think you can use the slider approach within the application for assessing mood without a clarification from a technical person?	(no)	☐				☐	(yes)
Is there some part of the functionality, which didn't work for you?							
Did you encounter any technical difficulties, while working with the application?							
Do you have any comments regarding the application?							


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 5

Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks

5.1 Introduction

Generating and recognizing facial expressions has numerous applications ranging from medical ones, such as inferring character [170], emotional states and intent [171], detection of diseases [172], authentication and biometrics [173], designing affective interfaces [1, 6, 15, 174], virtual avatars [124, 125], and computer graphics [175]. State of the art uses machine- and deep-learning methods to achieve impressive results [117, 118]. However, most approaches are able to recognize or synthesize facial expressions as a categorical state with few recent works being able to also account for varying emotion intensities [119, 120]. Correctly recognizing or generating expressions at varying emotion intensities is important as a nuanced expression result in changes in meaning [176] Approaches relying on generating facial expressions need to be able to account for variations in facial expressions intensities. Within the context of using facial expressions as a medium for self-reporting ones' mood, computer-generated facial expressions offer some significant advantages where they can: i) generate a gradient of emotion intensities, ii) be trained and subsequently generate an arbitrary set of expressions (i.e. provided suitable data exists) and iii) decouple the enactment of facial expressions from the portraying identity (e.g. being able to transfer facial expressiveness between identities).

In this chapter, an approach is presented that addresses i) and ii) of these challenges. This is accomplished by using a deconvolutional neural network, which accepts identity, expression and eye-gaze direction as input parameters. Herein, the use of Delaunay triangulation combined with simple morphing techniques to

blend images of faces is investigated as an augmentation method for existing facial expression dataset. The augmentation allows to create and automatically label facial expressions portraying controllable intensities of emotion. It has been applied on the Radboud Faces Database (RafD) dataset [8], consisting of 67 participants and 8 categorical emotions and evaluated in a facial expression generation and recognition tasks using deep learning models.

The novelty of this approach is the application of the augmentation for improving the quality of computer-generated facial expressions and their recognition in images. Due to the fact that generative approaches are harder to assess and quantify (e.g. the quality of generated facial expression images can only be reliably assessed by human raters), the approach has been evaluated within a facial-expression recognition task as well. The argumentation being that if in a recognition task a classification network achieves similar or better results with the augmentation, compared to omitting it, argues for its utility and subsequently – the output of the generative model. For the generation task, a deconvolution neural network is used, which encodes input images in a high-dimensional feature space and can generate expressions depicting increasing intensities of emotion by interpolating over this space. For the recognition task, pre-trained DenseNet121 and ResNet50 networks were evaluated with either the original or augmented dataset.

As described in Chapter 1, Section 2.4, the proposed method of self-reporting mood relies on using realistic-looking facial expressions. As a result of applying the augmentation for training the generative model, one can create realistic facial expressions with a smooth transition over categorical dimensions of emotion (e.g. neutral to happy) and between arbitrary emotions and intensities (e.g. somewhat happy to mildly surprised). In turn, expressions created by the network could be used for creating facial expression-based mood self-report tools. Additionally, those expressions would be granular enough to be able to portray subtle, and imperceptible changes in expressions when incrementing the intensity of emotion in a facial expression, a quality not available in either schematic or drawn representations of facial expressions or images which have a fixed granularity [27]. Creating computer-generated facial expressions depends on having suitable training images for the preferred ones. Using generative neural networks (e.g. or other similar methods) would allow crafting expression- and identity-agnostic scales, tailored to a users' background and their familiarity with particular facial expressiveness.

Content of this Chapter was published as a conference article [2], reproduced with permission from Springer Nature.

5.2 Dataset

In order to create the augmentation and train the generative and classification models, the Radboud Faces Database (RafD) is used [44]. The dataset consists of 8040 coloured images with 681x1024 resolution including 67 persons – 57 adults and 10 children. It is labelled for person's identity, gender, ethnicity, facial expression,

eye gaze direction, and camera angle. For the purpose of simplifying the method and evaluation described herein, only images of front-facing facial expressions were used, omitting angles different than 90°. Furthermore, only images of adults were used, which resulted in 1336 images being selected from the dataset.

5.3 Augmented dataset

In this section the pre-processing steps required to create the augmented dataset are described. Python libraries dlib, FaceAligner, imutils and OpenCV2 were used for the preprocessing steps of landmark detection, alignment, centring and cropping. OpenCV2 is used for computing the Delaunay triangulation and affine transforms.

5.3.1 Alignment, Centring & Cropping

In the first phase, 68 landmarks were detected and located. Those landmarks were then used to align all the images using rigid registration. Generally, the images in the RafD dataset are well aligned, however, there was still a benefit in using the aligner as a slight tilt was present in some of them. This yielded some improvements, particularly for features such as the eyes and mouth, since it ensures that those are stacked on the same spatial coordinates for every image. Finally, detected faces were cropped out of the original image with output dimensions of 550x550 pixels.

5.3.2 Computing Delaunay triangulation and Transform

To generate the augmented dataset, Delaunay triangulation was computed on the aligned and centred images by re-identifying landmark points using dlib after applying the preprocessing steps described earlier. Computing the Delaunay triangulation can be done in multiple ways [158]. For this purpose, OpenCV's implementation for *calculateDelaunayTriangles*, *similarityTransform*, *warpAffine* and *warpTriangle* were which allows to easily reproduce those steps. Point-to-point correspondence registration was trivial, since the landmarks descriptive of facial features were an ordered set. The similarity transform between the two point clouds was found, facilitating a rotation, translation and scaling for each triangle in the Delaunay triangulation. Subsequently, by computing the transformation for all triangles, images were morphed together, where the texture blending was controlled by a factor between $[0, 1]$ weighing each images' contribution to resulting pixel values. To create the augmented dataset, the method was applied per expression and within subjects. For each emotion, the neutral and a 'target' one were selected, where as a result of the blending, for example using a factor of 0.5, the output image would portray the 'target' expression at half intensity. The same approach can also be applied to produce varying levels of intensities by adjusting the blending factor.

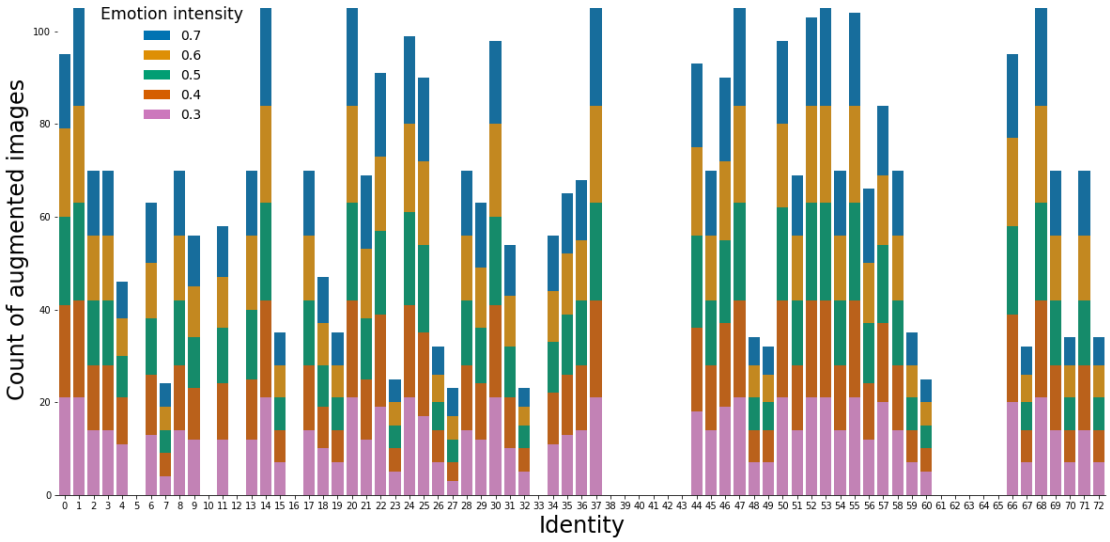


Figure 5.1: Augmented images without artefacts per subjects in the RafD dataset, split by emotion intensity. Optimally, 105 images per subject were expected for the $[0.3, 0.7]$ intensity range descriptive of 5 intensities, 7 emotions and 3 eye gaze directions.

Expressions were created for the intensity range $[0.3, 0.7]$ at increments of 0.1 as expressions with no (e.g. neutral expression) or maximum intensity were already part of the original dataset. By applying this augmentation per person and then per expression, while using the neutral expression as a baseline, the approach yielded 6840 augmented images. In some cases, mostly concerning specific expressions, the augmentation produced visual artefacts, which made them unsuitable for use as depicted in Figure 5.2.

Figure 5.2: Figure containing a successful and unsuccessful augmentation using Delaunay triangulation blending method. On the left is a successful augmentation portraying happiness with an 0.3 intensity. On the right is an unsuccessful augmentation.

Figure 5.1 depicts the number of augmented images created per subject. Noticeable were clusters with a near-similar amount of augmented images, due to the fact that those visual artefacts manifested on particular images, which rendered any blends thereof unusable. Excluding unsuccessful augmentations, the final augmented dataset consisted of 3848 images. Table 5.1 portrays the distribution of created augmentations according to expression and intensity of emotion. It consists of mostly evenly distributed number of augmentations for intensities. However, there are less samples for sad, angry and contemptuous facial expressions than

Table 5.1: Number of augmented images void of artefacts per incremental step.

Emotion\Label	0.3	0.4	0.5	0.6	0.7
happy	122	122	122	122	122
sad	98	98	100	99	98
angry	93	91	85	81	77
contemptuous	103	102	102	98	99
disgusted	122	122	122	122	122
fearful	119	118	119	119	119
surprised	122	122	122	122	122
total	779	775	772	763	759

happy, disgusted, fearful and surprised ones.

5.3.3 Artefacts

Visual artefacts were expressed as solid black regions and were mostly localized in the area around the mouth (see Figure 5.2), such that the subsequent blending between both sets of points fails. In this approach, expressions were blended per subject (i.e. and not across subjects), where artefacts were found present in nearly half of the augmented images. Those images were discarded from the final augmented dataset as they would severely hamper the performance in the subsequent generation and recognition tasks.

When blending two Delaunay triangulations, facial features are aligned and stacked together producing an augmented image, however, regions, which lie outside the face contours are simply added and averaged together. Consequently, hairstyles of all blended subjects are visible in the resulting images as seen in Figure 5.2. In this case, this effect is marginal, as the augmentation is performed within-subjects and the RafD dataset consists of subjects that mostly have the same hairstyle in their depictions of different expressions.

5.4 Facial expression generation task

The model used in the generation task is a deconvolutional decoder network, previously used for morphing of objects in Figure 5.3 [177].

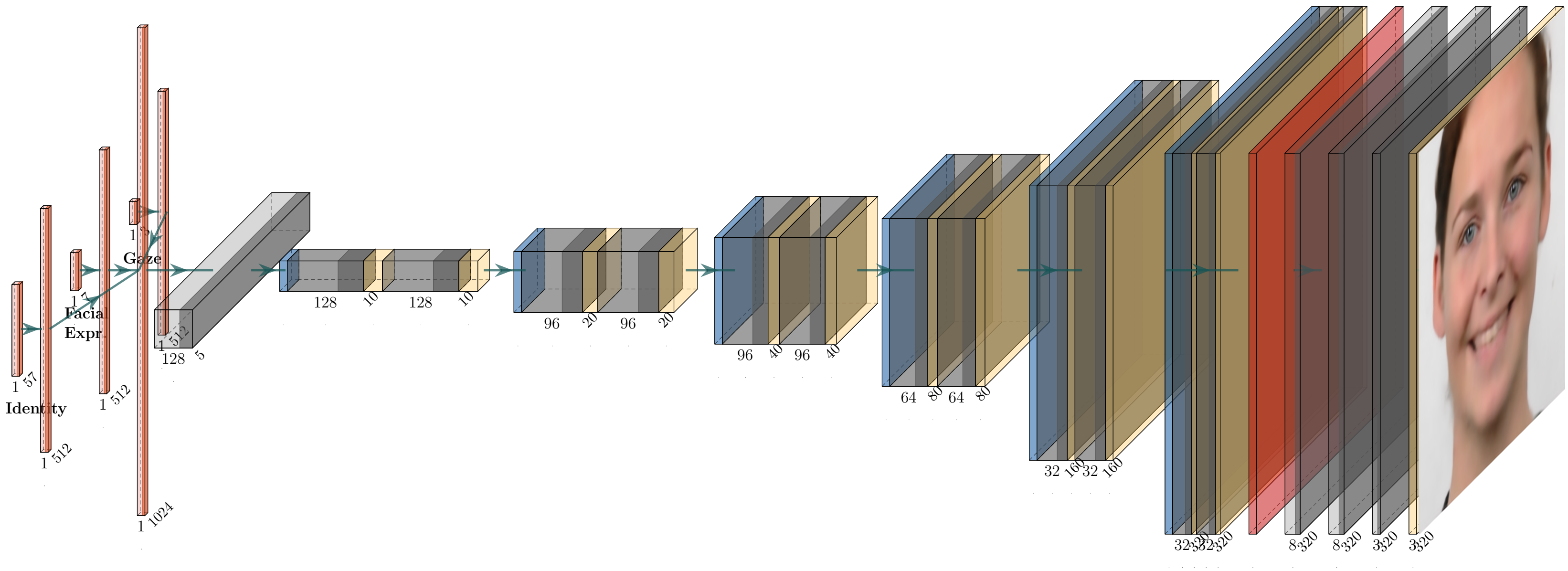


Figure 5.3: Topology of the generative deconvolutional neural network model. Each layers' volume can be found at its bottom right corner and correspond to the their width in the visualisation. Layers are colour coded by operation, where orange codes for dense layers, blue – for upsampling, grey – for convolution, yellow – for batch-normalization and red – for max-pooling. The first dense layer was rotated to improve the legibility of the plot.

It includes as an additional output a segmentation mask, which unnecessarily complicates the model and was removed [177]. The relative simplicity of the deconvolutional model allows better control over the input parameters and the subsequent interpretation of the output. In contrast, the more sophisticated Generative Adversarial Networks (GANs) [178] do not guarantee to converge [179] and therefore provide less consistency of their output. The model was implemented using the TensorFlow library [180] as TensorFlow allows more flexibility to fine-tune individual layers and apply minor optimizations. For training, both the original RafD dataset and the augmented one were used. The model loss was computed based on the labels from the original dataset and the soft labels created by the augmentation. An early-stopping criterion was used, which concluded the training after 96 epochs.

Figure 5.4: Output from the generative model trained on the original and augmented RafD dataset for a single identity. Images on the right depict the neutral expression while those on the left – the facial expressions for happiness, sadness, disgust, anger, fear, surprise and contempt at their maximum intensity. The intermediary images represent in-between intensities sampled at increments of 0.1.

Figure 5.4 displays the results from the generative model for the facial expressions of happiness, sadness, disgust, anger, fear, surprise and contempt each presented in successive rows. Images on the left depict expressions at their maximum intensity and each subsequent image to the right represents a decrement of 0.1 in portrayed emotion intensity towards the neutral expression.

The model was able to produce realistic-looking depictions of facial expressions for all emotion intensities with particular facial features appearing crisp and detailed. Instance normalization produced significantly better results over batch normalization. Furthermore, uneven class distribution, as expected, plays a role when the amount of augmented training samples outweighs the original dataset. The number of samples between the augmented and original dataset was at a proportion of approximately 3:1. Including both dataset in full resulted in artefacts from the augmentation being present in the generated images. Therefore, the training set was balanced to a uniform distribution by oversampling the originals. This approach also mediated the best results.

In contrast, faces generated only with the original RafD dataset appeared blurred and the intermediary expressions were rather ambiguous. Omitting the augmentation does not provide sufficient training data for the generative model to learn textures such as those for teeth. In this case, the model substitutes texture for teeth with that for lips or skin in most intermediary expressions featuring an open-mouth such as those for anger or happiness. This implies that the network does not learn the meaning behind facial expressions or rather the interaction between individual features, albeit it apparently learns low-level and in subsequent layers high-level feature representations in order to adequately generate them.

Using the augmentation alone does not create perfect synthetic facial expressions either, as it is prone to artefacts itself (e.g. shadows when blending images of people with different hair styles). Using the augmentation in combination with the original dataset appears to provide suitable synthetic images for the model to learn accurate approximations for facial expressions of intermediary intensities. This was also the case for subjects in the dataset, for which there were relatively few augmentations available due to artefacts in the augmented images (see Table 5.1).

5.5 Facial expression recognition task

To establish whether the augmentation can improve results for facial expression recognition classification, the quality of the augmentation was evaluated using a downstream task. In this manner the performance efficacy of networks trained on the original dataset and the augmentation could be compared. The DenseNet121 [181] and ResNet50 [182] networks were used as they have achieved impressive results in various classification tasks. A suitable metric for evaluating generative models is the Classification Accuracy Score (CAS) [183] which was subsequently used. Both DenseNet121 and ResNet50 were pre-trained on ImageNet [184]. Soft labels generated by the augmentation were used as ground truth for facial expressions of intermediary intensities. It is important to note that, while the soft labels (e.g. intensities of emotions) in the original dataset are either 1 or 0 as there are no intermediary expressions, the augmented ones only consists of ones in the intensity range $[0.3, 0.7]$ (see Table 5.1). For training, both original and augmented datasets were split in a 90% training- and 10% test-sets, stratified according to subject. For

testing, since the augmentation evaluation is done as a downstream task, only the test-set from the original dataset was used. As mentioned above, those included only binary labels. The DenseNet121 and ResNet50 models were adapted to predict 7 classes, representing expressions available in the dataset, in place of the default 1000 classes. The training parameters consisted of ADAM optimizer [185] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-8}$, learning rate $\alpha = 0.001$ and no weight decay. The model was evaluated using two additional strategies, where in A) no augmentation was used and B) further standard augmentations were used consisting of rotation (+-30), scaling with a random factor between 0.8 and 1.2, and shear random factor between 0.9 and 1.1. The experiment results can be seen in Table 5.2.

Table 5.2: Facial expression recognition accuracy scores from ResNet50 and DenseNet121 models where Augmented and Original refer to the RafD augmented and original dataset and A) no additional augmentation was applied and B) rotation $(+/-30)$, scaling with a random factor between 0.8 and 1.2, and shear random factor between 0.9 and 1.1 was applied.

	DenseNet121		ResNet50	
	Augmented	Original	Augmented	Original
A	0.965	0.965	0.972	0.965
B	0.993	0.958	0.958	0.972

While the performance for the RafD dataset are reaching human level of performance (recognizing simple emotions in high resolution images is a relatively simple task for deep learning models), the results hold true also for the augmented data. In addition, using only the augmented images performed better or on par, compared to the original images in every situation except for ResNet50 B). An interesting observation is that augmented images do not include facial expressions at maximum intensity but have been evaluated on such. This implies that training on augmented images featuring medium intensities of portrayed emotions was sufficient to classify those at maximum intensity as well. In this case, cross-entropy with soft labels was not symmetric due to the fact that 1) soft-labels in the augmentation weigh towards the neutral expression and 2) the augmentation for certain expressions is more prone to artefacts (e.g. sadness and anger), which resulted in a unbalanced training samples between classes. Contrary to expectations, however, this did not appear to hamper the CAS scores for the model trained only on augmented images. This is a positive result, as it implies that the networks learn class distribution on- or almost on-par and in some cases even better compared to using the original dataset, which makes both dataset practically interchangeable for training.

5.6 Discussion

For the purposes of evaluating its efficacy in the generation and recognition of nuanced expressions, augmentation were created only for nuanced expressions within-subjects. However, Delaney triangulation can also be applied in a person-agnostic manner as long as facial feature landmarks can be correctly identified to create novel morphed identities from existing faces. Those blended pseudo-identities would be of sufficient quality to be reliably used as additional labelled data from an existing dataset. A positive side effect of using multiple faces within one such blend is that with an increase of identities, artefacts caused by the augmentation method are reduced as the contribution of individual images to the blended image is reduced. Recently, it has also been a topic of discussion that many dataset are biased and are not representative for group diversity. This method can be particularly useful to correct or mitigate for this by generating morphed samples for under-represented groups. In turn, more balanced dataset would help produce less biased algorithms relying on facial expression generation or classification.

Holistic facial-expression recognition approaches can benefit from such augmentation as well. It can provide plausible samples for training, which can help classification models to learn more accurate decision boundaries between classes, thus reducing misclassification for nuanced facial expressions. Currently, more expressive features of the face are weighted more such that they disproportionately influence classification. Alternatively, using the presented method to generate augmented images for nuanced expressions can be applied to strengthen the interdependence of formant facial features contributing to distinct facial expressions in recognition tasks.

The use of the deconvolutional generative model allows to create gradients of emotion intensities between two arbitrary points in the latent space, expressed as a sequence of facial expression images morphing between arbitrary emotion classes and intensities. As such, facial expression-based scales can be created for and between arbitrary emotion classes. Additionally, the use of the augmentation improves the quality of the generated images where interfaces relying on realistic portrayal of emotional faces can leverage it and benefit from a realistic and truly continuous facial expression feedback.

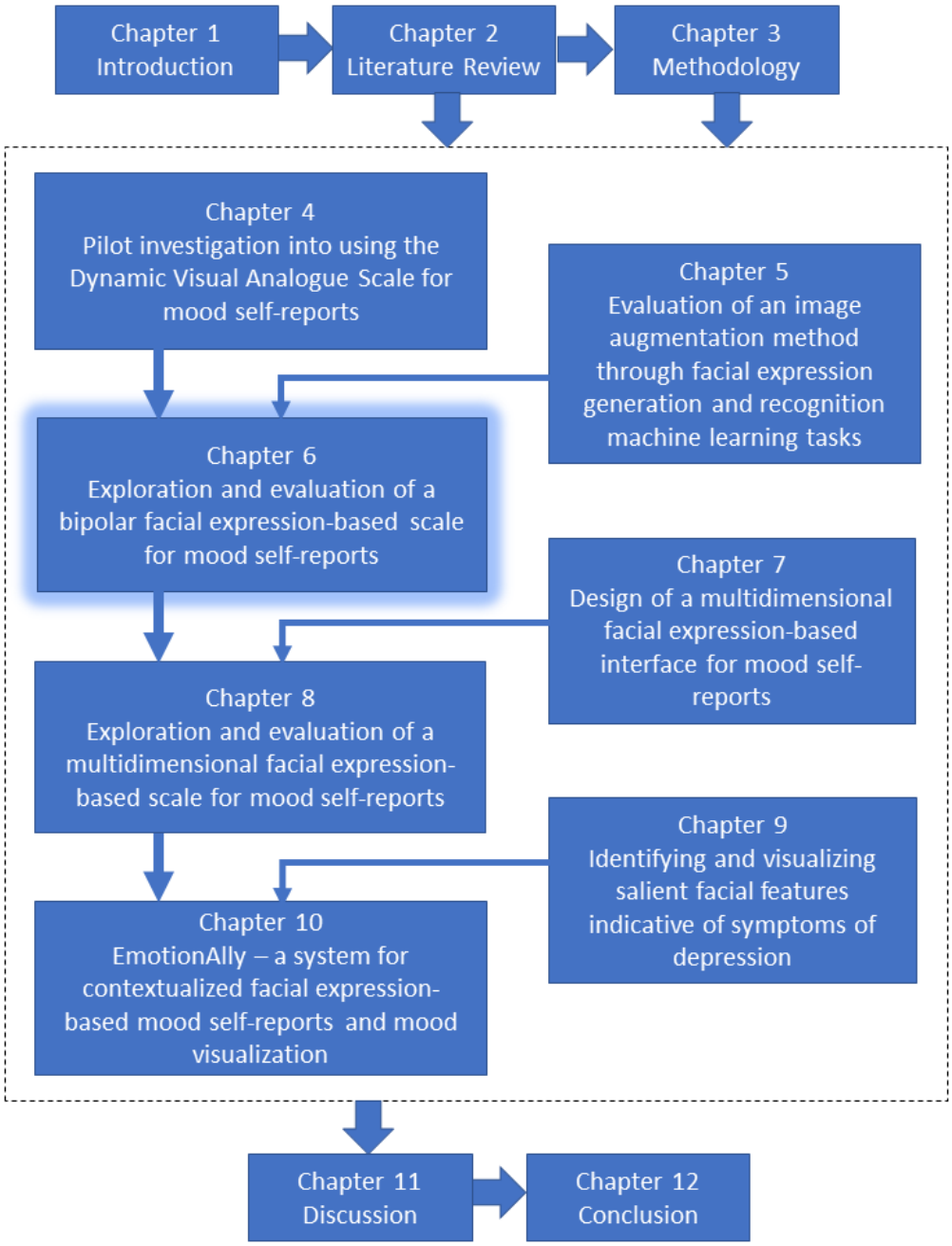
5.7 Conclusion

In this chapter, a known method for blending images of faces was evaluated in its use to create an augmentation for an existing facial expression dataset. The proposed approach was applied to create labelled facial expressions portraying varying intensities of emotion. The augmentation was evaluated in a facial expression generation and recognition tasks and the results indicate significant improvements in generating facial expressions when using the augmented dataset in combination with the original one. In a facial expression recognition task, results indicate

that the original and augmented dataset can be used interchangeably as both achieve near-similar levels of classification accuracy. There are benefits to using this augmentation for a wider set of tasks, in particular for networks that aim to recognize intensities of facial expression beyond simple categorical classifications.

5.8 Chapter Summary

In this chapter, an effective and inexpensive method to create a labelled augmented facial expression dataset consisting of expressions portraying varying intensities of emotion from a categorically labelled one was developed and evaluated in a facial expression generation and recognition tasks. In the generative task, it yielded images of sufficient quality and resolution, while in the latter, models trained on the augmented dataset only achieved near- or in some instances better results compared to those trained only on the original dataset. Additionally, potential caveats when working with the augmented data were highlighted such as uneven sample distribution of labelled classes, particularly when working with soft labels. The methods' simplicity makes it easily reproducible as well as applicable for further facial expression generation and recognition tasks. The application of this augmentation can address contemporary problems in deep learning algorithms related to improvements in nuanced facial expression recognition and generation, uneven sample distribution and de-biasing facial expression datasets. The use of machine learning approaches such as generative deconvolutional networks carry the benefit of being able to synthesize gradients of emotion intensities in facial expressions where the augmentation significantly improves the quality of the generated images. Consequently, interfaces using facial expressions to capture constructs such as emotion or mood can benefit greatly from this approach.


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 6

Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports

6.1 Introduction

In Chapter 4 the Dynamic Visual Analogue Scale (D-VAMS) [27] featuring photographs of real persons enacting the facial expressions of happiness and sadness was used to capture assessments of complex emotional states such as awe, guilt, compassion and others elicited by a number of vignettes. Those findings indicated that complex emotional states can be captured rather well using the basic emotions of happiness and sadness.

The goal of this study was to investigate whether a facial expression-based scale is applicable as a tool for self-reporting mood in daily life. For this purpose, the Ecological Momentary Assessment (EMA) framework [17], a widely-used approach for providing ecologically valid self-reports (Chapter 2, Section 2.3.4) was utilized as the basis for facilitating the mood self-assessments. Through the use of EMAs self-reports are provided close to time of experience and maximize ecological validity, while mitigating the effects of the retrospective recall bias [95, 96]. Chapter 2, Section 2.3.4 elaborates in detail EMAs, their use-cases and benefits.

In Chapter 5 a blueprint was presented for generating realistic facial expressions through the use of a generative deconvolutional neural network model. The model was trained on a the Radboud Faces Database (RafD) dataset [8] and an augmented dataset consisting of blended facial expressions consisting of intermediary intensities of emotion. Subsequently, in this Chapter, the facial expressions for sadness and happiness as created by the generative model were used as feedback within a bipolar facial expression based assessment scale. The scale borrows the smartphone application prototype design used in Chapter 4.

6.2 Methods

6.2.1 Study Design

The experiment used a within-subjects design with two independent variables – assessments provided with a facial expression-based happiness-sadness scale (FEAS) and an equivalent visual analogue scale (VAS) [19]. Each participant received 5 notifications per day at semi-fixed intervals on their smartphone for the duration of 2 weeks, prompting them to self-report their mood using both scales.

6.2.2 Participants

Participants were recruited through fliers distributed across the High Tech Campus in Eindhoven, The Netherlands and Technical University Eindhoven (TU/e). Participants were required to be between 18 and 65 years of age with no history of mental disorders, have a good proficiency in English and own an android smartphone. 37 participants took part in the experiment, where 32 out of 37 (12F/20M, M=29, SD=9) provided information about their age and gender. All participants were required to provide consent and were remunerated for their participation with a voucher for several online and physical stores in the Netherlands worth €15.

6.2.3 Materials

Figure 6.1: Figure containing computer-generated facial expressions for the emotions of happiness (top) and sadness (bottom) for the intensity range of no emotion (0) to peak emotion (1) at increments of 0.1 from right to left respectively. The expression on the utmost right position for sadness and happiness is the neutral expression, featured in the centre-point of the scale. Each dimension within FEAS was represented by 100 such images.

An android smartphone application was developed for the purpose of this study. It featured a facial expression-based assessment scale portraying the expressions of happiness, sadness and the neutral expression. Underneath, the scale was represented as a bipolar scale with 201 discrete points. The happiness and sadness dimensions each featured 100 distinct points, representing increasing intensities of emotion. The sadness dimension was allocated in the $[0 - 99]$ range while the happiness dimension in $[101 - 201]$ range, where the lowest and highest points

in those ranges portrayed those expressions at peak intensities respectively. The neutral expression was allocated to the centre-point of the scale. Figure 6.1 portrays a smaller subset of the range of intensities of emotion for the happiness and sadness dimensions presented in FEAS.

The images were created using the neural network model presented in the generative task described in Chapter 5. The model was sampled on the respective categorical dimension at increments of 0.01 in order to produce 100 images for each one. The underlying numerical mapping was obscured from participants, where as in the previous experiment in Chapter 4, the only visible element of FEAS was an image containing the facial expression feedback.

Navigating through FEAS was accomplished using a vertical gesture by sliding up or down on the image, increasing or decreasing the intensity of the portrayed emotion. For example, sliding upwards when at the neutral expression would display increasingly happier facial expressions and sliding downwards – sadder. This design rationale was made in order to prevent participants from simply transplanting their assessments from either VAS to FEAS or vice-versa.

The application also included a bipolar VAS scale, implemented as a pseudo-continuous horizontal slider featuring the emotion dimensions for happiness, sadness and neutral mapped onto 201 discrete points. VAS featured labels denoting the emotion categories for happiness and sadness on the right and left end of the slider respectively and the neutral dimension was represented by its centre-point. The application logged automatically each interaction with either FEAS or VAS as a timestamp and a numerical value.

A digitized version of the System Usefulness (SYSUSE) and Interface Quality (INTERQUAL) subsets of the Computer System Usability Questionnaire (CSUQ) [156] were administered separately for each assessment scale. Additionally, a semi-structured questionnaire comprising of a mixture of open- and closed-ended questions (see Appendix 6.A) was used to investigate aspects of the scale, such as the user experience with each interface, the utility of facial expressions as a feedback mechanism, and questions prompting participants to share their intuitions.

6.2.4 Procedure

Upon expressing interest in the study, prospective participants were sent an information letter detailing the experiments' aims, procedure, risks and burdens. An appointment was scheduled with willing participants and on the arranged date, they were accompanied to an on-site lab. There, they were handed a printed copy of an informed consent form and privacy notice detailing the content of the experiment, the task they are expected to accomplish and how their data will be used. Participants were given ample time familiarize themselves with the contents and were allowed to ask any questions pertaining to the experiment, their participation and the outcome of their data. After providing written consent, a mobile application was installed on their smartphone.

In the following two weeks, the application triggered notification prompts on

the users' device at semi-fixed intervals. A total of five notifications per day were delivered at 10, 12, 16, 18, and 20 o'clock including a randomization jitter of ±30 minutes. For each prompt, participants were asked to assess their mood using both the FEAS and VAS scales, where those were presented sequentially in a randomized order. Prior to each self-report, both scales were initialized on the neutral position – the center-point for VAS and the neutral expression for FEAS. Participants could also self-report using the application on their own initiative apart from responding to a notification should they chose to do so.

After two weeks had elapsed from the date the application was installed, a notification was triggered which directed participants to fill out a digital version of the CSUQ questionnaire [156] featuring the System Usability (SYSUSE) and Interface Quality (INTERQUAL) subsets. This questionnaire was filled out once for each assessment interface. Then, participants were directed to a screen allowing them to submit their self-reports and CSUQ data digitally. Upon receiving data from a participants, a user experience questionnaire was sent asking about their experience with the interfaces (see Appendix 6.A).

The study took 3 months to complete and took place between 1st of October and 31st December 2019 counted from the date of enrolment of the first participant to receiving data from the last.

6.2.5 Statistical analysis

Linear regression models were used to compare assessments between FEAS and VAS on each emotion dimension separately. This resulted in three types of assessment categories – 1) those rated on the happiness dimension, 2) those rated the sadness dimension and 3) neutral assessments. Those were clustered accordingly by assuming the following schema – 1) assessments rated on the happiness dimensions on both FEAS and VAS, 2) assessments rated on the sadness dimension on both FEAS and VAS, 3) assessments rated as neutral (or the centre-point) on either FEAS or VAS. This partition resulted in 1422 (77.6%) assessments on the happiness dimension, 249 (13.6%) – on sadness and 107 (5.8%) classified as neutral, where 47 were rated as such on both scales, 8 only on FEAS and 52 only on VAS. 55 (3%) remained undifferentiated as in the paired assessments between both FEAS and VAS, they featured an assessment on the happiness and sadness dimensions. Due to their small contribution relative to the total amount of assessments, it was assumed that those ambivalent responses mostly comprise of assessment errors and were excluded from the subsequent analyses.

Complementary to the LR models, the range utilization of both scales was compared by visualising the interdependence in the distribution of assessments for both FEAS and VAS by computing a kernel density estimation (KDE).

Additionally, both scales were compared by the time required to provide an assessment. Therein, median duration was taken, rather than mean as there were 51 assessments (43 on the happiness, 1 on the sadness and 7 on the neutral dimension) which took longer than a minute. Duration of assessments for each

emotion dimension were examined for significance using a t-test.

Quantitative results of the CSUQ and questionnaires were collated and examined for significance on each subscale using a t-test. Some answers were either missing, did not match the premise of the question or could not be interpreted as a distinct categorical response where those have been subsequently omitted. A thematic analysis was conducted on the open-ended questions.

The numerical data were analysed using Python 3.6 and numpy and pandas libraries [169]. The visualisations were created using the seaborn library [148]. 13 self-reports were not recorded correctly by the application and were discarded from subsequent analyses.

6.3 Results

In total 37 dataset were collected consisting of 1833 assessments (M=49.5, SD=17.05) or 70% of the expected 2590 assessments. Table 6.1 contains the

Dimension	Type	n	M (SD)
Happiness	FEAS	1422	63.11* (±22.62)
Happiness	VAS	1422	39.80* (±21.38)
Sadness	FEAS	249	44.59¶ (±25.27)
Sadness	VAS	249	36.04¶ (±22.31)
Neutral	FEAS	107(8)§	17.21 (±27.18)
Neutral	VAS	107(52)§	-0.51 (±4.76)

Table 6.1: Table containing the sample size (n), mean and standard deviation (M (SD)) for assessments provided with either FEAS or VAS grouped by emotion dimension. The neutral dimension includes assessments rated on the centre-point of either FEAS or VAS. Ambivalent responses (n=55) featuring assessments on both the happiness and sadness dimensions with either FEAS or VAS were excluded. *The happiness dimension was normalized to the [0-100] interval. ¶The sadness dimension was inverted within the [0-100] interval with 0 representing the lowest intensity of emotion and 100 – the highest, cohesive with the happiness dimension. §The numbers in brackets denote assessments lying on the neutral position for the respective assessment scale.

descriptive characteristics of assessments split by emotion dimension and used scale (e.g. FEAS or VAS).

6.3.1 Inferential statistics

Linear regression models (LR) were used to evaluate whether and to what extend assessments provided on FEAS and VAS are correlated where those were computed

Dimension	slope	intercept	r	p	95% CI [LL, UL]	SE
Happiness*	0.63	0.27	0.67	.000	[0.64, 0.69]	0.02
Sadness¶	0.61	8.87	0.69	.000	[0.62, 0.75]	0.04
Neutral	0.02	98.07	0.07	.48	[-0.12, 0.26]	0.02

Table 6.2: Parameters describing the linear regression models fitted on the emotion dimensions for happiness, sadness and neutral with both the FEAS and VAS. The columns indicate the linear regression slope (s), intercept (i), correlation coefficient (r), significance value (p), 95% confidence intervals (CI) and standard error (SE) *The happiness dimension was normalized to the [0-100] interval. ¶The sadness dimension was inverted within the [0-100] interval, such that 0 represents the lowest intensity of emotion and 100 – the highest, cohesive with the happiness dimension.

separately for each emotion dimension. Table 6.2 presents an overview of the linear regression model parameters for each dimension including linear regression slope (s), intercept (i), correlation coefficient (r), significance value (p), 95% confidence intervals (CI) and standard error (SE). The regression visualisations can be found in Figure 6.2. The results indicate a strong correlation between assessments on both – the happiness and sadness dimensions with FEAS ($r(1422) = 0.67$, $p < .001$) and VAS ($r(249) = 0.69$, $p < .001$). Those correlations were statistically significant with 95% confidence intervals for happiness [.64, .69] and sadness [.62, .75]. The LR model for neutral assessments showed no correlation. The slopes of both regression models indicate a proportional increase in the assessments within each emotion dimension.

Complementary to the LR models, the range utilization of both scales was investigated by visualising the interdependence in the distribution of assessments for both FEAS and VAS by computing a kernel density estimation (KDE, see Figure 6.2). While the LR models were informative of how assessments both scales capture the emotion dimensions of happiness and sadness, they tell us little about the underlying distribution of assessments on either FEAS or VAS. KDE reveals the underlying probabilistic distribution over a two-dimensional space spanned by assessments provided with both scales. This visualisation can reveal systematic inter-dependencies between ranges of values. Using this visualisation, a noticeable difference was the relative low density of assessments in the [70,90] range for the VAS scale on both emotion dimensions, which was not observed in those for FEAS. There, inversely, a higher spread of assessments was observed, where for example the [40,60] range on the VAS happiness scale roughly corresponds to a [50,100] range on the FEAS scale. Although the dimension for sadness features significantly less assessment points, a similar trend was observed for the same [70,90] VAS range.

The median time required to provide a self-report with each scale per emotion dimension was 1.3 seconds for VAS and 2.7 with FEAS (see Table 6.3). The difference in response times between scales and per emotion dimension were

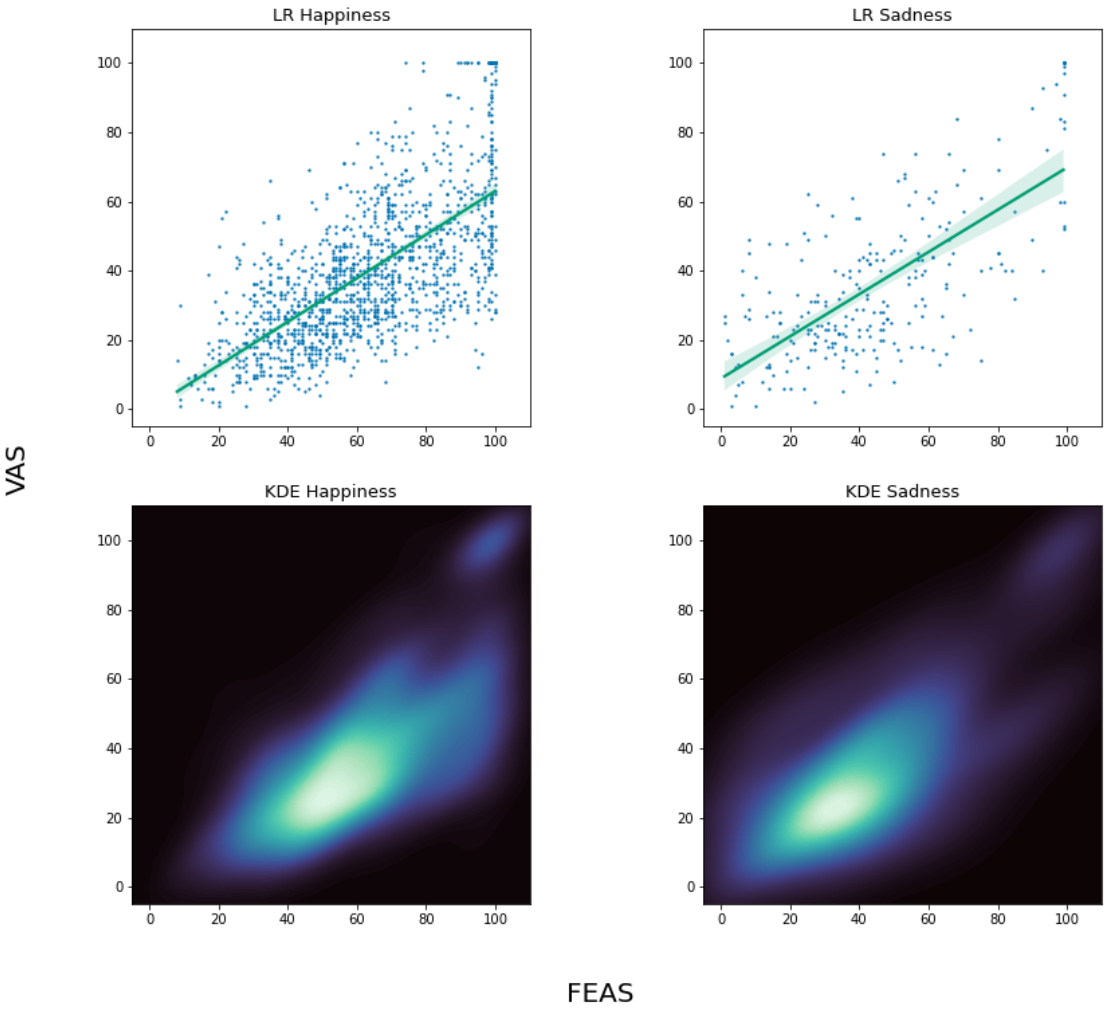


Figure 6.2: Linear regression models of the assessments provided with the VAS and FEAS scales separated by emotion dimension (top) and the respective KDE plot (bottom). The regression line visualisations uses jitter on the x-axis data-points for overlapping assessments.

significant ($p < 0.001$).

The scores obtained on the Computer System Usability Questionnaire (CSUQ) for both System Usefulness (SYSUSE) and Interface Quality (INTERQUAL) subsets (see Table 6.4) were not significant.

Participants rated the happiest and saddest expressions in FEAS as representing their respective emotion dimensions at their peak intensity on a scale between 0 and 10. The happiest expression was rated on average as 8.2 (n=32) and that for sadness as 9 (n=28).

Type	Happiness*	Sadness*	Neutral*
FEAS	M=2.7, SD=3.8	M=2.7, SD=3	M=2.5, SD=5.1
VAS	M=1.4, SD=2.6	M=1.3, SD=1.7	M=0.7, SD=2.6

Table 6.3: Average duration in seconds for providing a self-report measured as the period between an respective interface being displayed to the user and its last interaction. *All computed comparisons were significant ($p < 0.001$).

	CSUQ	
Scale	SYSUSE*	INTERQUAL*
FEAS	17.4	8.9
VAS	17.9	8.9

Table 6.4: CSUQ mean scores for FEAS and VAS on the System Usefulness (SYSUSE) and Interface quality (INTERQUAL) sub-scales, rated on a $[1, 7]$ Likert scale. Lower is better. *Neither subscale difference in scores was significant ($p > 0.05$).

6.3.2 Qualitative results

Two explicit mood-matching strategies were outlined – active and passive in order to discern how participants used a facial expression to indicate their mood. *Active* implies that participants explored how they felt prior to or while self-assessing their mood and actively tried to find a matching expression on the interface (*"does the person on the interface feel like I feel?"*). *Passive* implied that they browsed the interface, visiting multiple emotion intensities first until they found one whose intensity matched their mood (*"do I feel like the person on the interface?"*). 21 participants used the active strategy to report their mood, 7 – the passive one, and 4 indicated they used both – each at different instances.

Comparing FEAS and VAS as the method for self-reporting mood, 19 participants indicated a preference for the facial expression scale, 7 for VAS, while 4 preferred a combination of both.

