DISS. ETH NO. 27843

Multimodal Affective State Prediction in Mobile Settings

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
Rafael Fjodor Wampfler
Master of Science ETH in Computer Science, ETH Zurich
born on 16.06.1986
citizen of Diemtigen (BE), Switzerland

accepted on the recommendation of
Prof. Dr. Markus Gross, examiner
Prof. Dr. Christian Holz, co-examiner
Prof. Dr. Otmar Hilliges, co-examiner
Dr. Shamsi Iqbal, co-examiner

2021

Abstract

Gaining awareness of affective states enables leveraging emotional information as additional context in order to design emotionally sentient systems. Applications of such systems are manifold. For example, the learning gain can be increased in educational settings by incorporating targeted interventions that are capable of adjusting to affective states of students. Another application consists of enabling smartphones to support enriched interactions that are sensitive to the user's contexts. To accomplish the prediction of affective states in different contexts, multimodal data tailored to the domain need to be collected and adequately modeled. Research on such affective models mainly focused on expensive and stationary lab devices that are not well suited for everyday use, but recently, lightweight data collection in mobile settings gained interest. In this thesis, we present data-driven models for the prediction of affective states. We focus on models relying on lightweight data collection tailored to mobile settings. We further discuss the protection of privacy and the usability in real-world environments of the different data modalities.

First, we propose a pipeline for affective state prediction based on front camera recordings (i.e., action units, eye gaze, eye blinks, and head movement) during math-solving tasks (active) and emotional stimuli from pictures (passive) shown on a tablet. Using data from a study with 88 participants, we demonstrate that our setup provides comparable performance for affective state prediction to recordings taken with an external and more obtrusive GoPro camera. In addition, we present a neural inpainting pipeline and techniques for image reconstruction of partially occluded and skewed faces. In combination with our novel hardware setup consisting of a cheap and unobtrusive mirror construction, the neural inpainting pipeline improves the visibility of the face in recordings of built-in cameras of mobile devices.

Second, we present an automated pipeline capable of accurately predicting (AUC up to 0.86) the affective states of users solving tablet-based math tasks using signals from low-cost mobile biosensors. In addition, we show that we can achieve a similar classification performance (AUC up to 0.84) by only using handwriting data recorded from a stylus while users solved the math tasks. Given the emerging digitization of classrooms and increased reliance on tablets as teaching tools, we demonstrate that stylus data may be a viable alternative to biosensors for the prediction of affective states in educational settings.

Third, we propose a system that analyzes the user's text typing behavior on smartphones using a semi-supervised deep learning pipeline for predicting affective states. Using a

data collection study in a laboratory setting with 70 participants on text conversations designed to trigger different affective responses, we developed a variational autoencoder to learn efficient feature embeddings of two-dimensional heat maps generated from touch data while participants engaged in these conversations. Using the learned embedding in a cross-validated analysis, our system predicts affective states with an AUC of up to 0.84. We demonstrate the feasibility of our approach to accurately predict affective states based only on touch data collected on smartphones.

Fourth, we present an approach to expand affective state prediction to smartphone usage in the wild. We developed two-dimensional heat maps generated from keystroke and smartphone sensor data. Using data collected in the wild from 82 participants over 10 weeks, we demonstrate that by using a convolutional neural network we can achieve an AUC of up to 0.85 for the prediction of affective states. We also show that using less privacy-invasive sensor data alone, a similar performance (AUC up to 0.83) can be achieved. In addition, by personalizing the model to the user, the performance can be increased by up to 0.07 AUC. We exemplify the usability of our model for the prediction of affective states in real-world environments based on readily available smartphone data.

Finally, we describe two widgets for a compact and unobtrusive visualization of users' affective states on mobile devices. We test the widgets on intuitiveness and understandability based on a user study with 644 participants.

We conclude with a discussion of the advantages and limitations of our methods and possible future work. As we believe that the knowledge of affective states will become crucial for a variety of systems and in different domains in the next decade, we hope that our work provides an important contribution in such a direction.

Zusammenfassung

Das Erkennen affektiver Zustände ermöglicht es, emotionale Informationen als zusätzlichen Kontext zu nutzen, um emotional empfindsame Systeme zu entwerfen. Die Anwendungen solcher Systeme sind vielfältig. Zum Beispiel kann das Lernen in Schulen durch gezielte Interventionen, die sich an die affektiven Zustände von Schülern anpassen, verbessert werden. Eine andere Anwendung besteht aus Smartphone-Interaktionen, die auf den Kontext des Benutzers reagieren. Um die Vorhersage von affektiven Zuständen in verschiedenen Kontexten zu ermöglichen, müssen multimodale Daten, die auf die Domäne zugeschnitten sind, gesammelt und adäquat modelliert werden. Die Forschung zu solchen affektiven Modellen hat sich hauptsächlich auf teure und stationäre Laborgeräte konzentriert, die für den alltäglichen Gebrauch nicht gut geeignet sind. In letzter Zeit hat die leichtgewichtige Datenerfassung in mobilen Umgebungen an Interesse gewonnen. In dieser Arbeit stellen wir datengetriebene Modelle für die Vorhersage von affektiven Zuständen vor. Wir konzentrieren uns auf Modelle, die auf einer leichtgewichtigen Datenerfassung beruhen und auf mobile Umgebungen zugeschnitten sind. Außerdem diskutieren wir den Schutz der Privatsphäre und die Verwendbarkeit der verschiedenen Datenmodalitäten in realen Umgebungen.