Two questions aimed to establish whether participants had a preference for the appearance of the facial expression assessment model and to extract a specific desired identity for it respectively. 16 participants indicated a preference for a model of similar appearance to themselves: *"I would prefer a model of my own because it would allow me to more accurately project [...] my emotions on the interface."* From this group, 5 participants further emphasized on the importance of the age of the model, where the indication was for an preference in a range similar to theirs. Additionally, 3 participants indicated a preference for gender matching their own: *"I would prefer a male model simply because I would associate with the facial expressions more easily."* 6 persons shared that they have no particular

preference towards the assessment model. 4 people indicated that they would like to see a model of themselves rather than a generic or an otherwise person anonymous to them or a person of similar appearance: *"I think that, during the assessment, my mind is trying to empathize with face on the screen. Having photo of myself or a face that is very similar to mine could help me to empathize more."*

Identifying whether participants had a concrete identity they wished for the model to assume, 17 participants indicated they would like to see a generic or otherwise person unknown to them: *"No, the more neutral the better."* Those included the majority of people that wished for a facial expression assessment model similar to, but not a mirror-image of them: *"I would not like to have my own face but I would like the model to have the same ethnicity, age and gender."* 12 persons preferred a model of themselves: *"I would prefer myself rather than another person that I know."*

Establishing whether either assessment method allowed participants to gain insights about their own mood, generally most participants indicated that they were more aware of how they felt due to their participation in the experiment. In particular, simply issuing a prompt appeared to increase ones' awareness of their own moods: *"The daily reminder "Take a moment to assess your mood" actually works a bit as a reflection moment."* This 'reflection moment' was encapsulated by another participant as well that used this information to formulate actionable strategies to improve their mood: *"I've got a lot of useful insights. Being more aware of my mood is a first step to understand why and when do I feel sad and nervous and which activity helps to improve my mood."* Another participant appeared to intuitively engage in emotion regulation techniques such as a form of savouring in order to reassess their mood prior to making an assessment: *"Yes, sometimes after getting the pop-up rather than instant reply I thought about good things happening around and marked my response."* Additionally, self-assessments with the facial expression interface appeared to be more meaningful: *"The facial expression interface is useful since I can better identify my mood with the face."* On the question of perceived ease of use, all participants estimated unanimously both scales to be easy to use.

6.4 Discussion

Results indicated that assessments made with FEAS and VAS are strongly correlated nearing 70% for both dimensions of happiness and sadness. The strong correlation was an indication that both scales captured mood similarly to one another. Emphasizing the significance of the correlation was the fact that the navigation scheme implemented in FEAS did not feature any numerical quantification apart from the facial expression feedback as an indicator of emotion intensity. Although one could still approximate a numerical value for FEAS in relation to the subjectively perceived distance from the neutral face or either of the extremes, it was not as straightforward to do so precisely as the bipolar scale featured 100 images per emotion dimension.

Evidently, there was an imbalance in the sample sizes of assessments provided on dimensions for happiness and sadness. This is not unexpected, as participants were healthy adults with no self-reported mental-health issues. Despite this imbalance, however, in the juxtaposition between FEAS and VAS assessments in Figure 6.2, some discrepancies were seen in the spread of assessments over the range of both scales on both emotion dimensions. It is important to note that, when interpreting a KDE plot, one has to think in terms of probabilities derived from the observed sample. In this way, a generalization on the likelihood of future assessments following the trend of the observed sample can be made and subsequently used as a basis to evaluate the underlying systematic usage of the scales. In particular, it appears that assessments provided with VAS are less probable to be found in the $[70, 90]$ range.

An explanation for that may lie in the way people typically interpret and interact with analogue scales. Works in the literature examining VAS identify biases relating to disproportionate utilization of their range [23]. The authors, similarly to the observations herein, identified an end-aversion on a bipolar scale and conclude that VAS scales may be well suited for an ordinal arrangement of items, but were susceptible to biases when used to measure cardinal values. This effect occurs by making a comparative judgments according to the labelled portion of the scale, which leaves any unlabelled range thereof subject to subjective interpretation [20]. A comprehensive examination of different types and arrangements of VAS scales [22] further confirmed this observation, where it was found that assessments tended to coalesce towards labels placed alongside the range of the scale, while ranges intermediary to them remained depressed. A combination of those effects could offer an explanation why the extremes ends of the VAS scale in this experiment were still utilized, while ranges nearing those extremes were not. While a labelled portion of a scale, for example its extreme was easy to interpret, participants were uncertain on the difference in meaning between the point at the extreme end of the scale and those near it, subsequently resulting in the observed distribution of assessments.

On the other hand, the happiness dimension of FEAS appeared mostly underutilized in its lower range, such that assessments lying in its upper range appeared more spread out relative to their paired VAS counterparts. An explanation may lie in how facial expressions spanning FEAS were sampled. The images originate from a generative neural network model (Chapter 5) trained on enacted emotions, where the training data for both sadness and happiness consisted of 6 images – an original image portraying the emotion at its maximum valence and 5 artificially augmented ones representing the intensity interval [0.3,0.7] for 57 identities. Subsequently, the model cannot generate facial expression intensities beyond the range contained in the training dataset. This was anticipated and as part of the user experience survey, participants were asked to rate the happiest and saddest expressions on a 10-item scale for how descriptive those were for their respective subscale at peak intensities. There, participants rated the happiest facial expressions as 8.2, while the saddest as 9 which may be an indication that the

most expressive intensity for the happiness emotion dimension may have been less expressive and subsequently offered less ranges of values for self-reports in the higher range of the FEAS happiness subscale. This effect was not observed for assessments provided on the sadness dimension, however, it also featured less assessments and as the participant sample consisted predominantly of healthy adults, it was not expected to see many assessments on the extreme end of the sadness subscale. However, this is an indication that when designing a facial expression-based scale, some considerations need to be made with respect to the intensity range of emotion dimensions such that they cover a range of facial expression intensities aligned with the users' idea of how they should appear at peak intensity within an emotion dimension

Additionally, while FEAS appeared to capture mood in a more consistent and gradual manner over its span compared to its VAS counterpart, expressions of low intensity for the happiness dimension were also observed to be used less frequently, particularly so for the range of [0-15]. While this is not surprising as that range coincides with the neutral dimension, it could be an indicator that those intensities were subtle and not evocative enough to be categorically associated with happiness. As elaborated in Chapter 2, Section 2.4.1, there is a minimally perceptible threshold of emotion that needs to be present in order to categorically detect an emotion in a face [186]. Nevertheless, those results may be an indication that the granularity of the FEAS scale achieved using computer-generated facial expressions with a mere 100 images for an emotion dimension may be sufficient such that those could portray emotions at the boundary of what is perceptible to a user. This could be particularly useful in specific use-cases such as capturing biases exhibited in the ability of a patient living with depression to recognize specific nuanced expressive faces as described in Chapter 2, Section 2.5.2.

The lack of a significant correlation for the neutral dimension is not striking as the assessment space was effectively a single point on a bipolar scale. Hence, any deviations in ratings offset slightly in either direction could significantly affect those results.

The responses on the CSUQ questionnaire were not significant for either of the Interface Quality (INTERQUAL) or System Usefulness (SYSUSE) subscale. This is an indication that participants did not differ in a significant way in their ratings between the FEAS and CSUQ sale. However, while VAS scales are used in a plethora of use-cases and are pervasive in our societies, a FEAS scales consisting of computer-generated facial expressions is largely unexplored. In that sense, this can also be interpreted as a positive result, since it would indicate that both scales did not differ significant from one another on SYSUSE or INTERQUAL subscale of the CSUQ questionnaire. Furthermore, even though FEAS was on average slower than VAS, providing an assessment was sufficiently fast enough as to not discourage users from using FEAS.

Curiously, the simple act of triggering a notification was sufficient to increase some participant's awareness of their mood, where in one case a participant even, on their own initiative, engaged in actionable emotion-regulation strategies.

The general sentiment towards the FEAS scale from the qualitative user feedback positioned FEAS as the more preferred method for providing an assessment. It appeared to contain qualities which made it more personable through the evocative visual aspect of representing mood as a facial expression appears, which subsequently appeared to foster a deeper and more relatable experience. Additionally, the interaction with appeared to be the more engaging and facilitated a more meaningful connection.

Most participants indicated a preference in tailoring the FEAS assessment scale further in a way that would increase its appeal and make it more accessible to users. This wish was expressed mostly through altering the appearance of the person enacting the expressions on the interface, where most participants preferred to see a person that resembled them in their appearance. Specifically, more than half of the participants indicated that that person needs to only approximately mimics their appearance, while slightly less than half of the participants would have liked to see a digital doppelgänger of themselves. In sum, participants' feedback suggests that the prevalent majority would like to alter the appearance of the person enacting the facial expressions on the interface in some way, wherein participants appear to differ in how and to what extend those alterations are expressed. As self-reporting mood is inherently an introspective process, it is not surprising that the majority of participants either desired to see themselves or an identity resembling them and highlights the potential for such interfaces, where further customizing such tools remains largely unexplored.

6.5 Limitations

An exploration of the range utilization of both FEAS and VAS revealed intrinsic measurement patterns which are speculated to be attributable to how those scales were perceived. For VAS, those appeared to be associated with the interpretation of anchors on the scale which only featured two labels at each extreme end. Using further indicators along the range of the VAS scale might have resulted in more evenly distributed assessments or alternatively revealed that assessments tend to coalesce over indicators on a VAS range aligned with findings in prior work. For FEAS, the range of facial expressions, specifically those for the emotion dimension of happiness might have not featured sufficient ones representing high intensities of emotion. In turn, this discrepancy between persons may have affected its use negatively.

6.6 Conclusion

This study demonstrated that assessments provided on a facial expression-based happiness to sadness scale are strongly correlated to those provided on a VAS scale. This indicates that the FEAS assessment may be used reliably for assessing

mood. A novelty in this approach was the application of computer-generated facial expressions used to span the emotion dimensions for happiness and sadness. Furthermore, most participants indicated to extract more value from the FEAS' more evocative depiction of mood, and expressed an interest in personalizing the interface further highlighting the potential of future developments of facial expression-based self-report tools.

6.7 Chapter summary

This study aimed to investigate whether assessing mood in a monitoring context through the expressions of happiness and sadness compares to a VAS scale featuring the same emotion dimensions. Results indicated that both VAS and facial expression-based scales are strongly correlated and may be used interchangeably. Furthermore, most participants indicated to extract more value from the FEAS' more evocative depiction of mood, and expressed an interest in personalizing the interface further, highlighting the potential of future developments of facial expression-based self-report tools.

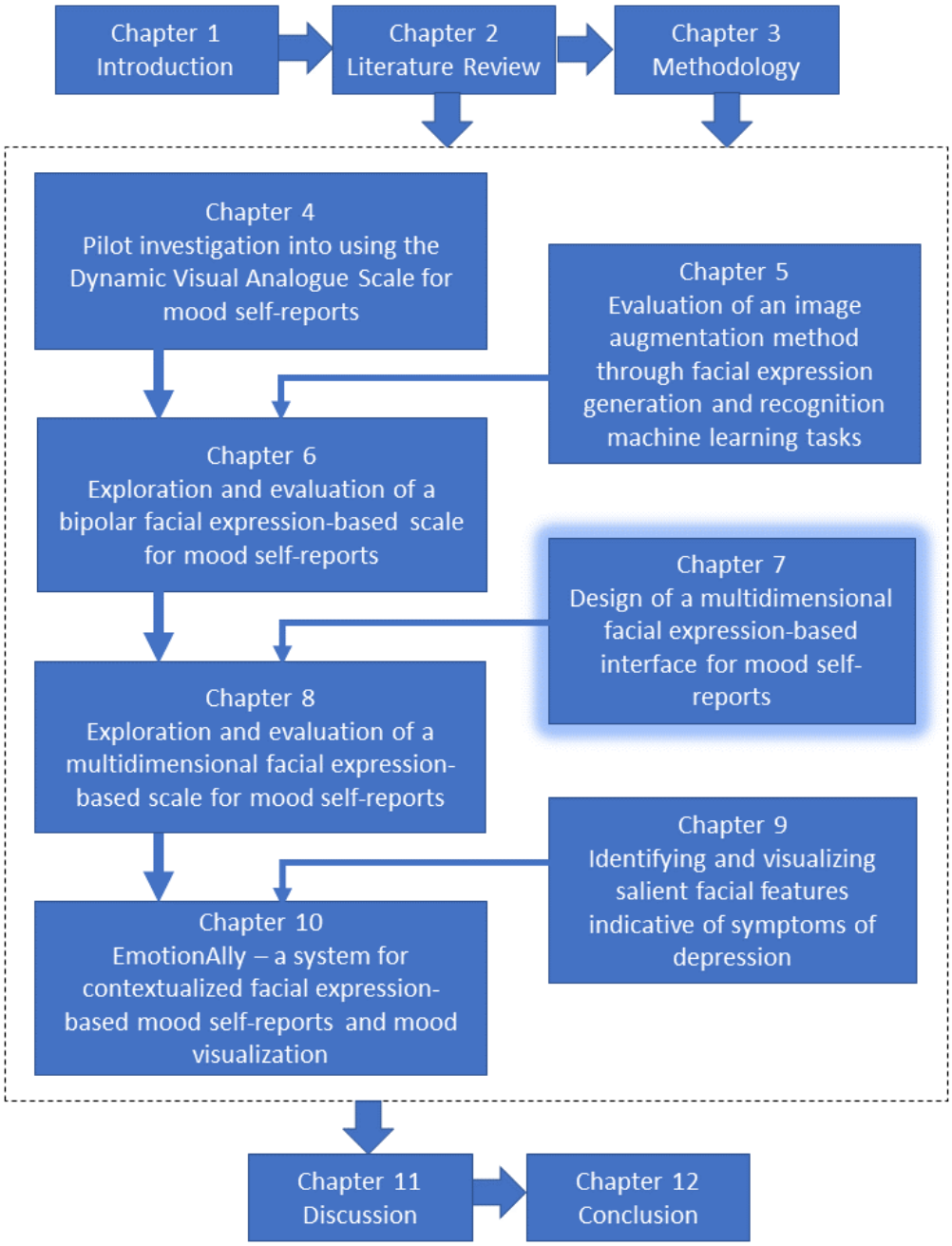
6.A Questionnaire

(a) The happiest expression on the FEAS interface.

(b) The saddest expression on the FEAS interface.

Figure 6.3: Happiest (left) and saddest (right) facial expression on the FEAS interface.

Participant information
What is your age? What is your gender?
User experience questions
Was it easy to find a facial expression intensity on the sadness-happiness scale, which corresponded to your mood? If no, why? Was it easy to find a spot on the sadness-happiness slider, which corresponded to your mood? If no, why? Does the happiest face on the scale (Figure 6.3a) represent extreme happiness in your opinion? If no, what value would you give to the happiest expression from 0 to 10 on a scale from neutral to extreme happiness? Does the sad face on the scale (Figure 6.3b) represent extreme sadness in your opinion? If no, what value would you give to the saddest expression from 0 to 10 on a scale from neutral to extreme sadness? When you were providing your assessments on the facial expression interface, (a) did you first try to figure out how you feel and then find a matching facial expression on the interface or (b) did you explore the facial expressions in the interface and try to compare it to how you felt to find your matching mood or (c) if neither of those, please elaborate Did you get any insights about your own mood, when providing the self-reports with the sadness-happiness facial expression interface? Did you get any insights about your own mood, when providing the self-reports with the sadness-happiness slider interface? Would you prefer a model of your own or a different gender, ethnicity or age when using a facial expression-based interface and why? Would you want to use a model of a particular person or yourself when assessing your mood? If you prefer a model of a particular person, who would that be? What did you like and what did you dislike about the facial expression interface? How can the facial expression interface be improved? What did you like and what did you dislike about the slider scale? How can the slider scale be improved? Were there any technical problems with the application? Do you have any other feedback about the use, looks, feel or anything you deem important regarding your experience with the application or the interfaces?


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 7

Design of a multidimensional facial expression-based interface for mood self-reports

7.1 Introduction

This chapter described a system design for a multidimensional facial expression-based scale. Chapter 4 investigated the use of a facial expression-based scale for capturing elicited emotions, where the scale comprised of the expressions for happiness and sadness portrayed by photographs of a real person. The emotions were elicited by a series of positive and negative vignettes featuring complex emotions such as joy, pride, disappointment and others. Assessments provided with the facial expression scale were compared to such provided on a visual analogue scale, where the indication was that both scales were strongly correlated.

Subsequently, Chapter 5 presented an approach relying on machine learning generative model to create a range of realistic human facial expressions at varying levels of intensities. Using the generative model also provided a numerical quantification of the emotion intensity for generated expressions. The computer-generated facial expressions were subsequently successfully applied within a similar happiness-sadness facial expression-based scale as the one used in Chapter 4 followed by an evaluation within a mood-monitoring study in Chapter 6. There, the correlation between the facial expression-based scale and its VAS counterpart were strongly correlated as well indicating that computer-generated facial expressions could also capture emotional content.

This provided sufficient evidence to believe that facial expression-scales 1) perform similarly to a visual analogue scale in the context of capturing elicited emotions and 2) the dimensions of happiness and sadness were able to adequately model and capture experienced positive and negative emotions. However, participant feedback from Chapters 4 and 6 indicated a desire for a facial expression-based interface, which features more emotion dimensions than those for happiness

and sadness alone. Since it was possible to use the generative model to create expressions beyond those for happiness and sadness, it was used to aid the creation of such as scale. This Chapter presented a practical implementation of one such multidimensional facial expression interface, whose design choices was motivated by leveraging the strengths and weaknesses of the machine learning generative model and existing theoretical knowledge on emotion theory. Moreover, this Chapter not only elaborated on the technical details of the design and implementation of this tool, but also provided the underpinnings for accommodating multiple discrete emotion dimensions and their intensities within an assessment interface. The end-goal was to create a robust method to accommodate an arbitrary number of facial expressions within a single assessment surface as part of a smartphone application, which can scale to include further novel expressions in the future. Additionally, due to the ubiquity of smartphones, such a tool would allow it to be used for self-reporting mood as ecological momentary assessments (EMA) [17] as well.

7.2 Designing a facial expression scale using computer-generated expressions

7.2.1 Generative model parameters

Conceptualizing a system which utilizes the facial expressions created by the generative model described Chapter 5 needed to adhere with the models' strengths and limitations. It is important to restate, that the dataset used for training the model consisted of the original and augmented The Radboud Faces Database (RafD) dataset [8]. Specifically, the generative models' input parameters consisted of three feature vectors (e.g. a vector of 1xN dimensions where N is the complete count of classes for a particular feature): those for eye-gaze direction, identity (i.e. person in the dataset), and facial expressions. The combination of those input features were used by the model to create an image which best fits those input parameters. As previously elaborated, the training dataset consisted of front-facing images only, portraying the facial expressions for anger, contempt, disgust, fear, happiness, neutral, sadness and surprise in 57 adults and with 3 eye gaze directions (i.e. left, right and centre). Therein, the RafD dataset consisted only of facial expressions at peak intensities, except that for the neutral expression, where the augmented one – as elaborated earlier in Chapter 5 – of emotion intensities in the range of [0.3,0.7] where 0 denotes absence of emotion and 1 – an emotion at its peak intensity.

Figure 5.3 in Chapter 5 depicts the architecture of the generative model, where the input feature vectors are presented on the left from which subsequent layers are used to encode feature variations and up-sample the user input, resulting in an image of a 550x550 pixels resolution. Input feature vectors were one-hot encoded (e.g. its elements are encoded within the $[0, 1]$ range), where, for example, each

element mapped for a distinct facial expression would represent absence of that emotion as 0 and its presence at peak intensity as 1. For example, a model that can synthesize 3 distinct categorical expressions is given as input a feature vector of [0 0.5 0], would in turn generate the facial expression represented by the element on the 2nd position at half-peak intensity. By modulating the feature vector values, expressions at varying intensities can be created. The feature vectors' data type is a 32-bit float, where 23-bits are reserved for the mantissa [187], which ascertained that the model can generate facial expressions intensities with high-granularity.

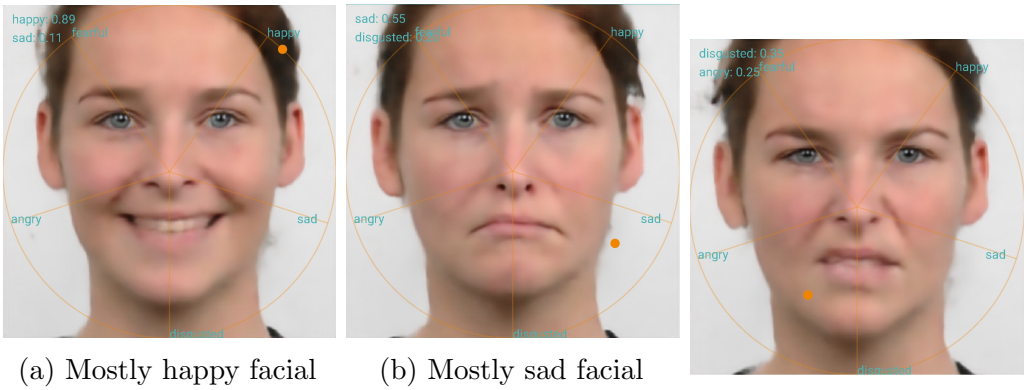
7.2.2 Facial expression navigation scheme

Designing an application for a mobile device allowed it to capitalize on the ubiquity of smartphones. Consequently, the navigation scheme employed to interact with the application needed to adhere to constraints imposed by a two-dimensional touchscreen most contemporary devices were equipped with. Thus, the organization of multiple facial expressions within a single surface needed to be i) *coherent*, such that each coordinate represents a distinct facial expression category and intensity and ii) *systematic*, such that facial expressions categories are separated according to their *between-class dissimilarity* and facial expression intensities clustered according to their *within-class similarity*. Or simply said, expressions belonging to the same class needed to be grouped and ordered such that their intensities gradually increment. Adhering to this organization would allow users, after a few interactions, to infer the expected facial expression category and intensity based on this cohesive spatial mapping even for previously unseen expressions, which would subsequently improve the tools' ease of use

A suitable choice to achieve this spatial organization was a polar coordinate system where the angle parameter determines the emotion dimension, while its radius – the intensity of the emotion. In a polar coordinate system, the space can be split infinitely by equidistantly placed radial lines. Using this organization, distinct emotions can be allocated on specific radial lines dividing the coordinate system. Those will be referred to as pivots from now on. On each pivot, the facial expression intensities were arranged in an incremental order such that a point laying on the circumference portrays an emotion at peak intensity and points close to the centre – of mild intensities. Over a pivot radial lines' coordinates images were arranged representing gradually increasing intensities of emotion. The polar coordinate system's centre-point represented the neutral expression as it does not posses a gradient of intensities.

Sectors between adjacent pivots represented blended expressions transitioning from one discrete emotion class to another. The intensity of emotion for those in-between emotions was fixed to the radius such that points closer to the centre also portray mild facial expressions and those closer to the circumference – those at high intensity of emotion. This arrangement satisfied the previously outlined conditions such that the interface is i) *coherent* – due to fact that no facial expression class or emotion intensity is present on the interface twice, ii) *systematic* – due

to the fact that each pivot is equidistant to its two adjacent pivots and upholds *between-class dissimilarity* and *within-class similarity* as expressions of the same class are incrementally arranged over their respective radial lines. Figure 7.1



(a) Mostly happy facial expression with an intensity of 0.89 for happiness and 0.11 for sadness.

(b) Mostly sad facial expression with an intensity of 0.55 for sadness and 0.20 for disgust.

(c) A mixture of anger and disgust with an intensity of 0.35 for disgust and 0.20 for anger.

Figure 7.1: Sample output from the application interface displaying facial expressions generated by the neural network model described in Chapter 5 as well as the underlying coordinate system. The top left corner features the quantification of emotion of the currently selected coordinate.

depicts the spatial arrangement for emotions utilizing the aforementioned scheme in an implemented mobile application prototype. Therein, the polar coordinate system's outlines as well as pivots are coloured in orange and overlayed over the facial expression feedback. The navigation dot descriptive of the currently selected coordinates is represented by a solid circle coloured in orange.

7.2.3 Facial expression quantification

The quantification for the neutral expression was relatively simple, since, as previously explained, it was zero-coded and would be positioned at the centre of the coordinate system. Similarly, mapping the discrete emotions on equidistantly placed pivots was also trivial, as the increments in the input parameter for the particular emotion class simply needed to be in proportion with the radius length corresponding to polar coordinates on the system. On the other hand, in-between expressions positioned in sectors between two pivots increased in intensity along their respective radial lines as well, consistent with how distinct facial expressions are organized. The proportion of each distinct pivots' emotion contributing to in-between ones was computed based on the distance of the currently selected coordinate to adjacent pivots, whereby an oblique projection was used to estimate this distance.

In sum, the closer a coordinate was to the polar coordinate systems' circumference, the higher the intensity of emotion was for that particular pivot or blended emotion and was subsequently reflected in the facial expression feedback. In-between expressions derived the proportion of contributing emotions based on a coordinates' distance from adjacent emotion pivots. Figure 7.1 features three images portraying distinct or blended emotions. In the upper left corner of either of the three images a quantification of the portrayed emotion ca be seen coloured in teal.

7.2.4 Allocation of emotions to pivots

Determining the spatial allocation of facial expressions to pivots on the polar coordinate system needed to be in concordance with a metric of emotion proximity to argument their adjacency. To motivate the their arrangement, the likelihood of emotion-pairs being co-experienced was considered. As sectors between each emotions consisted of intensities gradually transitioning from one emotion to another, such an approach ensured that those that are more likely to be co-experienced were placed adjacent to one another.

Two prominent models of organizing emotion have been widely accepted in the literature – the categorical and dimensional models of emotion [9, 188]. Both attempt to organize emotions from a different perspective and while there has been a debate on which model is a better reflection of reality, they are not necessarily at odds with each other as elaborated in Chapter 2, Section 2.2. Recent findings suggests, that a hybrid theory of emotion accommodating both views may be best suited [59]. Using facial expression to represent emotions inherently adheres to the Basic Emotion Theory (BET) model as it presupposes that emotions are distinct entities, whereas arranging them within a coordinate system is intrinsically suited to borrow from the dimensional model of emotion as it aligns with the Circumplex Model of Affect [9]. As the latter allocates emotions within the two-dimensional space of valence and arousal, the dimensional model allowed to derive a metric of emotion proximity based on the spatial allocation of emotions within the valence-arousal paradigm as seen in Figure 7.2.

As such, this valence-arousal organization of emotions served to inform the spatial allocation of facial expressions to pivots in the assessment interface. Another benefit of using the dimensional model to argument the arrangement of expressions was the classification of emotions whose expressions are not part of the six basic facial expressions of emotion [25]. Future iterations of this tool, which may consist of further expressions can borrow from the same heuristic and further expand the number of available emotion dimensions. In this practical use-case, a limitation was imposed by the number of emotions present in the training set used to train the generative model described in Chapter 5). In principle, however, the argumentation for spatial allocation of facial expression on a polar coordinate system can scale to an arbitrary set of emotions.

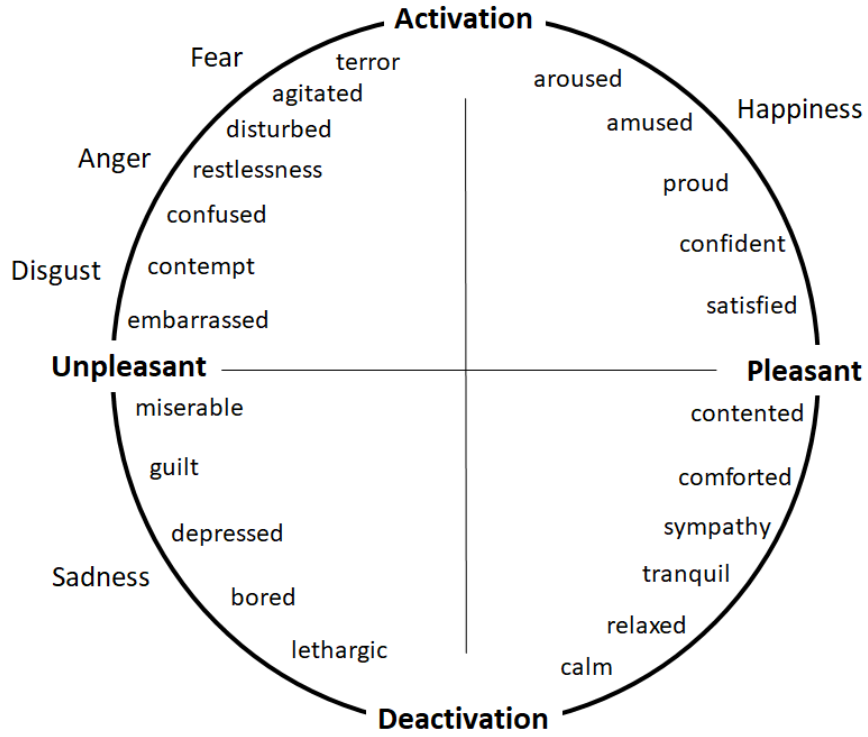


Figure 7.2: Dimensional models' Circumplex Model of Affect portraying emotions and their spatial allocation according to arousal-valence paradigm.

7.3 Prototype implementation benchmarks

A prototype was developed for the Android platform using Java and Kotlin. It was subsequently benchmarked in various configuration in order to establish the optimal parameters for a setup which allowed the facial expression feedback to be immediate. The benchmarks were performed as the platform, hardware and the neural network model themselves pose certain limitations on resources. The android platform imposes limitations on storage afforded to mobile applications. Additionally, smartphones' processors typically use the ARM (Advanced RISC Machines) architecture, featuring a reduced instruction-set aimed for low- or battery-powered devices. Four parameter variants were benchmarked according to three criteria: *response time* (e.g. latency of the facial expression feedback), *storage* and *granularity of facial expressions* contained in the interface.

The prototype featured two elements – a navigation interface representing the polar coordinate system and an image containing the facial expression feedback. First, the generative neural network model from Chapter 5 was converted from TensorFlow to TensorFlow-lite, which enables low-latency on-device inference for embedded and mobile devices. This is a necessary step as a neural network model trained with TensorFlow cannot be used directly on a mobile device with an ARM CPU architecture. For a second variation, the TensorFlow-lite model was quantized

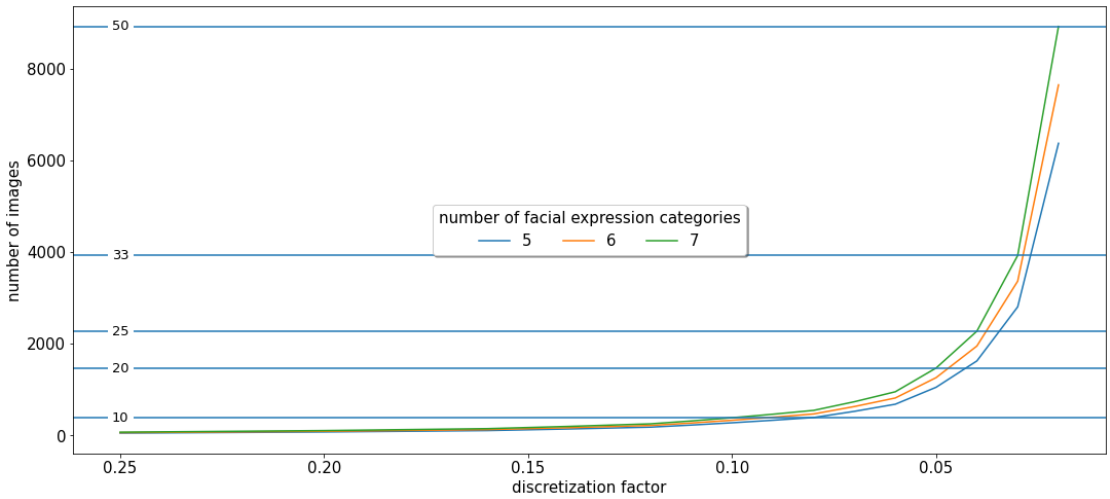
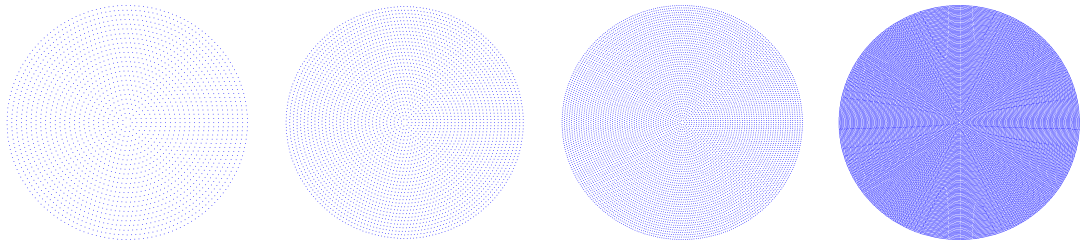


Figure 7.3: Number of discrete points in the polar coordinate system according to discretization factor and number of included facial expression classes. The horizontal lines show the number of distinct images representing facial expression intensities on each radial line. Discretization factor of 0.01 yields 25251, 30301 and 35351 images for 5, 6 and 7 number of emotions respectively, but has been omitted in favour of plot legibility.

[189], where the process of quantization reduces a neural network models' parameter precision resulting in faster inferences while sacrificing quality in the output (e.g. in this case, produced images were more noisy). The second pair of variants used pre-generated images from the neural network model. As the spatial arrangement of images within the coordinate system was known, doing so eliminated the need to do an inference on the device and could improve the responsiveness of the application. Within this variation, however, a range of discretization factors were selected which determined the number of unique image on the interface and subsequently the density of coordinates associated with a unique image on the polar coordinate system. In this investigation a discretization factor of 0.04 and 0.01 were used in the benchmark tests. In order to determine a suitable discretization factor for those pre-generated images, Figure 7.3 portrays the relationship between *number of emotions* (e.g. facial expression classes) and the resulting number of images for discretization factors in the range of [0.04,0.01]. Additionally, Figure 7.4 visualizes the same range of discretization factors and their effect on the number of distinct images on the interface.

Table 7.1 summarizes the benchmark test results, where each approach excelled at a particular metric. Using the TF model resulted in best granularity due to the fact that it can generate arbitrary many intensities per facial expression. However, the latency incurred for generating a single image was approximately 660 milliseconds, which rendered this approach unusable in practice. The quantization marginally improved inference times and moderately improved storage requirement



(a) Discretization factor of 0.04. (b) Discretization factor of 0.03. (c) Discretization factor of 0.02. (d) Discretization factor of 0.01.

Figure 7.4: Effects of applied discretization factors from 0.04 to 0.01 expressed as density of unique images on the interface. Each dot in each subfigure indicates a unique pre-generated image assigned to the respective coordinate on the polar coordinate system.

Table 7.1: Benchmarks for latency and size of the facial expression representation. For the approach relying on pre-generating images, the columns *Discretization*, *Latency*, *Size* and *Number of images* refer to the discretization factor, average response time required to update the facial expression feedback when choosing a new coordinate on the interface, required storage space for either the neural network model or pre-generated images and number of distinct facial expression feedback images contained within the application respectively. The device used in the benchmark tests was a OnePlus5, equipped with Qualcomm Snapdragon 835 processor.[*]Q refers to the quantized version of the TensorFlow model.[¶] n refers to the number of distinct emotion dimensions (e.g. happiness, sadness, and others) where $(n * 2^{23})^{\mathbf{n}}$ accounts for blended expressions (e.g. expressions portraying emotion intensities on multiple dimensions).

Type	Discretization	Latency (ms)	Size	Number of images
TF Model	–	663	44MB	$(n * 2^{23})^{n}$[¶]
TF Model (Q[*])	–	655	14MB	$(n * 2^{23})^{n}$[¶]
Discretization	0.04	70	134MB	1626
Discretization	0.01	70	2913MB	35351

by a factor of 3.14 for the TF model. In sum, inferences with a neural model on mobile devices took a considerably longer time.

Hence, the optimal approach leveraging latency was to use pre-generated images. Depending on the discretization factor, as expected, required storage space increases in proportion with its granularity. Considering the benchmark results, a discretization factor of 0.4 yielded sufficient granularity, as a trade-off for storage space needed to accommodate those pre-generated images. However, the required storage space was acceptable while still achieving an optimal latency in the facial expression feedback. Therefore, the application was developed using

pre-generated images with a discretization factor of 0.04.

7.4 Limitations

A potential caveat in this approach was that the distinction between two facial expression classes in sectors between two pivots is not made explicit to a user besides the facial expression feedback itself. This can further be exacerbated if the underlying coordinate system does not feature labels for the respective emotions on the interface to users. Furthermore, although the machine learning model can generate in-between facial expressions comprising of emotions defined by the two adjacent pivots, there is no guarantee that the expressions would be truly representative of a mixture of those emotions. Consequently, a user navigating away from one distinct facial expression section to another would, however, observe a smooth morph in the facial expression feedback.

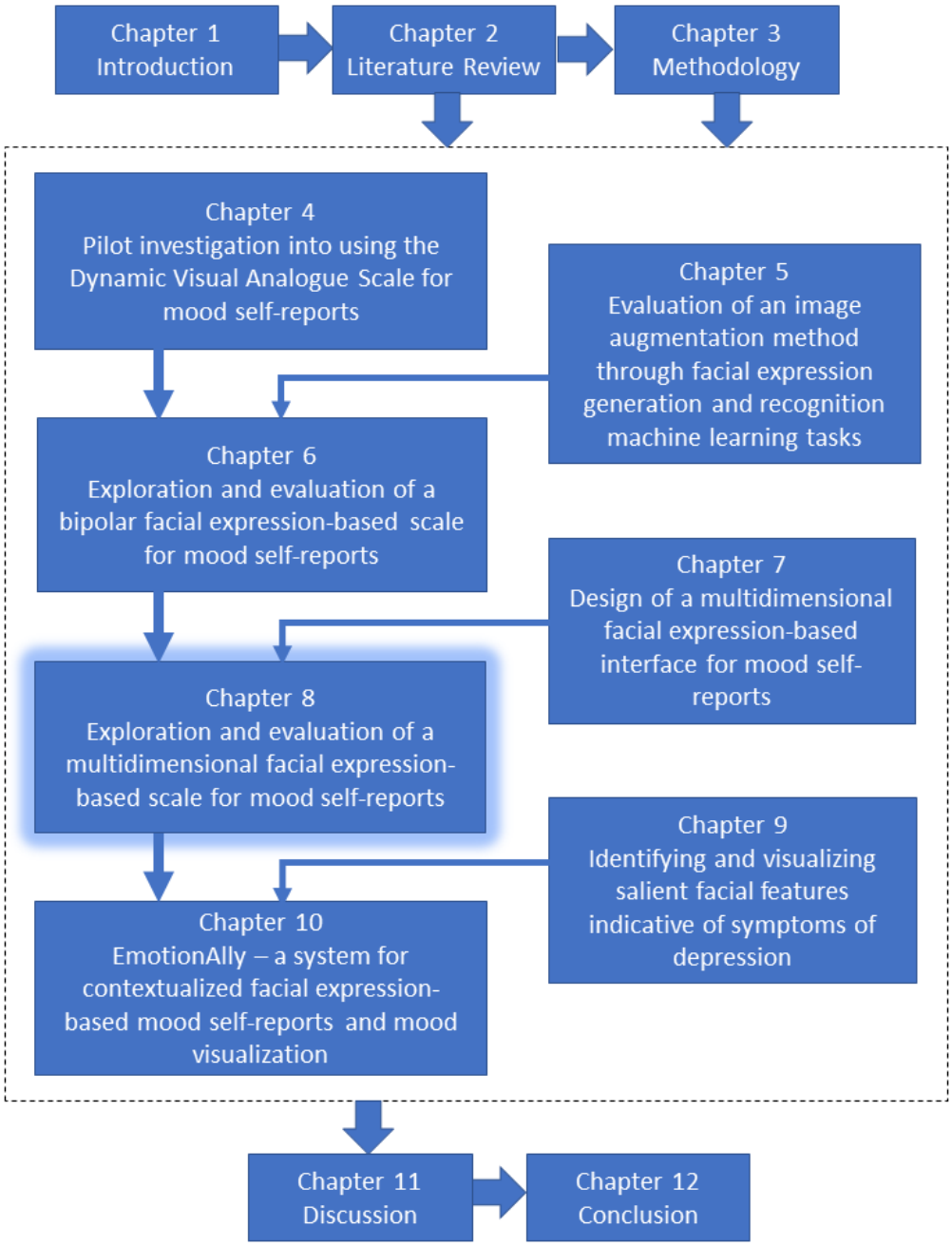
It is important to acknowledge that generative model is restricted to producing facial expressions or identities contained in the training set. However, there is no limitation to expressions or identities it would be able to generate. Similarly, creating augmented images is also not restricted to particular types of expressions (e.g. happy vs angry), albeit due to the arrangement of facial features associated to particular expressions, in some cases it may produce more artefacts. Thus, provided suitable training data, this approach can accommodate an arbitrary amount of facial expressions, identities or other features provided they are available and labelled for their respective classes in the training set. A limitation in the generated intermediary facial expression intensities is that those transform facial features *spatially*. Subsequently, those may not adhere to how facial expressions are enacted and unfold over time in reality [190, 191].

Finally, the choice of adjacency for facial expressions is fixed and cannot accommodate a user to choose how to arrange the expressions on the interface themselves. In this case, since the most viable approach was to use pre-generated images, it already presupposes this particular arrangement. Further improvements, however, could allow this degree of customisation, where the arrangement of expressions on the interface could be chosen by a user themselves. While at present, computation power does not allow to facilitate the realistic use of complex neural network models tools on mobile devices, in the future those considerations could become irrelevant.

7.5 Chapter summary

In this chapter, an approach was presented which organizes computer-generated facial expressions using a generative neural network model within an assessment interface. The assessment interface consists of a polar coordinate system, where each coordinate represented a distinct facial expression and intensity. The spatial

organization scheme of facial expressions was motivated by existing models of emotion and makes use of an emotion-proximity metric of the dimensional model of emotion [9]. The employed approach was generic and the same principle can be applied to allow such tools to accommodate arbitrary many facial expressions. Finally, leveraging contemporary technology, multiple practical implementations were investigated in facilitating a responsive interaction with the interface on a smartphone application for mobile devices.


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 8

Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports

8.1 Introduction

This Chapter will investigate the use of a multidimensional facial expression-based scale (MFEAS) for mood self-reports. Qualitative feedback obtained in Chapters 4 and 6 evaluating two-dimensional happiness-sadness facial expression scales indicates the participants' desire to use multiple expressions for their assessments. For that purpose, Chapter 5 described an approach to generate facial expressions of arbitrary intensities from images displaying categorical emotions. Subsequently, in Chapter 7 an approach was presented which accommodates multiple facial expressions within a two-dimensional interface and implemented in a prototype suitable for assessments on mobile devices.

The approach for creating the multidimensional facial expression-based scale employed a heuristic for placing and positioning facial expressions. Expressions are allocated within a polar coordinate system such that adjacently placed expressions are allocated according to their similarity defined by the dimensional model of emotion [9]. This method is preferred as it has practical implications consisting of restricting the space for providing assessments using blended emotions to only adjacent ones. Hence, the conceptualization of the scale utilizing existing knowledge on emotions allows to position more frequently co-occurring facial expressions in adjacency to one another. The benefit of using a two-dimensional surface in MFEAS is that it would allow users to quickly assess their mood where such a tool would allow users to provide quick one-touch self-assessments guided by facial expression feedback suitable for ecological momentary assessments (EMA) [17].

The aim of this Chapter is to perform an evaluation of a MFEAS featuring

the emotions of happiness, sadness, anger, fear, and surprise represented as facial expressions. Our research question specifically involves a quantitative and qualitative comparison to an alternative based on well-known visual analogue scales (VAS) Within this evaluation a MFEAS was contrasted to several visual analogue scales (VAS) featuring an equivalent number of emotions. A quantitative analysis will be performed on self-assessments of moods elicited by images from the International Affective Picture System (IAPS) [10]. In addition, a qualitative analysis will be detailed using the Computer System Usability Questionnaire (CSUQ) [156] and a thematic analysis of a semi-structured interview conducted with each participant investigating their experience with the interface.

8.2 Methods

8.2.1 Study Design

The experiment employed a between-subjects design, where each participant was randomly assigned to use either MFEAS or VAS. Each participant was asked to rate 60 images with on the respective scale.

8.2.2 Participants

In total 47 participants took part in the study (24 used MFEAS and 23 used VAS), contributing mood assessments for each of the 60 presented images in the stimulus set. They were recruited through flyers distributed across the High Tech Campus in Eindhoven, The Netherlands and Eindhoven University of Technology (TU/e). They were distributed in two groups: those that used MFEAS (14F/10M, Mean age=27 years, SD=6 years) and those that used VAS (12F/10M, Mean age=28 years, SD=5 years). The inclusion criteria required participants to be between 18 and 65 years of age, to be mentally healthy and not diagnosed with a mental disorder, have a good proficiency in English and own an android smartphone. All participants were provided with an informed consent form and privacy notice prior to their enrolment and have been supplied the same materials for their attention upon intake. All participants were required to provide consent and were remunerated for their participation with a voucher for several online and physical stores in the Netherlands worth €10.

8.2.3 Materials

Android assessment application An android application was developed with two distinct assessment interfaces – a multidimensional facial expression-based assessment interface (MFEAS) and one consisting of 5 unipolar VAS scales. The MFEAS represented the emotions for happiness, sadness, anger, fear, and disgust in addition to neutral as facial expressions and consisted of two elements – 1) an

image containing the facial expression feedback displayed in the upper portion of the screen and 2) a navigation interface which a user can use to modulate the facial expression feedback to select and indicate a facial expression for their assessment. The facial expressions feedback was sampled from the neural network model described in Chapter 5 and follows the design and rationale described in Chapter 7. The navigation interface consists of a polar coordinate system mapping the previously mentioned facial expressions. The included facial expressions were arranged along 5 radial lines at equidistant angles of 72 degrees, with 5 sectors in between. The coordinates over each radial line represent distinct facial expressions, whereby moving away from the centre increases the respective facial expression intensity. Similarly, when moving away from the centre within a sector increases the cumulative facial expression intensity by leveraging each adjacent facial expressions contribution to the displayed image. The neutral expression is represented as a single point in the centre of the coordinate system. No visual indicators were present in the navigation interface (Figure 8.1) indicative of the underlying coordinate system except for an outline of its boundaries and a navigation dot indicating the currently selected coordinate (i.e. facial expression).

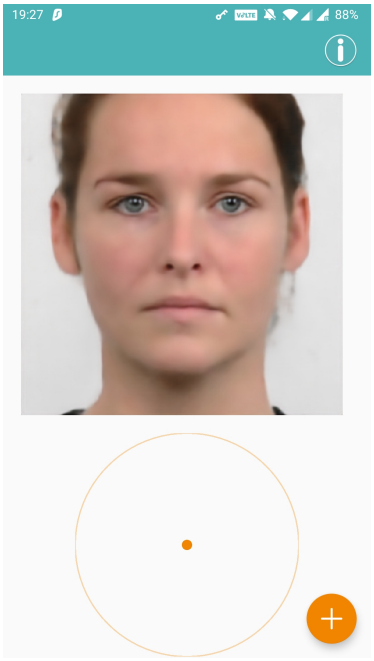


Figure 8.1: MFEAS interface used within the experiment.

The VAS interface consisted of 5 horizontal unipolar sliders, where each corresponded to either happiness, sadness, anger, fear, or disgust. Neutral is represented when all VAS scales are positioned at their lowest possible value. For each VAS scale two anchors positioned at each extreme denoted the emotion direction. The left side of each slider represented an emotion at no intensity and the right one at its peak intensity.

The assessment application (either MFEAS or VAS) was installed on the Android smartphone of each participant if they owned one. Motorola G2 smartphones were available for participants who did not own an Android smartphone.

Stimuli The emotion elicitation material was selected from IAPS [10], a well-known and documented image database, widely used in mood elicitation experiments. IAPS consists of 1193 images rated on the dimensions of valence, arousal, and dominance. Due to the original IAPS ratings not containing ratings on discrete emotion categories, a subset of IAPS [11] was used. It is comprised of 203 negative and 187 positive images and was used to select the emotion elicitation material. Therein, the ratings were divided in differentiated (i.e. an image considered to unambiguously elicit a particular emotion stronger than others it has been rated on) and non-differentiated images (i.e. images rated

to elicit a particular emotion, but also rated sufficiently high on others as well). Sixty images were selected such that each emotion dimension was represented by 10 images, except that for anger for which the total number of categorically rated images was 8. The 2 images missing from the anger elicitation material were replaced by ones from the happiness dimension due to the overwhelming representation of negative emotions as well as the limited number of differentiated images for happiness. As a rule, differentiated images were preferred in the selection of the stimulus material. Additionally, particularly graphic images, for example those containing mutilation or violence were excluded from the selection. There were no differentiated images unambiguously eliciting the emotion for anger, so undifferentiated ones were used. Additionally, for the emotions of disgust, fear and happiness, the number of differentiated images was limited as well, where those dimensions were supplemented with undifferentiated ones.

Table 8.1: Selected images from the IAPS dataset rated on categorical emotions for Anger (A), Disgust (D), Fear (F), Happiness (H), Sadness (S) and Neutral (N). *12 images eliciting happiness were selected, higher than the target of 10 images per emotion dimension to compensate for the lower representation of overall positive images in the selected stimulus set as the total number of images eliciting anger was 8 in the categorically rated IAPS subset [11]. ¶Number of differentiated stimulus images for the respective emotion category.

	A	D	F	H*	S	N
Total	8	77	43	36	62	/
Selected	8	10	10	12	10	10
Differentiated¶	0	7	7	4	10	/

The approach to select images for each emotion dimension aimed to maximize variance, such that images were chosen that are spread over the complete rating range. Neutral images were taken from the original IAPS image set, where images containing faces were avoided as prior research indicates that those supplied contextual information [192] which may induce ambiguity in the assessments. Table 8.1 describes the number of differentiated images for each emotion category.

A 21" desktop screen was used to present the stimulus material to participants.

Questionnaires & Interview Digital versions of the System Usefulness (SYSUSE) and Interface Quality (INTERQUAL) subsets of the Computer System Usability Questionnaire (CSUQ) [156] were used to asses the usefulness and interface quality of both the MFEAS and the VAS interfaces.

A semi-structured interview script was prepared to inquire about various aspects of the interfaces, the method of assessing emotion through facial expressions as well customisation options and improvements.

8.2.4 Procedure

Upon expressing interest in the study, prospective participants were sent an information letter detailing the experiments' aims, procedure, risks, and burdens. An appointment was scheduled with willing participants and on the date were accompanied to an on-site study site. There, they were handed a printed copy of an informed consent form and privacy notice accompanying the experiment and were given ample time to familiarize themselves with the contents. They were allowed to ask any questions pertaining to the experiment, their participation, and the outcome of their data. After providing consent, the mobile application was installed on their smartphone or on one provided to them. Each participant was randomly allocated to either the group designated to use MFEAS or VAS.

Initially, a subset of 5 images from IAPS were shown on an external monitor allowing users to familiarize themselves with the interface assigned to them. Therein, a stimulus image was displayed on the screen for 4 seconds and thereafter the assessment interface appeared on the participants' smartphone screen. There was no time-constraint in how much time was required to provide an assessment, however participants were instructed to assess using their first emotional response to the presented image. The protocol used to instruct participants on how to assess the images was based on the IAPS rating protocol [10]. The IAPS images were rated using the Self-Assessment Manikin (SAM) [33], where the instructions for SAM were substituted for instructions specific to the MFEAS or VAS tools used within this study. Similar to previous studies described in Chapters 4 and 6, any numerical or spatial indication of the underlying coordinate system in MFEAS was obscured to ensure that participants relied solely on the facial expression feedback to orient themselves when assessing the emotion elicitation material. That is, participants were not informed about any aspect of the content of MFEAS – the number of available discrete emotions, how they were positioned and organized within the polar coordinate system and how the navigation between different expressions worked. The same instructions were applied when instructing a participant that was assigned to use VAS.

Following each assessment, both the computer display portraying the emotion-elicitation images as well as the smartphone displayed a blank white screen for 4 seconds. Thereafter, the next image was presented on the monitor and followed the same process until the images were exhausted. After completing this pilot phase, participants were allowed to ask further questions before proceeding.

The experiment phase consisted of 60 images selected from a subset of the IAPS database [10] and followed the same procedure as in the pilot phase. Stimulus images eliciting the emotions of happiness, sadness, anger, fear, disgust, and neutral ones were presented in a randomized order to mitigate carry-over effects. Upon exhausting all images, a digital version of the SYSUSE and INTERQUAL subsets of the Computer System Usability Questionnaire (CSUQ) [156] was presented on the smartphone screen. Afterwards, a semi-structured interview was conducted with each participants about the MFEAS interface, where the conversation was

recorded to an audio file with an anonymized name.

The study took approximately 3 months to complete and took place between 1st of October and 31st December 2019. From intake to completion, the experiment in its entirety took approximately one hour. The audio files obtained during the semi-structured interview were transcribed within 2 weeks of their recording.

8.2.5 Statistical and qualitative analyses

To compare assessments provided on MFEAS and VAS scales, they have been pre-processed to obtain ratings on each discrete emotion category. For VAS, each emotion dimension was represented as a pseudo-continuous slider consisting of 100 discrete points. The pre-processing for assessments made on VAS was trivial where those were normalized to the $[0, 1]$ interval, where 0 corresponds to no intensity of affect and 1 – expressed at its maximum intensity.

For MFEAS, accurately translating polar coordinates to values on distinct categorical emotion dimensions will ensure that assessments made on MFEAS can be compared to those made on VAS. As elaborated in Chapter 7, the positioning of facial expressions on the interface is accomplished in such a way that each facial expression on the interface maps uniquely to a distinct or blended emotion category and intensity. Distinct facial expressions were represented by 5 distinct radii, called pivots along the polar coordinate system spaced at equivalent angles. Expressions located between each pair of pivots are a blend of the two distinct expressions. Increased distance from the coordinate systems' centre (i.e. radial length) translates to an increase in the intensity of the expression. Similar to the approach for VAS, the radial length was normalized within the range $[0, 1]$, where 0 corresponds to no emotion and 1 to an emotion at its peak intensity. For blended expressions, each contributing expression was quantified by computing a parallel projection to each of the adjacent pivots resulting each in a value for a categorical emotion's intensity.

Contingency tables were created, visualising the degree of overlap between the prevalent emotion dimension, i.e. emotion with the highest rating, in the assessed images with VAS and MFEAS and the prevalent emotion dimension in the elicitation material ratings. The contingency tables were grouped by emotion dimension. Linear regression models were used to compare assessments between MFEAS and VAS as well as between each MFEAS and VAS and the image ratings, where for all models assessments with overlapping prevalent emotion dimensions were used. Quantitative results of the CSUQ questionnaire were collated into their respective SYSUSE and INTERQUAL sub-scales and examined for significance.

Thematic analysis was used to analyse participant reposes provided during the semi-structured interview where anonymized transcriptions of the original recordings were used. To analyse the data, a combination of inductive and deductive coding was used, where the predefined questions in the semi-structured interview served as deductive codes and during the transcript analysis new codes emerged through inductive coding.

The numerical data were analysed using Python 3.6 and numpy and pandas

libraries [169] and visualisations were created using the seaborn library [148]. Datasets of 3 participants that used VAS were discarded, where 1 dataset was not recorded in full, another was corrupted, and 1 participant provided neutral or near-neutral responses to all images.

Atlas.TI 8 was used to group and analyse qualitative data [193].

8.3 Results

8.3.1 Quantitative comparison between MFEAS and VAS

Table 8.2: Contingency tables for MFEAS and VAS allocating the distribution of the prevalent emotion in the assessments to the prevalent emotion in the stimulus images' ratings. Each column contains the proportion of classifications for stimulus images per emotion dimension as assessed with the respective interface.

(a) MFEAS contingency table containing the proportion of classified emotions according to the prevalent emotion dimensions in the stimulus material.

MFEAS

	A	D	F	H	S
A	0.4	0.22	0.18	0.07	0.23
D	0.32	0.53	0.17	0.09	0.06
F	0.11	0.11	0.57	0.11	0.07
H	0.03	0.03	0.0	0.6	0.12
S	0.14	0.11	0.08	0.14	0.52

(b) VAS contingency table containing the proportion of classified emotions according to the prevalent emotion dimensions in the stimulus material.

VAS

	A	D	F	H	S
A	0.6	0.18	0.16	0.09	0.23
D	0.12	0.69	0.11	0.08	0.06
F	0.14	0.05	0.67	0.05	0.1
H	0.02	0.0	0.02	0.74	0.02
S	0.12	0.07	0.05	0.05	0.59

Table 8.2 presents the contingency tables for assessments provided on MFEAS and VAS according to the prevalent emotion dimension, i.e. the emotion dimension with highest rating for each image, regardless of whether they were differentiated or not. Each column contains the proportion of classifications for stimulus images per emotion dimension. Naturally, the diagonal contains the proportion of classified images that overlapped with the prevalent emotion rating of the ground truth image ratings [11]. Those outside of the diagonal contain the proportion of images which did not match the dominant primary emotion in the stimulus material (e.g. for images eliciting fear no assessments were provided with MFEAS that had happiness as prevalent dimension). As the LR models were computed only on assessments whose primary category overlapped with that of the ground truth image ratings, it is important to interpret the results considering both scales' ability to categorically classify an emotion.

To compare assessments provided on MFEAS and VAS, Linear Regression models (LR) were computed for each emotion category: MFEAS and VAS assessments were regressed on ground truth ratings (e.g. original stimulus image

ratings provided on categorical emotion dimensions) and MFEAS assessments were regressed on VAS assessments. Table 8.3 depicts the LR model regression slope (s), intercept (i), correlation coefficient (r), significance value (p) and standard error (SE) parameters. Figure 8.2a visualizes the regression model for MFEAS assessments regressed on ground truth image ratings and Figure 8.2b that for VAS assessments regressed on ground truth ratings.

The results indicate a moderate correlation between the images' ground-truth ratings and the assessments provided with MFEAS, particularly for the dimensions of sadness ($r = .33$), anger ($r = .32$) and disgust ($r = .46$) where those correlations were significant ($p < 0.01$). Assessments provided on the dimensions for happiness and fear were not significant ($p > 0.05$).

Assessments provided through VAS were moderately correlated with images' ground truth ratings for the emotion dimensions of sadness ($r = .28$), fear ($r = .33$) and disgust ($r = .51$), where those correlations were significant ($p < 0.01$). Assessments provided on the dimensions for happiness and anger were not significant ($p > 0.05$).

Comparing assessments provided on MFEAS and VAS, the respective LR model indicates a strong correlation between assessments provided on the emotion dimensions of sadness ($r = .78$), fear ($r = .69$) and disgust ($r = .86$), which all were significant ($p < 0.05$). The correlation for the dimension of happiness was approaching significance ($p \sim 0.07, r = .54$). The correlation on the emotion dimension of anger was not significant ($p > 0.05$).

All neutral images were rated as such on both MFEAS and VAS interfaces, with no emotion dimension assessments exceeding 0.1. No models have been computed for those, as the neutral dimension was therefore more or less represented by a single point on the MFEAS and VAS scales.

Table 8.3: Parameters of the linear regression correlation models for assessments made with MFEAS and VAS regressed on the ground truth (GT) ratings as well as between each other where MFEAS was regressed on VAS. The columns indicate the linear regression slope (s), intercept (i), correlation coefficient (r), significance value (p) and standard error (SE). Emotion dimensions are encoded as follows: H – happiness, S – sadness, F – fear, A – anger, D – disgust.

MFEAS - GT LR model

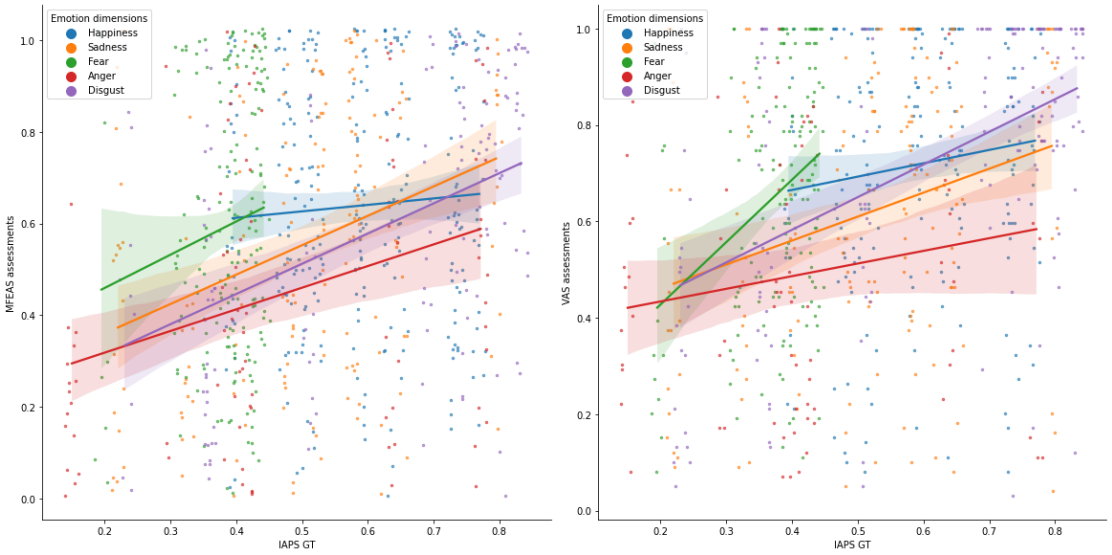
	slope	i	r	p	SE
H	0.14	0.56	0.06	0.33	0.15
S	0.64	0.23	0.33	0.0	0.14
F	0.73	0.31	0.14	0.06	0.38
A	0.47	0.22	0.32	0.0	0.15
D	0.66	0.18	0.46	0.0	0.11

VAS - GT LR model

	slope	i	r	p	SE
H	0.28	0.55	0.13	0.07	0.15
S	0.5	0.36	0.28	0.0	0.13
F	1.29	0.17	0.33	0.0	0.29
A	0.26	0.38	0.18	0.11	0.16
D	0.68	0.31	0.51	0.0	0.09

MFEAS - VAS LR model

	slope	i	r	p	SE
H	1.09	0.02	0.54	0.07	0.54
S	0.66	0.25	0.78	0.01	0.19
F	0.89	0.13	0.69	0.03	0.33
A	0.11	0.44	0.15	0.72	0.3
D	0.97	0.14	0.86	0.0	0.21



(a) MFEAS assessments regressed on ground truth ratings

(b) VAS assessments regressed on ground truth ratings

Figure 8.2: Linear Regression Correlation models for MFEAS and VAS assessments regressed on ground truth ratings as well as MFEAS assessments regressed on VAS.

On the System Usefulness (SYSUSE) CSUQ subscale MFEAS scored M=21.3 and VAS scored M=16.69, a difference that was not significant in a t-test ($t = 1.53, p = 0.13$).; On the Interface Quality (INTERQUAL) CSUQ subscale MFEAS scored M=9.6 and VAS scored M=7.375, and this difference was not significant either in a t-test ($t = 1.64, p = 0.11$).

8.3.2 Qualitative experiences with the MFEAS interface

The qualitative results in this section elaborate on aspects of the approach of representing and assessing emotional content through facial expressions as well as the practical implementation and use of the prototype through the user's perspective.

Emotion coverage Most participants were satisfied with the number of emotions present on the interface as most participants did not indicate additional emotions which should be added in the interface. A few participants, however, pointed out emotions which would be important for tools for self-reporting mood which were not present in the interface, especially positive emotions besides that for happiness: *"More extreme joy and I was looking for something like content, [...] but I think there is also like a calmness maybe. I would see those as different dimensions, independent of the happiness expression."*

Additionally, some users indicated that they would further wish to see facial

expression-based representations for the expressions of excitement, curiosity, happy-surprise and interestingly a different variation of neutral representing a neutral mundane feeling: *"In terms of mood, I think trying to find out emotions a little more subtle [...]. So sometimes you're not feeling either excited or sad. You're not feeling any specific emotion, but [...] a mundane feeling."*

Emotion intensity, granularity and range The interface granularity is reflected in how transitions between distinct coordinates on the coordinate system is perceived as reflected in the update of the facial expression feedback. Most participants found the granularity of the facial expressions in the interface (e.g. range of intensities) to be sufficient: *"I like that it was smooth to move it and that I didn't have to wait for the face to change."*

With respect to the range of available intensities for the facial expression in the interface, a few participants highlighted that the expressions for happiness and sadness both lacked more expressive representations of higher intensities of emotion.

Emotion blends Interestingly, some of the blended expressions (i.e. those consisting of a mixture of two distinct facial expressions) appeared to be recognized as such: *"I [...] could tell this person is upset, but also slightly mad or angry about something."* Nevertheless, providing assessments with others was more difficult: *"I had the feeling that [...] some in-betweens [blended expressions were] [...] hard to tell, which expression should I pick."* Some participants indicated a desire to assess using blended emotions of their choice in contrast to the fixed order presented in the interface.

Facial expression realism With respect to the realism of the expressions provided within the application, two concepts emerge: 1) realism pertaining to how close to reality the person portraying the facial expressions is, i.e. how close to a 'photograph-like' the image of a person conveying the facial expressions on the interface is, and 2) how realistic or otherwise plausible to encounter the portrayed expressions in real life are.

For the former, an overwhelming majority of participants agreed that the generated images were realistic enough to convey that a real person is enacting the expressions: *"They are based off a real human being, they are pretty accurate."* Nevertheless, it was also obvious to participants that the images were also not real.

In some instances, there were some slight but discernible visual artefacts in the generated facial expression images. Those artefacts were noticeable enough, but not sufficient to break the 'realness' of how the expressions were perceived: *"It was obviously not a realistic face. [...] Sometimes there were [...] slight artefacts, especially towards this angry part. [...] But in general, I think it was fairly realistic."*

In the latter aspect of realism, some participants differentiated between 'archetypal' expressions, also known as forced or enacted expressions, and real

as in spontaneous or natural expressions encountered in everyday communication in day-to-day life. Some participants remarked on the distinction between enacted and spontaneous expressions: *"It was a bit built up. [...] Realistic, but artificial."* This was particularly noticeable for the expressions for fear: *"The ones where she is scared, that felt a little off for me [...] because with those expressions you don't really come across genuinely a lot. [...] But otherwise it feels very good, very real."* The expressions for happiness appeared to be the most natural-looking ones: *"I think the happy expression was realistic. The others struck me as acted"*.

The fact that expressions were considered to consist of enacted and genuine expressions did not appear to affect how the application was used: *"It's something that in the first time you notice it's acted or not genuine, but once you do it [interact with the interface], it doesn't matter anymore"*. Additionally, the fact that the facial expression feedback appeared realistic resulted in conveying emotional content more clearly and subsequently made the interaction with the interface more relatable: *"If it was less realistic, I wouldn't relate to it as much"*, *"I think the ease of use comes with how realistic it is. I feel like if it was not as realistic it would probably be a lot harder to use or at least harder to express what you wanted to show."*.

Facial expressions as interface feedback The use of facial expression feedback to represent emotions was indicated by the majority of participants as helpful in guiding them in providing their assessments: *"She [the model on the interface] was a way for me to show how I felt"*. Interestingly, the presence of a tangible facial expression feedback appeared to prompt participants to fine-tune their responses: *"Sometimes I moved to a too intense expression or something that made me think twice. [...] So this feedback kind of helped me with the assessments."*

Interface labels for emotions As elaborated on earlier, no instructions were provided to participants how MFEAS works except that it uses facial expressions to represent emotions. Additionally, apart from an outline of the polar coordinate system and a navigation dot denoting the currently selected coordinate, no further visual or textual indicators were provided to signify the number of positions of emotions. Despite that, however, most participants implied that the inclusion of textual labels to signify which emotions and their location on the interface is not needed: *"I think [the lack of labels] makes the experience easier, rather than with just words or descriptions."*

Conversely, several participants indicated that including labels to denote the distinct emotions on the interface for a few initial assessments could allow users to locate and familiarize themselves with the interface quicker: *"Maybe the first few tries to sort of prompt you or [...] visually give you a cue. But other than that I liked the fact that there were no labels, that I could freely go around and rotate it."* A few participants also noted that they need to provide assessments on several images to be competent at using the interface and learn all the available expressions. Those participants also implied that the inclusions of labels might

lower the learning curve required to use MFEAS. Reflecting on that statement, those participants reinforced the idea that including labels would in some way dilute the interaction of users with the interface and detract them from exploring the facial expression intensities and using that feedback to fine-tune their assessments.

Interface responsiveness and ease of use All participants found the interface to be very responsive and subsequently also easy and quick to learn to use: *"Generally it feels pretty good. For the distinct ones [distinct facial expressions] it was very good. After a minute or two you figure it out and it was very easy"*. Additionally, the facial expression feedback was immediate which is something that most participants found valuable: *"The [facial expression] feedback was quick, almost immediate, so I could very quickly say or try to find the emotion"*. The simplicity of ease of use was also mentioned, where the use of a mobile application for the prototype also facilitated simpler and subsequently quicker interaction to provide assessments: *"I liked how simple it was. I liked that I could learn it in a couple of goes and could use it with one hand."*

Customizing the appearance of the facial expression model Regarding customisation, participants were split almost in equal parts in how they would like to customize it further should they would use the tool themselves.

Slightly under half of participants liked for the avatar to remain generic, i.e. the model to be of an identity unknown to them: *"I think it helps that it [the model] isn't someone specific. [...] The fact that it is just a regular woman or man it's fine"*.

Slightly over half of the participants indicated that they would prefer a model that resembled them, and there were varying interpretations of how this resemblance was defined. Within that sample, approximately half of the participants defined this resemblance as a similarity of appearance, but felt uncomfortable if the model was a digital mirror image of themselves: *"If it is someone that looks quite close to me it's okay. If it is me, I would prefer that it is someone else quite close to me."* The other half preferred the identity to resemble them as closely as possible: *"[The model should be] of me. The point is that if I see somebody else there, I would try to imagine how that person would feel."* Here the indication is that a user could provide more accurate assessments if the model resembled them.

Within both groups – those that desired either a resemblance or likeness and those that preferred an anonymous identity for the person on the interface – the majority of participants preferred this person to somewhat match their demographic characteristics, with age as the most important feature, second to gender: *"I think the closer it is to your own [age and gender] the better. The easier it is to relate."*

Interface use in daily life The majority of participants indicated that for assessing mood on a daily basis, MFEAS would be a suitable tool to do so as it facilitates the ability to provide a self-report quickly: *"On one hand I want 10*

sliders, [...] but then obviously that takes very long, where this is super quick. So I feel like this I would be more likely to do it every single day than the sliders." Here the contrast to an equivalent VAS scale is made where VAS could capture data in a more detailed fashion at the expense of an extra time investment.

8.4 Discussion

8.4.1 Quantitative results

The quantitative results indicate two interesting observations. First, LR models containing MFEAS and VAS regressed on the ground truth (i.e. IAPS image ratings) in Table 8.3 show a weak to moderate positive correlation between assessments provided on both scales and the underlying ground truth image ratings apart from the dimensions for fear and happiness for MFEAS and anger and happiness for VAS. Additionally, the contingency tables for MFEAS and VAS assessments in Table 8.2, portray how well participants were able to categorically disambiguate the primary emotion (e.g. the highest rated emotion in the stimulus images), where VAS appears to outperform MFEAS, since the correct rates (on the diagonal) are clearly higher. In summary, these observations indicate that in terms of detecting a primary emotion categorically, VAS is better than MFEAS. Additionally, the correlation of assessments on both scales regressed onto ground truth image ratings is moderate when it exists.

Second, the LR regression model regressing MFEAS assessments on VAS ones shows significant correlations for the emotions of sadness, fear, and disgust, with happiness approaching significance ($p \sim 0.07$), and these were strong to very strong positive correlations.

Both observations indicate some inconsistencies in the obtained results. The following sections will attempt to elucidate on how and why assessments provided with MFEAS and VAS are strongly correlated with each other, but less so to the ground truth image ratings as well as delve into individual emotion dimensions to explore whether there were factors which may account for a particularly strong or weak correlation for the computed LR regression models.

Emotion elicitation material Addressing the lower correlation of assessments with MFEAS and VAS to ground truth ratings relative to that between MFEAS and VAS, it is important to reiterate that the original IAPS image ratings have been provided for the dimensions of valence, arousal, and dominance. As such, those ratings cannot be interpreted as ratings on discrete emotion dimensions, which necessitated the use of a IAPS subset [11], where two such subsets were created consisting of images rated on either 5 positive or 5 negative discrete emotion dimensions. Based on those ratings, images were tagged as differentiated, i.e. an image considered to elicit a particular emotion stronger than others, when the averaged rating on a particular emotion was prevalent such that its confidence

intervals did not overlap with those of ratings provided on complementary emotion dimensions. As seen in Table 8.1, the emotion dimensions for anger and happiness featured the lowest proportion of differentiated images (e.g. none for anger and 4 out of 12 eliciting happiness), compared to other emotion dimensions. Additionally, as specifically stated by the authors of the categorical IAPS image ratings [11], the happiness dimension contained many blended emotions. It is important to restate that as a general rule, undifferentiated images were selected only when insufficient differentiated images were available, but this was not the case for anger and happiness. Naturally, undifferentiated images introduce ambiguity in the emotion elicitation material and may affect how those images were subsequently interpreted by participants. Coincidentally, both the dimensions for anger and happiness did not reach significance in the model regressing MFEAS assessments on VAS assessments, which could be attributed to the stimulus material selection. Assuming there were more differentiated images eliciting anger and happiness, those emotion dimensions might have resulted in a significant and substantial correlation as well.

In Table 8.2, MFEAS appeared to capture the prevalent emotion less accurately than VAS. This could be attributed to a slower learning curve associated with MFEAS: As some participants shared during the semi-structured interview, they required up to 10 additional images beyond the 5 training ones to explore and find all facial expressions and intensities. Hence, initial ratings provided through MFEAS might have been assessed on another emotion dimension due to a participant's lack of awareness for all present emotion dimensions.

Alternatively, image ratings of the IAPS subset [11] might have varied such that ratings across participants differed in the prevalent emotion dimension each image was eliciting. Unfortunately, the image ratings did not feature data on individual participants but rather as averages. Consequently, it is not possible to compare a contingency table for the original image ratings to those for MFEAS and VAS. Therefore, it might be the case that scores in the original image ratings contingency table could be similar to those of the assessments provided with MFEAS or VAS.

An often addressed topic in psychological, psycho-physiological and neuroscience research is the use of standardized emotion elicitation material [194]. The benefits of using standardized emotion elicitation sets allows empirical results to be compared between studies, where IAPS has filled this gap and its images have been used in a wealth of emotion research studies. However, an often-encountered statement is that IAPS is conceived and developed more than two decades ago. Within a changing socio-cultural context, an argument can be made that those images are outdated [194]. This could be an alternative explanation why some emotion dimensions did not yield a significant correlation.

Additionally, since the categorically rated IAPS subset [11] did not specify or elaborate on the protocol used to instruct participants, an adaptation of the original IAPS protocol was used. Therefore, a discrepancy between the protocols may have also influenced assessments, subsequently impacting the results for the LR models regressed on the ground truth image ratings.

MFEAS-VAS regression model and study design implications The strong to very strong positive correlation of the LR model regressing MFEAS assessments on VAS ones is a strong argument vouching for the ability of a facial expression-based scale to capture emotions compared to a traditional VAS scale. Compounding on that argument is largely the employed between-subjects study design which precluded carry-over effects should participants have rated images using both scales. As previously elaborated, the emotion dimensions for happiness and anger contained the lowest amount of differentiated images, where subsequently those two dimensions coincide with the absence of significant correlations in the model ($p > 0.05$). The results also included a very strong and statistically significant correlation in the LR model regressing MFEAS assessments on VAS outside of the dimensions for happiness and anger. Apparently, the LR model comparing MFEAS and VAS scales directly shows higher correlations than the two LR models regressing MFEAS or VAS assessments on ground truth image ratings. This might be due – in line with previous reasoning – to a combination of protocol differences between our rating instructions and those used to rate IAPS images categorically, in combination with a lack of sufficient differentiated images for the emotion dimensions of happiness and anger.

Using MFEAS or VAS to disambiguate the primary emotion in the stimulus material An interesting observation is the difference in both scales' ability in allowing users to categorically detect the primary emotion in the elicitation material, where VAS appears to outperform MFEAS on all emotion dimensions. Elaborating on that, it is important to highlight a distinction in how both scales were presented. For VAS, textual labels were present indicating the available emotion dimensions, where MFEAS did not feature any discernible landmarks indicating the number or type of emotions portrayed as facial expressions. Conversely, MFEAS featured only the polar coordinate system's circumference and a navigation dot linking the currently selected coordinates to the displayed facial expression. Also, participants were not instructed which expressions were present on the MFEAS interface and also how those were organized within the polar coordinate system. Therefore, participants that used VAS received more information about the potential emotions (in the form of clearly visible labels) than participants that used MFEAS, which could, in part, explain the lower categorical disambiguation ability of MFEAS as seen in Table 8.2.

Alternatively, there is an important distinction in how facial expressions are processed. Prior research identified that detecting an emotion in the face and its intensity are two separate, co-occurring processes [57, 195]. This division between categorical and dimensional processing may also affect the accuracy of provided assessments negatively as both processes would be in competition when selecting a suitable categorical facial expression and intensity on the MFEAS scale. On the VAS scale, the emotions categories were labelled, so this competition is not expected to occur to the same extent, and this difference with MFEAS might

explain why disambiguation appears to be easier with the VAS scale. However, without further research it is unclear whether this competitive processing has a lasting impact to an experienced user, already familiar with the interface.

Comparing between-emotion assessments provided with MFEAS and VAS An interesting observation pertains to the variation in regression line slopes for the LR models regressing MFEAS and VAS assessments on the ground truth image ratings in Figure 8.2. Therein, the regression line slopes for MFEAS model appear to be more similar in size to one another than those for the VAS model. This may be an indication that a facial expression-based interface may be able to capture emotions in a way that allows assessments provided on different emotion dimensions to be compared to one another. In contrast, as VAS records assessments on a numerical scale, and ratings provided on different emotion categories may be interpreted differently. However, further research is required to establish whether this is indeed the case.

CSUQ SYSUSE and INTERQUAL results The results obtained from the CSUQ questionnaire were not statistically significant, which implies that ratings provided on both System Usefulness (SYSUSE) and Interface Quality (INTERQUAL) sub-scales for MFEAS and VAS respectively did not statistically differ from one another. This may be due to the relatively lower sample size inherent to between-subjects study designs. However, it does indicate that MFEAS and VAS are similar in the way they are useful to users for assessments and in the quality of their interfaces. This could be interpreted as promising in the light of comparing MFEAS, a novel interface that has not been seen or evaluated in previous studies, in contrast to VAS which is an established and well-known assessment instrument.

8.4.2 Qualitative results

This section highlights important aspects of the facial expression-based method for assessing mood and qualitative characteristics of the evaluated prototype as perceived by users.

Facial expression coverage A pertinent feature of the MFEAS prototype is the distinct emotion categories available on the interface. While most participants did not indicate an emotion that was missing for them, the experimental context might have influenced those results. Some participants, however, did indicate expressions and associated emotions which they would like to have access to for self-reporting mood. Contemporary research found at least 28 distinct categories of emotions (such as amusement, anger, awe, concentration, confusion, embarrassment, surprise) represented as facial-bodily expressions (i.e. where the predominant feature is variations in expressiveness or the inclusion of small bodily cues such as head pose or hand-gestures complementing the expression) frequently occurring in everyday life

[196]. The authors argue that those emotion categories are mapped onto an emotion space not as distinct entities, but rather as a gradients that cross categorical emotion boundaries [196]. Additionally, the expressions' continuous intensity variations were considered to translate into continuous variations in meaning [196], akin to how facial expressions were represented in MFEAS. Naturally, incorporating additional expressions within new versions of MFEAS is trivial, since the employed approach to generate facial expressions (Chapter 5) and the heuristic to organize them (Chapter 7), both can accommodate arbitrary many expressions. Increasing the number of expressions by including ones apart from the basic facial expressions of emotion [25] would inadvertently increase the potential of facial expression-based assessment tools.

Facial expressions' granularity Most participants have indicated that the variation of facial expression intensities in the interface is good as indicated by reporting that transitions between facial expression intensities were smooth. That is not surprising, since the discretization factor defining how many images were allocated on the interface (see Chapter 7) was 0.04. Subsequently, the intensity variations for facial expressions in the interface were captured within 25 images.

Recognition of blended expressions Interestingly, some participants were able to recognize some of the blended expressions in MFEAS such as sad-angry, while others were not. This is not surprising as morphologically dissimilar expressions (e.g. expressions which contain variations of the same facial features) such as happiness and sadness are difficult to be represented simultaneously within a static image. It is known that facial expression processing consists of a holistic and parts-based processing [197], where either a congruent face (i.e. face for which all parts are cohesive in the expression they portray) or individual facial features associated with specific expressions are used to recognize the true underlying emotion. Conversely, non-congruent combination of facial-feature parts as expected reduces the accuracy of recognizing the underlying expression [197]. Within this interface, blended facial expressions are averaged across facial features such that both expressions are, in essence, morphed together. Prior work has investigated the cross-categorical boundaries between such morphed expressions [198], however further work is needed to explore whether and which blended emotions at the categorical boundaries of both expressions can also be recognized as such.

Facial expression realism Interestingly, with respect to facial expression realism, most participants considered the facial expression feedback to be realistic enough, such that participants were aware that the images were computer-generated, however they were of a high-enough fidelity to convey a realistic portrayal of a real human being. The realism of the human being was preferred by most participants where some indicated that they could map their own emotions onto and emphasize with the expressions portrayed by the computer-generated model.

Participants were also able to distinguish that most expressions on the interface were acted or posed rather than spontaneous expressions. It is known that the Radboud Faces Database (RafD) [44] consists of posed expressions. While posed expressions are feigned facial enactments of an emotion without the presence of a stimulus eliciting that response, spontaneous are used in every day situations as a reaction to events occurring in our daily lives. Generally, posed expressions are often exaggerated expressions of emotion with high expressiveness in the face. Conversely, spontaneous expressions are typically subtle and are interpreted in concordance with body posture and other cues in the environment. As a consequence, in the context of providing self-reports through facial expressions, posed expressions possess a larger variance in expressiveness, while spontaneous smiles were found to have smaller amplitude and slower onset than posed ones [191, 199]. However, their use and in turn interpretation varies depending on the social context, where posed expressions are perceived as unauthentic [200]. In this sense, examining which type of facial expressions is more suitable for a particular use case is very important as the difference in interpretation can influence the interaction of the user with the interface. However, in the scope of using facial expressions for mood self-assessments, both types in essence represent the same underlying emotion, regardless of whether they were feigned or genuine.

In order to adequately be able to portray spontaneous (e.g. authentic) expression, those inherently require relatively higher fidelity that is able to portray sufficient detail, such that a user may be able to indeed recognize the presented images as spontaneous emotions. It can be argued that the use of posed over spontaneous expressions may be beneficial in the use case for self-reporting mood, where a broader range of facial expressiveness is desired or when the fidelity of the digital representation of a human face is insufficient to adequately portray subtle or nuanced expressions.

The use of posed over spontaneous facial expressions may also yield a benefit in being able to portray a broader range of intensities for each expression, even though this is achieved at the cost of not appearing as genuine expressions of emotion. Inversely, spontaneous expressions will provide a smaller variation in facial expressiveness, at the benefit of being a closer approximation to reality.

Interface responsiveness and ease of use The qualitative feedback pertaining to the interface was overall positive as in participants rated unanimously the interface to be responsive with immediate facial expression feedback and easy to use. In fact, a few participants indicated that it took between 5 to 10 images to familiarize themselves with the organization of facial expressions within the interface. Subsequently, assessments allowed users to go to regions representative of distinct emotions, where the facial expression feedback appeared to have assisted participants in fine-tuning their responses. In the scope of assessment scales, providing real-time feedback is considered to result in the most accurate self-reports [22]. The confirmation that the facial expression feedback was utilized as intended to

allow participants to fine-tune their responses strengthens the utility and usefulness of the approach.

Interface labels for emotions An interesting observation resulting from the qualitative feedback is that the majority of participants did not want to include labels denoting the regions where distinct emotions can be found within the facial expression-based interface permanently. The main arguments presented were that the expressions themselves were sufficient to convey the emotion content in terms of categorical emotions and emotion intensities and including labels might prompt users to use the interface differently and would needlessly clutter the interface. It was acknowledged, however, that although MFEAS was easy to learn, the temporary inclusion of labels might allow persons to use the interface more quickly. This may indeed be a beneficial approach since quantitative results revealed an overall lower categorical disambiguation for MFEAS compared to VAS. As such the initial inclusion of labels for emotions or allowing the option to include those temporary or for all assessments could serve the purpose of allowing for more precise assessments to be made. Subsequently, an interesting point to investigate in the future is whether a temporary inclusion of labels would bridge the gap in detecting the prevalent emotion from the stimulus material.

Interface model appearance customisation Participant feedback was most polarized in customizing the appearance of the model enacting the facial expression on MFEAS. Therein, participants were split between using a person anonymous to them or having the model's appearance resemble or mirror them.

Within the first group, persons were generally averse to see a digital doppelgänger of themselves. However, it is not clear whether that is influenced by the context of the alternative consisting of using their own face to provide mood self-reports or rather as due to a more fundamental principle. In the latter group consisting of persons that wished to see the model resemble them, those were split in those that wanted to see a mirror image of themselves and those that preferred a resemblance but not likeness.

Technologies already exist which can swap one persons' face onto that of someone else in photographs, which could prove to be useful method to personalize such scales [121, 201]. Interestingly, prior research investigating user preferences for 3D avatars found a similar distribution of users which were found to prefer and be satisfied with an approximate representation of themselves versus a mirror image [202]. This implies that technologies used to create virtual avatars or digital representations of persons do not need to be photo-realistic in order to address the needs of users. In fact, they only need to approximate reality to the point of providing a plausible and realistic depiction of a human face and sufficient fidelity to adequately portray facial expressions where some leeway in the expected quality is not detrimental [202]. Similarly, within this investigation persons which used the facial expression-based scale could discern that the model on the interface was

computer-generated, but not a single participant reported an uncanny-valley effect [122] while using it or indicated that a more realistic depiction would have somehow improved the interaction experience or provided assessments. In conclusion, an interesting avenue for investigation would be to conduct a contemporary exploration of technologies capable of creating digital human faces and identify a 'sweet-spot' in terms of realism, fidelity and facial expression granularity contrasted to user acceptance and self-report accuracy.

Finally, it is important to note that facial expression technologies which can be customized by a user also pose a privacy risk to users. Particularly so when the underlying technology can create high-fidelity faces. Provided the customisation options allow users to create digital identities of high fidelity, where those mimic the appearance of a real person, those may either make the user identifiable or the specificity of the customisation options themselves may be misused. While this aspect has not be explicitly addressed within this research, it is still an important consideration to be made in light of facial expression-based tools.

8.5 Limitations

Potential limitations of the current study pertain to the stimulus image choice, which was limited in the number of differentiated images (e.g. images found to elicit predominantly a single emotion), especially for anger and happiness. This was reflected in the results, where consistently those dimensions were insignificantly correlated in the LR models to the ground truth values. Additionally, the absence of the instruction protocol used to instruct participants in rating the IAPS image subset on categorical dimensions of emotion and lack of contingency tables for each image may have shed light on whether the results achieved by either MFEAS or VAS in detecting the primary emotion in an image are comparable to those found in the original study [11]. A future study may address those deficits by, for example, using a categorically rated emotion elicitation dataset [194].

8.6 Conclusion

This study evaluated the application of a multidimensional facial expression-based scale (MFEAS) to assess the presence and intensity of emotion in emotion eliciting images [11]. The scale was contrasted to an equivalent VAS scale, where assessments provided with both were strongly correlated. Also, the usability of the MFEAS application was evaluated to not significantly differ from that featuring a series of VAS scales. Additionally, semi-structured interviews indicated an overall positive reception towards the technology and its application. In addition, multiple avenues to further improve the method were outlined drafting future research directions.