Als erstes schlagen wir eine Pipeline zur Vorhersage des affektiven Zustands vor, die auf Aufnahmen der Frontkamera (d.h. Bewegungseinheiten, Blick, Augenblinzeln und Kopfbewegung) während dem Lösen mathematischer Aufgaben (aktiv) und dem Betrachten emotionaler Stimuli von Bildern (passiv), die auf einem Tablet gezeigt wurden, basiert. Anhand von Daten aus einer Studie mit 88 Teilnehmern zeigen wir, dass unser Setup eine vergleichbare Leistung für die Vorhersage des affektiven Zustands bietet wie Aufnahmen mit einer externen und sichtbaren GoPro-Kamera. Darüber hinaus stellen wir eine neuronale Inpainting-Pipeline und Techniken zur Bildrekonstruktion von teilweise verdeckten und schiefen Gesichtern vor. In Kombination mit unserer neuartigen, günstigen und unauffälligen Spiegelkonstruktion, verbessert die neuronale Inpainting-Pipeline die Sichtbarkeit des Gesichts in Aufnahmen der eingebauten Kameras von mobilen Geräten.

Zweitens stellen wir eine automatisierte Pipeline vor, die in der Lage ist, die affektiven Zustände von Nutzern, die Tablet-basierte Matheaufgaben lösen, anhand von Signalen kostengünstiger mobiler Biosensoren genau vorherzusagen (AUC bis zu 0,86). Darüber hinaus zeigen wir, dass wir eine ähnliche Klassifikationsleistung (AUC bis zu 0,84) erreichen können, wenn wir nur die Handschriftdaten verwenden, die von einem Stift aufgezeichnet

wurden, während der Benutzer die Matheaufgaben löste. In Anbetracht der zunehmenden Digitalisierung von Klassenzimmern und dem verstärkten Einsatz von Tablets als Lehrmittel zeigen wir, dass Stiftdaten eine brauchbare Alternative zu Biosensoren für die Vorhersage von affektiven Zuständen in Schulen sein können.

Drittens schlagen wir ein System vor, welches das Tippverhalten von Text auf Smartphones mit einer teilüberwachten Deep-Learning-Pipeline zur Vorhersage affektiver Zustände analysiert. Anhand einer Datenerhebungsstudie in einer Laborumgebung mit 70 Teilnehmern basierend auf Textkonversationen, die unterschiedliche affektive Reaktionen auslösen sollten, entwickelten wir einen Variational Autoencoder, um effiziente Datenrepräsentation von zweidimensionalen Heat Maps zu lernen, die aus Touchscreen Daten generiert wurden, während die Teilnehmer Konversationen führten. Unter Verwendung der gelernten Einbettung in einer kreuzvalidierten Analyse sagt unser System affektive Zustände mit einem AUC von bis zu 0,84 voraus. Wir demonstrieren die Machbarkeit unseres Ansatzes zur genauen Vorhersage affektiver Zustände, welcher nur auf mit Smartphones gesammelten Touchscreen Daten basiert.

Viertens präsentieren wir einen Ansatz, um die Vorhersage affektiver Zustände auf die Smartphone-Nutzung im Alltag auszuweiten. Wir entwickelten zweidimensionale Heat Maps, welche aus Tastendruck- und Smartphone-Sensordaten generiert werden. Anhand von Daten, die im Alltag von 82 Teilnehmern über einen Zeitraum von 10 Wochen gesammelt wurden, zeigen wir, dass wir mit einem gefalteten neuronalen Netzwerk einen AUC von bis zu 0,85 für die Vorhersage von affektiven Zuständen erreichen können. Wir zeigen auch, dass allein durch die Verwendung von weniger datenschutzrelevanten Sensordaten eine ähnliche Leistung (AUC bis zu 0,83) erzielt werden kann. Darüber hinaus kann durch benutzerspezifische Personalisierung des Modells die Leistung um bis zu 0,07 AUC gesteigert werden. Wir veranschaulichen die Anwendbarkeit unseres Modells für die Vorhersage von affektiven Zuständen in realen Umgebungen auf der Basis von Smartphone-Daten.