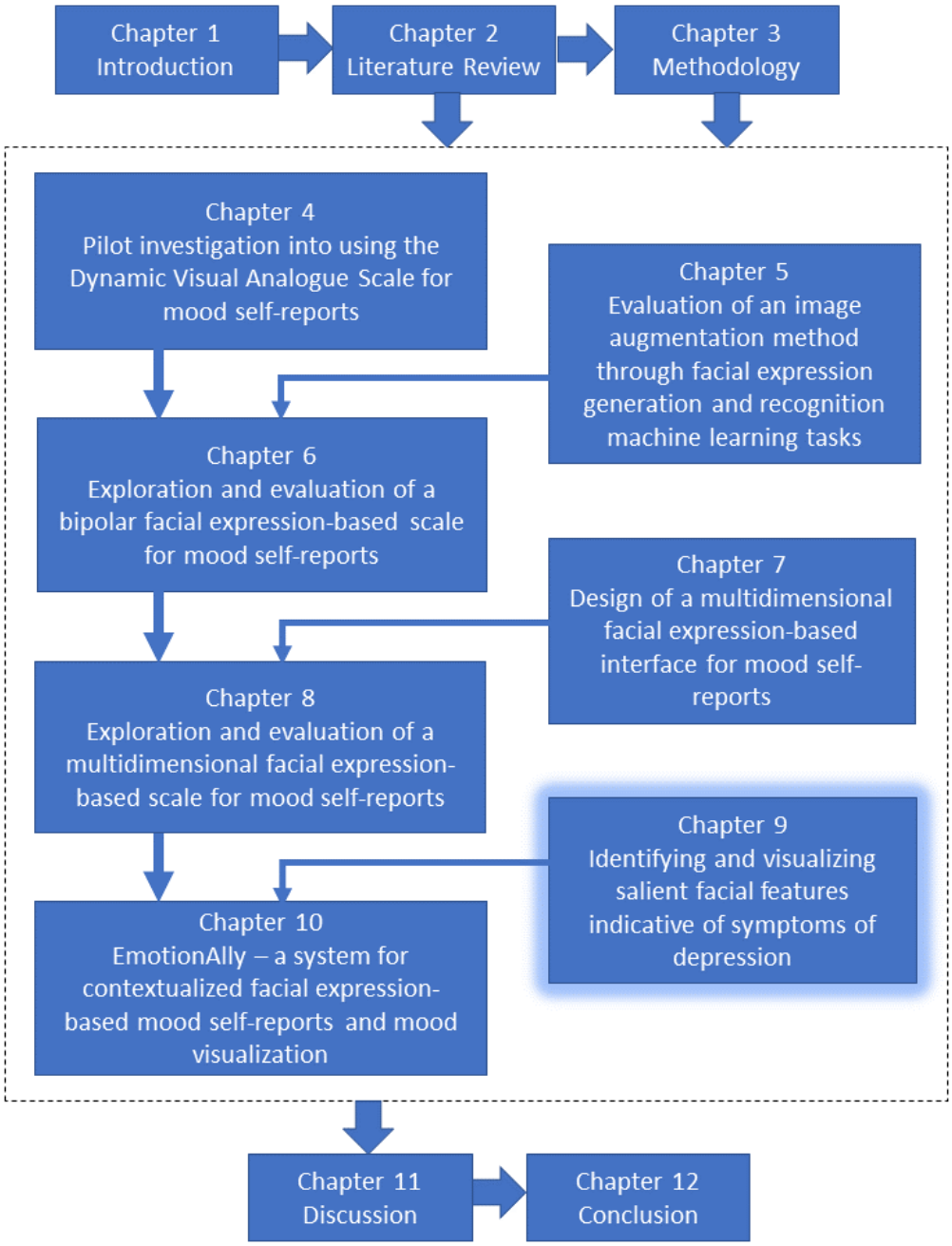
8.7 Chapter summary

This study aimed to investigate whether assessing mood through a multidimensional facial expression-based interface – containing facial expressions for happiness, sadness, anger, fear, and disgust – is comparable to that of a set of VAS scales for the same set of emotions. As stimuli emotion eliciting images [11] were used. The study employed a between-subjects design, where neither the instructions given to participants nor the MFEAS interface informed users a-priori the content and variety of expressions. Results indicate that ratings provided on both scales are strongly correlated to one another, at least for all emotions for which sufficient images with an unambiguously rated primary emotion were present in the stimulus material. In addition, both MFEAS and VAS scales were also moderately correlated with the stimulus material's ground truth ratings on these emotion dimensions, although to a lesser extent. Usability questionnaire results did not indicate any significant difference between both scales, and this finding was supported by positive user reactions in a post-use interview session. Participants also indicated a strong desire to customize the appearance of the model enacting the facial expressions on the interface and suggested the inclusion of further expressions beyond the basic facial expressions of emotion [25, 196].

Therefore, these results lead us to conclude that facial expressions can be used to represent emotions and subsequently enable assessments provided through facial expression-based technology, and that it can be a feasible alternative for VAS-based self-reports.

8.A Questionnaire

Participant information
What is your age? What is your gender?
Semi-structured interview questions on the facial-expression interface
What was your experience when exploring the facial expression interface? What did you like about the facial expression interface? What did you dislike about the facial expression interface? What would you like to see improved about the facial expression interface? What do you think about the realism of the available facial expressions? Did you have any comments about a particular facial expression's realism? Did the presence or lack of realism affect your interaction with the interface? If so, how? How did you find moving around the interactive area in order to find a suitable facial expression? Was the precision with which individual facial expressions changed fine enough to provide your assessments? If no, which facial expressions did not provide a sufficient variety in precision? Was the diversity of the available facial expressions sufficient to provide your assessments? If no, which facial expressions would you have liked to see? Would you have preferred for the interface to include labels for the positions of the available facial expressions? How did the lack of labels for emotional categories affect you in your usage of the interface? Did the visualisation of a facial expression influence you in your assessments? If yes, in what way? Would you use this interface is you had to report your mood on a daily basis? Would you prefer a model of your own or different age, when using such an interface and why? Would you prefer a model of your own or a different gender, when using such an interface and why? Would you prefer a model of your own or a different ethnicity, when using such an interface and why? Would you want to use a model of a particular person when assessing your mood? If so, who would that be? Would you want to see a similar interface with your own facial expressions for reporting your mood? Please elaborate.
Semi-structured interview questions on both interfaces
Were there any technical problems using the scale, which hampered your user experience? Do you have any other feedback about the use, looks, feel or anything you deem important regarding your experience with the application and the interface?


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 9

Identifying and visualising salient facial features indicative of symptoms of depression

9.1 Introduction

Faces encode a plethora of information about a person. We are well versed in recognizing and associating features by reading a person's face. A variety of inferences can be made about a person from facial characteristics such as their identity, age, mood or insights about their health and well-being [30, 31]. By observing and learning variations of facial features, we can associate those to descriptive information about a person. Facial expressions in particular, are reliable indicators of emotion [25]. They encode internal affective states in reaction to social, environmental or internal emotional stimuli. Contemporary understanding of the interplay between emotions and facial expressions identifies at least 4 distinct culturally-universal facial expressions of emotion (Chapter 2, Section 2.4). However, there is also evidence for the existence of expressions associated with further distinct emotions [196, 203].

While facial expressions are descriptive of short-lived emotional experiences, a question remains whether faces could also portray longer-lasting states. Facial features can encode information about a persons' health where, for example, genetic conditions such as Williams syndrome or Wilson's disease have reliable persistent markers encoded in the facial appearance of a person [204]. It is known that depressions also appears to have an effect on facial appearance expressed as differences between persons experiencing low and high depression [205]. However, no extensive evaluation has been done on individual symptoms of depression and their effects on facial appearance. In this chapter, twelve symptoms of depression will be investigated for having a distinct fingerprint, discernible from a face. The symptoms are derived from the Patient Health Questionnaire (PHQ-9), a commonly used screening tool aiding depression diagnosis [42]. In order to obtain a visual

representation for each distinct symptom, the reverse correlation classification images (CI) method was used [43]. It relies on augmenting a base-image of a face with patterned noise, which introduces transformations in the image altering facial features. It is a powerful tool, used in perceptual research for decades that allows to draw out a persons' visual representation of a wide range of mental constructs, where prior work has used it to visualize how laypeople perceive ethnicity [32], social traits [206] and others. For this experiment, the use of the CI method will generate visual models of faces, perceived to depict distinct symptoms common to depressions' symptomatology. In sum, this Chapter investigated whether and which symptoms of depression have discernible facial features and created a visualisation thereof where salient features emerged for symptoms that did, thus delineating them from one-another or the base, non-augmented image.

9.2 Methods

9.2.1 Study Design

The experiment was conducted online on the Amazon mTurk platform. Therein, the nomenclature used to denote a participant is mTurker, where a task assigned to an mTurker is named a Human Intelligence Task (HIT). It allows access to a significant participant pool and its users can deliver data with good quality [207, 208]. Initially, 600 participants were recruited to achieve the target of 50 persons per symptom of depression. This decision was based on prior research which applied the CI method successfully, wherein 20 to 35 persons per criterion were used [32, 209]. However, following an intermittent analysis, the data of 33 was discarded due to poor quality. Additional participants were then recruited to reach the target of 50 people per symptom of depression. Participants were required to provide information about their age, gender and fill out a digital version of the Patient Health Questionnaire (PHQ-9) [42].

9.2.2 Participants

Recruitment was limited to residents of the USA in order to ensure sufficient English language proficiency. Participants were required to be between 18 and 65 years of age and were allowed to take part in the experiment once only. mTurkers with Masters Qualification (i.e. a designation awarded to persons with long-standing history of delivering data of good quality) were initially recruited. However, due to the fact that this population was limited, that constraint was removed after recruiting 357 people. Subsequently recruited mTurkers were only required to have a at least 95% HIT success rate (i.e. 95% of all submitted tasks were accepted). In total, 633 participants were recruited.

Table 9.1 depicts the population demographics split be gender and PHQ-9 scores respectively. The average age of participants was 39.34 (SD=9.36). The

Table 9.1: Demographic characteristics of the participant sample. It consisted of 324 (54%) males and 275 (46%) females, both aged (M=39.34 years, SD=9.36) and with PHQ-9 scores (M=4.76, SD=5.46).

Age	Gender	N	PHQ- Scores				
			0-4	5-9	10-14	15-19	20-27
18-24	M	6	6	0	0	0	0
	F	2	1	0	0	0	1
25-29	M	39	23	11	3	1	1
	F	29	13	6	5	2	3
30-34	M	97	58	22	7	5	5
	F	45	24	12	5	3	1
	Other	1	0	1	0	0	0
35-39	M	68	37	15	12	3	1
	F	60	29	18	10	3	0
40-44	M	42	31	7	2	1	1
	F	52	36	8	6	2	0
45-49	M	34	19	11	2	0	2
	F	30	20	6	4	0	0
50-65	M	38	30	5	1	2	0
	F	57	38	14	2	3	0
		600	365	136	59	25	15

average PHQ-9 scores for participants were 4.76 (SD=5.46). The data collection took place during the first wave of the Covid-19 pandemic between March and August 2020.

9.2.3 Stimuli

The stimuli presented to participants consisted of 500 pairs of images consisting of the same greyscale base-image of a neutral face and a unique noise pattern augmenting it. The base-image was created using the RADIATE dataset, a dataset featuring 1,721 asian, black, hispanic and white adults [14]. An image was initially created for each gender and ethnicity (e.g. asian male, black female, etc.) using only the neutral expressions in the dataset by applying the Delaunay Triangulation [158] method for blending faces. Subsequently, those averaged representations for gender and ethnicity were blended together to create the final base-image. The reason being was that the RADIATE dataset is intentionally imbalanced in its sample sizes for ethnicity, reflective as they reflect birth rates in the USA aiming to address algorithmic biases. It is also unintentionally imbalanced in its gender representation. The CI method, however, works best with a face which is as impersonal and androgynous as possible. That is, a face which does not posses features characteristic of a particular gender, ethnicity or any other defining

features. By performing the intermediary step of creating average representations for each combination of gender and ethnicity, the final base-image is ensured to be comprised of equal parts of each group.

The classification images (CI) [43] approach was employed within a two alternative force choice task [159] (2AFC), which implies that a participant was presented with two choices and was asked to select one. In this experiment, each participant was presented with 500 pairs of noise-augmented images of the base-image, and had to select one of each pair that best resembled someone with a specific symptom of depression (e.g. being down, depressed or hopeless).

(a) Base-image (b) Noise pattern (c) Inverse noise pattern

Figure 9.1: Base-image generated from the RADIATE dataset (a); Pair of images presented in the 2AFC experiment task (b), (c).

Figure 9.1 portrays the base-face image as well as one of the 500 image-pairs presented to participants. Each image-pair in the 2AFC task was created by once augmenting the base-image with a randomized sinusoidal noise. For this purpose the rcicr package available for R [13] was used, where the noise was created in patches by varying parameters for scale, orientation, phase and contrast. Initially, multiple patterns were generated for 6 predefined orientations (0° up to 150° in increments of 30°) and 2 phases (0 and $\pi/2$) [209]. For each generated pattern, a random contrast coefficient was used [209]. This pattern was generated in 5 spatial scales (2, 4, 8, 16, 32) and was then repeated over 2 spatial frequency cycles for each image, generating the final noise pattern [209]. In the experiment, cohesive with prior research, the original and inverse pattern of the generated noise were used to augment the images in each pair respectively [32, 209].

9.2.4 Tasks

Facial expression 2AFC Task Symptoms of depression were derived from PHQ-9 screener questionnaire [42], widely-used to assess a persons' overall depression severity. Table 9.2 contains paraphrased questions in the form of "Which face

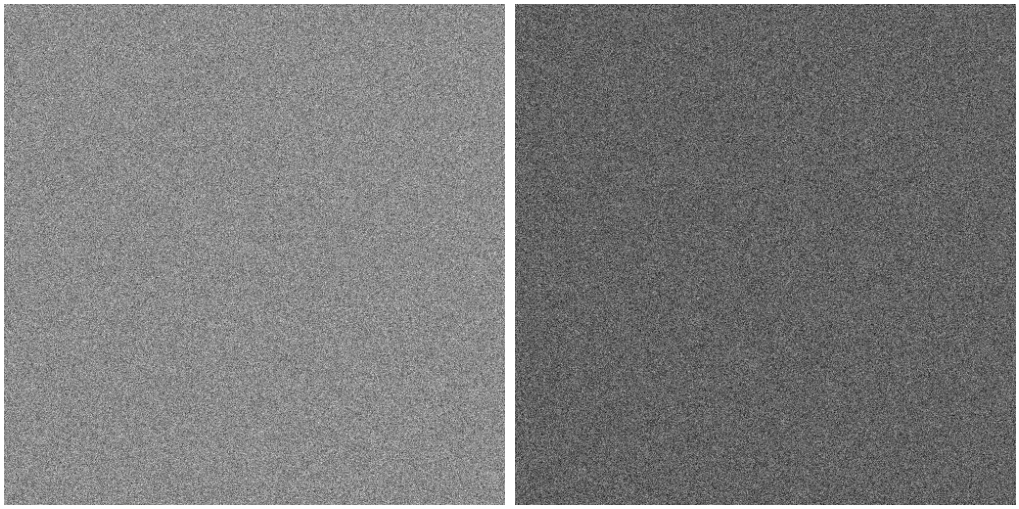
[1]The numerical enumeration corresponds to the question order in the PHQ-9 questionnaire. Sub-questions (a) and (b) were created for PHQ-9 items measuring both polarities of a symptom.

Table 9.2: Depression symptoms used as criterion for the experiment as derived from the PHQ-9 item list in the form of questions.

Index[1]	Question
1	Which face resembles more someone with little interest or pleasure in doing things?
2	Which face resembles more someone who is down, depressed or hopeless?
3[a]	Which face resembles more someone who has trouble falling or staying asleep?
3[b]	Which face resembles more someone sleeping too much?
4	Which face resembles more someone who is tired or having little energy?
5[a]	Which face resembles more someone with a poor appetite?
5[b]	Which face resembles more someone who is overeating?
6	Which face resembles more someone who is feeling bad about themselves?
7	Which face resembles more someone having trouble concentrating on things?
8[a]	Which face resembles more someone moving or acting slowly?
8[b]	Which face resembles more someone being fidgety or restless?
9	Which face resembles more someone thinking of hurting themselves or thinking that they are better off dead?

resembles more someone –" concatenated by the respective PHQ-9 item. Three questions in the PHQ-9 questionnaire inquire about the presence of a symptom on both polarities (e.g. trouble falling or staying asleep and sleeping too much). For this reason, those were split into two univalent variants, resulting in a total of 12 questions. For the duration of the experiment, a participant would see one of the questions from Table 9.2 on top of their screen. Below, they would be presented with a pair of images augmented by noise. Figure 9.1 portrays one such pair.

Visual sensitivity 2AFC Task To ensure each participant was able to distinguish the noise components in the experimental task, a visual sensitivity task was designed. It consisted of selecting the brighter of two 512x512 pixel white-noise images. Those were drawn from a set of 600 pre-generated ones with random cumulative pixel intensities. Figure 9.2 portrays the brightest and darkest ones in the dataset. The cumulative pixel intensity was used to measure the distance between two images in the data set. Initially, two images were presented, whose distance were at half-length of the whole dataset. In a staircase procedure, each correct choice in selecting the brighter of the two images halved the distance between subsequently presented image pairs, while a mistake doubled it. When presenting a subsequent image pair based on the expected distance, the space was sampled randomly over its full range such that overall brighter as well as darker pairs would appear. The staircase procedure terminated after reaching 5 turning points (e.g. making a mistake after a streak of correct answers or recovering from one) [210]. Alternatively, the task ended after 30 pairs were shown. The distance



(a) The brightest image in the dataset. (b) The darkest image in the dataset.

Figure 9.2: Two 512x512 pixel images containing white noise used in the visual sensitivity task. Image on the left is that with the most cumulative pixel intensity and that on the right – with the least cumulative pixel intensity.

(i.e. difference in cumulative pixel intensity) between pairs of images at each turning point was taken and averaged to yield an indication of the visual sensitivity of the participant. If a participant scored below a predefined threshold, they were allowed to continue. If not, their participation ended and they were not allowed another admission. General advice was given to participants prior to attempting the visual sensitivity task, which included emphasizing on the use of a visual aid if needed and avoiding exposing the screen to direct sunlight.

9.2.5 Procedure

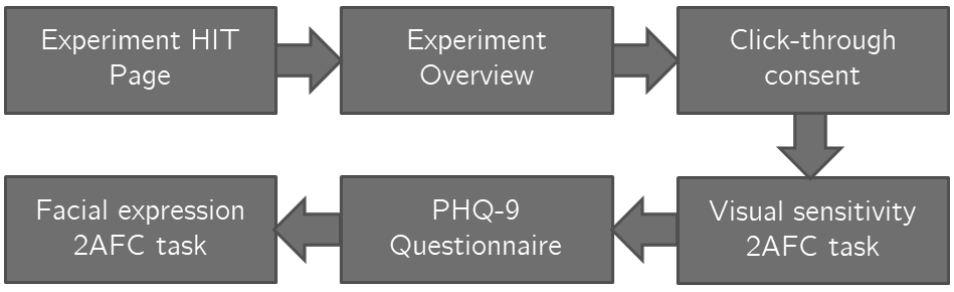


Figure 9.3: Experiment procedure flowchart.

Figure 9.3 provides an overview of study procedure. Upon enrolment participants were redirected to a web-page which elaborated on all steps involved in the experiment procedure with a brief description thereof. Next, a click-through consent data form was shown. It contained information about which data would be recorded

and how data will be anonymized to respect their privacy. Due to the fact that the experiment involved observing images and identifying features characteristic of depression, a recommendation was included to withdraw ones' participation should they have been feeling mentally unwell. This page also included information and self-help material about depression. After obtaining consent, participants were redirected to the visual sensitivity task. If successful, they were then asked to fill out a digitized version of the PHQ-9 questionnaire and provide their age and gender [42]. The tasks' order ensured that data was collected only for participants admitted to the experimental task. Upon providing this information, they were redirected to the experiment task. Based on collected population statistics, each participant was assigned to assess a single symptom of depression, typically the most under-powered one. Upon selecting the face that most resembled the assigned symptom to their task, a participant would advance to the next pair. This process would repeat until all 500 pairs of images have been seen. The order in which the 500 pairs were presented as well as the relative position (e.g. left/right) of images within each pair was randomized. On the 150^{th}, 275^{th} and 400^{th} image-pair, participants were given an optional 30 second break, which they could voluntarily skip.

Data collection seized after reaching the target of 50 participants per symptom. It resumed with an additional 33 participants after an intermittent analysis (described in depth in Section 9.2.6) was completed and dataset of low information quality were discarded.

On average, from intake to completion the experiment took 38 (SD=15) minutes. Participants were remunerated with 10$ for their effort. No partial remuneration was offered to participants not meeting the inclusion criteria or disqualified in the visual sensitivity task. This was done in order to not incentivise mTurkers, who are known to be tech savvy, to optimize their strategy to acquire the potential partial remuneration instead.

All participants that successfully completed the experiment were paid, regardless of whether their data was used or discarded.

9.2.6 Data Analysis

As the experiment was conducted online, it was not possible to oversee participants and ensure their attentiveness during the task. In order to mitigate that, response time and choice selection pattern were analysed for disingenuous responses.

Response time was a fairly simple and straightforward criteria to use. Some participants were unreasonably quick in completing the task, while others took significantly longer than the average. Longer times could have been caused by participants forcing a break themselves and did not warrant suspicion as the experiment did not impose a time-limit on completing the task. Extremely quick completion times, however, were unfeasible to be associated with an earnest response.

To estimate information quality of a selection choice pattern, aberrant responses

which are uncommon for a genuine selection and indicative of inattentiveness needed to be defined. Those can manifest in two ways – a repetition of a single character, or a repetition of a combination of characters. The probability of a participants' selection choice pattern containing those frequently was extremely low, since image positions (e.g. left/right) within pairs and the order of pairs within the 2AFC task were randomized. To analyse selection choice data, two types of analysis were considered – compression algorithms and entropy. Those methods would be able to quantify the frequency of occurrence of aberrant patterns within an input sequence and output a score, representative of its information quality.

Compression Compression algorithms are able to transform information into a more efficient, shorter representation. Compression inherently requires the presence and substitution of repeated elements within the input sequence with a shorter variant. The difference between uncompressed input and compressed output reveals its efficacy. In lossless compression, this process is also reversible, such that the original data can be reconstructed from a compressed output with no loss of information. This is important as lossless compression algorithms are also deterministic, which means that for any given input, a compression thereof would result in one and the same output. In this context, only lossless algorithms were considered suitable as they ascertain that the resulting score from performing a compression can be explained, interpreted and replicated.

Two algorithms were suitable for detecting aberrant responses – Run length encoding (RLE) [211] and DEFLATE compression [212]. RLE is a lossless data compression algorithm where sequences of the same value are stored as that value and a count number. RLE is ideal for identifying single character repetitions. DEFLATE is a well-known lossless encoding scheme which compresses subsets of characters based on their frequency of occurrence. This algorithm is suited to identify recurrent repeated sequences consisting of multiple characters. Those methods were scored based on their compression efficiency by dividing the length of the compressed sequence by the length of the original. A lower score indicates a more efficient compression and in turn – more repeated sub-sequences found in the input. In the subsequent analysis, DEFLATE is coded using the DFL acronym.

Entropy Another way of identifying repetitions in an input sequence was to compute the entropy within a response's distribution. To achieve that, a scheme was needed to break down an input sequence and represent it as a multiple fixed-length sub-patterns. The selection choice pattern of the 500 image-pairs was encoded as L and R, where L coded for images positioned on the left side of the screen and R – on the right. Numerically, L and R were represented as 0 and 1 respectively, however, in this section the nomenclature of L and R will be used. An more elaborate description of this method can be found in Appendix 9.A.

From the resulting encoding, permutation tables of different lengths were built, consisting of all selection combinations. Since the experiment gave participants

a dichotomous choice (e.g. choosing the left or right image), those permutations only feature combinations of L and R in varying order. Short permutation tables, for example, with a fixed-length of 2 consisted of the combinations LL, LR, RL, RR and were able to detect short-length recurrent sequences. Larger permutations tables (e.g. LLLL, LLLR, LLRL, etc..) are able to detect longer repeated sequences. Calculating the entropy using permutation tables implies creating a score representative of the difference between the distribution of sub-sequences of length of the permutation table in a participants' selection choice pattern and the normal distribution. As the image positions within pairs and order between pairs were randomized (thus uncorrelated), those scores assume that single or multiple character repetitions are uncommon (e.g. observing LRRL is more probable than LLLL). For the subsequent analysis, entropy was computed using permutation tables of length of 1, 2, or 4 coded as ENT1, ENT2 and ENT4 respectively.

Data Exclusion The analysis consisted of using statistical methods performed on the distribution of scores for response time, and compression and entropy metrics [213]. Statistical methods are most suitable to analyse this type of data as they do not rely on knowledge of what the underlying data represents, but rather simply flag tail-ends of a distribution. Due to some participants presumably enforcing self-enforced breaks, the distribution of response times had a heavy upper tail. In order to normalize response times, a Box-Cox power transformation was applied first. The resulting distribution was normal, which allowed to apply the 95% threshold rule on the lower distribution tail containing the quicker responses. This metric alone ruled out 22 participants. For the entropy and compression metrics, a more conservative 99.7% threshold was applied. This was due to the fact, that those methods are not standardized and were applied as an information quality metric specifically for data obtained in this experiment. Each of the metrics resulted in the following number of dataset exclusions DFL: 12, RLE: 9, ENT1: 7, ENT2: 10, ENT4: 13. As expected, there was a large degree of agreement between metrics, resulting in a total of 17 unique exclusions. Finally, combining the response time and selection pattern metrics resulted in excluding dataset of 33 participants. A more elaborate analysis of the employed data exclusion approach, methods, scoring mechanism as well as some further considerations on the metrics' applicability for similar experiments can be found in Appendix 9.A.

9.2.7 Image Analysis

As elaborated on in Section 9.2.3, using reverse correlation CIs, irrelevant noise is filtered out, while salient one is retained or reinforced. This promotes that, within the 2AFC task one or the other image in each image-pair might posses some facial features, indicative of how the symptom is reflected in the face. Averaging the noise component over multiple such choices (e.g. participant selections in each image-pair) draws out salient facial features, while suppressing irrelevant noise. The average image obtained from multiple selections by multiple participants for one

specific depression symptom then visualizes what facial features are characteristic for that symptom according to the population. This was done for all nine PHQ-9 symptoms of depression.

The rcicr R package [13] was used to analyse the retained dataset. First, for each symptom of depression, all selections made by participants allocated to that particular condition were summed up and averaged pixel-wise to create a single image. Thus, each symptom visualisation was in effect comprised of 25,000 contributing images derived from the 500 selections made by each of the 50 participants. Hence, each resulting image for a specific symptom depicted an aggregated visualisation of how that symptom was perceived to be reflected in the face by all participants allocated to that condition. Thus, this resulted in a visualisation of 12 images descriptive of the symptoms of depression as defined in Table 9.2.

In total, two additional visualisations were created from the symptom images, which highlighted further interesting or contrasting features. Those visualisations consisted of: 1) the aggregated visual representations of a symptom as derived by participants; 2) A symmetric representation, where the resulting image was created by averaging of the original symptom image and its mirrored-variant and 3) a vertical-split visualisation, where the right half of the face (from the perspective of the reader) were mirrored horizontally. To delineate those visualisation from one another in future sections, those were henceforth referred to as 1) representation, 2) mirror-average and 3) vertical-split. Figures 9.4, 9.5 and 9.6 display the results from this analysis and featured all three types of visualisations for each symptom.

The reason for creating additional visualisations in addition to the aggregated representations created by participants was that those are able to draw out particular facial features better. Those are also able to partly compensate for some inherent characteristics of the reverse correlation CI method. For example, due to the fact that the augmenting noise was pseudo-random, alterations in facial features were not symmetric across both left and right sides of the face. To create a more uniform representations for symptoms, a symmetric variant was created. Additionally, for the same reason, some symptoms possess features emphasized more strongly on either the left or right sides of the face. As we are primed to process faces holistically [214], to highlight those discrepancies the vertical-split representation was created.

Finally, Appendix 9.E, and Appendix 9.D contains facial representations of symptoms of depression subtracted from the base-image as well as those of a subset of the population which scored 7 or more on the PHQ-9 questionnaire. Those have been included here for completions' sake and will not be referred to in the rest of this chapter. However, they will be later referred to in the Thesis Discussion in Chapter 11 highlighting their potential application.

9.3 Results

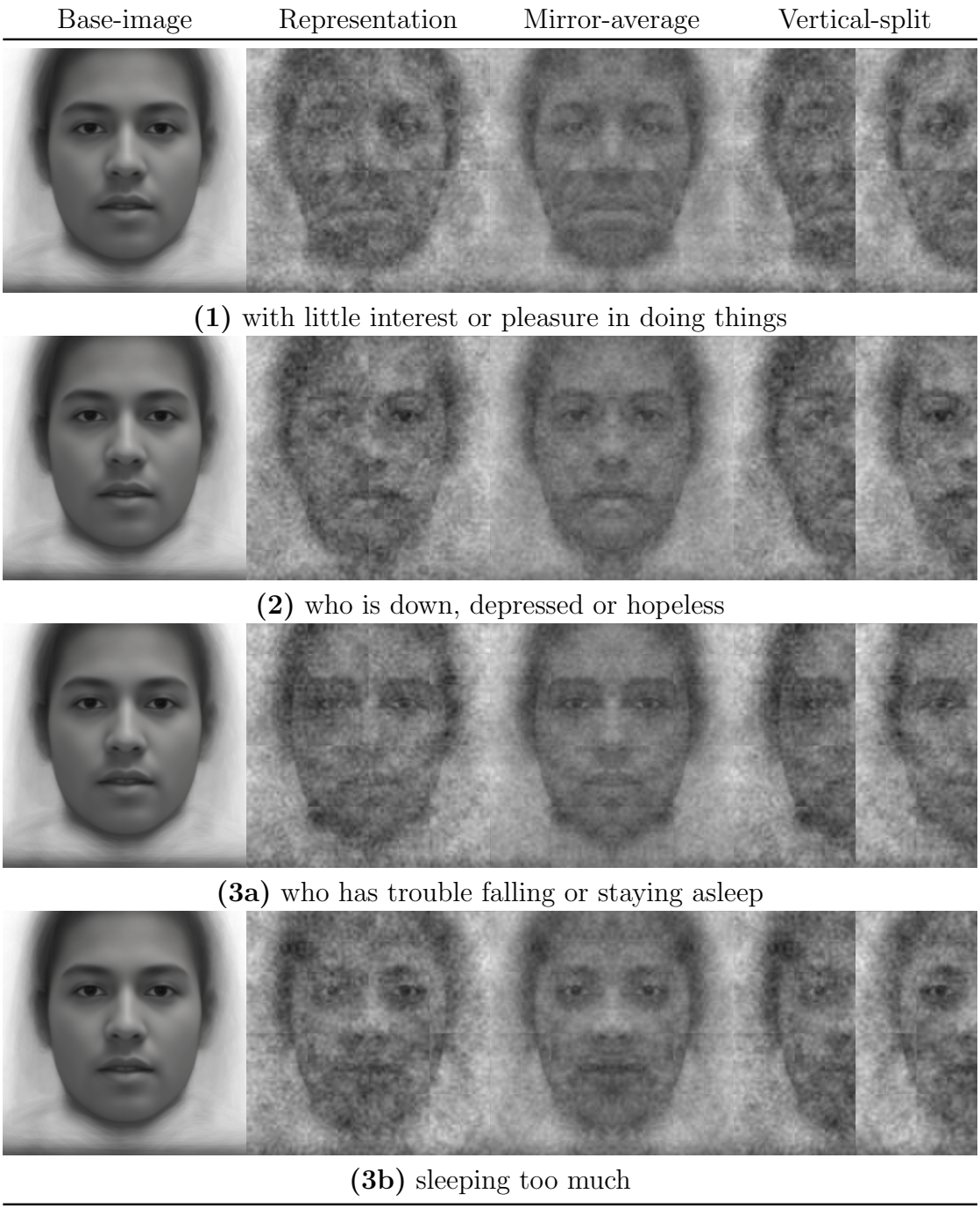


Figure 9.4: Figure containing the base-image, created representations as well as their mirror-average and vertical-split variants for symptoms (1), (2), (3a) and (3b). The labels below indicate the exact criterion participants used for their selection: "Which face resembles more someone –"

The symptom representations for *having little interest or pleasure in doing things* (1) (Figure 9.4) was characterized by noticeably low contrast in the eyes, portrayed a somewhat vacant expression. Additionally, the mouth displays a down-open curve, similar to the facial expression for a mouth frown. *Feeling down, depressed or hopeless* (2) showed only one distinguishing feature, namely more arched eyebrows, which are particularly noticeable in the vertical-split. The pair of symptoms for *under-* (3a) and *oversleeping* (3b) exhibited a strong difference, particularly in the area of the eyes, as expected. What was clearly noticeable in (3a) was that the eyes appear more closed and there was an overall shading around them. Conversely, (3b) also features shading primarily concentrated under the eyes, where the eyes appear to be more wide-open. In addition, the eyebrows were raised conveying a more alert and overall awake expression. This becomes more apparent when looking at the vertical-split and mirror-average representations. There, (3a) portrayed an overall shading around the eyes, while (3b) – only below them.

Feeling tired or having little energy (4) (see Figure 9.5) portrayed a down-open curve of the mouth, arched eyebrows and lowered outer corners of the eyes. *Poor appetite* (5a) and *overeating* (5b) did not differ in the shape of the face, but predominantly in the facial expression they portrayed. The face for *poor appetite* portrayed a rather sad expression characterized by a down-open curve of the mouth and lowered eyebrows, where that for *overeating* did not. *Feeling bad about oneself* (6) also did not stand out with any clearly distinguishing features, delineating it from the base-image.

The face for *having trouble concentrating on things* (7) (see Figure 9.6) resembled the base-image used for the experiment. There were no particular facial features standing out, which was also be observed in the mirror-average representation. *Moving or acting slowly* (8a) displayed a lighter shading of the eyes as well as a faint smile. *Being fidgety or restless* (8b) displayed a darker shading in the eyes, slightly inwards raised eyebrows, typically associated with the facial expression for surprise or worry. Additionally, a down-open curvature of the mouth as well as shading under the eyes was observed. Those features were particularly noticeable in the mirror-average representation. The symptom for *thinking of hurting oneself or that one is better off dead* (9) featured lowered eyebrows, a down open curvature of the mouth as well as lowered outer corners of the eyes. We also noticed an overall darker shading of the eyes.

Regarding facial expression symmetry, some symptom representations did not exhibit features consistently across both halves of the face. Those differences were somewhat obscured as intrinsically we interpret faces holistically [214]. The symptoms for someone: *with little interest or pleasure in doing things* (1), *who is down, depressed or hopeless* (2), *sleeping too much* (3b) in Figure 9.4; *who is tired or having little energy* (4), *who is feeling bad about themselves* (6) in Figure 9.5, and *who has trouble concentrating on things* (7) and *moving or acting slowly* (8a) in Figure 9.6 portrayed consistent facial features across both halves of the face and exhibited a variation in tone between both halves. Conversely, the expression for *having trouble falling or staying asleep* (3a) (Figure 9.4), *poor appetite* (5a),

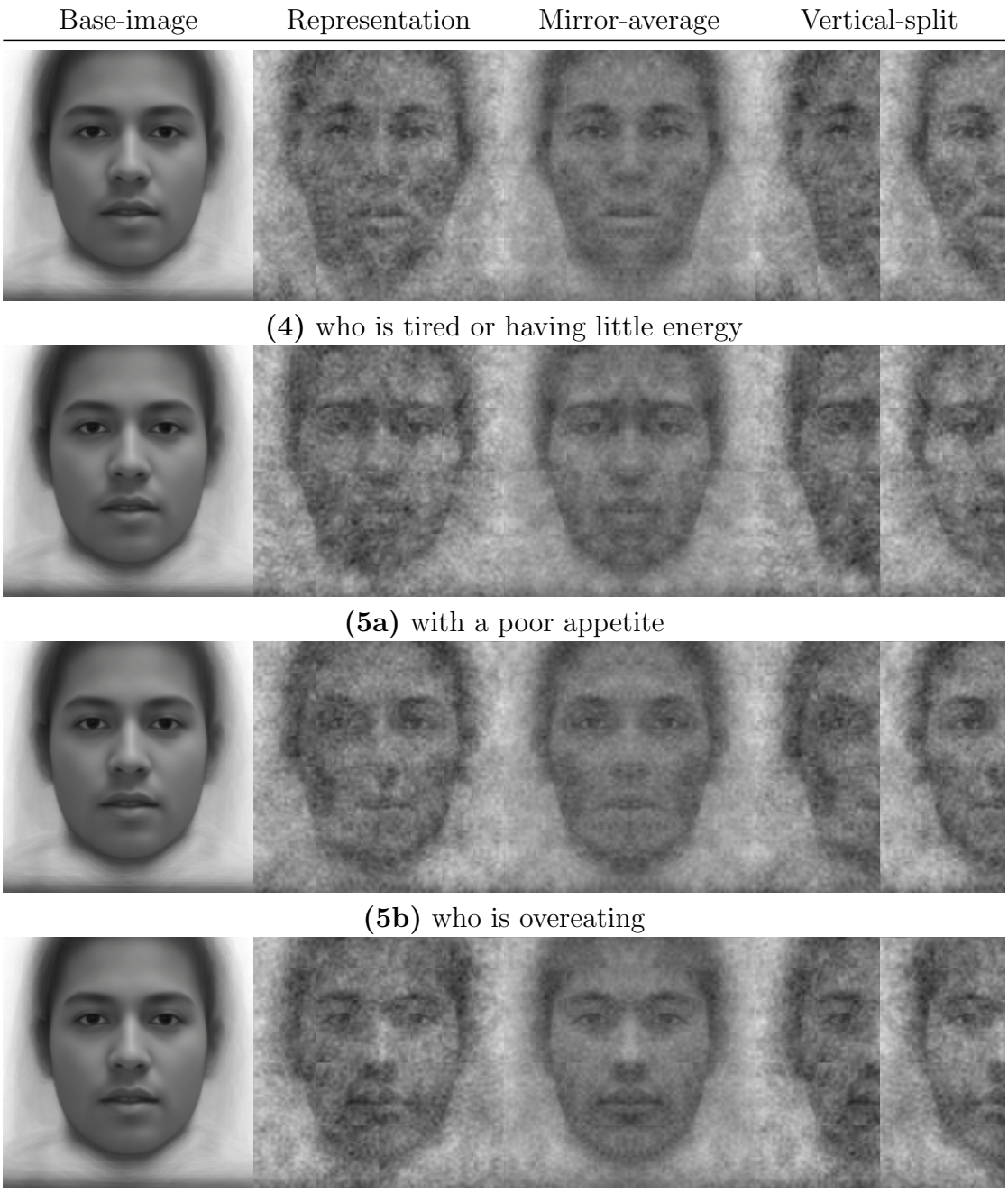


Figure 9.5: Figure containing the base-image, created representations as well as their mirror-average and vertical-split variants for symptoms (4), (5a), (5b) and (6). The labels below indicate the exact criterion participants used for their selection: "Which face resembles more someone –"

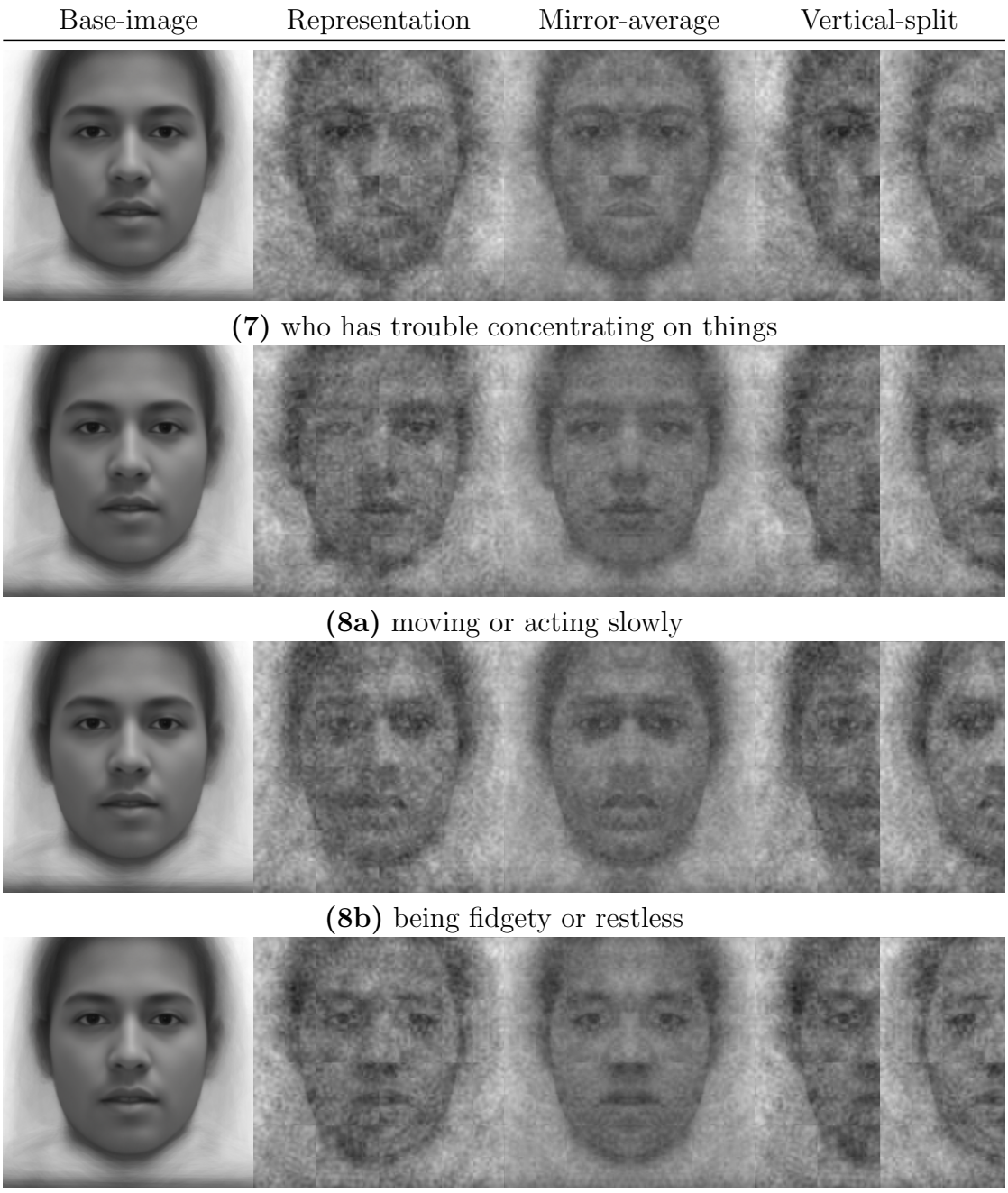


Figure 9.6: Figure containing the base-image, created representations as well as their mirror-average and vertical-split variants for symptoms (7), (8a), (8b) and (9). The labels below indicate the exact criterion participants used for their selection: "Which face resembles more someone –"

overeating (5b) (Figure 9.5), *being fidgety or restless* (8b) and *thinking of hurting oneself or thinking that one is better off dead* (9) (Figure 9.6), however, portrayed different expressions in both halves of the face. In the vertical-split for *someone who has trouble falling or staying asleep* (3a), the corners of the mouth were misaligned resembling simultaneously a slight smile on the right side of the face and a frown on its left side. A similar observation was made for the symptoms for *poor appetite* (5a) and *overeating* (5b).

The representations for *being fidgety or restless* (8b) and *thinking of hurting oneself or thinking that one is better off dead* (9) diverged in some interesting ways. In the vertical-split representation for *being fidgety or restless* (8b) the eye on the left side of the face appeared to portray a down outwards-oriented gaze direction, akin to averting one's eyes. For *thinking of hurting themselves or thinking that they are better off dead* (9), the right side of the face portrayed an almost neutral-looking facial expression, while in contrast, the right side was distinctly characterized by a raised inner corner of the eyebrow, and narrowed and downwards-oriented corner of the eye. The combination of those features portrayed a sad and almost tearful-looking facial expression.

9.4 Discussion

9.4.1 Symptoms with distinct facial features

Symptoms which had distinct features present were found for the symptoms of *having little interest or pleasure in doing things* (1), *has trouble falling or staying asleep* (3a), *sleeping too much* (3b), *being tired or having little energy* (4), *poor appetite* (5a), *overeating* (5b), *being fidgety or restless* (8b), *thinking of hurting oneself or thinking that one is better off dead* (9) which delineated them from the base face.

The symptom of *little interest or pleasure in doing things* (1) can also be seen as a proxy for anhedonia. In its representation in Figure 9.4, particularly noticeable was contrast of the eyes, which appears to be lowest from all symptom representations. In addition, the slight down-open curvature of the mouth was associated with a mouth-frown and could have conveyed the notion of being displeased. Previous research indicated that anhedonia is not only associated with deriving less joy from positive experiences, but also with less facial expressiveness [215, 216] which was consistent with the obtained representation.

The symptoms of *has trouble falling or staying asleep* (3a), *sleeping too much* (3b), also possess distinguishing facial features, whereas those were expressed primarily in the eyes (Figure 9.4). *has trouble falling or staying asleep* (3a) featured narrowed eyes, and somewhat lowered eyelids portraying tiredness, while that for *sleeping too much* (3b) – wide-open eyes. In addition, a darker tone was be noticed around the eyes for *has trouble falling or staying asleep* (3a), consistent with skin discolouration, a known effect of this symptom [217] and a darker tone concentrated

solely under the eyes for *sleeping too much* (3b). Both of those descriptions were considered characteristic consistent for the symptoms they described.

The symptoms for *Being tired or having little energy* (4) and *poor appetite* (5a), both faces possessed features, which would normally not be implicitly associated to them. Namely, both exhibited characteristics similar to that of sadness and namely a down-open curved mouth, arched eyebrows, and lowered corners of the eyes. For *being tired or having little energy* (4), it was unclear whether the symptom is perceived to have incorporated a socially attributed component of sadness without it being explicitly stated in the question. Prior research indicates that patients who are feeling tired all the time perceive the symptom to be somatic, whereas doctors view it as psychologically driven [218]. This hinted that there may be discrepancies within the formal classification of the symptom and the lived experiences of patients. The social attribution of features, which participants might have associated as accompanying a symptom might explain the resulting features in this representation.

Similarly, the symptom for *poor appetite* (5a) exhibits similar characteristics as *being tired or having little energy* (4). A potential explanation may also lie in the social perception of this symptom. First, the item phrasing for symptoms in the PHQ-9 questionnaire for *poor appetite* (5a) and *overeating* (5b) differ such that the first explicitly states a change in appetite, while the latter – the act of overeating. Although both *poor appetite* (5a) and *overeating* (5b) could be caused by a disease or ailment, contemporary perception of *overeating* (5b) is typically not associated with those as a cause [219]. That of *poor appetite*, however, is less prevalent in developed countries and could carry socially imbued characteristics. Those may be implicitly attributed to the symptom itself without having been explicitly stated in the question. As such, the symptom for *overeating* (5b), did not exhibit the same features, characteristic of sadness as *poor appetite* (5a). Additionally, prior work indicates that changes in appetite does not result in weight difference for depressed persons [220]. However, it is feasible to expect that laypersons might not posses in-depth knowledge of the course-trajectory of symptoms of depression and may rather have oriented themselves according to common knowledge and social associations.

The symptom for *being fidgety or restless* (8b) exhibited noticeable down- and sideways eye gaze direction, particularly visible in the vertical-split representation in Figure 9.6. Furthermore, in its the mirror-averaged representation, lowered eyebrows and narrowed eyes were observed, a combination of features known to convey the feeling of being worried [221]. It is uncertain, whether that is indeed the case, however, depression and anxiety are a known comorbidity, hence a certain degree of overlap in the experience of certain symptoms is expected. In fact, restlessness is a symptom which is used to contribute to an anxiety score in the Hamilton rating scale for anxiety (HAM-A) [222].

The symptom for *thinking of hurting oneself or thinking that one is better off dead* (9) was perhaps the most expressive and evocative of all representations. In the vertical-split model in Figure 9.6, the right and left half of the face, however,

displayed very different expressions. The right image featured lowered corners of the eyes and eyebrows and a downwards-curved mouth, is significantly more expressive compared to the left side, and bears a stronger association with the symptom. In contrast to the left image resembled the neutral expression instead. The expressiveness in the right half of the face, however was strong enough such that in the mirror-average representation, the expression still strongly resembled that of sadness. Prior works have identified that laypeople are able to detect suicidality from yearbook photos [223], a finding that supports that this symptom may have sufficient influence to be manifested in the face and subsequently recognized within a facial expression.

9.4.2 Symptoms absent of distinct facial features

A number of symptoms did not possess any distinguishing facial features. Those were the representations for *being down, depressed or hopeless* (2), *overeating* (5b), *feeling bad about oneself* (6), *having trouble concentrating on things* (7) and *moving or acting slowly* (8a).

Surprisingly, for the symptom of *being down, depressed or hopeless* (2), the expectation was to see a facial expression resembling that of sadness. However, there were no particular features that distinctly conveyed that emotion. A possible explanation may lie in how the question was phrased. It featured three emotionally-charged words that on their own describe similar, but distinct emotional states. Thus, the complexity introduced by including all three terms may have hindered participants from being able to accurately associate images presented during the task to this symptom. Therefore, decoupling the question into three distinct single-word variants might have yielded expressions with more pronounced features.

Another symptom, which does not appear to deviate significantly from the base-image was that for *overeating* (5b). When regarding this symptom in the context of its opposite – the symptom for *poor appetite* (5a), the expectation was that those would diverge primarily in the shape of the face (e.g. face contours), where overeating would have depicted a generally broader face, while that for poor appetite – the inverse. However, although the symptom for *poor appetite* (5a) possessed facial features more characteristic of sadness, neither of those two symptoms differed much in the shape of the face. As previously mentioned, the symptom of *overeating* in particular was defined as the act of overeating rather than as an increase in appetite, which further reinforces this unexpected result. A plausible explanation for this observation may lie in how the reverse correlation method works. As the base-image was augmented by patterned pseudo-random noise, altering spatially compact areas in an image is relatively easy. One only needs to be able to identify and consistently select images which portray a particular alteration in a feature (e.g. eyes, mouth corners, eyebrows). For spatially smaller features such as the eyes, the method appears to work. However, to significantly affect a spatially distributed feature such as face-contours, one needs select images which offset the face contours along its complete circumference within a single

image. Hence, even if participants were trying to select images which did alter face-contours, it was unfeasible to do so consistently for its complete circumference as that would have required a significant portion of images to have offset the contours of the face consistently. This was very unlikely due to the randomness inherent to the method of creating the augmenting noise. Subsequently, through the averaging process of selected images, those features were diluted in the resulting representation.

For the symptom of *feeling bad about oneself* (6), the only distinguishing feature was the arrangement of the lips where they appeared sealed, a feature not observed in any of the other symptoms. However, it was not clear whether was of any significance to this representation or was an artefact caused by the method. The symptoms *having trouble concentrating on things* (7) and *moving or acting slowly* (8a) did not appeared to deviate in a noticeable way from the base-image either. The common denominator for those three symptoms is that they feature a distinct cognitive component [224], where typically, cognitive symptoms are expressed through changes in behaviour, which is difficult to express or identify in a static image. Hence, it was not unexpected to observe a lack of strongly emphasized facial features. In fact, recent findings confirm that cognitive traits did not appear to have an effect in the development of particular facial features [225], which may translate to those symptoms having little influence in facial expressiveness later in life.

9.4.3 On facial expression symmetry

Regarding the symmetry of facial expressions, there are some visible differences that delineated the left and right sides of the face.

In the vertical-split of the symptoms for *having trouble falling* or *staying asleep* (3a) in Figure 9.4 and *poor appetite* (5a) and *overeating* (5b) in Figure 9.5, both halves of the face portrayed a divergent shape of the corners of the mouth. In either the left or right side of the face, one half of the mouth appeared to express a light smile, while the other – a slight frown. A reasonable explanation for those differences may be due to lack of informativeness of certain facial features to an expression or symptom. For example, the shape of the mouth would not be typically associated with either of the above-mentioned symptoms, where as previously explained, sleep-related symptoms would be expected to be reflected through the eyes, while those for changes in appetite – through the shape of the face. As a consequence, lightly expressed features may be a by-product of noise-induced randomness and may not necessarily be informative to the respective visualisations.

The vertical-split representations for the expressions of *being fidgety or restless* (8b) and *thinking of hurting oneself or thinking that one is better off dead* (9) in Figure 9.6 exhibit some stark contrasts. Therein, the right image was significantly more expressive than the left. It is known that the augmenting noise in not distributed symmetrically across the complete image and consequently, it is unlikely to see the equivalent alterations in both paired symmetric facial features such as

eyes, mouth corners and others. Hence, participants may have tried to identify salient features locally in for one of the paired facial features, rather than perceiving the face holistically. In turn, over the course of the experiment, they may have chosen to focus on a particular side of the face or facial feature of interest. Prior work highlights that the left side of the face (right image) is more emotionally expressive [226]. Additionally, in an experiment where faces were vertically split in two halves and subsequently presented to participants, they judged the left side as portraying emotions more intensely [227]. This could, in part, explain the significant differences for *being fidgety or restless* (8b) and *thinking of hurting oneself or thinking that one is better off dead* (9) in their vertical-split representation.

9.5 Limitations

Certain limitations to the approach employed within this chapter pertain to the conditions used to define the symptoms as well as the reverse correlation classification images method itself. In the former, the symptoms were crafted from the PHQ-9 questionnaire, a well-known instrument for measuring depression severity, as complete sentences describing the presence of a symptom. However, over the course of the experiment, more succinct phrasing, ideally a single word variants may have isolated some confounding socially attributed characteristics which were observed for some of the symptoms. In the latter, the reverse correlation method was found lacking in its ability to augment spatially distributed features such as face-contours.

9.6 Conclusion

In this Chapter a visual representation of what each symptom of depression according to the PHQ-9 questionnaire looks like in the face were presented. The results from this study indicate that laypeople are able to recognize a number of symptoms of depression and those were reflected in the created representations for the symptoms of depression. This implies that depression may indeed influence facial expressiveness or impacting facial features in some distinct ways which makes them recognizable. The novelty in this approach is within the visualisation of those symptoms and the implications those may have for creating facial expression-based scales for assessing symptoms of depression frequently through the use of smartphone applications or in recognizing distinct symptoms of depression from a face in an image using, for example, machine learning methods.

9.7 Chapter Summary

The present study explored whether laypeople have an established perceptual model to recognize a range of symptoms of depression in a face. Additionally, through the

reverse correlation classification images method, it was possible to visualize a group of participants' mental representation of how a particular symptom was reflected in the face. Most symptoms, albeit not all, had distinct facial features, which delineated them from the base-image and from one another. Those observations were discussed in the context of existing literature.

9.A On compression and entropy data quality metrics

This appendix provides a more elaborate analysis of the data exclusion approach described in Section 9.2.6.

In an experiment, when distracted, disengaged or not interested, persons which deceptively aim to portray genuineness in their responses would often rely on a rhythmic pattern. It serves the purpose of diversifying their choices thus giving them credibility and speeding up their response times aimed to minimize time spent on the task. It's a strategy optimizing for the least amount of work and time spent while still receiving the full benefit from doing the task. The present approach attempts to identify such patterns. Aberrant responses found within a selection pattern can be defined as repeatable sequences, indicative of automatic engagement with the experiment without attending to its conditions. They can manifest in two ways – a repetition of a a) single character or b) combination of characters. In the following sections, further information will be provided on how both, compression algorithms and entropy, can be used to detect both a) and b) type of aberrant responses and why they are not interchangeable but rather work complementary to one another.

9.A.1 On compression

Run-length-encoding (RLE) [211] is a lossless data compression algorithm where sequences of the same data value are stored as a value and count number. This algorithm is perfectly suited for detecting a) single character repetitions. The DEFLATE algorithm [212] uses Huffman codes to encode repeatable structures in the input data, based on their frequency of occurrence. In essence the algorithm minimizes the cost of representing a recurrent sequence with a shorter placeholder sequence. Therein the more frequent the co-occurring pattern is, the shorter the substituting placeholder sequence. This algorithm is suited to identify b) recurrent structures. We will omit a more in-depth analysis of DEFLATE as this appendix intends to provide a functional understanding of its functionality. It is also possible to compress data with DEFLATE using multiple passes (e.g. that is performing compression on already compressed data). However, although effective for reducing required storage space further, this hurts interpretability as it is difficult to predict what patterns could emerge in the compressed data after the first pass. Furthermore, multiple passes can also reduce variability, thus increasing the difficulty of defining a threshold delineating genuine from disingenuous responses. Compression algorithms' score on an input sequence is computed by dividing the compressed output by the uncompressed input.

9.A.2 On entropy

Another way of identifying repetitions in an input sequence is to quantify its entropy. The selection pattern obtained from a participant in the experiment task described in Section 9.2.4 is in essence a series of dichotomous choices. As such, it can be represented in binary such that selected images positioned on the left portion of the screen are coded as 0 and those on the right – as 1. The a-priori randomization over 1) the order of image-pairs and 2) positions of images within pairs ensures that the resulting left/right or 0/1 choices are uncorrelated. Next, using this encoding allows to build permutation tables, which can be used to detect repeated sub-sequences within an input pattern. Those contain all possible combinations of choices according to the length of the permutations. For dichotomous data encoded as 0 and 1, a permutation table of length two would contain the sequences 00, 01, 10 and 11. The input sequence is parsed using an envelope which equals the length of the permutations.

A stride parameter is a number which dictates how many characters are skipped between iterations during parsing. As an example, for the input sequence 1010, a permutation table of length 2 and stride parameter of 2, the sub-sequence for 10 is identified twice. All other permutation bins will be empty. Alternatively, in the same conditions but with a stride parameter of 1, the sub-sequence 10 is detected twice and 01 – once. To evaluate information quality in this experiment, permutation tables of length 1, 2 and 4 were chosen and coded as ENT1, ENT2 and ENT4 respectively. The stride parameter was chosen to be equal to the length of the permutation tables in each variant, which ensures that each choice will contribute to a single detection. In this case ENT1 can detect abnormalities in the distribution between 0s and 1s in the input string, descriptive of type a) aberrant response. ENT2 and ENT4 would be able to capture repeated sequences of length 2 and 4, descriptive of type b) aberrant responses. The entropy score is computed per permutation table variant by comparing its distribution table to that of the normal distribution.

Longer permutation tables could also be used, however, they come at a trade-off. It consists of being able to capture longer sub-sequences more effectively at the expense of reducing the number of observations. For example, ENT4 with a stride parameter of 4 for an input sequence of 500 nets only 125 observations. Alternatively, shorter variants such as ENT2 and ENT4 could still detect longer sequences as a combination as long as there is a common denominator between permutation and sequence length. For example, the repeated sequence 010100 could easily be captured by ENT2 as three separate observations, while ENT4 would be able to detect its double (e.g. 010100010100). Hence, a longer permutation table would feature less data, yield diminishing returns, but can be useful in very specific situations. Alternatively, one could opt for a shorter stride parameter implying that choice(s) on the edge of the moving window will be counted and contribute to multiple detections. That, however, artificially inflates the input data. As a consequence, it is anticipated that the difference in score between sequences

containing repeated patterns and those that do not, would increase. However, further experimentation may be needed to confirm whether that is indeed the case.

9.A.3 Compression vs Entropy

Table 9.6 portrays the correlation between each employed compression and entropy metric as well as their Interquartile ranges (IQR) (i.e. Q3 - Q1) as a measure of variability. DFL, ENT2 and ENT4 have a moderate to strong correlation. Those metrics were also intended to be sensitive to type b) aberrant responses. RLE and ENT1, sensitive to type a) aberrant responses, are moderately correlated, however, less so than DFL, ENT2 and ENT4 are. The relatively lower correlation between RLE and ENT1 can be explained by a fundamental difference in how they score an input sequence. ENT1 processes an input holistically and in essence measures the ratio between 0s and 1s, while RLE substitutes single character repetitions as the character and its repetition count. As such, they are not easily comparable or interchangeable. Another observation was that IQR for the more complex DEFLATE compression algorithm noticeably lower compared to other metrics. This is due to the fact that it is very efficient in compressing data, however, its low variability warrants caution in its use as it gives less flexibility in delineating aberrant from genuine responses. Appendix 9.B and 9.C display a part of the encoded input sequence and metrics scores for a portion of the excluded and included participants respectively. The table also features the ENTR metric, which is a variation of ENT1 described as the ratio between 0s and 1s. Therein, a ratio of 0.5 is optimal as it implies a uniform distribution of 0s and 1s in a participants' selection choice pattern.

Table 9.6: Correlation table between the RLE, DFL, ENT1, ENT2 and ENT4 and their Interquartile ranges (IQR) (i.e. Q3 - Q1).

	RLE	DFL	ENT1	ENT2	ENT4
RLE	1.0	0.194	0.448	0.337	0.273
DFL	0.194	1.0	0.607	0.875	0.921
ENT1	0.448	0.607	1.0	0.861	0.745
ENT2	0.337	0.875	0.861	1.0	0.97
ENT4	0.273	0.921	0.745	0.97	1.0
IQR	0.086	0.01	0.0158	0.0259	0.0351

9.A.4 Final considerations

This described approach assumes that there will be a number of dishonest responses in a dataset and attempts to identify the most likely ones based on scoring participants' selection patterns. The experiment described in Chapter 9 (p. 118) featured a large sample and applying the 99.7% threshold resulted in relatively low count of exclusions. Applying the same procedure in low-powered experiments could introduce or reinforce a bias or obscure an effect. Using the entropy metrics

also requires a good understanding of the underlying data structure and expected distribution of selection patterns. Those metrics, however, are perfectly suited for unstructured data, where items are known to be uncorrelated with each other. Structured formats, such as questionnaires need to have the order of questions as well as responses randomized before they are administered. The effect of randomization should effectively decorrelate responses on items, however without further investigation it is not known whether that alone would be sufficient.

Entropy scores are normalized by default since the reference is the normal distribution and their thresholds can be compared between experiments. The entropy-based approach was applied, in this case, on dichotomous data. However, it should also be applicable to multidimensional data (e.g. multiple-choice tasks), where as a result permutation tables will be longer as there would be more combinations between elements. Conversely, compression-based ones can only be evaluated relative to experiment-specific input sequence length. In conclusion, compression and entropy-based techniques can be a powerful instrument to measure data quality and discard disingenuous results.

9.B Selection pattern and metric scores sample of excluded participants

Table 9.7: Binary-encoded selection pattern data of the last 148 choices for a sample of excluded participants. The green shading indicates a confidence rating for a selection being a genuine response, while red – aberrant. Due to extreme outliers, the second-worst score was taken as the lowest boundary in coloring aberrant responses to provide a better disambiguation. Similarly, for the ENT1R metric the second highest and second lowest were taken.

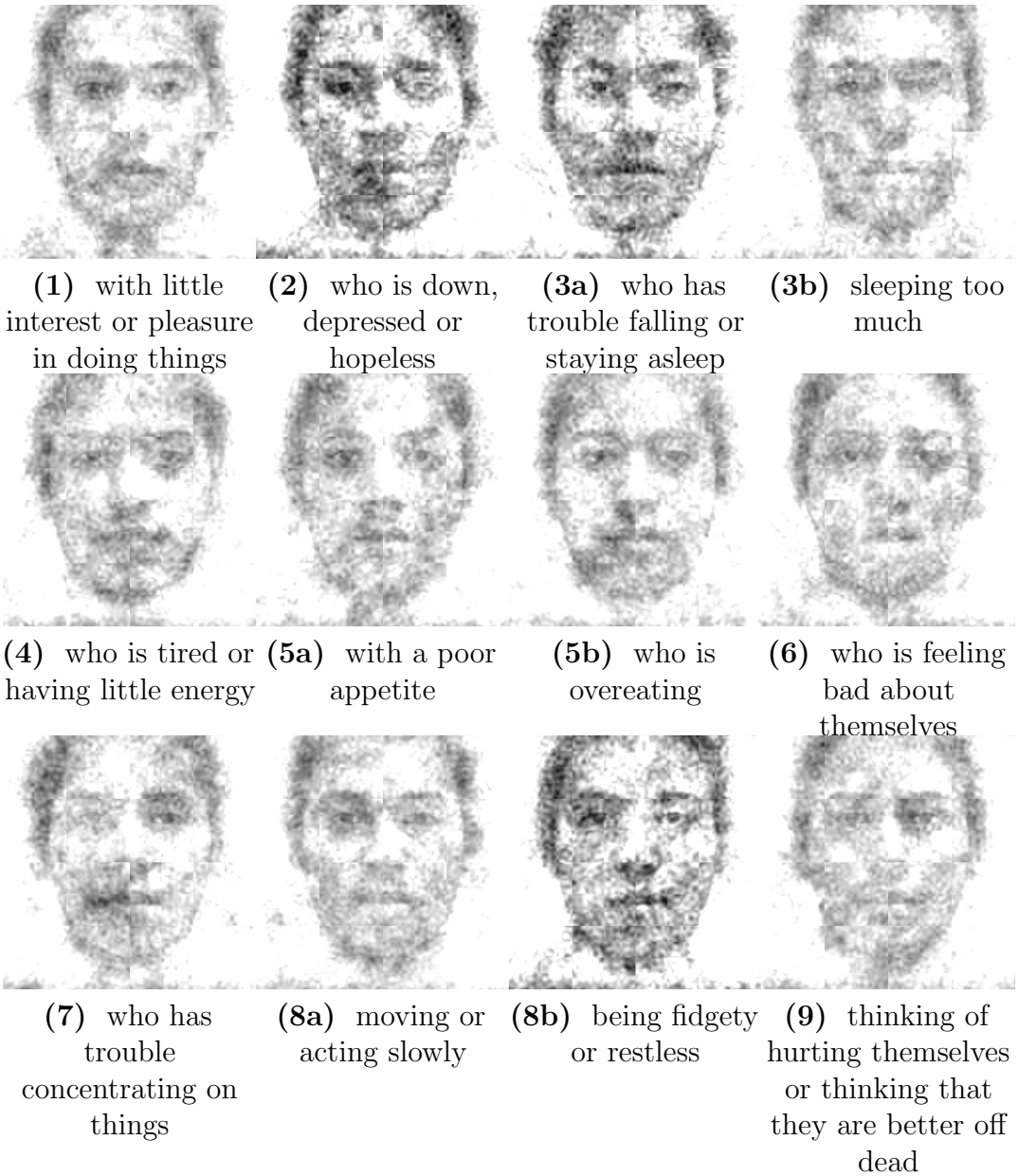
id	pattern	DFL	RLE	ENT1	ENT2	ENT4	ENTR
42	000001011001000000000000000000010000001001110100100110000001100101100011000 1 0011011000001110100100100000000100000000000010000010000010000000100001000000	0.212	0.336	0.814	0.808	0.788	0.748
307	000000000100101001000001100010001001011100111110000000100100000000000000000 00101000110000001100011110000101100110000000001011000010000000100000100100 0	0.2	0.31	0.738	0.737	0.714	0.792
342	01010010100101001010010100101001010010100101001011001010001001010010100101 00101001010010100101001010010100101001010010100101001010010100101001010010	0.066	0.73	0.983	0.766	0.668	0.576
500	00000000000000000000000000000110100100011111111000000000000000001001111110 10 00011111111111001111111000000000000000000001000000011111111111100000011110111	0.188	0.18	0.982	0.792	0.703	0.578
511	00001111111111111111111111111111111110000000000000000000000000000000000000000 00011111111111000000000000000000000000000000011111111100000000000000000 0	0.146	0.102	0.9	0.707	0.553	0.684
562	101101010101010101010101110101000101011110001111111010111101111101010111110 1 0101010101010101110110101010111010111101010101010100011100100111110110101 0	0.152	0.824	0.986	0.767	0.619	0.43
563	11111100011100010000000000000000000100000100000000110000000000000000101000001 0 0101110111100010010000100001000000000000000000000010000000100010000100000000	0.198	0.344	0.785	0.784	0.742	0.766
565	00011111111100000000111111111111000000000000001111111000000000000001111111110000001 1 1110000001111111000000000000000111111000000001111100000000000000000011111111110000	0.15	0.126	0.98	0.727	0.611	0.584
569	0000000000000000000000000000000000000000000000111111111111111111111111111 11111111111111111111111111111111111111111111111111111111111111111111111111	0.06	0.072	0.869	0.603	0.369	0.29
575	00000000000000000001100001000000001000000000000000000000000000000000000000111 000000000000000000000000000000000000000000000000001111101101111111111110111	0.134	0.104	0.482	0.443	0.402	0.896
583	0101010011001011001011010010100110101101001010101100100100101010101010 1001 0101010101100101010100110101010101001101010101001010101100110101001010101 1	0.186	0.786	1.0	0.776	0.686	0.5
597	0011111111111111111111111111111111111111111111111111111111111111111111111111 11111111111100001111111111111111111111111111111111111111111111111111101111111	0.17	0.138	0.718	0.646	0.543	0.198
623	00000000000000000000000000000000000000000000000000000000000000000000000000 00000000000000000000000000000000000000000000000000000000010111111111111111	0.036	0.008	0.194	0.111	0.058	0.97
624	0000000000000000000000000000000111111111111111000000000000000000000000000000000 0000000000000000000000000000000000000000000000000000000000000000000000000000	0.076	0.03	0.488	0.324	0.233	0.894
629	1010101010101010100101010100101001010101000101010110100101001010101010101010 1100101011101111010101010101010101110101010101011010011111010101010101010101 01	0.168	0.856	0.999	0.796	0.662	0.514
636	1111110101010101011110101101010101010101010101000101111101010101001010101 0 10100011001011101101101101001010101101010101010101010101101010101010101010	0.168	0.84	0.997	0.82	0.684	0.47
637	11111111111111111111111111111111111111000000000000000000111111111111111110000000000 000000000000000000000000000000000000000000000000000000000000000000000000000	0.126	0.13	0.98	0.775	0.547	0.584

9.C Selection pattern and metric scores sample of a sample of included participants

Table 9.8: Binary-encoded selection pattern data of the last 148 choices for a sample of included participants. The green shading indicates a confidence rating for a selection being a genuine response, while red – aberrant. Due to extreme outliers, the second-worst score was taken as the lowest boundary in coloring aberrant responses to provide a better disambiguation. Similarly, for the ENT1R metric the second highest and second lowest were taken.

id	pattern	DFL	RLE	ENT1	ENT2	ENT4	ENTR
264	00001000100110111100111101000001101111100111111011101100110111111001001100 01111111100100101111111111110110110011100111110011000111001101111001100 1010	0.234	0.39	0.946	0.933	0.915	0.364
265	11001001101110010011110001100111000111001100000011011100110010001110100110 00111001001100001011100110111010010111011001100001101101101101100101111111	0.226	0.45	0.999	0.996	0.937	0.48
266	00011111011000000111111111111000010101111110111111011101111011111100100111 11110000100000110011101110111100111010010011101101110111111011110111010011	0.242	0.468	0.996	0.99	0.975	0.462
267	00100011111110000111111110001011110111001111011000010000010010101101100000 00110010011001000000000101001010111001000111100010001100110011010111000 1110	0.232	0.442	0.996	0.995	0.96	0.538
268	00001011001011001111111000110011110101010000000101111110001100100100110100 1 10110010011110011000101100011000101101100110001001110100111010111011001110	0.242	0.494	1.0	0.998	0.977	0.498
269	10111001010101110100101100001110101011000101011111011101100111010110111010 00100010100010110110111001000101111000110100110010101111000100000000100100	0.24	0.498	0.997	0.997	0.97	0.468
270	01011111111001010011110001101000000101000110111000100010110010010101001010 1 11011100101010100100010111110010101101110100011001100011010101001010001010	0.236	0.574	1.0	0.98	0.964	0.51
271	11001100001001100111011110101101000000101110111100010011001101111101000011 1 00101110101011001110101111000001010110001101111101101011100100001011001011	0.246	0.52	1.0	0.999	0.99	0.498
272	11110000100111100011010100010101100000000110111111111010001000111110010110 1 00111001100001011100010001101101001101000011111000001101011101101010011001	0.252	0.514	0.999	0.995	0.987	0.48
273	10011110010110000000001000101111100010001001100101100111110100110011001111 0 11001011111111111110000100111111111101100011100110111101110111011010111111 1	0.242	0.434	0.992	0.987	0.959	0.448
274	10000101111100111101000100000000000001100000010101001000010001000000101011 10 11000111100011100111001011101110011110000011010101111110011111110000001101	0.232	0.468	0.993	0.986	0.97	0.45
275	11000010111011111010011000110101011011111011111100010110100101011101111101 10101111011000110111011111110100011010101101100111100011000011011111011001	0.244	0.538	0.979	0.971	0.956	0.414
276	00010110010110101001011100111100101111010110101111111001110011010011011001 00101001110101110010011001101001011001110100100111001110011111101111011011	0.238	0.538	0.995	0.985	0.959	0.46
277	11111011011111110000111110111111000011111110011101111101001110111100111100 11100001111111111110111011111111001111010111100011000111111101100001111011 11	0.21	0.278	0.951	0.88	0.841	0.37
278	10011110010101101111111111110100110101111110001111101001011010010111111110 11 01111000011010011111000011111011111101101111001011010101011101101000101101 11	0.232	0.512	0.98	0.977	0.961	0.416
279	10100000111000000101000100111000010010110001110101001100011001111010100010 10110001000000000101100000110000010001101100010011000000011111001001000101	0.24	0.51	0.993	0.993	0.962	0.548
280	01010000000001100001100001111101010010010001010010101011001100111111110000 0 01011101000111010010101111110011101010100110010011000001011110111000101100	0.242	0.484	0.994	0.994	0.974	0.546
281	01110011000000101110001000010110101111110000101111111111111110011111111010 1 00100100110001011110100100000010110100011001001011101000010000110000010110	0.224	0.39	0.971	0.949	0.918	0.4
282	10110101001100111001110110000011101110110101100110111011110001010010100011 10011111111110111110010110100010110110110110000111111101011111011110100111 0	0.244	0.512	1.0	0.998	0.984	0.512
283	00001111111011011110001100011100011001100101001100100001011011111001011111 1 00000011100111000000000011100100111111000111110111011101000111101000101111 1	0.242	0.388	0.998	0.969	0.961	0.472
284	00110001110101011010001101000000101000100101010010000111001100000010010000 00 10001001110110100000111100001101011000011000000110001100011100000111110110	0.238	0.49	0.98	0.978	0.955	0.584
285	11111001100110000000100101000101010111011101111011111100010000101100001001 1 11011010110011110101011011000100001000111100001011010110011010001100100110	0.236	0.516	1.0	0.988	0.971	0.492
286	00011100100111111111110000000111110000000101111001110010100000101110011000 1 11111110001011010101110111111100100010000000001011011001011100010010011011	0.238	0.416	0.996	0.988	0.967	0.462
287	01010111000010110110111101010101001101111110111011101011001111010010111010 1 01101000001110000000000110111011010101001111010001011100100000001101111110 0	0.242	0.502	1.0	0.999	0.983	0.51
288	10011100011111101000101110010110001011100000110001001101011010000000111001 01011111100111010110000101110000011101101001001101011010011011011110001011	0.252	0.522	1.0	0.998	0.984	0.496
289	01001000011110100010110110011101001000101100010111001101011011110001011101 10111110100100110010111011100010101111011011100011011110101110000101011111	0.24	0.504	0.999	0.994	0.972	0.478

9.D Symptoms of depression subtracted from the base-image



(1) with little interest or pleasure in doing things

(2) who is down, depressed or hopeless

(3a) who has trouble falling or staying asleep

(3b) sleeping too much

(4) who is tired or having little energy

(5a) with a poor appetite

(5b) who is overeating

(6) who is feeling bad about themselves

(7) who has trouble concentrating on things

(8a) moving or acting slowly

(8b) being fidgety or restless

(9) thinking of hurting themselves or thinking that they are better off dead

Figure 9.7: Symptoms of depression subtracted from the base-image. The labels indicate which exact criterion the participants used for their selection: "Which face resembles more someone –":

9.E Facial representations of symptoms of depression for participants that scored 7 or more on the PHQ-9 questionnaire

(1) with little interest or pleasure in doing things

(2) who is down, depressed or hopeless

(3a) who has trouble falling or staying asleep

(3b) sleeping too much

(4) who is tired or having little energy

(5a) with a poor appetite

(5b) who is overeating

(6) who is feeling bad about themselves

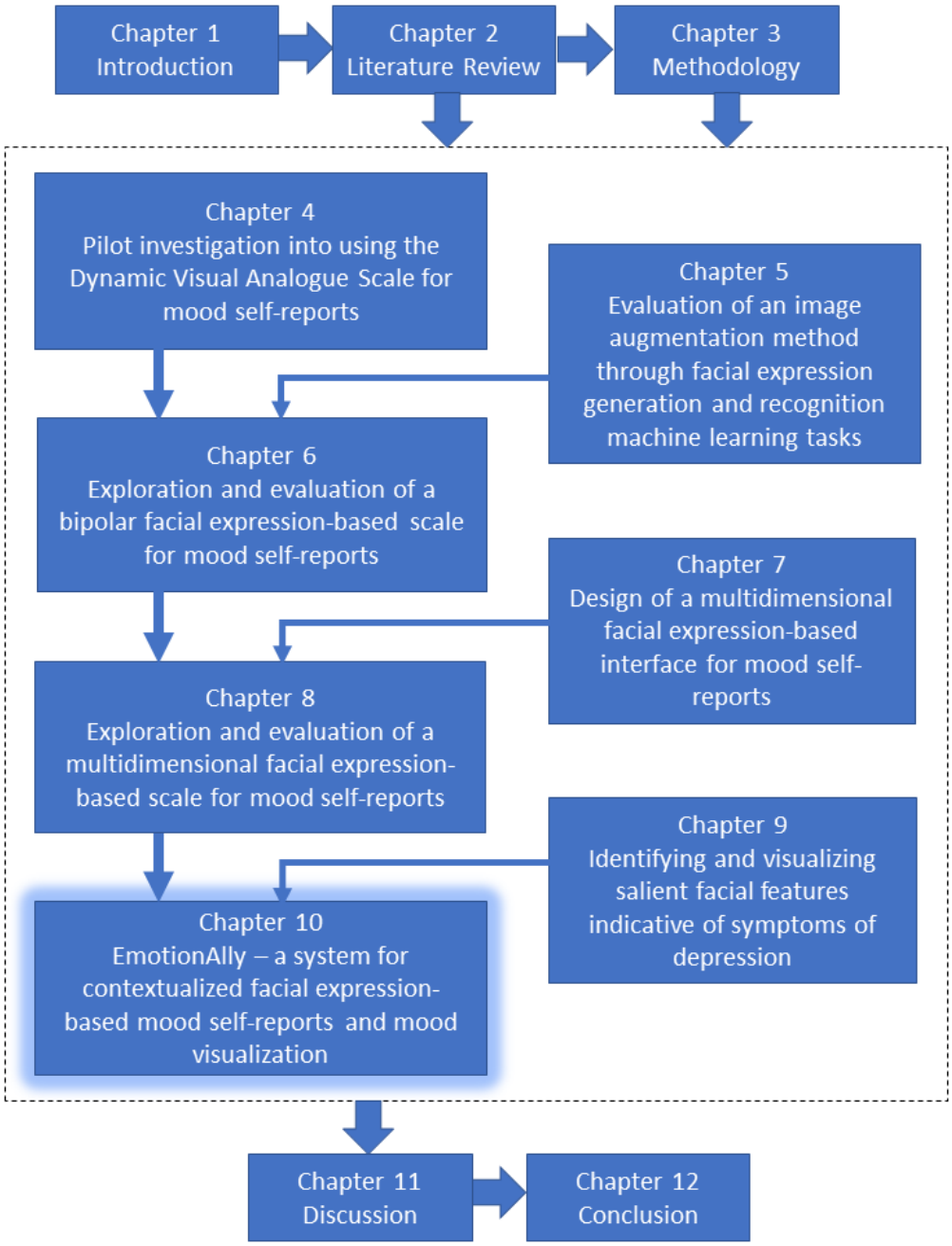
(7) who has trouble concentrating on things

(8a) moving or acting slowly

(8b) being fidgety or restless

(9) thinking of hurting themselves or thinking that they are better off dead

Figure 9.8: Facial representations of symptoms of depression for participants that scored 7 or more on the PHQ-9 questionnaire. The labels indicate which exact criterion the participants used for their selection: "Which face resembles more someone –". The sample sizes for those visualisations are: **(1)**:N(11), **(2)**:N(18), **(3a)**:N(9), **(3b)**:N(19), **(4)**:N(14), **(5a)**:N(14), **(5b)**:N(8), **(6)**:N(10), **(7)**:N(8), **(8a)**:N(11), **(8b)**:N(11), **(9)**:N(13).


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 10

EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualisation

10.1 Introduction

This chapters describes EmotionAlly – a smartphone application for self-reporting and visualising mood using facial expression-based technologies in a context-aware manner. In Chapters 4, 6 and 8 three mood self-report systems were evaluated in a variety of experimental contexts consisting of assessing complex emotions using vignettes (Chapter 4), in the wild for self-tracking mood (Chapter 6) and in the lab using photographic material to elicit emotions (Chapter 8). Therein, prototypes in Chapters 4 and 6 investigated a bipolar happiness and sadness scale, while Chapter 8 investigated a multidimensional facial expression-based scale. Additionally, the prototype in Chapter 4 used real photographs of persons, while those in Chapters 6 and 8 used computer-generated facial expressions.

In this chapter, details guiding the design and development of features for EmotionAlly were elaborated. The design was, in-part, utilizing user-centered design principles [228] and specifically – user stories. As in previous prototypes, the aim was to create an application that can be used ubiquitously on mobile devices, that is self-contained, and does not rely on online services or information.

First, user stories were identified from feedback provided on previous iterations of the assessment tool, compiled as a list of actionable features, and incorporated within the application. Although users were not involved throughout the iterative process of co-creating and co-designing EmotionAlly, they were involved in its conceptualization, i.e. by providing qualitative feedback as part of questionnaires (Chapters 4 and 6) or interviews (Chapter 8) during the evaluation of previous prototypes. Thereafter, the strengths and weakness of the prototype will be explored

as well as alternative configurations of various elements in relation to their effect on the use of the application and their informative value.

10.2 Feature identification

The design of EmotionAlly was informed by findings from qualitative user-provided feedback conducted in studies conducted in Chapters 4, 6 and 8, provided as comments to questionnaires in Chapters 4 and 6 or during the semi-structured interview conducted in Chapter 8. Excerpts were extracted from the combined user feedback, summarized into user stories and translated into actionable features.

Table 10.1: Identified user stories from qualitative feedback provided in user-studies conducted in Chapters 4, 6 and 8 used to aid the design of EmotionAlly. Each user story begins with *I would like to...* followed by a short feature description assuming the perspective of what the user wants to accomplish. Each feature description begins with *Through...* followed by a concrete description of the user-requested feature and subsequently (following the arrow →) a reformulation tying into the original user-provided feedback. The feedback was slightly rephrased to fit into this representation.

№	Component	User Story	Feature
1	Self-report interface	I would like to... *"be able to assess my mood using different facial expressions"*	through... using the multidimensional facial expression-based self-report tool (Chapter 8) → multiple expressions can be provided to users to select from.
2	Mood feedback interfaces	I would like to... *"see a visualisation of my mood"* I would like to... *"be able to identify my mood through facial expressions"*	through... using the facial expression-based method for generating facial expressions → a user can be shown a historical aggregate of their mood self-reports.

3	Self-report interface	I would like to... *"in the beginning see a description of what expressions are available in the interface"* I would like to... *"use the interface as it is, but it can be helpful to know initially where the expressions are (on the interface)"*	through... including labels for facial expressions on the interface in the initial assessments → a user will be able to associate quicker regions where distinct facial expressions are and focus on selecting the correct intensity instead. through... including labels for facial expression → it will be possible for a user to arrange expressions on the polar coordinate system according to their preference.
4	Self-report and mood feedback interfaces	I would like to... *"see the history of my emotions....it might be interesting to see how my emotions towards something change"* I would like to... *"see my history throughout the day and I might want to have something that kind of indicates a reason for an emotion"* I would like to... *"know which activity helps to improve my mood"*	through... providing the ability to supplement context information to self-reports → those could be associated and grouped based on that context.
5	Application	I would like to... *"have an application which is more colorful and catchier"* I would like to... *"see some visual flare"*	through... improving the appearance of the application → careful attention to visual interface design.

Table 10.1 contains a list of the user stories as well as their mapping into concrete and actionable features that were subsequently implemented as part of EmotionAlly.

10.3 EmotionAlly

EmotionAlly is an application developed for the Android Platform. It includes a mood self-report interface as conceptualized and designed in the system design for the multidimensional facial expression-based assessment tool in Chapter 7. An improvement to its design is the ability for a user to provide contextual information enhancing the informativeness of self-reports. Additionally, EmotionAlly also allows a user to see a visualisation of their mood as a facial expression by aggregating multiple self-reports. Furthermore, this visualisation also allows a user to filter historical mood feedback by time or contextual information provided during providing a self-report. The resulting average mood for that context or for that time interval is then visualized utilizing facial expressions. The design and details of those functionalities will be elaborated on below.

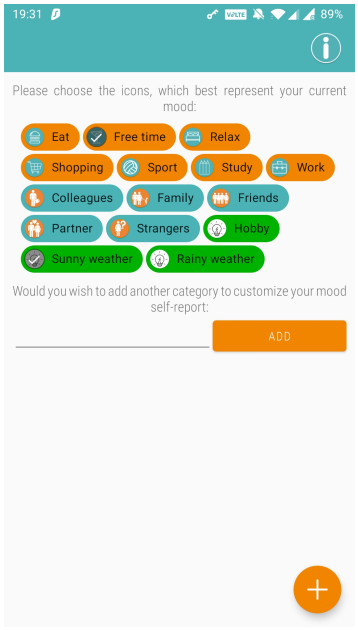
10.3.1 Mood self-report interface

The basics of the mood self-report interface The self-report interface in EmotionAlly uses the multidimensional facial expression-based interface described in Chapter 8 as it allows the tool to portray multiple facial expressions at once. Figure 10.1a shows two screenshots of the mood assessment interface. There, a user can navigate between different emotions or emotion-pairs arranged on a polar coordinate system. It features two elements: a) an image containing the facial expression feedback on top displaying an expression corresponding to the currently selected coordinate and b) a polar coordinate system on the bottom where a user may navigate between different coordinates corresponding to distinct expressions. The coordinate system itself consists of five elements: 1) a point indicating the currently selected coordinate, 2) pivot radii corresponding to the number of distinct facial expressions available in the interface over which emotions at increasing intensities are located 3) labels for distinct emotions, 4) the outline of the coordinate system and 5) the numerical translation from the selected coordinate to emotion intensities displayed in the top left corner.

The interface initially presents elements 1)-4) to the user for a first few initial assessments. This is done in order to help users identify which emotions are available in the interface and quickly familiarize themselves with the application. Thereafter, elements gradually disappear after a number of self-reports are provided with only the point indicative of the currently selected coordinate remaining. This is done as participants indicated that they preferred a blank screen which featured no indication of the underlying coordinate system which allowed them to explore and discover the different facial expressions.

Adding context to self-reports Adhering to user provided feedback outlined in Section 10.2, a user may add contextual information to augment a self-report. This is accomplished by using a second interface portrayed in Figure 10.1b, which is presented after a user provides a self-report using the interface in Figure 10.1a.

(a) Mood assessment interface. It features two elements – a) facial expression feedback (top) displaying a facial expression located for the currently selected coordinate and b) the coordinate system, where for b) the following elements are visible: 1) the currently selected coordinate (purple) 2) pivot radii where discrete emotions are located and 3) their labels (green) and 4) the translation of the selected coordinate into an emotion intensity quantification (teal).



(b) Screen containing contextual tags to augment mood self-reports. Contextual tags can be of 3 types: (in orange) physical activity tags, (in teal) social activity tags and (in green) personalized tags created by the user themselves. Each assessment may accommodate an arbitrary number of tags which allow to attach particular activities to mood self-reports.

Figure 10.1: Mood self-report interface.

This context is provided by selecting between multiple predefined tags, where those consist of 3 colour-coded categories: 1) physical activity (in orange), 2) social activity (in teal) and 3) user-created tags (in green). Naturally, physical activity tags describe active experiences such as working, eating, doing sports and others, while social activity – the social environment that may have influenced the assessment, such as interactions with family members, partner, friends and colleagues. User-defined tags can be of arbitrary nature, pertinent to what the user is interested in tracking. There is no limit to how many tags can be assigned to each self-report.

10.3.2 Mood visualisation interface

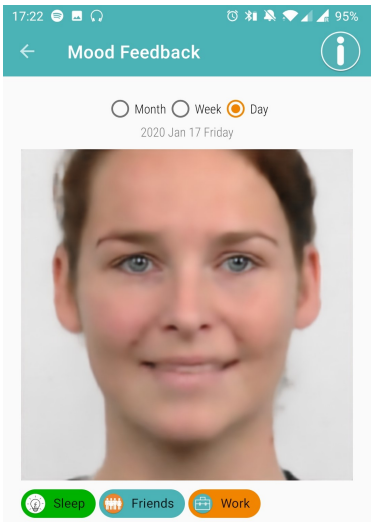


Figure 10.2: Historical facial expression-based mood feedback.

The basics of the mood visualisation interface Figure 10.2 presents an image of a facial expression descriptive of a number of self-reports provided with the self-report interface. This component utilizes the generative neural network model described in Chapter 5 and specifically, the ARM-compatible (e.g. suitable for mobile device processor architectures) TensorFlow-lite version, evaluated in Chapter 7. Using the model in this interface allows to combine and visualize an arbitrary number of assessments as a single image. First, self-reports are polled from the applications' database and subsequently aggregated and averaged as a single combined assessment. In the aggregation step assessments are summed and subsequently averaged for each emotion dimension, such that the their results consist of a single number representing their average from all selected assessments. Thereafter, those values are used as input to the neural network model, which generates the facial expression feedback. Therein, as established in Chapter 7 the generated image requires approximately 600ms, which is an adequate response time for this visualisation as in essence this visualisation needs to only provide one inference per selected filtering criteria (e.g. contextual tags, or time range). It is important to note that since the mood self-report interface only allows a user to select between a mixture of up to two emotions, the mood self-report visualisation may feature a combination of arbitrary facial expressions and intensities.

Filtering historical mood by context or time interval Historical mood self-reports can be filtered by time interval (e.g. day, week or month), by contextual tags, or using both – a time interval and contextual tags. By default, mood history for the most recent day is visualized along with all provided contextual tags displayed below the image. If a user selects one or more contextual tags or a different time interval, self-reports matching the new criteria are fetched and aggregated and subsequently the steps outlined previously are executed, resulting in the model computing a new facial expression visualisation according to the those new criteria. Using those filters allows a user to visualize historical mood for a particular period in the past or according to an activity and visualize its effects on their mood.

10.3.3 Customization options

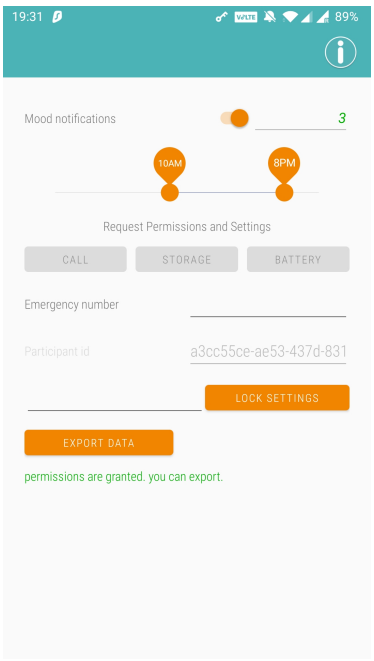


Figure 10.3: EmotionAlly settings page.

EmotionAlly allows a user to control and customize its functionality by defining the number of prompts they will receive during the day, delivered as notifications on the user's device. Additionally, a user can set a time-interval for when notifications are allowed to be delivered or even turn off notifications altogether. For example a user may want to avoid receiving a prompt during active during active hours of the day. When enabled, EmotionAlly spreads the number of notifications over the allowed time-range at semi-fixed equidistant intervals, modulated by a small randomization jitter. Figure 10.3 displays a screenshot of the settings interface.

In addition, a user can export all data collected by the application by clicking the *export data* button. A user may also provide an emergency number which can be dialed from the application itself. Finally, the settings screen can be locked (i.e. all interactive elements are visible, but cannot be interacted with) using a alphanumeric key embedded in the source code. This is particularly useful when is it not desired for a user to be able to control the application settings (e.g. if a user is participating in a study). It is important to note that although this key could be retrieved and does not yield perfect security, this is not a trivial task and requires some technical know-how.

10.4 Discussion

10.4.1 Mood self-report interface

Labels denoting the emotion dimensions on the self-report interface Guided by participant feedback, the facial expression-based interface for providing self-reports features labels denoting emotion categories as well as a visualisation of the underlying coordinate system for a few initial assessments. The study conducted in Chapter 8 utilized the same facial expression-based assessment interface, however, it featured neither labels denoting the expressions available on the interface nor any indication of the underlying polar coordinate system organizing those expressions and their intensities within sectors. As identified in Chapter 8, a labelled visual analogue scale (VAS) was better at capturing the prevalent emotion in the stimulus material. This was an indication that the presence of labels could initially help users to identify the full content in the interface. Additionally, prior work established that categorical identification of an emotion in a face appears to co-occur with

estimating its intensity [195]. This is relevant, since in the multidimensional facial expression-based assessment interface, those processes occur simultaneously as well. Taking this into account, an indication or knowledge of the underlying emotional categories in a facial expression-based interface may allow users to quickly identify the emotional content categorically and use the navigation to precisely select a suitable intensity for their assessment. By large, most participants that used the multidimensional facial expression-based interface evaluated in Chapter 8 preferred the absence of labels, while some indicated that they would prefer their inclusion for a few initial assessments. Therefore, the presence of the labels was implemented to appear only during the first few assessments – when the user is getting acquainted with the application – to accommodate the wishes of both types of users.

Alternative interface configurations At present the interface arranges facial expressions in a fixed-order within the self-report interface. This is done due to limited computing power available on mobile devices preventing the direct use of the neural network model. To circumvent those limitations, output from the model is pre-generated according to those fixed locations for emotion categories on the polar coordinate system. Both of those arrangement were elaborated on in greater detail in Chapter 7. However, some participants indicated a desire to make a selection of a combination of emotions not available on the interface. A variation of this interface could be designed, which instead of using a polar coordinate system allows a user to first select two emotions and then uses a semi- or quarter-circle to display a blend only between those two emotions. This would effectively make the assessment interface customizable, such that any pairing between any two expressions is possible. The implications of such a design choice would result in higher storage costs for this application on the device.

10.4.2 Mood feedback interface

Qualitative feedback collected by participants in studies described in Chapters 6 and 8 indicated that the facial expression feedback appears to be more engaging and evocative. Thus, the same computer-generated facial expressions were used to visualize a collection of mood self-reports. However, visualising average mood as single image based on the aggregation of multiple self-reports inherently has certain implications.

Aggregating mood When aggregated, self-reports that feature a single prevalent emotion will be reflected accordingly in their respective visualisation. However, when self-reports feature assessments on multiple prevalent emotion dimensions, where their respective facial expressions are morphologically inverse or dissimilar (e.g. the shape of the mouth for the expressions of happiness and sadness), the resulting visualisation those features may 'cancel out' portraying a more neutral-looking or averaged feature. While that may be exactly what a participant would

expect, it is inherently prone to obscure information and its correctness is dependent on whether a user views particular emotion dimensions as opposites of one another or as independent dimensions. When visualising emotions through facial expressions, which are morphologically similar (e.g. fear and surprise) facial features will be represented as an accurate blend between both or alternatively, in the case when two expressions affect a different set of facial features, those will also be reflected as a blended combination of both.

This differential outcome in visualising a combination of expressions may lead to an inconsistency in the expected output which is also not evident or made obvious to a user. It is important to note that those particularities are as a result of the employed approach relying on a generative neural network to generate those expressions. Alternative underlying technologies applied to accomplish this task may or may not necessarily behave in a similar way.

Alternative visualisations Alternative ways to visualize historical moods may be accomplished through the use of animations. Therein, instead of aggregating a selection of self-reports as a single image, the mood feedback may be realized as an animation cycling between expressions aggregated on shorter time intervals. This would allow each individual assessments to be accurately represented in the feedback as, for example, a mood summary of the day. However, visualising longer spans of time would still be impractical, where in that case interpolating between an aggregation of assessments within a single day or month may be more suitable when, for example, representing weekly or yearly summaries respectively. Technically, using the method for generating faces from Chapter 5 is ideal as it allows to easily interpolate between any two points in the encoded facial expression space and allows to create smooth animations cycling from one expression or aggregation of assessments to another. The animation will also be smooth as practically an infinite number of points can be sampled from each respective subspace such that a suitable transition speed between individual frames can be identified. A benefit of this mood visualisation approach is that it provides more detail into the mood of a user over a span of time and allows them to peek beneath a simple average represented as a single-image. Subsequently, such a visualisation may also allow a person to identify a periodicity or trend in their long-term mood which may be indicative or attributable to, for example, an affective disorder or a period marked by a significant life event.

10.4.3 Modular system design

EmotionAlly is designed to be modular, such that common functionality (e.g. input methods or visualisations) can be inherited. This means that further modules can be added in addition to ones already available in a way that can borrow from the existing code-base. For example, applying this principle to contextual tags augmenting mood self-reports, those can be easily extended to feature other types of categories besides the physical and social activity ones. This allows to potentially

include further instruments and subsequently allow a user to choose between their preferred configurations. In a similar manner, enriching self-reports through passive sensing can be accomplished as well. Prior art suggests that activity tracking is a potent source of information for assessing a persons' mental-health state [229, 230]. For example, physical activity has been successfully used to estimate depression severity as it is part of depressions' symptomatology [231]. In the context of smartphones or other wearable devices such as smartwatches, activity tracking can be used to unobtrusively detect periods of physical activity and its intensity [67, 232, 233]. Alternatively, passive sensing monitoring application and overall smartphone usage could also be used to supplement contextual information to mood self-reports, i.e. suggest applications which have been used for non-trivial amount of time prior to providing a self-report or a detected physical activity from another connected device. Future work may incorporate such features, integrating passive sensing from smartphone sensors or other devices.

10.4.4 Open vs closed-loop systems

Open and closed-loop systems are two system design paradigms, where a distinctive feature of the latter is the ability of the system to self-correct based on incoming data. Within the digital healthcare domain, data is typically input from the user either through actively providing a self-report or passively obtained from various connected sensors. Therein, the self-correction mechanism may be reflected in the applications' functionality (e.g. reducing the number of self-report prompts) or towards the user in the form of an intervention (e.g. by notifying the user of a detected trend, suggesting an activity or another recommendation). Hence, there are two different types of self-correction mechanisms within closed-loop system which can be used to alter the systems' state: one targeted at the user, and the other – an adjustment of the systems' response towards the user.

EmotionAlly could be extended to include interventions as a closed-loop system. For example, there are various emotion regulation techniques which could be included within the application and served to the user based on provided self-reports or their aggregation. Those techniques could be applied on self-reports on negative or positive emotion dimensions. Echo is one such example of a smartphone applications using emotion regulation techniques to reinforce positive mood [234]. It allows a user to upload photos associated with positive memories and write short descriptions for them, where the applications stores those within a database that can be played back to a user when needed in the future [234]. Alternatively, interventions which aim to help a person manage negative emotions can be used as well. Some well-known emotion regulation techniques are temporal and spatial distancing [235, 236], distraction [237] and positive emotion [238]. Based on user self-reports a suitable emotion regulation technique, or a combination of techniques can be suggested empowering the user to gain more control over their affective state.

10.4.5 Privacy

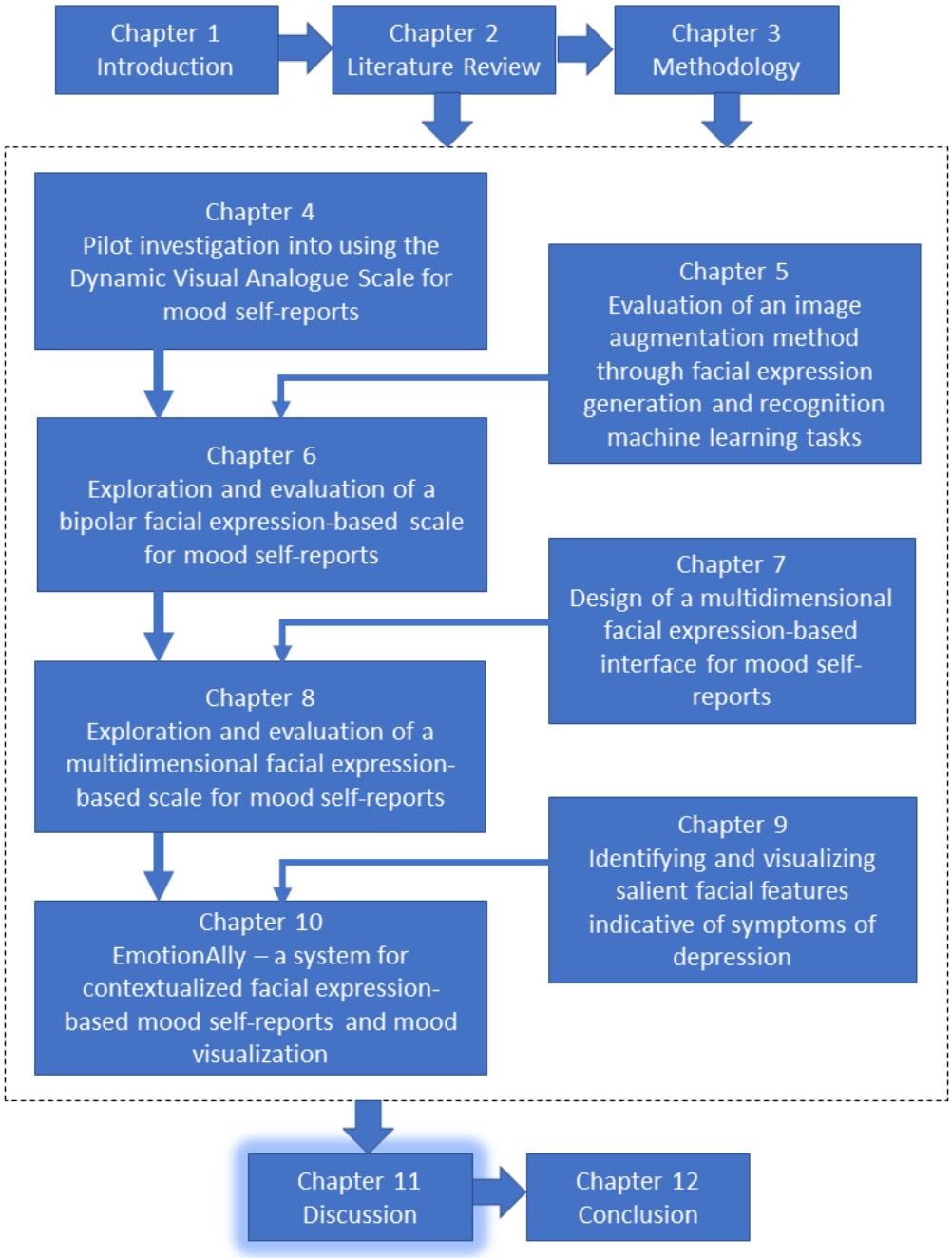
Part of the core-functionality of EmotionAlly is the option to allow a user control over the complete functionality of the application, for example by allowing whether, when and how many notification prompts will be delivered to a user within a day. Additionally, the user is empowered by giving them ownership of their data through the option of exporting the local database in full. Thus, a user can review their own data before sharing it with others or even analyse it themselves. In this manner, data that could potentially be shared is not obscured through broad uninformative descriptions, albeit this is admittedly a less user-friendly way [239]. In addition, EmotionAlly collects and processes data on-device and does not require an internet connection for any of its functionality.

10.5 Conclusion

EmotionAlly is a digital health application for smartphones designed for the Android platform, aiming to assist persons in recording and monitoring their mood and thus gaining awareness of circumstances or events which may have an impact on it. It is a modular-application, which makes extending existing functionality easy and is privacy-aware such that each component in the application can be turned on or off by the user. It was developed as a proof of concept for digital health applications featuring contextualized mood self-reports and historical visualisation using facial expression feedback.

10.6 Chapter summary

In this Chapter EmotionAlly, a digital heath application aimed to help persons self-report, track and subsequently gain awareness of their mood and factors which may contribute to it, was described. The system and its functionality was described in full and contextualized to prior research.


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 11

Thesis Discussion

11.1 Introduction

To start with, it is worth refreshing the reader's memory with the key work completed across Chapters 4 to 10.

In Chapter 4 the Dynamic Visual Analogue Mood Scale (D-VAMS) [27] happiness-sadness facial expression scale, developed as a smartphone application, was contrasted to an equivalent Visual Analogue Scale (VAS) [19] for its ability to capture positive and negative emotions in an experiment with 11 participants using vignettes featuring complex emotions such as awe, pride, disappointment, among others. Subsequently, Chapter 5 presented a data augmentation method for blending faces applied to the Radboud Faces Database (RafD) dataset [8] in a generative and classification machine learning tasks, subsequently improving the quality of generated facial expressions and recognition results. Thereafter, the generative machine learning model was used to create facial expressions portraying emotions at varying intensities, which were used as facial expression feedback in two prototypes, subsequently evaluated in Chapters 6 and 7. Chapter 6 explored a happiness-sadness scale for providing self-reports within a mood-monitoring experiment with 37 participants. Chapter 7 described the design and development of a multidimensional facial expression-based scale (MFEAS) featuring a heuristic to accommodate multiple distinct and blended facial expressions within a 2D tactile surface as an android smartphone application. As before, this prototype also utilized the generative model from Chapter 5 for creating the facial expressions feedback. MFEAS featured the expressions for happiness, sadness, anger, disgust, fear, and neutral and was compared to an equivalent VAS scale in a mood elicitation experiment with 47 participants which rated emotion eliciting images from the International Affective Picture System (IAPS) [10] in Chapter 8. Thereafter, Chapter 9 explored whether and which symptoms of depression possess distinguishable facial features as perceived by healthy adults in an online-conducted experiment with 600 participants. Finally, Chapter 10 presented EmotionAlly, a prototype featuring contextualized facial expression-based

mood self-report and historical mood visualisation functionality. Its development was guided by participant feedback obtained from previous studies conducted in Chapters 4, 6 and 8.

Hereafter, the main findings in the thesis will be discussed in the context of the research questions posed at the beginning of this thesis in Chapter 1. This chapter will highlight the importance, novelty, and limitations of those findings, and provides recommendations for their practical application or recommend future work.

11.2 Reflection on Thesis' Research Questions

This thesis explored four research questions, which are individually addressed and discussed within this section as follows:

Research Question A) How can facial expression-based methods be evaluated as a valid way to self-report mood? How would such tools compare to an established method, such as traditional numerical-based scales?

The studies relevant for this research questions comprised of evaluating quantitatively a photograph-based bipolar happiness-sadness Dynamic Visual Analogue Scale (D-VAMS) in Chapter 4, a bipolar facial expression-based scale (FEAS) for mood self-reports in Chapter 6, and a multidimensional facial expression-based scale (MFEAS) consisting of the expressions for happiness, sadness, fear, anger and disgust in Chapter 8. The latter two relied on a machine learning model for generating the facial expression feedback.

As elaborated in Chapter 2, Section 2.2.4, emotions and mood are two different constructs intertwined, such that they constantly mutually influence one another. Hence, due to the way they are understood, it is difficult to isolate an emotion from the background mood of a person or vice-versa. Typically, for the purpose of self-reporting mood, visual analogue scales (VAS) [19], Likert scales [18], or low-fidelity schematic or drawn faces [21, 33], or photographs [27] have been used. While, the approach employed within this thesis relied on a novel application of computer-generated facial expressions. As facial expressions are an accurate reflection of internal emotional states [25], the use of realistic facial expressions could improve the state of the art and allow for better, customizable and closer to mood aligned representations such as facial expressions to be used. Nevertheless, due to the difficulties in being able to measure emotions or empirically and as ground truth, in order to explore and evaluate whether a novel tool can successfully capture those modalities, the empirical investigation was conducted by contrasting assessments provided with a facial expression-based scale to an established VAS one.

The findings identified strong significant correlations between assessments provided on both happiness-sadness facial expression-based scales and their VAS counterparts apart from the multidimensional facial expression-based scale, for which only the dimensions of sadness, fear, and disgust resulted in strong significant correlations. Supporting those results were the variations in experimental contexts, where the studies described in Chapters 4 and 8 relied on using emotion elicitation material in the form of vignettes and images respectively, while the study described in Chapter 6 required participants to monitor and report their own mood over the course of 2 weeks. Additionally, the studies described in Chapters 4 and 6 employed a between-subjects design where participants used both a facial expression-based scale and VAS. On the other hand, the study described in Chapter 8 employed a between-subject study design where participants used either MFEAS or its VAS equivalent, thus precluding the possibility for carry-over effects influencing the comparison of assessments. Both, the bipolar happiness-sadness facial expression-based scales and MFEAS obscured any numerical feedback, which prevented participants from replicating their assessments across scales. The consistent results obtained in a variety of study designs, prototypes, and emotion elicitation material, and variations in employed scales suggest that the facial expression-based approach can be used for mood self-reports. It needs to be acknowledged that the prototypes described in this study were not developed with the intent to become de-facto scales for self-reporting mood, but rather explore the feasibility of such an approach and generate knowledge on technologies which can be used within the facial expression-based design space for mood self-report instruments.

The main findings pertain to the iterative development of three prototypes and their quantitative evaluation for assessing mood. Over three separate studies, the indication is that facial expressions are a useful and valid modality for self-reporting mood.

Limitations Concerning the evaluation of the bipolar happiness-sadness scales as well as the multidimensional facial expression-based scale, some limitations were acknowledged with respect to the mood elicitation material.

In Chapter 4, positive and negative vignettes have been captured within the emotion dimensions for happiness and sadness respectively. While the positive vignettes have been validated and rated for their elicitation content [167], the negative vignettes were not. Nevertheless, assessments provided through the facial expression-based scale and its VAS equivalent were strongly correlated, where in addition, the stimulus material was captured well within the expected emotion dimensions (e.g. positive vignettes within the happiness dimension and negative ones within the sadness dimension).

Furthermore, in Chapter 8 MFEAS was compared to a VAS equivalent, where a limited number of images rated unambiguously on the categorical emotion dimensions of happiness and anger were available. Coincidentally, the correlations for those two emotion dimensions were also not significant. In contrast, the

dimensions for fear, disgust, and sadness for which the elicitation material featured more images unequivocally rated to elicit those emotions resulted in significant strong correlations between assessments provided on MFEAS and VAS.

Future research directions Affective disorders have a unique fingerprint in how they influence the interpretation of facial expressions. Depression specifically, appears to affect the processing of facial expressions in distinct ways as elaborated in Chapter 2, Section 2.5.2, where the extend of and manner in which those perceptual perturbations are expressed appear to correspond to depression severity (Chapter 2, Section 2.5.3). Those biases are not observed in other representations conveying emotional content, such as schematic or drawn faces, textual descriptions or others [106, 107]. Additionally, since those perceptual biases are expressed as a subtle offset in the ability of a person living with depression to detect a nuanced expression or accurately estimate its intensity, technologies which can create facial expressions with a wide range of intensities at high granularity could capture and quantify those biases. The further development of facial expression-based methods positions those technologies in a unique position to capture those perceptual biases and subsequently those technologies can be used to quantify depression severity. A benefit of using facial expression-based tools for this particular use-case is that depression severity estimation could be computed on top of mere mood self-reports as those biases will be reflected by deviations or drifts over time in the use of intensity ranges for specific emotion dimensions associated with depression's symptomatology course-trajectory (e.g. as a patient goes through mild, moderate, or severe depression or remission). This highlights the potential of facial expression-based tools for clinical applications beyond the scope of them as being an alternative to traditional self-report tools [3].

Research Question B) Which aspects of facial expression-based tools are valuable to users and which further capabilities are desired?

The studies evaluating facial expression-based prototypes quantitatively also explored those qualitatively. In Chapter 4, Dynamic Visual Analogue Scale (D-VAMS) [27], a known facial expression scale was evaluated by collecting user feedback by 11 participants on an administered user experience questionnaire (Appendix 4.A). Subsequently, in Chapter 6 a facial expression-based scale (FEAS) using computer-generated facial expressions was evaluated used a structured questionnaire allowing 37 participants to rate the method of using facial expressions for mood self-reports and the practical implementation of the prototype on a number of characteristics such as performance, ease of use, preference, among others. They could also provide additional commentary in free form to a few open ended questions as well (Appendix 6.A). In Chapter 8, a multidimensional facial expression-based scale (MFEAS) was evaluated using a semi-structured interview conducted with 47 participants (Appendix 8.A). The results were transcribed,

aggregated, and analysed aiding in the identification of features important for the future development and improvement of facial expression-based mood self-report tools.

At present, technologies used to generate human or human-like faces and expressions are either of low-fidelity, or high-fidelity but offer little in terms of customisation, or are not easy to use without specialized knowledge [121, 240]. In the context of using facial expression-based tools for self-reporting mood, assessing the users' preference regarding the appearance of the model, its fidelity and other qualitative characteristics is of important for such tools to be more easily accepted and adopted.

Findings suggest that participants were able to successfully use the facial expression feedback to provide assessments through each of the developed prototypes. A key finding, is that most participants preferred the facial expression-based feedback over the visual analogue scale for self-reporting mood. Generally, participants were also more engaged in their feedback towards the facial expression-based prototypes in contrast to VAS. This finding may partly be attributed to their ability to emphasize with the emotions portrayed by a realistic-looking human face. Additionally, this facial expression feedback was reported to be helpful in guiding users to provide and fine-tune their responses by navigating through different expressions and intensities presented on the interfaces. It is known from studies investigating visual analogue scales that real-time feedback facilitates the most accurate self-reports [22], where this same principle was applied within the facial expression-based scale, naturally featuring expressions at different intensities as feedback.

In the context of using expressions to indicate mood, most participants mentioned that the depiction of a real person helped them map their own emotions onto those presented in the prototypes, where a more commonly used strategy was to browse expressions until one was found that matched their mood.

Although, the images of facial expressions were recognized to be computer-generated, participants acknowledged that they were of high-enough fidelity to be considered realistic and did not elicit the uncanny valley effect [122]. Additionally, most participants recognized that the expressions on the interface were enacted, which is in line with the Radboud Faces Database (RafD) dataset [8] used for training the generative machine learning model. However, the fact that the expressions were enacted did not appear to impact how the prototypes were used. The indication for such tools was that a degree of realism is preferred such that those images provide realistic portrayals of facial expressions, although they do not necessarily need to conform to how expressions are enacted and used in daily-life.

Additionally, participants were split on their preference for the model's appearance between an anonymous identity and one that resembles their own. Therein, an anonymous person was seen as a person unknown to them which is used as a canvas to provide assessments. Conversely, persons that preferred to customize the model indicated that they would prefer a model that resembles

them in appearance, where approximately half of those participants wished to see a digital doppelgänger of themselves, while the others half – a lookalike which only approximates their appearance. For both groups – those that preferred an anonymous identity or one that resembles them – most participants indicated that age and gender were the most important characteristics of the model that ideally would match their own. Similar to prior works investigating user preference for 3D avatars, a substantial portion of users preferred avatars to only resemble them in appearance rather than be their mirror image [202]. Hence, summarizing on those two points, technological tools relying on human faces should provide sufficient customisation options to be able to create a lookalike of oneself and posses sufficient fidelity, do not necessarily need to be photo-realistic and do not have to reflect reality in by enacting realistic spontaneous expressions. This hints at potential future developments of such tools, where those may not necessarily need to excel in either of those categories except for being flexible enough to allow users to customize their appearance to a degree where they could identify themselves. Within the context of tools for providing mood self-reports an interesting exploration would be to identify the optimal balance between realism, fidelity and detail in facial expressiveness contrasted to user acceptance and self-report accuracy.

Interestingly, most participants preferred solely the facial expression feedback as an indicator for experienced emotions and emotion intensity, in contrast to complementary information such as text labels. This observation is intrinsically related to the employed navigation strategies, where those may have been sufficiently intuitive in the way facial expressions were organized within the explored prototypes as to not require additional cues elaborating on their content. The use of familiar gestures or interactions such as the vertical sliding gesture and the polar coordinate system appeared to transfer rather well to a previously unseen context such as an facial expression-based interface for mood self-reports. Moreover, those interactions were considered by most participants to be quick to learn and subsequently the prototypes – easy to use.

The main findings pertain to the iterative development of three prototypes and their qualitative evaluation for assessing mood. Over three separate studies, the indication is that facial expressions are a useful modality to represent mood. In general, a facial expression-feedback elicited more through and thoughtful responses, were considered to be more engaging, and elicited more elaborate responses with respect to their future customisation and improvements. Just as importantly, they were preferred to traditional graphical scales for providing self-reports.

Limitations A limitation of how the facial expression-based scales were perceived was expressed as a lack of further emotion dimensions expressed as distinct expressions. While blended expressions were available in MFEAS, those only portrayed blended emotions rather than new emotion classes. Contemporary understanding of facial expressiveness and facial expressions identifies six such ones [25], which have been subsequently focused on in academic work. In this sense,

facial expression-based interfaces may be limited to the range of emotion dimensions they can portray and can be expressed in the face. While technologically, the employed machine learning method could accommodate further expressions, at present dataset containing expressions beyond the universal facial expressions of emotion do not exist.

Future research directions Recent works have identified facial-bodily expressions which apart from facial expressiveness are complemented by variations in head pose and the inclusion of hand gestures [196]. Therein, the authors have identified 28 naturalistic displays of emotion including ones such as pride, compassion, or others. Additionally, those are believed to be mapped onto an emotion space not as distinct entities, but rather as a gradients that cross categorical emotion boundaries, similar to how the distinct adjacent facial expressions in MFEAS blended into one another in the prototype evaluated in Chapter 8. By combining those representations with a method similar to the one used to generate facial expressions at different intensities in Chapter 5, or a similar one, and the heuristic developed to accommodate multiple facial expressions within a two-dimensional polar coordinate system in Chapter 7, may allow to create an equivalent interface containing each of those 28 representations. The identification of those novel facial-bodily representations shows promise, as expanding the assessment space to include expressions beyond the universal facial expressions of emotion would inadvertently enhance the potential and applicability of facial expression-based mood self-reports tools.

Research Question C) Which technologies can generate expressions depicting a range of emotion intensities using images of arbitrary expressions? Could those be used within applications on commodity hardware, such as smartphones?

Historically the development of facial expression-based scales has been accomplished using low-fidelity facial expression representations such as schematic or drawn faces [21, 33]. As elaborated in Chapter 2, Section 2.3.3, those representations are inherently limited in the range of emotion intensities they can portray. Additionally, creating plausible and realistic high-fidelity faces has traditionally been a challenge in the fields of computer-graphics, computer-vision, and machine learning. However, recent advancements in those fields [118, 120, 241] has resulted in the development of technologies which pushed the boundaries of realism in allowing the creation of human-like computer-generated faces.

One of the most important aspects in designing a facial expression interface is accuracy of the mapping to an underlying numerical value for the intensity of emotion in a facial expression. Within this thesis, the relationship between facial expressions, their intensity, and a numerical equivalent has been explored using a generative neural network trained to create facial expressions at different intensities.

Therein, the input parameters for the generative model, used to generate a desired facial expression and its intensity were assumed as ground truth for what those expressions were supposed to represent. The results from those studies indicate that the mechanism through which the network learned to create facial expression intensities from numerical input are reliable in creating a gradient of expression intensities coherent with the true underlying intensity as perceived by participants. This quantification has been explored in two quantitative studies (Chapters 6 and 8) whose results indicate the validity of using computer-generated facial expressions at varying intensities for mood self-assessments. Although similar scales have evaluated the use of photographs of facial expressions [27], a novelty in this approach is the exploration of computer-generated facial expressions and their potential for mood self-report tools. While static photographs are indeed useful, machine learning methods can be further improved in the number of different expressions that they can create, their fidelity and granularity of facial expression intensities [242–244].

The enactment of a facial expression consists of a time-dependent activation sequence of facial-musculature [190, 191, 245]. As this spatio-temporal sequence unfolds, a map is created between the person enacting that expression and the observer. Therein, the observers' attribution of meaning towards the seen expression and its intensity follows the enactment progression of an expression. This quality also appeared to be well-captured within the applied method for generating faces of varying expressions and sufficient in representing a range of their intensities. Additionally, as some participants noticed, the expressions contained within the investigated prototypes consisted of enacted expressions in contrast to spontaneous ones. When facial expressions are computer-generated using an image dataset, intermediary expressions (e.g. those between a neutral expression and one at its peak intensity) consist of interpolated variants where they vary only over the spatial domain. Posed expressions are characterized by higher amplitudes of emotion and subsequently a broader range of intensities, while spontaneous ones – of lower amplitudes and smaller range of intensities. For spontaneous expressions, even more precise methods may be needed to accurately create and reflect them which are sensitive to subtle changes in expressions and can adequately generate them. To achieve this, machine learning methods are a suitable tool they can create a gradient of facial expression intensities with sufficient granularity.

Granularity, within the context of facial expressions-based interfaces, describes the amount of variability within an emotions' intensity range. Therefore, optimal granularity can be defined as the smallest perceptible variation in expressiveness, while the lowest – when an emotion is either present or absent in a face. Specifically, optimal granularity in the intensity variations of facial expressions lies at the border of when a user can barely distinguish a difference between two facial expressions of adjacent intensities. As findings in Chapter 8 indicate, using a generative machine learning model for generating facial expressions does allow for such high granularity in the created emotion intensities. Nevertheless, there are diminishing returns to interfaces which exceed this practical perceptual limit. Within this thesis, participant feedback indicated that the number of intermediary expressions in all

investigated prototypes was sufficient to provide their assessments and allow them to fine-tune their responses.

An important usability criteria of facial expression-based assessment tools for smartphone applications was the ability to provide quick assessments. As this was an anticipated requirement, the developed prototypes were designed using pre-generated expressions from the generative machine learning model, as generating those on the fly was found to be too costly (Chapter 7). Subsequently, as remarked by multiple participants, this approach allowed for the facial expression feedback to be immediate and facilitated a smoother interaction when navigating between expressions or their intensities and this property of the prototypes received a positive reception.

The usability of the evaluated prototypes was affected by how the interaction between a user and the interface is accomplished and through the way facial expressions are organized within an interface. Those entail the navigation scheme employed to allow a user to alter between different expressions or intensities in an easy and intuitive manner. Two different navigation strategies have been employed within the developed prototypes. First, a vertical sliding gesture allowed users to navigate a bipolar scale consisting of two emotion dimensions. Second, a polar coordinate system organized facial expressions and their intensities within areas on a polar coordinate system. Therein, the radius coded for emotion intensity, while angle from the centre separated distinct expressions categorically. Regions between distinct expressions contained blended expressions comprised of both adjacent categorical emotions. Both navigation strategies were unanimously rated as easy to learn and use as they promoted a cohesive arrangement of categorical expressions and intensities.

One of the main challenges, when designing facial expression tools using computer-generated facial expressions is the lack of dataset containing nuanced spontaneous expressions. At present, most consist of categorical depictions of enacted facial expressions [8, 14]. This is predominantly caused by difficulties in correctly labelling expressions beyond the binary representation of an emotion in a face. However, the task of assigning a numerical value to an expression is prone to disagreements between human-raters as outlined in Chapter 2, Section 2.4.2. Hence, the use of generative machine learning methods is restricted to expressions available in the training dataset. Although they are flexible in their ability to generate expressions portraying varying intensities of emotion, those methods cannot yet create novel expressions. An potential application and improvement to those challenges may be the use of augmented images which improve the variability of nuanced expressions in categorical dataset such as the one proposed in Chapter 5. Although in that particular use-case an augmentation was created from a categorical dataset it did yield improvements in generating expressions of intermediary intensities as well as performed on par as a training dataset for models evaluated on recognizing categorical expressions. Therefore, a potential application for the approach could also be in nuanced facial expression dataset. Consequently, facial expression-based interfaces may be able to be build using

spontaneous expressions resembling closer how facial expressions are encountered and used in day-to-day life. It would be interesting to evaluate whether spontaneous expressions, albeit known to contain less variations in expressiveness in their range from mild to peak intensities would achieve similar results as in the evaluation of prototypes herein relying on enacted ones.

The main findings pertaining to this research question relate to the iterative development of four prototypes, a data augmentation method applied to a generative machine learning model for creating highly-granular facial expressions, and a heuristic that maps multiple facial expressions and their intensities numerically onto a coordinate system which allowed users to navigate between them.

Figure 11.1: Renders at various angles of Digital humans created with MetaHumans [246].

Limitations Devices such as smartphones are somewhat limited in computing resources and subsequently also limited in being used to train or use more complex models. As evaluated in Chapter 7, those factors still pose challenges in their application on mobile devices apart from more trivial cases such as object detection or recognition [247]. Although various techniques, such as quantization [189] and pruning [248] exist, their application reduces the image quality and as explored in Chapter 7, quantization yielded marginal improvements to inference latency. While the development of a heuristic for arranging facial expressions within an interface using a predefined schema which in turn also allowed to pre-compute the model's output those come at a the cost of increased storage requirements and rigidity in

not allowing users to choose emotion dimensions relevant to them in their preferred order (Chapter 8). While successful in making the prototypes responsive and easy to use, future improvements in machine learning methods targeted at edge devices would inadvertently also benefit facial expression generation approaches as well.

Future research directions Alternative technologies for creating high-fidelity faces and facial expressions such as virtual 3D avatars could be used within facial expression-based technologies. One such technology is MetaHumans, a modelling tool designed to work with Unreal Engine [246], a framework allowing the creation of real-time animations using full-body 3D avatars, among others. Figure 11.1 portrays a collection of such avatars, which display an impressive level of detail. The practical realization of this technology relies on representing faces and bodies through landmarks in a 3D space, which can be manipulated to create body movements, or facial expressions. Curiously, those landmarks are identical between models created with their framework, which implies that animations created for one 3D avatar could be transferred to another. As a result, facial expression animations created for one model of a particular appearance could be used within another – of a different appearance – resulting in an identical re-enactment of an expression. In contrast to machine learning methods, which are susceptible to within-class differences, e.g. variations in the enactment of facial expressions between different persons, a 3D model-based approach would allow users to create or customize an avatar on their smartphone which, however, would use the same facial expressions as the avatars of others. Therefore, an empirical evaluation of a hypothetical prototype using this technology for self-reports would allow users to alter the appearance of the model while still, in essence, using the same expressions, which in turn could be validated as a facial expression-based scale. Or alternatively, provided the technology is easy to use, users can create their own expressions for different affective states and even share them with one another. Such an approach can capitalize on identifying facial expression-based representations of emotions beyond the universal facial expressions of emotion [25] and could be used in an approach to identify and define novel ones. Therefore, an empirical evaluation of a hypothetical prototype using this technology in a facial expression-based tool self-reports would allow users to alter the appearance of the model while still, in essence, using the same validated expressions. Furthermore, using the full-body 3D models can allow creating bodily representations for various affective states using a combination of facial and bodily cues (e.g. posture, gait) thus enriching the assessment space. For example, it is known that depression has an effect on gait, and posture [249], among others. Those effects could be conveyed only using a 3D model and a facial expression-based only representation is not sufficient.

Research Question D) Are there specific facial expression-based representations descriptive of affective states? If so, could those be used within self-report tools?

It is known that depression appears to affect facial expressiveness and those differences appear to allow differentiating between persons experiences low and severe depression [205]. Additionally, prior work has explored the ability of laypersons to detect specific symptoms, part of depressions' symptomatology, such as suicidal ideation [223] from the face. However, no extensive evaluation has been done on individual symptoms of depression and their effects on facial appearance. This research question aimed to explore whether and which symptoms did possess distinct facial characteristics expressed in the face. To answer this research question an online study was conducted on the Mechanical Turk (mTurk) platform described in Chapter 9. Therein, 12 symptoms were derived from the Patient Health Questionnaire (PHQ-9), a widely-used depression screener questionnaire, and for each of those symptoms an aggregated perceptual model was created comprised of the individual models of 50 people to visualize how each symptom appears in a face. Each persons visualisation was created by using the reverse correlation classification images (CI) method, used in prior works to mental constructs such as perception of ethnicity, social traits, and others [32, 79]. The method works by augmenting an image of a neutral face (e.g. androgynous and with the neutral expression) with patterned noise. Within a Two Alternative Force Choice (2AFC) experiment a user chose one of two noise-augmented from 500 image-pairs and subsequently, due to their choice, features of relevance were reinforced, while inconsequential ones were smoothed out. The results and visualisations obtained in this study are novel as the method has never been applied within the domain of affective disorders to visualize trait-like symptoms of depression.

The results indicated that most symptoms of depression did have discernible features delineating them from the neutral face. In particular, those were the symptoms of *having little interest or pleasure in doing things* (1), *trouble falling or staying asleep* (3a), *sleeping too much* (3b), *being tired or having little energy* (4), *poor appetite* (5a), *overeating* (5b), *being fidgety or restless* (8b), *thinking of hurting oneself or thinking that one is better off dead* (9) were found to possess distinct characteristics which delineated them from the neutral face.

The visual representations for *little interest or pleasure in doing things* (1) (e.g. proxy for anhedonia) were consistent with what is known in the literature where anhedonia is expressed as deriving less joy from positive experiences, and dampened facial expressiveness [215, 216]. The symptoms of *having trouble falling or staying asleep* (3a), *sleeping too much* (3b) both portrayed features affecting primarily the the area around and of the eyes resembling skin discolouration, an effect of dysregulation of sleep [217]. Surprisingly, the symptoms for *being tired or having little energy* (4) and *poor appetite* (5a), both exhibited characteristics akin to sadness. The expectation, for being tired or having little energy was that of a droopy-looking expressions and that for poor appetite to reflect a more bare

face. While different from the neutral face, those symptoms were not found to be expressed through features normally associated with their effects on the face.

The symptom for *being fidgety or restless* (8b) is characterized by a combination of features known to convey the feeling of being worried [221]. The symptom for *thinking of hurting oneself or thinking that one is better off dead* (9) was perhaps the most evocative of all representations. Prior work has established that suicidality can be detected by lay persons from photographs with an accuracy greater than chance [223]. The presence of distinct features for suicidal ideation served as further evidence for the symptom being reflected in the face in addition to providing a concrete visualisation of the appearance of a person experiencing this symptom.

However, not all symptoms yielded an expression with distinct facial features. In this respect, a key outcome was that symptoms which feature a strong cognitive component such as *feeling bad about oneself* (6), *having trouble concentrating on things* (7) and *moving or acting slowly* (8a) [224] did not possess facial features, delineating them from the neutral face. Prior work established that cognitive traits do not appear to have an effect in the development of particular facial features [225], which may also translate to how such symptoms affect facial expressiveness or appearance later in life. Typically, symptoms with a cognitive component are expressed through changes in behaviour, which is difficult to identify within a static image.

With respect to facial expression symmetry, this study found that on average the left side of the face contained on more evocative features, albeit not exclusively so. Additionally, for most visualisations, both halves of the face would sometimes possess stark differences in features and expressiveness from one another. In those observations, our findings appeared to align with prior work which identified the left side of the face to be more expressive and portray emotions more intensely [227]. However, this may also be an artefact caused by the noise augmentation since all participants made their selections using the same 500 image-pairs, where the noise generation may have influenced the left portion of the images stronger than others.

The main findings in this study pertained to identifying and visualising salient features of individual symptoms of depression. The indication is that most symptoms of depression possess characteristics which makes them identifiable in the face. Those results are novel, since no exploratory work has attempted to probe and visualize the perception of laypeople in how depression affects facial appearance on the symptom level.

Limitations Interestingly, as the symptoms for *being down, depressed or hopeless* (2) and *overeating* (5b) did not resemble the expected expression for sadness and a broader face contours respectively. This may be an indication for limitations of the reverse correlation method, where for the former, the symptom was described by three affective states which each may have had its distinct representation. Subsequently, the complexity introduced by including all three terms may have

hindered participants from accurately associating the images presented during the task to the symptom. For the latter, the indication is that the reverse correlation method may not be suitable for altering spatially distributed features such as face contours. While spatially compact features need less noise to result in conveying a different impression, the randomness of how the noise is generated may be unsuitable to consistently reinforce features spread over a broader area in an image. Reinforcing this observation was also the symptom for *poor appetite* (5a) which also did not significantly differ in its face contours from the neutral face.

Additionally, the symptoms for *being tired or having little energy* (4) and *poor appetite* (5a) were represented by facial features resembling the expression for sadness. While not completely unexpected, participants may have been influenced in their choice by their social understanding of stereotypes of depression, where this attribution of socially imbued characteristics may have altered or shifted an underlying 'true' representation.

Future research directions While creating visual representations for individual symptoms of depression yielded distinct facial features for most symptoms, further validation work may be required to establish whether inversely, those can be matched to their symptom descriptions. This might reveal whether and to what extend those visualisations are representative of the popular perception of depression's symptomatology and how reliable those are in identifying a symptom in a face. For example, as previous investigation found the presence of suicidality to be discernible in the face [223], this approach may further confirm those findings.

The primary objective within the context of this thesis is to use those representations as visual cues for self-reporting on one's symptoms of depression. This can be accomplished using tools similar to those explored in Chapters 4, 6 and 8 in this thesis. Naturally, this poses some challenges, particularly so when combined with a machine learning-based approach to potentially create a gradient of intensities associated to a symptom severity. As the images created with the reverse correlation are inherently noisy, this may severely hamper the ability of a neural network to adequately learn the underlying facial features as those models are known to incur significant penalties when dealing with noisy data.

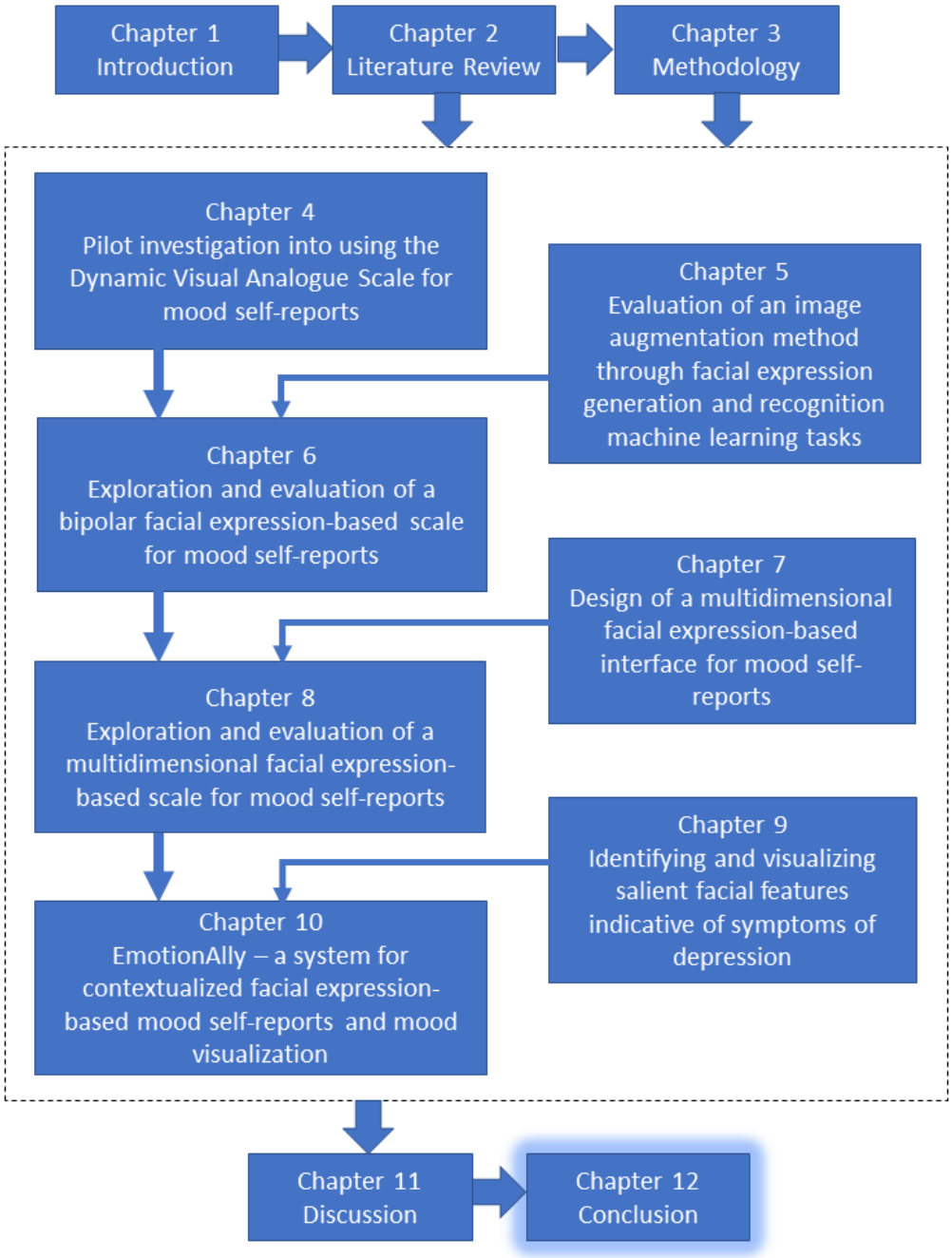
Another related practical implication of these findings may lie within the facial expression recognition domain, where those visualisations could be used as training data for facial expression recognition machine learning models to recognize the degree to which symptoms of depression are present in a face. Besides the difficulties of such models to handle noisy data, such a use-case also poses significant risks to the privacy of persons in a vulnerable position. While theoretically this use-case may be useful in monitoring patient at risk of relapse, the impact of developing such technologies need to be carefully considered.

Besides using those visualisations for creating depression symptom-specific self-report scale, this research also carries wider implications concerning depression's effects on one's appearance and our innate ability to detect depression's symptoms

in a person. Contrasting laypersons perception of symptoms of depression as investigated in Chapter 9 to a clinical sample may yield some insights as to whether there are particular differences in either group in their ability to recognize symptoms of depression or even in their ability invoke a mental representation. In Chapter 9, Appendix 9.E visualisations were created only for participants that scored 7 or higher on the PHQ-9 screener questionnaire [42]. Although the sample size is small, there is noticeably less expressiveness in the models contained therein. However, further work is needed with a sufficient and balanced sample sizes of patients living with depression at varying severity to establish and contrast those representations.

Additionally, as elaborated in Chapter 2, Section 2.5.2, depression alters the way facial expressions are perceived. Depression appears to affect the way emotional faces are processed and alters the degree of impairment appears to be correlated to its severity. Further work may reveal how depression impacts facial expression processing and whether those impairments affects the recognition of more atomic parts of facial features rather than distinct categorical expressions.

Finally, the method employed to investigate this research question could also be extended to other affective disorders such as anxiety. In turn, this would allow to identify further mental constructs associated with their symptomatology. Affective disorders such as depression and anxiety are a known comorbidity [250], which implies that a certain degree of overlap between their symptom representations would be expected, however, those representations in an experiment contrasting both groups could also highlight specific, consistently-expressed differences in particular facial features.


Chapter 1
Introduction
Chapter 2
Literature Review
Chapter 3
Methodology
Chapter 4
Pilot investigation into using the Dynamic Visual Analogue Scale for mood self-reports
Chapter 5
Evaluation of an image augmentation method through facial expression generation and recognition machine learning tasks
Chapter 6
Exploration and evaluation of a bipolar facial expression-based scale for mood self-reports
Chapter 7
Design of a multidimensional facial expression-based interface for mood self-reports
Chapter 8
Exploration and evaluation of a multidimensional facial expression-based scale for mood self-reports
Chapter 9
Identifying and visualizing salient facial features indicative of symptoms of depression
Chapter 10
EmotionAlly – a system for contextualized facial expression-based mood self-reports and mood visualization
Chapter 11
Discussion
Chapter 12
Conclusion

Chapter 12

Conclusion

This thesis attempted to investigate an application of computer-generated facial expressions for self-reporting mood within mobile applications and explored alternative facial expression-based representations relevant to depression symptomatology. This was accomplished in three experimental studies described in Chapters 4, 6, and 8 and an exploratory study described in Chapter 9.

The thesis attempted to guide the reader through a gradual increase in complexity of ideas and their evaluation, beginning with a simple happiness-to-sadness scale to measure positive and negative affect, following up with using multiple basic facial expressions of emotion and building up to identify facial features characteristics for the symptomatology of depression. In this manner, the leitmotif of representing constructs from short-lived emotion-states and longer-lasting moods to trait-like symptoms lasting from months to years permeates the thesis. As the duration of those affective states increases, inversely the information content in the face decreases. The overall thesis conclusion is that facial expressions are a valuable modality through which mood can be represented.

Concluding on research question A) *"How can facial expression-based methods be evaluated as a valid way to self-report mood? How would such tools compare to an established method, such as traditional numerical-based scales?"*, using facial expressions were found to be a feasible modality to represent mood and a valid construct for self-reporting mood. Over three separate studies, variants of facial expression-based scales were contrasted to a more traditional measurement instrument such a visual analogue scale, where assessments were found to be strongly correlated. While this comparison only indicated whether both a facial expression-based scale and a VAS one assess similarly, while it is known that VAS scales are susceptible to systematic errors in their measurements more extensive evaluation is needed to establish whether one captures mood better than the other.

Concluding on research question B) *"Which aspects of facial expression-based tools are valuable to users and which further capabilities are desired?"*, is that facial expressions were a preferred method to provide mood self-reports, were more evocative and had a greater appeal to users. This was expressed through a wide range of preferences and customisations in the collected user feedback specifically

applicable to facial expression-based tools. In general, the facial expression-feedback elicited more through and thoughtful responses, were considered to be more engaging, and elicited more elaborate responses with respect to their future customisation and improvements. Just as importantly, they were also preferred when contrasted to traditional visual analogue scales for mood self-reports.

Concluding on research question C) *"Which technologies can generate expressions depicting a range of emotion intensities using images of arbitrary expressions? Could those be used within applications on commodity hardware, such as smartphones?"*, is that computer-generated facial expressions using machine learning models were very valuable in creating highly granular realistic faces used within facial expression-based scales. The underlying numerical encoding of facial expression classes and intensities did serve as a de-facto ground truth for their numerical quantification within multiple mood self-assessment prototypes, and within empirical analyses those were found to be a representative feedback. Additionally, their use within a smartphone application was feasible, although, due to resource constraints, those could not fully benefit from their direct use on an end-user's device.

Concluding on research question D) *"Are there specific facial expression-based representations descriptive of affective states? If so, could those be used within self-report tools?"*, most symptoms of depression appeared to possess distinct facial characteristics distinguishing them from a neutral face. In particular those were symptoms which did not feature a strong cognitive component, where conversely, those that did were not found to be distinguishable from a neutral face.

12.1 Limitations

A main point this thesis acknowledges is the limitations of what can be expressed through facial expressions and within the provided content attempts to map out the boundaries of that space. Although the research herein concentrated predominantly on self-reporting mood, there are broader implications in the space of perceptual research.

From the broader observations of this thesis, it is acknowledged that there states in the scope of emotions, moods as well as affective states, which cannot be represented through facial expressions. Prior work on facial expressions of emotion identified six expressions, considered to be universal across cultures [25]. Those alone do not provide sufficient variety to represent the plethora of distinct emotions or moods which accompany the human experience. In line with the core-concept of using facial expressions or features as feedback to self-assess mood, the conclusion is that facial expressions have certain limitations which do not position them as a substitute for traditional self-assessment methods.

There are also limitations to how facial expressions are perceived by different demographics, which need to be taken into account when designing such scales. Not only do some expressions have a distinct enactment across cultures, but the

range of intensity while enacting different expressions may vary. Although, there are many tangible benefits in using facial expressions for representing mood.

12.2 Future Work

Extending this work, there is evidence for the existence of more facial-bodily expressions, which appear to have distinct characteristics to portray 28 naturalistic expressions [196]. By reconciling the existing taxonomy imposed by existing models of emotions and integrating novel experimental fieldwork within the field of social face perception, further research could integrate those novel findings within a new generation of facial expression interfaces that cover a broader spectrum of emotions.

Alternatively, future work could integrate research on bodily expressiveness [251, 252], where the body can be another informative modality in conveying emotional information. By identifying bodily metaphors representing affective states that cannot be expressed through the face (e.g. posture for pride), such approach could further improve upon the method. Bodily cues are also relevant in conveying states associated with depression symptomatology [253, 254]. Thus, by using those two complementary modalities, robust methods can be created to portray a wider range of affective states. As elaborated on in Chapter 11, novel techniques in 3D modeling can already animate expressions and body movements at a great level of detail. Furthermore, those are computationally viable to be used even on mobile devices, albeit with a penalty to quality. The eased accessibility to such technologies could in turn also allow users to create their own metaphors for their own subjective states.

In Chapter 5, an augmented dataset was created using Delaunay triangulation [158] which was applied to a facial expression generation and recognition tasks. The augmentation yielded promising results in improving the quality expressions created in the generative task. Within the recognition task, when evaluating the performance of the augmented dataset, used for training on the original one it achieved near-similar performance to simply using the original dataset. It would be interesting to apply this technique to a dataset consisting of facial expressions portraying subtle intensities of emotion and evaluate whether the augmentation consisting of facial expressions at intermediary intensities improves classification results for ambiguous faces and specifically morphologically similar ones (e.g. surprise and fear).

As elaborated on in Chapter 2, Section 2.5.2, depression influences the perception and processing of facial expressions expressed as biases. Among others, the most reliably replicated bias is expressed as a dampening in sensitivity to identify subtle facial expressions except for that of sadness, where inversely, a heightened sensitivity is observed. Those, as proposed by other authors as well [116], appear to have a diagnostic utility in assessing depression severity through a set of perceptual biases observed in patients living with depression (Chapter 2, Section 2.5.2). Specifically, those biases are manifested only when observing realistic facial expressions and are

absent in other types of emotional content such as text or representations such as emojis, or schematic or drawn faces [106–108]. This observation positions facial expression-based technologies in a unique spot to capture and quantify them. Hence, facial expression-based self-assessment instruments, similar to those explored and used in this thesis that possess sufficient granularity to represent a wide range of nuanced emotion intensities could be explored clinically for their diagnostic utility. A benefit of this approach would be that the quantification of those biases can be accomplished using regular self-reports at no additional cost to a user (Chapter 2, Section 2.5.2).

In Chapter 9 an exploratory study visualized each symptom of depression as defined by the Patient Health Questionnaire (PHQ-9) and found that some symptoms had distinguishing features identifiable in a face. Therein, a future research could validate the expressions that did, whether those can be correctly attributed by unbiased observers to their respective textual formulations, thus consolidating that those representations indeed represent a shared perception of reality. Another potential research direction would be to explore how a depressed population perceives symptoms of depression to be reflected in a face. Knowing that depression has an effect on processing of emotional faces, it would be interesting to know whether those representations match or differ from the ones obtained by healthy laypersons in Chapter 9. This may be an indication whether those biases affect the recognition or processing of distinct expressions or more atomic parts of facial features contributing to facial expressiveness. Should such an effect be found, identifying its direction and magnitude (e.g. increased or decreased sensitivity) could shed some light into how depression affects the identification of information, salient to detect depression symptomatology in oneself or others and whether this effect possesses a distinct symptom-to-symptom relationship (e.g. whether the presence of a symptom increases or decreases a persons' sensitivity to detect it in a face). Alternatively, since some symptoms of depression had distinct features reflected in the face, this observation may be an indication of neurological changes in facial musculature-activation accompanied by the onset of a depressive symptom.

Bibliography

[1] Hristo Valev, Tim Leufkens, Corina Sas, Joyce Westerink, and Ron Dotsch. "Evaluation of a Self-report System for Assessing Mood Using Facial Expressions." In: *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer International Publishing, 2019, pp. 231–241. DOI: 10.1007/978-3-030-25872-6_19 (Cited on pages iv, 46, 57).

[2] Hristo Valev, Alessio Gallucci, Tim Leufkens, Joyce Westerink, and Corina Sas. "Applying Delaunay Triangulation Augmentation for Deep Learning Facial Expression Generation and Recognition." In: *Pattern Recognition. ICPR International Workshops and Challenges*. Springer International Publishing, 2021, pp. 730–740. DOI: 10.1007/978-3-030-68796-0_53 (Cited on pages iv, vi, 58).

[3] Hristo Valev, Tim Leufkens, Corina Sas, and Joyce Westerink. "On the perception of facial expressions in affective disorders and potential technological uses." In: *25th annual international CyberPsychology, CyberTherapy & Social Networking Conference*. 2020 (Cited on pages iv, 162).

[4] Hristo Valev, Tim Leufkens, Joyce Westerink, and Corina Sas. "An Interface with Computer-Generated Facial Expressions as an Alternative for Mood Self-Reports in an EMA Context." In: *Annual Conference Symposium*. Society for Affective Science, 2020, p. 47 (Cited on pages iv, vi).

[5] Hristo Ventzeslavov Valev, Timmy Robertus Maria Leufkens, Joanne Henriëtte Desirée Monique Westerink, Dooren Marieke Van, Ee Raymond Van, Willem Huijbers, Benito Maria Estrella Mena, and Adrianus Johannes Maria Denissen. "Apparatus for Determining and/or Assess Depression Severity of a Patient." Pat. WO2022101108A1. May 2022 (Cited on page iv).

[6] Desirée Colombo, Javier Fernández-Álvarez, Carlos Suso-Ribera, Pietro Cipresso, Hristo Valev, Tim Leufkens, Corina Sas, Azucena Garcia-Palacios, Giuseppe Riva, and Cristina Botella. "The need for change: Understanding emotion regulation antecedents and consequences using ecological momentary assessment." In: *Emotion* 20.1 (Feb. 2020), pp. 30–36. ISSN: 19311516. DOI: 10.1037/emo0000671 (Cited on pages 16, 57).

[7] Asmae Doukani, Robin van Dalen, Hristo Valev, Annie Njenga, Francesco Sera, and Dixon Chibanda. "A community health volunteer delivered problem-solving therapy mobile application based on the Friendship Bench 'Inuka Coaching' in Kenya: A pilot cohort study." In: *Global Mental Health* 8 (2021). DOI: 10.1017/gmh.2021.3.

[8] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. "Presentation and validation of the Radboud Faces Database." In: *Cognition & Emotion* 24.8 (Dec. 2010), pp. 1377–1388. ISSN: 02699931. DOI: 10.1080/02699930903485076 (Cited on pages vi, 33, 35, 42, 58, 69, 85, 159, 163, 167).

[9] James A. Russell. "A circumplex model of affect." In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178. DOI: 10.1037/h0077714 (Cited on pages vi, 6, 12, 88, 93, 95).

[10] Margaret M. Bradley and Peter J. Lang. "International Affective Picture System." In: (2017), pp. 1–4. DOI: 10.1007/978-3-319-28099-8_42-1 (Cited on pages vii, 7, 96, 97, 99, 159).

[11] 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." In: *Behavior Research Methods* 37.4 (Nov. 2005), pp. 626–630. ISSN: 1554351X. DOI: 10.3758/BF03192732. arXiv: NIHMS150003 (Cited on pages vii, 97, 98, 101, 107, 108, 114, 115).

[12] Jason M. Gold. *Reverse Correlation*. 2013. DOI: 10.4135/9781412972000.n280 (Cited on pages vii, 32).

[13] Ron Dotsch. *rcicr: Reverse correlation image classification toolbox*. R package version 0.3.0. 2014 (Cited on pages vii, 121, 127).

[14] May I. Conley, Danielle V. Dellarco, Estee Rubien-Thomas, Alexandra O. Cohen, Alessandra Cervera, Nim Tottenham, and BJ Casey. "The racially diverse affective expression (RADIATE) face stimulus set." In: *Psychiatry Research* 270.March (Dec. 2018), pp. 1059–1067. ISSN: 18727123. DOI: 10.1016/j.psychres.2018.04.066 (Cited on pages vii, 33, 35, 42, 120, 167).

[15] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and Corina Sas. "HCI and Affective Health: Taking stock of a decade of studies and charting future research directions." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, May 2019. DOI: 10.1145/3290605.3300475 (Cited on pages 2, 15, 57).

[16] Lisa Barrett Feldman, Michael Lewis, and Jeannette Haviland-Jones M. *Handbook of Emotions (4th edition)*. Vol. 1. 2016, pp. 742–756. ISBN: 0-89862-988-8 (Hardcover) (Cited on page 2).

[17] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. "Ecological Momentary Assessment." In: *Annual Review of Clinical Psychology* 4.1 (Apr. 2008), pp. 1–32. ISSN: 1548-5943. DOI: 10.1146/annurev.clinpsy.3.022806.091415 (Cited on pages 2, 16, 69, 85, 95).

[18] Rensis Likert. *A technique for the measurement of attitudes*. eng. Archives of psychology ; no. 140. New York: [s.n.], 1985 - 1932 (Cited on pages 2, 14, 16, 160).

[19] Donald D. Price, Patricia A. McGrath, Amir Rafii, and Barbara Buckingham. "The validation of visual analogue scales as ratio scale measures for chronic and experimental pain." In: *Pain* 17.1 (Sept. 1983), pp. 45–56. DOI: 10.1016/0304-3959(83)90126-4 (Cited on pages 2, 14, 16, 33, 37, 46, 70, 159, 160).

[20] Keith J Holyoak. "Comparative judgments with numerical reference points." In: *Cognitive Psychology* 10.2 (Apr. 1978), pp. 203–243. ISSN: 00100285. DOI: 10.1016/0010-0285(78)90014-2 (Cited on pages 2, 15, 78).

[21] Christopher D. Lorish and Richard Maisiak. "The face scale: A brief, nonverbal method for assessing patient mood." In: *Arthritis & Rheumatism* 29.7 (July 1986), pp. 906–909. ISSN: 15290131. DOI: 10.1002/art.1780290714 (Cited on pages 2, 3, 15, 17, 24, 45, 160, 165).

[22] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. "The Effect of Visual Appearance on the Performance of Continuous Sliders and Visual Analogue Scales." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, May 2016, pp. 5421–5432. ISBN: 9781450333627. DOI: 10.1145/2858036.2858063 (Cited on pages 2, 14, 78, 112, 163).

[23] G.W. Torrance, D. Feeny, and W. Furlong. "Visual Analog Scales: Do They Have a Role in the Measurement of Preferences for Health States?" In: *Medical Decision Making* 21.4 (Aug. 2001), pp. 329–334. ISSN: 0272989X. DOI: 10.1177/02729890122062622 (Cited on pages 2, 15, 78).

[24] Norbert Schwarz, Barbel Knauper, Hans-J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark. "Rating Scales: Numeric Values May Change the Meaning of Scale Labels." In: *Public Opinion Quarterly* 55.4 (1991), p. 570. ISSN: 0033362X. DOI: 10.1086/269282 (Cited on pages 2, 15).

[25] Paul Ekman. "Universal Facial Expressions of Emotion." In: 8 (4).4 (1970), pp. 151–158 (Cited on pages 3, 11, 13, 18, 20, 22, 88, 111, 115, 118, 160, 164, 169, 176).

[26] Paul Ekman. "Facial expression and emotion." In: *American Psychologist* 48.4 (1993), pp. 384–392. ISSN: 1935-990X. DOI: 10.1037/0003-066X.48.4.384 (Cited on pages 3, 18).

[27] Paul D Barrows and Shirley A Thomas. "Assessment of mood in aphasia following stroke: validation of the Dynamic Visual Analogue Mood Scales (D-VAMS)." In: *Clinical Rehabilitation* 32.1 (June 2017), pp. 94–102. ISSN: 14770873. DOI: 10.1177/0269215517714590 (Cited on pages 3, 6, 16, 17, 45–47, 58, 69, 159, 160, 162, 166).

[28] Carlos Crivelli and Alan J. Fridlund. "Facial Displays Are Tools for Social Influence." In: *Trends in Cognitive Sciences* 22.5 (May 2018), pp. 388–399. ISSN: 1879307X. DOI: 10.1016/j.tics.2018.02.006 (Cited on pages 3, 17).

[29] Theo Vos et al. "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015." In: *The Lancet* 388.10053 (Oct. 2016), pp. 1545–1602. ISSN: 1474547X. DOI: 10.1016/s0140-6736(16)31678-6. arXiv: arXiv:1011.1669v3 (Cited on page 3).

[30] Elizabeth Brown and David I Perrett. "What Gives a Face its Gender?" In: *Perception* 22.7 (July 1993), pp. 829–840. ISSN: 03010066. DOI: 10.1068/p220829 (Cited on pages 3, 118).

[31] Emily J. Cogsdill, Alexander T. Todorov, Elizabeth S. Spelke, and Mahzarin R. Banaji. "Inferring Character From Faces: A Developmental Study." In: *Psychological Science* 25.5 (Feb. 2014), pp. 1132–1139. ISSN: 14679280. DOI: 10.1177/0956797614523297 (Cited on pages 3, 118).

[32] Ron Dotsch, Daniël H.J. Wigboldus, Oliver Langner, and Ad van Knippenberg. "Ethnic Out-Group Faces Are Biased in the Prejudiced Mind." In: *Psychological Science* 19.10 (Oct. 2008), pp. 978–980. ISSN: 09567976. DOI: 10.1111/j.1467-9280.2008.02186.x (Cited on pages 3, 119, 121, 170).

[33] Margaret M. Bradley and Peter J. Lang. "Measuring emotion: The self-assessment manikin and the semantic differential." In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (Mar. 1994), pp. 49–59. ISSN: 00057916. DOI: 10.1016/0005-7916(94)90063-9. arXiv: 0005-7916(93)E0016-Z (Cited on pages 3, 15, 24, 99, 160, 165).

[34] Joost Broekens and Willem-Paul Brinkman. "AffectButton: A method for reliable and valid affective self-report." In: *International Journal of Human-Computer Studies* 71.6 (June 2013), pp. 641–667. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2013.02.003 (Cited on pages 3, 16, 45).

[35] Iyubanit Rodriguez, Valeria Herskovic, Carolina Fuentes, and Mauricio Campos. "B-ePain: a wearable interface to self-report pain and emotions." In: *UbiComp Adjunct* (Sept. 2016), pp. 1120–1125. DOI: 10.1145/2968219.2972719 (Cited on pages 3, 16, 45).

[36] Sky Tien-Yun Huang, Akane Sano, and Chloe Mun Yee Kwan. "The moment." In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct* (Sept. 2014), pp. 235–238. DOI: 10.1145/2638728.2638784 (Cited on pages 3, 15).

[37] Muhammad Umair, Muhammad Hamza Latif, and Corina Sas. "Dynamic Displays at Wrist for Real Time Visualization of Affective Data." In: *DIS 2018 - Companion Publication of the 2018 Designing Interactive Systems Conference* (May 2018), pp. 201–205. DOI: 10.1145/3197391.3205436 (Cited on pages 3, 15).

[38] John P. Pollak, Phil Adams, and Geri Gay. "PAM: A Photographic Affect Meter for frequent, in situ measurement of affect." In: *Conference on Human Factors in Computing Systems - Proceedings* (May 2011), pp. 725–734. DOI: 10.1145/1978942.1979047 (Cited on pages 3, 15).

[39] Ingrid Brdar. "Positive and Negative Affect Schedule (PANAS)." In: (2014), pp. 4918–4920. DOI: 10.1007/978-94-007-0753-5_2212 (Cited on pages 3, 15).

[40] John R. Crawford and Julie D. Henry. "The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample." In: *British Journal of Clinical Psychology* 43.3 (Sept. 2004), pp. 245–265. ISSN: 01446657. DOI: 10.1348/0144665031752934 (Cited on pages 3, 15).

[41] Garreth W. Tigwell and David R. Flatla. "Oh that's what you meant!" In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. Cmc. New York, New York, USA: ACM, Sept. 2016, pp. 859–866. ISBN: 9781450344135. DOI: 10.1145/2957265.2961844 (Cited on pages 3, 16).

[42] Kurt Kroenke and Robert L Spitzer. "The PHQ-9: A New Depression Diagnostic and Severity Measure." In: *Psychiatric Annals* 32.9 (Sept. 2002), pp. 509–515. ISSN: 00485713. DOI: 10.3928/0048-5713-20020901-06. arXiv: 217052580 (Cited on pages 7, 20, 34, 118, 119, 121, 124, 173).

[43] R. F. Murray. "Classification images: A review." In: *Journal of Vision* 11.5 (May 2011), pp. 2–2. ISSN: 1534-7362. DOI: 10.1167/11.5.2 (Cited on pages 7, 35, 119, 121).

[44] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. "Presentation and validation of the Radboud Faces Database." In: *Cognition & Emotion* 24.8 (Dec. 2010), pp. 1377–1388. DOI: 10.1080/02699930903485076 (Cited on pages 7, 58, 112).

[45] Robert W. Levenson. "Basic Emotion Questions." In: *Emotion Review* 3.4 (Sept. 2011), pp. 379–386. ISSN: 17540739. DOI: 10.1177/1754073911410743 (Cited on page 11).

[46] Carroll E. Izard. "Forms and Functions of Emotions: Matters of Emotion–Cognition Interactions." In: *Emotion Review* 3.4 (Sept. 2011), pp. 371–378. ISSN: 17540739. DOI: 10.1177/1754073911410737 (Cited on page 11).

[47] Robert Plutchik. "The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice." In: *American Scientist* 89.4 (2001), pp. 344–350. ISSN: 00030996 (Cited on page 11).

[48] Rachael E. Jack. "Culture and facial expressions of emotion." In: *Visual Cognition* 21.9-10 (Sept. 2013), pp. 1248–1286. ISSN: 13506285. DOI: 10.1080/13506285.2013.835367 (Cited on pages 11, 18, 19).

[49] Rachael E. Jack, Wei Sun, Ioannis Delis, Oliver G. B. Garrod, and Philippe G. Schyns. "Four not six: Revealing culturally common facial expressions of emotion." In: *Journal of Experimental Psychology: General* 145.6 (June 2016), pp. 708–730. ISSN: 00963445. DOI: 10.1037/xge0000162 (Cited on pages 11, 25).

[50] Simeng Gu, Fushun Wang, Tifei Yuan, Benyu Guo, and Jason H Huang. "Differentiation of Primary Emotions through Neuromodulators: Review of Literature." In: *International Journal of Neurology Research* 1.2 (2015), pp. 43–50. ISSN: 2313-5611. DOI: 10.17554/j.issn.2313-5611.2015.01.19 (Cited on page 11).

[51] Alfredo Pereira and Fushun Wang. "Neuromodulation, Emotional Feelings and Affective Disorders." In: *Mens Sana Monographs* 14.1 (2016), p. 5. ISSN: 19984014. DOI: 10.4103/0973-1229.154533 (Cited on page 11).

[52] Wilhelm Wundt. *Ethics: An investigation of the facts and laws of the moral life, Vol 1: Introduction: The facts of the moral life.* Swan Sonnenschein Co, May 1897. DOI: 10.1037/12898-000 (Cited on page 11).

[53] James A Russell and Albert Mehrabian. "Evidence for a three-factor theory of emotions." In: *Journal of Research in Personality* 11.3 (Sept. 1977), pp. 273–294. DOI: 10.1016/0092-6566(77)90037-x (Cited on page 11).

[54] Jonathan Posner, James A. Russell, Andrew Gerber, Daniel Gorman, Tiziano Colibazzi, Shan Yu, Zhishun Wang, Alayar Kangarlu, Hongtu Zhu, and Bradley S. Peterson. "The neurophysiological bases of emotion: An fMRI study of the affective circumplex using emotion-denoting words." In: *Human Brain Mapping* 30.3 (Mar. 2009), pp. 883–895. ISSN: 10659471. DOI: 10.1002/hbm.20553 (Cited on page 12).

[55] Mary Katsikitis. "The Classification of Facial Expressions of Emotion: A Multidimensional-Scaling Approach." In: *Perception* 26.5 (May 1997), pp. 613–626. ISSN: 03010066. DOI: 10.1068/p260613 (Cited on page 12).

[56] Takuma Takehara and Naoto Suzuki. "Differential processes of emotion space over time." In: *North American Journal of Psychology* 3.2 (2001), pp. 217–228. ISSN: 1527-7143 (Print) (Cited on page 12).

[57] Yoshi-Taka Matsuda, Tomomi Fujimura, Kentaro Katahira, Masato Okada, Kenichi Ueno, Kang Cheng, and Kazuo Okanoya. "The implicit processing of categorical and dimensional strategies: an fMRI study of facial emotion perception." In: *Frontiers in Human Neuroscience* 7.SEP (2013). ISSN: 16625161. DOI: 10.3389/fnhum.2013.00551 (Cited on pages 12, 109).

[58] Georgia Panayiotou. "Emotional dimensions reflected in ratings of affective scripts." In: *Personality and Individual Differences* 44.8 (June 2008), pp. 1795–1806. ISSN: 01918869. DOI: 10.1016/j.paid.2008.02.006 (Cited on page 12).

[59] Tomomi Fujimura, Yoshi-Taka Matsuda, Kentaro Katahira, Masato Okada, and Kazuo Okanoya. "Categorical and dimensional perceptions in decoding emotional facial expressions." In: *Cognition & Emotion* 26.4 (June 2012), pp. 587–601. ISSN: 02699931. DOI: 10.1080/02699931.2011.595391 (Cited on pages 12, 88).

[60] Simeng Gu, Fushun Wang, Nitesh P. Patel, James A. Bourgeois, and Jason H. Huang. "A Model for Basic Emotions Using Observations of Behavior in Drosophila." In: *Frontiers in Psychology* 10.APR (Apr. 2019), p. 781. ISSN: 16641078. DOI: 10.3389/fpsyg.2019.00781 (Cited on page 12).

[61] Paul Ekman. "Expression and the Nature of Emotion." In: *Approaches to Emotion*. 1984, pp. 319–344 (Cited on page 12).

[62] Lisa Barrett. *How emotions are made : the secret life of the brain*. Boston: Houghton Mifflin Harcourt, 2017, p. 425. ISBN: 9780544133310. DOI: 10.1037/teo0000098 (Cited on page 13).

[63] Paul Ekman and Harriet Oster. "Facial Expressions of Emotion." In: *Annual Review of Psychology* 30.1 (Jan. 1979), pp. 527–554. ISSN: 0066-4308. DOI: 10.1146/annurev.ps.30.020179.002523 (Cited on pages 13, 17).

[64] Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, and Klaus Scherer. "Relative importance of face, body, and speech in judgments of personality and affect." In: *Journal of Personality and Social Psychology* 38.2 (1980), pp. 270–277. ISSN: 00223514. DOI: 10.1037/0022-3514.38.2.270 (Cited on pages 13, 17).

[65] Karleyton C. Evans, Christopher I. Wright, Michelle M. Wedig, Andrea L. Gold, Mark H. Pollack, and Scott L. Rauch. "A functional MRI study of amygdala responses to angry schematic faces in social anxiety disorder." In: *Depression and Anxiety* 25.6 (June 2008), pp. 496–505. ISSN: 10914269. DOI: 10.1002/da.20347 (Cited on page 15).

[66] Noam Sagiv and Shlomo Bentin. "Structural encoding of human and schematic faces: Holistic and part-based processes." In: *Journal of Cognitive Neuroscience* 13.7 (Oct. 2001), pp. 937–951. ISSN: 0898929X. DOI: 10.1162/089892901753165854 (Cited on pages 16, 21).

[67] Katrin Hänsel, Akram Alomainy, and Hamed Haddadi. "Large scale mood and stress self-assessments on a smartwatch." In: *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Sept. 2016), pp. 1180–1184. DOI: 10.1145/2968219.2968305 (Cited on pages 16, 45, 156).

[68] Desirée Colombo, Carlos Suso-Ribera, Javier Fernandez-Álvarez, Isabel Fernandez Felipe, Pietro Cipresso, Azucena Garcia Palacios, Giuseppe Riva, and Cristina Botella. "Exploring Affect Recall Bias and the Impact of Mild Depressive Symptoms: An Ecological Momentary Study." In: *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST* 288 (2019), pp. 208–215. ISSN: 18678211. DOI: 10.1007/978-3-030-25872-6_17 (Cited on page 16).

[69] Jinhyuk Kim, Toru Nakamura, Hiroe Kikuchi, Kazuhiro Yoshiuchi, Tsukasa Sasaki, and Yoshiharu Yamamoto. "Covariation of Depressive Mood and Spontaneous Physical Activity in Major Depressive Disorder: Toward Continuous Monitoring of Depressive Mood." In: *IEEE Journal of Biomedical and Health Informatics* 19.4 (July 2015), pp. 1347–1355. ISSN: 21682194. DOI: 10.1109/JBHI.2015.2440764 (Cited on page 16).

[70] Arthur A. Stone and Saul Shiffman. "Ecological Momentary Assessment (Ema) in Behavioral Medicine." In: *Annals of Behavioral Medicine* 16.3 (Jan. 1994), pp. 199–202. ISSN: 08836612. DOI: 10.1093/abm/16.3.199 (Cited on pages 16, 17).

[71] Arthur A. Stone, Joseph E. Schwartz, John M. Neale, Saul Shiffman, Christine A. Marco, Mary Hickcox, Jean Paty, Laura S. Porter, and Laura J. Cruise. "A comparison of coping assessed by ecological momentary assessment and retrospective recall." In: *Journal of Personality and Social Psychology* 74.6 (June 1998), pp. 1670–1680. ISSN: 00223514. DOI: 10.1037/0022-3514.74.6.1670 (Cited on page 16).

[72] Paul Ekman, Wallace V. Freisen, and Sonia Ancoli. "Facial signs of emotional experience." In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1125–1134. DOI: https://doi.org/10.1037/h0077722 (Cited on page 17).

[73] Rachael E. Jack, Caroline Blais, Christoph Scheepers, Philippe G. Schyns, and Roberto Caldara. "Cultural Confusions Show that Facial Expressions Are Not Universal." In: *Current Biology* 19.18 (Sept. 2009), pp. 1543–1548. ISSN: 09609822. DOI: 10.1016/j.cub.2009.07.051 (Cited on page 17).

[74] Klaus R. Scherer, Marcello Mortillaro, and Marc Mehu. "Understanding the Mechanisms Underlying the Production of Facial Expression of Emotion: A Componential Perspective." In: *Emotion Review* 5.1 (Jan. 2013), pp. 47–53. ISSN: 17540739. DOI: 10.1177/1754073912451504 (Cited on page 17).

[75] Alaa Althubaiti. "Information bias in health research: definition, pitfalls, and adjustment methods." In: *Journal of Multidisciplinary Healthcare* 9 (May 2016), p. 211. ISSN: 11782390. DOI: 10.2147/jmdh.s104807 (Cited on page 18).

[76] Magdalena Rychlowska, Rachael E. Jack, Oliver G. B. Garrod, Philippe G. Schyns, Jared D. Martin, and Paula M. Niedenthal. "Functional Smiles: Tools for Love, Sympathy, and War." In: *Psychological Science* 28.9 (July 2017), pp. 1259–1270. ISSN: 14679280. DOI: 10.1177/0956797617706082 (Cited on pages 18, 26).

[77] John T. Cacioppo, Bert N. Uchino, Stephen L. Crites, Mary A. Snydersmith, Gregory Smith, Gary G. Berntson, and Peter J. Lang. "Relationship between facial expressiveness and sympathetic activation in emotion: A critical review, with emphasis on modeling underlying mechanisms and individual differences." In: *Journal of Personality and Social Psychology* 62.1 (1992), pp. 110–128. ISSN: 00223514. DOI: 10.1037/0022-3514.62.1.110 (Cited on pages 18, 19).

[78] Fabien Trémeau, Dolores Malaspina, Fabrice Duval, Humberto Corrêa, Michaela Hager-Budny, Laura Coin-Bariou, Jean-Paul Macher, and Jack M. Gorman. "Facial Expressiveness in Patients With Schizophrenia Compared to Depressed Patients and Nonpatient Comparison Subjects." In: *American Journal of Psychiatry* 162.1 (Jan. 2005), pp. 92–101. ISSN: 0002953X. DOI: 10.1176/appi.ajp.162.1.92 (Cited on page 18).

[79] L. Brinkman, A. Todorov, and R. Dotsch. "Visualising mental representations: A primer on noise-based reverse correlation in social psychology." In: *European Review of Social Psychology* 28.1 (Jan. 2017), pp. 333–361. ISSN: 1479277X. DOI: 10.1080/10463283.2017.1381469 (Cited on pages 18, 170).

[80] Christopher Y. Olivola, Friederike Funk, and Alexander Todorov. "Social attributions from faces bias human choices." In: *Trends in Cognitive Sciences* 18.11 (Nov. 2014), pp. 566–570. ISSN: 1879307X. DOI: 10.1016/j.tics.2014.09.007 (Cited on page 18).

[81] Manuel C. Voelkle, Natalie C. Ebner, Ulman Lindenberger, and Michaela Riediger. "Let me guess how old you are: Effects of age, gender, and facial expression on perceptions of age." In: *Psychology and Aging* 27.2 (2012), pp. 265–277. ISSN: 08827974. DOI: 10.1037/a0025065 (Cited on page 18).

[82] Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E. Ricci-Bitti, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. "Universals and cultural differences in the judgments of facial expressions of emotion." In: *Journal of Personality and Social Psychology* 53.4 (1987), pp. 712–717. ISSN: 00223514. DOI: 10.1037/0022-3514.53.4.712 (Cited on page 19).

[83] David Matsumoto and Paul Ekman. "American-Japanese cultural differences in intensity ratings of facial expressions of emotion." In: *Motivation and Emotion* 13.2 (June 1989), pp. 143–157. ISSN: 01467239. DOI: 10.1007/BF00992959 (Cited on page 19).

[84] David Matsumoto, Tsutomu Kudoh, Klaus Scherer, and Harald Wallbott. "Antecedents of and Reactions to Emotions in the United States and Japan." In: *Journal of Cross-Cultural Psychology* 19.3 (Sept. 1988), pp. 267–286. ISSN: 15525422. DOI: 10.1177/0022022188193001 (Cited on page 19).

[85] Paula M. Niedenthal, Magdalena Rychlowska, Adrienne Wood, and Fangyun Zhao. "Heterogeneity of long-history migration predicts smiling, laughter and positive emotion across the globe and within the United States." In: *PLOS ONE* 13.8 (Aug. 2018). Ed. by Alex Mesoudi, e0197651. ISSN: 19326203. DOI: 10.1371/journal.pone.0197651 (Cited on page 19).

[86] Hillary Anger Elfenbein. "Learning in emotion judgments: Training and the cross-cultural understanding of facial expressions." In: *Journal of Nonverbal Behavior* 30.1 (Mar. 2006), pp. 21–36. ISSN: 01915886. DOI: 10.1007/s10919-005-0002-y (Cited on page 19).

[87] Xiaoqian Yan, Timothy J. Andrews, and Andrew W. Young. "Cultural similarities and differences in perceiving and recognizing facial expressions of basic emotions." In: *Journal of Experimental Psychology: Human Perception and Performance* 42.3 (2016), pp. 423–440. ISSN: 19391277. DOI: 10.1037/xhp0000114 (Cited on page 19).

[88] Hillary Anger Elfenbein, Martin Beaupré, Manon Lévesque, and Ursula Hess. "Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions." In: *Emotion* 7.1 (Feb. 2007), pp. 131–146. ISSN: 15283542. DOI: 10.1037/1528-3542.7.1.131 (Cited on page 19).

[89] Ann M. Kring and Albert H. Gordon. "Sex differences in emotion: Expression, experience, and physiology." In: *Journal of Personality and Social Psychology* 74.3 (1998), pp. 686–703. ISSN: 00223514. DOI: 10.1037/0022-3514.74.3.686 (Cited on page 19).

[90] Judith A. Hall. "Gender effects in decoding nonverbal cues." In: *Psychological Bulletin* 85.4 (1978), pp. 845–857. ISSN: 00332909. DOI: 10.1037/0033-2909.85.4.845 (Cited on page 19).

[91] Harold Kaplan. *Kaplan and Sadock's synopsis of psychiatry : behavioral sciences, clinical psychiatry*. Baltimore: Williams & Wilkins, 1998. ISBN: 9780683303308 (Cited on page 19).

[92] D. Healy and J. M. Williams. "Dysrhythmia, dysphoria, and depression: The interaction of learned helplessness and circadian dysrhythmia in the pathogenesis of depression." In: *Psychological Bulletin* 103.2 (1988), pp. 163–178. ISSN: 00332909. DOI: 10.1037/0033-2909.103.2.163 (Cited on page 19).

[93] William M. Reynolds and Kenneth A. Kobak. "Reliability and validity of the Hamilton Depression Inventory: A paper-and-pencil version of the Hamilton Depression Rating Scale Clinical Interview." In: *Psychological Assessment* 7.4 (1995), pp. 472–483. DOI: 10.1037/1040-3590.7.4.472 (Cited on page 20).

[94] R. P. Snaith, F. M. Harrop, D. A. Newby, and C. Teale. "Grade Scores of the Montgomery-Åsberg Depression and the Clinical Anxiety Scales." In: *The British Journal of Psychiatry* 148.5 (May 1986), pp. 599–601. DOI: 10.1192/bjp.148.5.599 (Cited on page 20).

[95] Karen Raphael. "Recall Bias: A Proposal for Assessment and Control." In: *International Journal of Epidemiology* 16.2 (1987), pp. 167–170. ISSN: 03005771. DOI: 10.1093/ije/16.2.167 (Cited on pages 20, 69).

[96] Eman Hassan. "Recall bias can be a threat to retrospective and prospective research designs." In: *The Internet Journal of Epidemiology* 3.2 (2006), pp. 339–412 (Cited on pages 20, 69).

[97] B. Thomas Longwell and Paula Truax. "The differential effects of weekly, monthly, and bimonthly administrations of the beck Depression Inventory-II: Psychometric properties and clinical implications." In: *Behavior Therapy* 36.3 (2005), pp. 265–275. ISSN: 00057894. DOI: 10.1016/s0005-7894(05)80075-9 (Cited on page 20).

[98] Shinichiro Tomitaka, Yohei Kawasaki, Kazuki Ide, Maiko Akutagawa, Hiroshi Yamada, Yutaka Ono, and Toshiaki A. Furukawa. "Distributional patterns of item responses and total scores on the PHQ-9 in the general population: data from the National Health and Nutrition Examination Survey." In: *BMC Psychiatry* 18.1 (Apr. 2018), pp. 1–9. ISSN: 1471244X. DOI: 10.1186/s12888-018-1696-9 (Cited on page 20).

[99] W. A. Eeden, A. M. Hemert, I. V. E. Carlier, B. W. Penninx, and E. J. Giltay. "Severity, course trajectory, and within-person variability of individual symptoms in patients with major depressive disorder." In: *Acta Psychiatrica Scandinavica* 139.2 (Dec. 2018), pp. 194–205. ISSN: 16000447. DOI: 10.1111/acps.12987 (Cited on pages 20, 21).

[100] Brooke Levis et al. "Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis." In: *Journal of Clinical Epidemiology* 122 (June 2020), 115–128.e1. ISSN: 18785921. DOI: 10.1016/j.jclinepi.2020.02.002 (Cited on page 21).

[101] Ian H. Gotlib, Elena Krasnoperova, Dana Neubauer Yue, and Jutta Joormann. "Attentional Biases for Negative Interpersonal Stimuli in Clinical Depression." In: *Journal of Abnormal Psychology* 113.1 (Feb. 2004), pp. 127–135. DOI: 10.1037/0021-843x.113.1.121 (Cited on page 21).

[102] Jutta Joormann and Ian H. Gotlib. "Selective attention to emotional faces following recovery from depression." In: *Journal of Abnormal Psychology* 116.1 (Feb. 2007), pp. 80–85. ISSN: 0021843X. DOI: 10.1037/0021-843x.116.1.80 (Cited on pages 21, 22).

[103] Qin Dai and Zhengzhi Feng. "More excited for negative facial expressions in depression: Evidence from an event-related potential study." In: *Clinical Neurophysiology* 123.11 (Nov. 2012), pp. 2172–2179. ISSN: 13882457. DOI: 10.1016/j.clinph.2012.04.018 (Cited on pages 21–23).

[104] Jukka M. Leppänen, Maarten Milders, J.Stephen Bell, Emma Terriere, and Jari K. Hietanen. "Depression biases the recognition of emotionally neutral faces." In: *Psychiatry Research* 128.2 (Sept. 2004), pp. 123–133. ISSN: 01651781. DOI: 10.1016/j.psychres.2004.05.020 (Cited on pages 21–23).

[105] M. N. Dalili, I. S. Penton-Voak, C. J. Harmer, and M. R. Munafò. "Meta-analysis of emotion recognition deficits in major depressive disorder." In: *Psychological Medicine* 45.6 (Nov. 2014), pp. 1135–1144. ISSN: 14698978. DOI: 10.1017/s0033291714002591 (Cited on pages 21, 22).

[106] Thomas Suslow, Klaus Junghanns, and Volker Arolt. "Detection of Facial Expressions of Emotions in Depression." In: *Perceptual and Motor Skills* 92.3 (June 2001), pp. 857–868. DOI: 10.2466/pms.2001.92.3.857 (Cited on pages 21, 162, 178).

[107] Stanislava Petkova Karparova, Anette Kersting, and Thomas Suslow. "Disengagement of attention from facial emotion in unipolar depression." In: *Psychiatry and Clinical Neurosciences* 59.6 (Dec. 2005), pp. 723–729. ISSN: 1323-1316. DOI: 10.1111/j.1440-1819.2005.01443.x (Cited on pages 21, 162, 178).

[108] Reza Pishyar, Lynne M. Harris, and Ross G. Menzies. "Attentional bias for words and faces in social anxiety." In: *Anxiety, Stress & Coping* 17.1 (Mar. 2004), pp. 23–36. ISSN: 10615806. DOI: 10.1080/10615800310001601458 (Cited on pages 21, 178).

[109] Amit Lazarov, Ziv Ben-Zion, Dana Shamai, Daniel S. Pine, and Yair Bar-Haim. "Free viewing of sad and happy faces in depression: A potential target for attention bias modification." In: *Journal of Affective Disorders* 238.3 (Oct. 2018), pp. 94–100. DOI: 10.1016/j.jad.2018.05.047 (Cited on page 22).

[110] Almudena Duque and Carmelo Vázquez. "Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study." In: *Journal of Behavior Therapy and Experimental Psychiatry* 46 (Mar. 2015), pp. 107–114. ISSN: 18737943. DOI: 10.1016/j.jbtep.2014.09.005 (Cited on page 22).

[111] Lemke Leyman, Rudi De Raedt, Rik Schacht, and Ernst H.W. Koster. "Attentional biases for angry faces in unipolar depression." In: *Psychological Medicine* 37.3 (2007), pp. 393–402. ISSN: 00332917. DOI: 10.1017/S003329170600910X (Cited on page 22).

[112] Scott A. Langenecker, Linas A. Bieliauskas, Lisa J. Rapport, Jon-Kar Zubieta, Elisabeth A. Wilde, and Stanley Berent. "Face Emotion Perception and Executive Functioning Deficits in Depression." In: *Journal of Clinical and Experimental Neuropsychology* 27.3 (Apr. 2005), pp. 320–333. ISSN: 1380-3395. DOI: 10.1080/13803390490490515720 (Cited on page 22).

[113] Elena S. Mikhailova, Tatjana V. Vladimirova, Andre F. Iznak, Emily J. Tsusulkovskaya, and Nataly V. Sushko. "Abnormal recognition of facial expression of emotions in depressed patients with major depression disorder and schizotypal personality disorder." In: *Biological Psychiatry* 40.8 (Oct. 1996), pp. 697–705. ISSN: 00063223. DOI: 10.1016/0006-3223(96)00032-7 (Cited on pages 22, 23).

[114] Simon A. Surguladze, Andrew W. Young, Carl Senior, Gildas Brébion, Michael J. Travis, and Mary L. Phillips. "Recognition Accuracy and Response Bias to Happy and Sad Facial Expressions in Patients With Major Depression." In: *Neuropsychology* 18.2 (2004), pp. 212–218. ISSN: 08944105. DOI: 10.1037/0894-4105.18.2.212 (Cited on page 22).

[115] Jutta Joormann and Ian H. Gotlib. "Is this happiness I see? Biases in the identification of emotional facial expressions in depression and social phobia." In: *Journal of Abnormal Psychology* 115.4 (Nov. 2006), pp. 705–714. ISSN: 0021843X. DOI: 10.1037/0021-843x.115.4.705 (Cited on page 22).

[116] Jackie K. Gollan, Michael McCloskey, Denada Hoxha, and Emil F. Coccaro. "How do depressed and healthy adults interpret nuanced facial expressions?" In: *Journal of Abnormal Psychology* 119.4 (Nov. 2010), pp. 804–810. ISSN: 19391846. DOI: 10.1037/a0020234 (Cited on pages 22, 177).

[117] Casey Chu, Andrey Zhmoginov, and Mark Sandler. "CycleGAN, a Master of Steganography." In: *arXiv preprint arXiv:1712.02950* (2017). DOI: 10.48550/ARXIV.1712.02950 (Cited on pages 24, 57).

[118] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (June 2018), pp. 8789–8797. ISSN: 10636919. DOI: 10.1109/cvpr.2018.00916. arXiv: 1711.09020 (Cited on pages 24, 57, 165).

[119] Hui Ding, Kumar Sricharan, and Rama Chellappa. "ExprGAN: Facial Expression Editing with Controllable Expression Intensity." In: *arXiv preprint arXiv:1709.03842* (Sept. 12, 2017). arXiv: 1709.03842 [cs.CV] (Cited on pages 24, 57).

[120] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. "GANimation: Anatomically-Aware Facial Animation from a Single Image." In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 835–851. DOI: 10.1007/978-3-030-01249-6_50. arXiv: 1807.09251 [cs.CV] (Cited on pages 24, 57, 165).

[121] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. "Deep Learning for Deepfakes Creation and Detection: A Survey." In: *researchgate.net* (Sept. 25, 2019). arXiv: 1909.11573 [cs.CV] (Cited on pages 24, 113, 163).

[122] Masahiro Mori. "The Uncanny Valley: The Original Essay by Masahiro Mori." In: *IEEE Robotics Automation Magazine* 12.Figure 1 (2012), pp. 1–6 (Cited on pages 24, 114, 163).

[123] Raymond Blanton and Darlene Carbajal. "Not a Girl, Not Yet a Woman." In: *Advances in Media, Entertainment, and the Arts*. IGI Global, 2019, pp. 87–103. DOI: 10.4018/978-1-5225-8535-0.ch006 (Cited on page 24).

[124] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. "Avatar digitization from a single image for real-time rendering." In: *ACM Transactions on Graphics* 36.6 (Nov. 2017), pp. 1–14. DOI: 10.1145/3130800.31310887 (Cited on pages 24, 57).

[125] Alessio Gallucci, Dmitry Znamenskiy, and Milan Petkovic. "Prediction of 3D Body Parts from Face Shape and Anthropometric Measurements." In: *Journal of Image and Graphics* 8.3 (2020), pp. 67–77. DOI: 10.18178/joig.8.3.67-74 (Cited on pages 24, 57).

[126] Yuji Kamashita, Tomomi Sonoda, Yumiko Kamada, Yasuhiro Nishi, and Eiichi Nagaoka. "Reliability, Validity, and Preference of an Original Faces Scale for Assessing the Mood of Patients with Dentures." In: *Prosthodontic Research & Practice* 6.2 (2007), pp. 93–98. ISSN: 1347-7021. DOI: 10.2186/prp.6.93 (Cited on pages 24, 45).

[127] Sharon McKinley, Katherine Coote, and Jane Stein-Parbury. "Development and testing of a Faces Scale for the assessment of anxiety in critically ill patients." In: *Journal of Advanced Nursing* 41.1 (Jan. 2003), pp. 73–79. ISSN: 03092402. DOI: 10.1046/j.1365-2648.2003.02508.x (Cited on page 24).

[128] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. "Facial expressions of emotion are not culturally universal." In: *Proceedings of the National Academy of Sciences* 109.19 (Apr. 2012), pp. 7241–7244. ISSN: 0027-8424. DOI: 10.1073/pnas.1200155109 (Cited on page 25).

[129] Dario Galati, Klaus R. Scherer, and Pio E. Ricci-Bitti. "Voluntary facial expression of emotion: Comparing congenitally blind with normally sighted encoders." In: *Journal of Personality and Social Psychology* 73.6 (1997), pp. 1363–1379. ISSN: 00223514. DOI: 10.1037/0022-3514.73.6.1363 (Cited on page 26).

[130] Felipe González Castro, Joshua G. Kellison, Stephen J. Boyd, and Albert Kopak. "A Methodology for Conducting Integrative Mixed Methods Research and Data Analyses." In: *Journal of Mixed Methods Research* 4.4 (Sept. 2010), pp. 342–360. ISSN: 15586898. DOI: 10.1177/1558689810382916 (Cited on page 28).

[131] Philippe Kruchten. *The rational unified process*. Reading, Mass: Addison-Wesley, 1999. ISBN: 9780201604597 (Cited on pages 28, 37, 38).

[132] David Cohen, Mikael Lindvall, and Patricia Costa. "An Introduction to Agile Methods." In: *Advances in Computers*. Vol. 62. 03. Elsevier, 2004, pp. 1–66. DOI: 10.1016/s0065-2458(03)62001-2 (Cited on page 28).

[133] David L. Morgan. "Pragmatism as a Paradigm for Social Research." In: *Qualitative Inquiry* 20.8 (Feb. 2014), pp. 1045–1053. ISSN: 15527565. DOI: 10.1177/1077800413513733 (Cited on pages 28, 30).

[134] Peter Dalsgaard and Christian Dindler. "Between theory and practice: Bridging concepts in HCI research." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2014, pp. 1635–1644. ISBN: 9781450324731. DOI: 10.1145/2556288.2557342 (Cited on page 28).

[135] Göran Goldkuhl. "Pragmatism vs interpretivism in qualitative information systems research." In: *European Journal of Information Systems* 21.2 (Mar. 2012), pp. 135–146. ISSN: 14769344. DOI: 10.1057/ejis.2011.54 (Cited on pages 28, 30).

[136] Steve Bruce and Herbert Blumer. "Symbolic Interactionism: Perspective and Method." In: *The British Journal of Sociology* 39.2 (June 1988), p. 292. ISSN: 00071315. DOI: 10.2307/590791 (Cited on page 28).

[137] Charles Sanders Peirce. "How to Make our Ideas Clear." In: *Pragmatism*. Vol. 12. January. Routledge, Nov. 2020, pp. 37–49. ISBN: 9780691137056. DOI: 10.4324/9781003061502-4 (Cited on page 28).

[138] Herbert Marcuse and. "Logic, The Theory of Inquiry." In: *Zeitschrift für Sozialforschung* 8.1 (1939), pp. 221–228. DOI: 10.5840/zfs193981/28 (Cited on page 29).

[139] Vernon E. Cronen. "Practical theory, practical art, and the pragmatic-systemic account of inquiry." In: *Communication Theory* 11.1 (Feb. 2001), pp. 14–35. ISSN: 10503293. DOI: 10.1111/j.1468-2885.2001.tb00231.x (Cited on page 29).

[140] John Dewey. "The Development of American Pragmatism." In: *Scientiae Studia* 5.2 (2007), pp. 227–243 (Cited on page 29).

[141] Tim Goles. "The paradigm is dead, the paradigm is dead... long live the paradigm: the legacy of Burrell and Morgan." In: *Omega* 28.3 (June 2000), pp. 249–268. ISSN: 03050483. DOI: 10.1016/S0305-0483(99)00042-0 (Cited on page 29).

[142] Shelia R. Cotten, Abbas Tashakkori, and Charles Teddlie. "Mixed Methodology: Combining Qualitative and Quantitative Approaches." In: *Contemporary Sociology* 28.6 (Nov. 1999), p. 752. ISSN: 00943061. DOI: 10.2307/2655606 (Cited on page 29).

[143] Denise F. Polit and Cheryl Tatano Beck. "Generalization in quantitative and qualitative research: Myths and strategies." In: *International Journal of Nursing Studies* 47.11 (Nov. 2010), pp. 1451–1458. ISSN: 00207489. DOI: 10.1016/j.ijnurstu.2010.06.004 (Cited on page 29).

[144] Debra Wetcher-Hendricks. *Analyzing Quantitative Data*. Wiley, July 13, 2011, p. 414. 416 pp. ISBN: 0470526831 (Cited on page 29).

[145] Robert Alan Stebbins. *Exploratory Research in the Social Sciences*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE PUBN, May 31, 2001. 80 pp. ISBN: 0761923993. DOI: 10.4135/9781412984249 (Cited on page 30).

[146] Michael C Neale, Steven M Boker, Gary Xie, and H Mx Maes. "Statistical modeling." In: *Richmond, VA: Department of Psychiatry, Virginia Commonwealth University* (1999). DOI: 10.13140/RG.2.1.5004.3366 (Cited on page 31).

[147] Manuela Aparicio and Carlos J. Costa. "Data visualization." In: *Communication Design Quarterly* 3.1 (Jan. 2015), pp. 7–11. DOI: 10.1145/2721882.2721883 (Cited on page 31).

[148] Michael Waskom. "seaborn: statistical data visualization." In: *Journal of Open Source Software* 6.60 (Apr. 2021), p. 3021. DOI: 10.21105/joss.03021 (Cited on pages 31, 49, 73, 101).

[149] H. Michael Chung and Paul Gray. "Special Section: Data Mining." In: *Journal of Management Information Systems* 16.1 (June 1999), pp. 11–16. ISSN: 07421222. DOI: 10.1080/07421222.1999.11518231 (Cited on page 31).

[150] Alan Fielding. *Machine Learning Methods for Ecological Applications*. Springer US, Aug. 31, 1999. 284 pp. ISBN: 0412841908 (Cited on page 31).

[151] Jürgen Schmidhuber. "Deep learning in neural networks: An overview." In: *Neural Networks* 61 (Jan. 2015), pp. 85–117. ISSN: 18792782. DOI: 10.1016/j.neunet.2014.09.003. arXiv: 1404.7828 (Cited on page 32).

[152] Melissa Johnston. "Secondary Data Analysis: A Method of which the Time Has Come." In: *Qualitative and Quantitative Methods in Libraries* 3.3 (2017), pp. 619–626. ISSN: 2241-1925 (Cited on page 32).

[153] Florian Lettner and Clemens Holzmann. "Automated and unsupervised user interaction logging as basis for usability evaluation of mobile applications." In: *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia - MoMM '12*. ACM Press, 2012, pp. 118–127. ISBN: 9781450313070. DOI: 10.1145/2428955.2428983 (Cited on page 33).

[154] Delroy L. Paulhus and Simine Vazire. "The self-report method." In: *Handbook of research methods in personality psychology* 1.2007 (2007), pp. 224–239 (Cited on page 33).

[155] Charles J. Cooper, Sharon P. Cooper, Deborah J. Del Junco, Eva M Shipp, Ryan Whitworth, and Sara R Cooper. "Web-based data collection: Detailed methods of a questionnaire and data gathering tool." In: *Epidemiologic Perspectives & Innovations* 3.1 (Jan. 2006), p. 1. ISSN: 17425573. DOI: 10.1186/1742-5573-3-1 (Cited on pages 33, 34).

[156] James R. Lewis. "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use." In: *International Journal of Human-Computer Interaction* 7.1 (Jan. 1995), pp. 57–78. ISSN: 15327590. DOI: 10.1080/10447319509526110 (Cited on pages 34, 71, 72, 96, 98, 99).

[157] Bill Gillham. *The Research Interview*. Bloomsbury Publishing PLC, June 1, 2000. 108 pp. ISBN: 082644797X (Cited on page 34).

[158] D. T. Lee and B. J. Schachter. "Two algorithms for constructing a Delaunay triangulation." In: *International Journal of Computer & Information Sciences* 9.3 (June 1980), pp. 219–242. ISSN: 00917036. DOI: 10.1007/BF00977785 (Cited on pages 34, 59, 120, 177).

[159] Michael Hautus. *Signal Detection Theory*. Vol. 21. Elsevier, 2015, pp. 946–951. ISBN: 9780080970875. DOI: 10.1016/b978-0-08-097086-8.43090-4. arXiv: arXiv:1011.1669v3 (Cited on pages 35, 121).

[160] Mike Cohn. *User Stories Applied*. Addison Wesley, Mar. 31, 2004. ISBN: 0321205685 (Cited on page 36).

[161] Mordecai Ezekiel. "Methods of Correlation Analysis." In: *Revista Mexicana de Sociología* 23.1 (Jan. 1961), p. 309. DOI: 10.2307/3538352 (Cited on page 36).

[162] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. Vol. 326. Wiley, Apr. 1998. DOI: 10.1002/9781118625590 (Cited on page 37).

[163] Virginia Braun and Victoria Clarke. "Thematic analysis." In: *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*. American Psychological Association, Mar. 2012, pp. 57–71. DOI: 10.1037/13620-004 (Cited on page 37).

[164] Victor R. Basili. "The experimental paradigm in software engineering." In: *Experimental Software Engineering Issues: Critical Assessment and Future Directions*. Vol. 706 LNCS. Springer Berlin Heidelberg, 1993, pp. 1–12. ISBN: 9783540570929. DOI: 10.1007/3-540-57092-6_91 (Cited on page 38).

[165] Philippe Kruchten. "Software architecture and agile software development." In: *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - ICSE '10*. Vol. 2. ACM Press, 2010, pp. 497–498. ISBN: 9781605587196. DOI: 10.1145/1810295.1810448 (Cited on page 38).

[166] Paul D Barrows. *D-VAMS Dynamic Visual Analogue Mood Scales* (Cited on page 47).

[167] Andrea Elizabeth Lagotte. "Eliciting Discrete Positive Emotions With Vignettes And Films: A Validation Study." Doctoral dissertation. 2014 (Cited on pages 47, 161).

[168] Friedrich Pukelsheim. "The Three Sigma Rule." In: *The American Statistician* 48.2 (May 1994), pp. 88–91. ISSN: 15372731. DOI: 10.1080/00031305.1994.10476030 (Cited on page 49).

[169] Python Development Team Guido van Rossum. *Python Tutorial: Release 3.6.4*. Amsterdam: 12TH MEDIA SERV, Feb. 3, 2018. 156 pp. ISBN: 1680921606 (Cited on pages 49, 73, 101).

[170] Emily J. Cogsdill, Alexander T. Todorov, Elizabeth S. Spelke, and Mahzarin R. Banaji. "Inferring Character From Faces." In: *Psychological Science* 25.5 (Feb. 2014), pp. 1132–1139. DOI: 10.1177/0956797614523297 (Cited on page 57).

[171] Ross Buck. "Social and emotional functions in facial expression and communication: the readout hypothesis." In: *Biological Psychology* 38.2-3 (Oct. 1994), pp. 95–115. DOI: 10.1016/0301-0511(94)90032-9 (Cited on page 57).

[172] Paul Kruszka et al. "22q11.2 deletion syndrome in diverse populations." In: *American Journal of Medical Genetics Part A* 173.4 (Mar. 2017), pp. 879–888. DOI: 10.1002/ajmg.a.38199 (Cited on page 57).

[173] Kyong I. Chang, K.W. Bowyer, and P.J. Flynn. "Multiple Nose Region Matching for 3D Face Recognition under Varying Facial Expression." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.10 (Oct. 2006), pp. 1695–1700. DOI: 10.1109/tpami.2006.210 (Cited on page 57).

[174] Miquel Alfaras, Vasiliki Tsaknaki, Pedro Sanches, Charles Windlin, Muhammad Umair, Corina Sas, and Kristina Höök. "From Biodata to Somadata." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Apr. 2020, pp. 1–14. DOI: 10.1145/3313831.3376684 (Cited on page 57).

[175] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. "Deep appearance models for face rendering." In: *ACM Transactions on Graphics* 37.4 (Aug. 2018), pp. 1–13. DOI: 10.1145/3197517.3201401 (Cited on page 57).

[176] Magdalena Rychlowska, Rachael E. Jack, Oliver G. B. Garrod, Philippe G. Schyns, Jared D. Martin, and Paula M. Niedenthal. "Functional Smiles: Tools for Love, Sympathy, and War." In: *Psychological Science* 28.9 (July 2017), pp. 1259–1270. DOI: 10.1177/0956797617706082 (Cited on page 57).

[177] Alexey Dosovitskiy, Jost Springenberg, Maxim Tatarchenko, and Thomas Brox. "Learning to Generate Chairs, Tables and Cars with Convolutional Networks." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.4 (2016), pp. 1–1. ISSN: 01628828. DOI: 10.1109/TPAMI.2016.2567384 (Cited on pages 61, 63).

[178] Sheng-Min Shih, Pin-Ju Tien, and Zohar Karnin. "GANMEX: One-vs-One Attributions Guided by GAN-based Counterfactual Explanation Baselines." In: *arXiv* (Nov. 11, 2020), pp. 1–15. ISSN: 23318422. arXiv: 2011.06015 [cs.LG] (Cited on page 63).

[179] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. "Which Training Methods for GANs do actually Converge?" In: 8 (Jan. 2018), pp. 5589–5626. DOI: 10.48550/ARXIV.1801.04406. arXiv: 1801.04406 (Cited on page 63).

[180] TensorFlow Developers. *TensorFlow*. Nov. 2021. DOI: 10.5281/ZENODO.5637331 (Cited on page 63).

[181] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017, pp. 4700–4708. DOI: 10.1109/cvpr.2017.243 (Cited on page 64).

[182] 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)*. IEEE, June 2016, pp. 770–778. DOI: 10.1109/cvpr.2016.90 (Cited on page 64).

[183] Suman Ravuri and Oriol Vinyals. "Classification Accuracy Score for Conditional Generative Models." In: *Advances in Neural Information Processing Systems*. arXiv, 2019, pp. 12268–12279. DOI: 10.48550/ARXIV.1905.10887 (Cited on page 64).

[184] 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." In: *International Journal of Computer Vision* 115.3 (Apr. 2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y (Cited on page 64).

[185] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In: *arXiv preprint arXiv:1412.6980* (2014). DOI: 10.48550/ARXIV.1412.6980 (Cited on page 65).

[186] R. A. Weale. "The Absolute Threshold of Vision." In: *Physiological Reviews* 35.1 (Jan. 1955), pp. 233–246. ISSN: 00319333. DOI: 10.1152/physrev.1955.35.1.233 (Cited on page 79).

[187] David Hough. "Applications of the Proposed IEEE 754 Standard for Floating-Point Arithetic." In: *Computer* 14.3 (Mar. 1981), pp. 70–74. ISSN: 00189162. DOI: 10.1109/c-m.1981.220381 (Cited on page 86).

[188] Paul Ekman, David Matsumoto, and Wallace V. Friesen. "Facial Expression in Affective Disorders." In: *What the Face RevealsBasic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, Apr. 2005, pp. 429–440. DOI: 10.1093/acprof:oso/9780195179644.003.0020 (Cited on page 88).

[189] Antonio Polino, Razvan Pascanu, and Dan Alistarh. "Model compression via distillation and quantization." In: (Feb. 15, 2018). arXiv: 1802.05668 [cs.NE] (Cited on pages 90, 168).

[190] Laura C. Trutoiu, Elizabeth J. Carter, Nancy Pollard, Jeffrey F. Cohn, and Jessica K. Hodgins. "Spatial and Temporal Linearities in Posed and Spontaneous Smiles." In: *ACM Transactions on Applied Perception* 11.3 (Oct. 2014), pp. 1–15. ISSN: 15443965. DOI: 10.1145/2641569 (Cited on pages 92, 166).

[191] J. F. Cohn and K. Schmidt. "The timing of facial motion in posed and spontaneous smiles." In: *Proceedings of the International Conference on Active Media Technology* 2.2 (2003), pp. 57–69. DOI: 10.1142/9789812704313_0005 (Cited on pages 92, 112, 166).

[192] Zhicha Xu, Rongsheng Zhu, Chanchan Shen, Bingren Zhang, Qianqian Gao, You Xu, and Wei Wang. "Selecting pure-emotion materials from the International Affective Picture System (IAPS) by Chinese university students: A study based on intensity-ratings only." In: *Heliyon* 3.8 (Aug.

2017), e00389. ISSN: 24058440. DOI: 10.1016/j.heliyon.2017.e00389 (Cited on page 98).

[193] T Muhr. *Scientific software development's Atlas.TI. The knowledge workbench. Short user's manual*. 1997 (Cited on page 101).

[194] Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. "Introducing the Open Affective Standardized Image Set (OASIS)." In: *Behavior Research Methods* 49.2 (Feb. 2016), pp. 457–470. ISSN: 15543528. DOI: 10.3758/s13428-016-0715-3 (Cited on pages 108, 114).

[195] Philip A. Kragel and Kevin S. LaBar. "Multivariate neural biomarkers of emotional states are categorically distinct." In: *Social Cognitive and Affective Neuroscience* 10.11 (Mar. 2015), pp. 1437–1448. ISSN: 17495024. DOI: 10.1093/scan/nsv032 (Cited on pages 109, 154).

[196] Alan S. Cowen and Dacher Keltner. "What the face displays: Mapping 28 emotions conveyed by naturalistic expression." In: *American Psychologist* 75.3 (Apr. 2020), pp. 349–364. ISSN: 0003066X. DOI: 10.1037/amp0000488 (Cited on pages 111, 115, 118, 165, 177).

[197] James W. Tanaka, Martha D. Kaiser, Sean Butler, and Richard Le Grand. "Mixed emotions: Holistic and analytic perception of facial expressions." In: *Cognition & Emotion* 26.6 (Sept. 2012), pp. 961–977. ISSN: 02699931. DOI: 10.1080/02699931.2011.630933 (Cited on page 111).

[198] Andrew J. Calder, Andrew W. Young, David I. Perrett, Nancy L. Etcoff, and Duncan Rowland. "Categorical Perception of Morphed Facial Expressions." In: *Visual Cognition* 3.2 (June 1996), pp. 81–118. ISSN: 13506285. DOI: 10.1080/713756735 (Cited on page 111).

[199] Karen L. Schmidt, Yanxi Liu, and Jeffrey F. Cohn. "The role of structural facial asymmetry in asymmetry of peak facial expressions." In: *Laterality: Asymmetries of Body, Brain and Cognition* 11.6 (Nov. 2006), pp. 540–561. ISSN: 1357650X. DOI: 10.1080/13576500600832758 (Cited on page 112).

[200] Emily A. Butler and James J. Gross. "Hiding feelings in social contexts; Out of sight is not out of mind." English (US). In: *The Regulation of Emotion*. Lawrence Erlbaum Associates, June 2004, pp. 103–128. ISBN: 1410610896. DOI: 10.4324/9781410610898 (Cited on page 112).

[201] Yuval Nirkin, Yosi Keller, and Tal Hassner. "FSGAN: Subject Agnostic Face Swapping and Reenactment." In: (Aug. 16, 2019). arXiv: 1908.05932 (Cited on page 113).

[202] Kristopher J. Blom, Anna I. Bellido Rivas, Xenxo Alvarez, Ozan Cetinaslan, Bruno Oliveira, Verónica Orvalho, and Mel Slater. "Achieving Participant Acceptance of their Avatars." In: *Presence: Teleoperators and Virtual Environments* 23.3 (Oct. 2014), pp. 287–299. DOI: 10.1162/PRES_a_00194 (Cited on pages 113, 164).

[203] Shichuan Du, Yong Tao, and Aleix M. Martinez. "Compound facial expressions of emotion." In: *Proceedings of the National Academy of Sciences* 111.15 (Mar. 2014), E1454–E1462. ISSN: 10916490. DOI: 10.1073/pnas.1322355111 (Cited on page 118).

[204] Sofia Mizuho Miura Sugayama, Cláudio Leone, Maria de Lourdes Lopes Ferrari Chauffaille, Thelma Suely Okay, and Chong Ae Kim. "Williams Syndrome: development of a new scoring system for clinical diagnosis." In: *Clinics* 62.2 (2007), pp. 159–166. ISSN: 18075932. DOI: 10.1590/s1807-59322007000200011 (Cited on page 118).

[205] Naomi Jane Scott, Robin Stewart Samuel Kramer, Alex Lee Jones, and Robert Ward. "Facial cues to depressive symptoms and their associated personality attributions." In: *Psychiatry Research* 208.1 (June 2013), pp. 47–53. ISSN: 01651781. DOI: 10.1016/j.psychres.2013.02.027 (Cited on pages 118, 170).

[206] Alexander Todorov, Peter Mende-Siedlecki, and Ron Dotsch. "Social judgments from faces." In: *Current Opinion in Neurobiology* 23.3 (June 2013), pp. 373–380. ISSN: 09594388. DOI: 10.1016/j.conb.2012.12.010 (Cited on page 119).

[207] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk." In: *Behavior Research Methods* 46.4 (Dec. 2013), pp. 1023–1031. ISSN: 15543528. DOI: 10.3758/s13428-013-0434-y (Cited on page 119).

[208] Gabriele Paolacci and Jesse Chandler. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." In: *Current Directions in Psychological Science* 23.3 (June 2014), pp. 184–188. ISSN: 14678721. DOI: 10.1177/0963721414531598 (Cited on page 119).

[209] Ron Dotsch and Alexander Todorov. "Reverse Correlating Social Face Perception." In: *Social Psychological and Personality Science* 3.5 (Dec. 2011), pp. 562–571. ISSN: 19485506. DOI: 10.1177/1948550611430272 (Cited on pages 119, 121).

[210] Richard M. Rose, Davida Y. Teller, and Paula Rendleman. "Statistical properties of staircase estimates." In: *Perception & Psychophysics* 8.4 (July 1970), pp. 199–204. ISSN: 00315117. DOI: 10.3758/BF03210205 (Cited on page 122).

[211] A.H. Robinson and C. Cherry. "Results of a prototype television bandwidth compression scheme." In: *Proceedings of the IEEE* 55.3 (1967), pp. 356–364. ISSN: 15582256. DOI: 10.1109/proc.1967.5493 (Cited on pages 125, 138).

[212] P. Deutsch. *DEFLATE Compressed Data Format Specification version 1.3*. Tech. rep. May 1996, pp. 1–15. DOI: 10.17487/rfc1951 (Cited on pages 125, 138).

[213] Eric R. Ziegel. "Understanding Statistical Process Control." In: *Technometrics* 35.1 (Feb. 1993), pp. 101–102. ISSN: 00401706. DOI: 10.1080/00401706.1993.10485025 (Cited on page 126).

[214] Jessica Taubert, Deborah Apthorp, David Aagten-Murphy, and David Alais. "The role of holistic processing in face perception: Evidence from the face inversion effect." In: *Vision Research* 51.11 (June 2011), pp. 1273–1278. ISSN: 00426989. DOI: 10.1016/j.visres.2011.04.002 (Cited on pages 127, 129).

[215] Winnie W Leung and Jack J Blanchard. "Experience and Expression of Emotion in Social Anhedonia: An Examination of Film-Induced Social Affiliative State in Schizotypy." Doctoral dissertation. Aug. 2006 (Cited on pages 132, 170).

[216] Winnie W. Leung, Shannon M. Couture, Jack J. Blanchard, Stephanie Lin, and Katiah Llerena. "Is social anhedonia related to emotional responsivity and expressivity? A laboratory study in women." In: *Schizophrenia Research* 124.1-3 (Dec. 2010), pp. 66–73. ISSN: 09209964. DOI: 10.1016/j.schres.2010.06.012 (Cited on pages 132, 170).

[217] Min Ah Kim, Eun Joo Kim, Byung Young Kang, and Hae Kwang Lee. "The Effects of Sleep Deprivation on the Biophysical Properties of Facial Skin." In: *Journal of Cosmetics, Dermatological Sciences and Applications* 07.01 (Jan. 2017), pp. 34–47. ISSN: 2161-4105. DOI: 10.4236/jcdsa.2017.71004 (Cited on pages 132, 170).

[218] L Ridsdale, A Evans, W Jerrett, S Mandalia, K Osler, and H Vora. "Patients who consult with tiredness: Frequency of consultation, perceived causes of tiredness and its association with psychological distress." In: *British Journal of General Practice* 44.386 (1994), pp. 413–416. ISSN: 09601643 (Cited on page 133).

[219] Eileen Kennedy. "The Obesity Crisis." In: *The Road to Good Nutrition*. Basel: KARGER, 2013, pp. 51–63. DOI: 10.1159/000355993 (Cited on page 133).

[220] Brian Harris, John Young, and Bill Hughes. "Appetite and weight change in patients presenting with depressive illness." In: *Journal of Affective Disorders* 6.3-4 (June 1984), pp. 331–339. ISSN: 01650327. DOI: 10.1016/s0165-0327(84)80011-7 (Cited on page 133).

[221] Paul Rozin and Adam B. Cohen. "High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans." In: *Emotion* 3.1 (2003), pp. 68–75. ISSN: 15283542. DOI: 10.1037/1528-3542.3.1.68 (Cited on pages 133, 171).

[222] Euan Thompson. "Hamilton Rating Scale for Anxiety (HAM-A)." In: *Occupational Medicine* 65.7 (Sept. 2015), pp. 601–601. ISSN: 14718405. DOI: 10.1093/occmed/kqv054 (Cited on page 133).

[223] Sela Kleiman and Nicholas O. Rule. "Detecting Suicidality From Facial Appearance." In: *Social Psychological and Personality Science* 4.4 (Nov. 2012), pp. 453–460. ISSN: 19485506. DOI: 10.1177/1948550612466115 (Cited on pages 134, 170–172).

[224] Cherie L Marvel and Sergio Paradiso. "Cognitive and neurological impairment in mood disorders." In: *Psychiatric Clinics of North America* 27.1 (Mar. 2004), pp. 19–36. ISSN: 0193953X. DOI: 10.1016/s0193-953x(03)00106-0 (Cited on pages 135, 171).

[225] Sahin Naqvi, Yoeri Sleyp, Hanne Hoskens, Karlijne Indencleef, Jeffrey P. Spence, Rose Bruffaerts, Ahmed Radwan, Ryan J. Eller, Stephen Richmond, Mark D. Shriver, John R. Shaffer, Seth M. Weinberg, Susan Walsh, James Thompson, Jonathan K. Pritchard, Stefan Sunaert, Hilde Peeters, Joanna Wysocka, and Peter Claes. "Shared heritability of face and brain shape distinct from cognitive traits." In: (Aug. 2020). DOI: 10.1101/2020.08.29.269258 (Cited on pages 135, 171).

[226] Joan C. Borod, Cornelia Santschi Haywood, and Elissa Koff. "Neuropsychological aspects of facial asymmetry during emotional expression: A review of the normal adult literature." In: *Neuropsychology Review* 7.1 (Mar. 1997), pp. 41–60. ISSN: 10407308. DOI: 10.1007/bf02876972 (Cited on page 136).

[227] Harold A. Sackeim, Ruben C. Gur, and Marcel C. Saucy. "Emotions Are Expressed More Intensely on the Left Side of the Face." In: *Science* 202.4366 (Oct. 1978), pp. 434–436. ISSN: 00368075. DOI: 10.1126/science.705335 (Cited on pages 136, 171).

[228] Margit Lillemaa. "User-centered design." In: *Encyclopedia of Human-Computer Interaction* (2004) (Cited on page 147).

[229] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers. "Personal tracking as lived informatics." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. May. ACM, Apr. 2014, pp. 1163–1172. ISBN: 9781450324731. DOI: 10.1145/2556288.2557039 (Cited on page 156).

[230] Maxime Taquet, Jordi Quoidbach, Yves-Alexandre de Montjoye, Martin Desseilles, and James J. Gross. "Hedonism and the choice of everyday activities." In: *Proceedings of the National Academy of Sciences* 113.35 (Aug. 2016), pp. 9769–9773. ISSN: 10916490. DOI: 10.1073/pnas.1519998113 (Cited on page 156).

[231] Mohammed T. Masud, Mohammed A. Mamun, K. Thapa, D.H. Lee, Mark D. Griffiths, and S.-H. Yang. "Unobtrusive monitoring of behavior and movement patterns to detect clinical depression severity level via smartphone." In: *Journal of Biomedical Informatics* 103.February 2019 (Mar. 2020), p. 103371. ISSN: 15320464. DOI: 10.1016/j.jbi.2019.103371 (Cited on page 156).

[232] Claire McCallum, John Rooksby, and Cindy M Gray. "Evaluating the Impact of Physical Activity Apps and Wearables: Interdisciplinary Review." In: *JMIR mHealth and uHealth* 6.3 (Mar. 2018), e58. ISSN: 22915222. DOI: 10.2196/mhealth.9054 (Cited on page 156).

[233] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study." In: *Journal of Medical Internet Research* 17.7 (July 2015), e175. ISSN: 14388871. DOI: 10.2196/jmir.4273 (Cited on page 156).

[234] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. "Echoes from the past: How technology mediated reflection improves well-being." In: *Conference on Human Factors in Computing Systems - Proceedings* (Apr. 2013), pp. 1071–1080. DOI: 10.1145/2470654.2466137 (Cited on page 156).

[235] S. P. Ahmed, L. H. Somerville, and C. L. Sebastian. "Using temporal distancing to regulate emotion in adolescence: modulation by reactive aggression." In: *Cognition and Emotion* 32.4 (Aug. 2017), pp. 812–826. ISSN: 14640600. DOI: 10.1080/02699931.2017.1358698 (Cited on page 156).

[236] John P. Powers and Kevin S. LaBar. "Regulating emotion through distancing: A taxonomy, neurocognitive model, and supporting meta-analysis." In: *Neuroscience & Biobehavioral Reviews* 96.April 2018 (Jan. 2019), pp. 155–173. ISSN: 18737528. DOI: 10.1016/j.neubiorev.2018.04.023 (Cited on page 156).

[237] Philipp Kanske, Janine Heissler, Sandra Schönfelder, André Bongers, and Michèle Wessa. "How to Regulate Emotion? Neural Networks for Reappraisal and Distraction." In: *Cerebral Cortex* 21.6 (Nov. 2010), pp. 1379–1388. ISSN: 10473211. DOI: 10.1093/cercor/bhq216 (Cited on page 156).

[238] Jenna R. Carl, David P. Soskin, Caroline Kerns, and David H. Barlow. "Positive emotion regulation in emotional disorders: A theoretical review." In: *Clinical Psychology Review* 33.3 (Apr. 2013), pp. 343–360. ISSN: 02727358. DOI: 10.1016/j.cpr.2013.01.003 (Cited on page 156).

[239] John Rooksby, Parvin Asadzadeh, Alistair Morrison, Claire McCallum, Cindy Gray, and Matthew Chalmers. "Implementing ethics for a mobile app deployment." In: *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI '16*. Vol. 51. September. New York,

New York, USA: ACM Press, 2016, pp. 406–415. ISBN: 9781450346184. DOI: 10.1145/3010915.3010919 (Cited on page 157).

[240] Rolandos Alexandros Potamias, Jiali Zheng, Stylianos Ploumpis, Giorgos Bouritsas, Evangelos Ververas, and Stefanos Zafeiriou. "Learning to Generate Customized Dynamic 3D Facial Expressions." In: *Computer Vision – ECCV 2020*. Vol. 12374 LNCS. Springer International Publishing, July 2020, pp. 278–294. ISBN: 9783030585259. DOI: 10.1007/978-3-030-58526-6_17. arXiv: 2007.09805 (Cited on page 163).

[241] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. "ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement." In: (Apr. 6, 2021). arXiv: 2104.02699 (Cited on page 165).

[242] Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. "Realistic Dynamic Facial Textures from a Single Image Using GANs." In: *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob (Oct. 2017), pp. 5439–5448. ISSN: 15505499. DOI: 10.1109/iccv.2017.580 (Cited on page 166).

[243] Naima Otberdout, Mohamed Daoudi, Anis Kacem, Lahoucine Ballihi, and Stefano Berretti. "Dynamic Facial Expression Generation on Hilbert Hypersphere With Conditional Wasserstein Generative Adversarial Nets." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.2 (Feb. 2022), pp. 848–863. ISSN: 0162-8828. DOI: 10.1109/tpami.2020.3002500. arXiv: 1907.10087 (Cited on page 166).

[244] Yong Zhao, Le Yang, Ercheng Pei, Meshia Cédric Oveneke, Mitchel Alioscha-Perez, Longfei Li, Dongmei Jiang, and Hichem Sahli. "Action Unit Driven Facial Expression Synthesis from a Single Image with Patch Attentive GAN." In: *Computer Graphics Forum* 40.6 (Mar. 2021), pp. 47–61. ISSN: 14678659. DOI: 10.1111/cgf.14202 (Cited on page 166).

[245] Ziheng Wang, Shangfei Wang, and Qiang Ji. "Capturing Complex Spatio-temporal Relations among Facial Muscles for Facial Expression Recognition." In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (June 2013), pp. 3422–3429. ISSN: 10636919. DOI: 10.1109/CVPR.2013.439 (Cited on page 166).

[246] Epic Games. *Unreal Engine* (Cited on pages 168, 169).

[247] Carole-Jean Wu et al. "Machine Learning at Facebook: Understanding Inference at the Edge." In: *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, Feb. 2019, pp. 331–344. ISBN: 9781728114446. DOI: 10.1109/hpca.2019.00048 (Cited on page 168).

[248] Michael Zhu and Suyog Gupta. "To prune, or not to prune: exploring the efficacy of pruning for model compression." In: (Oct. 5, 2017). arXiv: 1710.01878 (Cited on page 168).

[249] Carissa Wilkes, Rob Kydd, Mark Sagar, and Elizabeth Broadbent. "Upright posture improves affect and fatigue in people with depressive symptoms." In: *Journal of Behavior Therapy and Experimental Psychiatry* 54 (Mar. 2017), pp. 143–149. ISSN: 18737943. DOI: 10.1016/j.jbtep.2016.07.015 (Cited on page 169).

[250] Zoltán Rihmer, Erika Szádóczky, János Füredi, Kitty Kiss, and Zsuzsa Papp. "Anxiety disorders comorbidity in bipolar I, bipolar II and unipolar major depression: results from a population-based study in Hungary." In: *Journal of Affective Disorders* 67.1-3 (Dec. 2001), pp. 175–179. ISSN: 01650327. DOI: 10.1016/s0165-0327(01)00309-3 (Cited on page 173).

[251] Paula M. Niedenthal. "Embodying Emotion." In: *Science* 316.5827 (May 2007), pp. 1002–1005. ISSN: 00368075. DOI: 10.1126/science.1136930 (Cited on page 177).

[252] Johannes Michalak, Jan Burg, and Thomas Heidenreich. "Don't Forget Your Body: Mindfulness, Embodiment, and the Treatment of Depression." In: *Mindfulness* 3.3 (May 2012), pp. 190–199. ISSN: 18688527. DOI: 10.1007/s12671-012-0107-4 (Cited on page 177).

[253] Johannes Michalak, Nikolaus F. Troje, Julia Fischer, Patrick Vollmar, Thomas Heidenreich, and Dietmar Schulte. "Embodiment of Sadness and Depression—Gait Patterns Associated With Dysphoric Mood." In: *Psychosomatic Medicine* 71.5 (June 2009), pp. 580–587. ISSN: 00333174. DOI: 10.1097/psy.0b013e3181a2515c (Cited on page 177).

[254] Corina Sas, Kobi Hartley, and Muhammad Umair. "ManneqKit cards: A kinesthetic empathic design tool communicating depression experiences." In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. ACM, July 2020, pp. 1479–1493. ISBN: 9781450369749. DOI: 10.1145/3357236.3395556 (Cited on page 177).