Schließlich beschreiben wir zwei Widgets für eine kompakte und unauffällige Visualisierung der affektiven Zustände von Benutzern auf mobilen Geräten. Wir testen die Widgets auf Intuitivität und Verständlichkeit anhand einer Nutzerstudie mit 644 Teilnehmern.

Wir schließen mit einer Diskussion über die Vorteile und Grenzen unserer Methoden und möglichen zukünftigen Arbeiten. Da wir glauben, dass das Wissen über affektive Zustände im nächsten Jahrzehnt für eine Vielzahl von Systemen und in verschiedenen Domänen entscheidend sein wird, hoffen wir, dass unsere Arbeit einen wichtigen Beitrag in diese Richtung leistet.

Acknowledgments

First of all, I would like to express my deepest gratitude to Prof. Dr. Markus Gross for letting me work on this exciting topic at the Computer Graphics Laboratory. I appreciated his vision in exploring domains that are at the edge of computer graphics. His unconditional support, guidance, and help throughout my Ph.D. were invaluable. The very inviting atmosphere made me feel welcome in the team right from the start.

Next, I would like to thank my co-advisors Dr. Barbara Solenthaler and Dr. Severin Klingler for their support, guidance, and fruitful discussions during my Ph.D. The time they invested in our weekly meetings was invaluable. Without their support, my Ph.D. would not have been possible.

Furthermore, I would like to thank all my collaborators for their support during my journey. Special thanks go to Prof. Dr. Christian Holz for the helpful discussions and guidance. I am very thankful for his help with the Institutional Review Board applications, setting up the experiments, and designing and evaluating the models, even outside usual office hours. I would like to thank Prof. Dr. Victor Schinazi for the offered expertise in study design and the tremendous contribution in writing the papers. I am also grateful to Prof. Dr. Tobias Günther, who was always available for discussions and providing helpful input. Further, I would like to thank Dr. Romann Weber for the insightful discussions. I also thank all the participants for their help in collecting high-quality data in the experiments.

I am also thankful to all current and former members of CGL, IGL, MTC, GTC, and DRZ for making the time during my Ph.D. that enjoyable and creating a nice working environment. I would like to extend my thanks to all my colleagues for their support during the ups and downs over the past four years.

Finally, I am deeply thankful to my family, especially my parents, Anita and Willi, for their unconditional support and trust. A special thanks goes to Franziska for her love, support, and understanding for sometimes very long working hours and lack of time.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In this thesis, we investigate different data modalities for predicting affective states for the masses. Awareness of the affective state of users can enhance the quality of the interaction making systems more usable, enjoyable, and effective for the users. Such affect-aware systems are useful in different domains such as education and health. For example, a learning environment that can detect and react to the frustration of the students can increase motivation and learning gain by adapting the environment (e.g., the difficulty level) [Sidney et al., 2005]. Recognizing the affective state of a person can also help with the treatment of mental health problems such as depression (e.g., as part of a therapeutic chatbot) [Riva et al., 2015].

The proliferation of smartphone and sensor-based technologies enables systems to recognize and process human affective states in real-world situations and in real-time by harnessing the properties of these mobile and sensing technologies. As such, our data-driven models presented in this thesis are based on novel data representations and features extracted from different data modalities (i.e., video data, biosensor data, stylus data, and smartphone data). We then use machine learning techniques to predict the affective states. Our models enable innovative applications in different fields. Two of the key challenges of designing such affect-aware models are privacy protection and real-world usability. Besides properties of the data modalities (e.g., intrusiveness and motion artifacts), privacy protection also has an implication on the usefulness of systems in real-world settings. People are usually sensitive when personal data is captured and depending on the degree of privacy invasion such systems are disliked and real-world applicability is degraded. Thus, we discuss also implications on the privacy and usage in the real world of our models and underlying data modalities.

1.1 Affective Computing

The main goal of affective computing is the development of systems that recognize and respond to affective states of users [Picard et al., 2004]. Affective states comprise two main types: emotions and moods [Politou et al., 2017]. Emotions are short in duration (ranging from seconds to minutes), are directed at something, and are triggered by events (e.g., seeing a bear). In contrast, moods last longer than emotions (from hours to days), are not directed towards a particular object, and have combined causes. In this thesis, we focus on emotions as it was shown that emotions have a big impact on different domains, such as education [Baker et al., 2012; Csikszentmihalyi, 2008; Miserandino, 1996] and health [Breazeal, 2011].

There exist three types of affective computing applications [Picard, 2000]. The first type comprises systems recognizing the emotions of the users. The second type is related to systems expressing emotions (e.g., an animated conversational agent). Finally, the third type comprises systems that calculate (i.e., feel) an emotion. Based on these three types of applications, an affective component consisting of emotion recognition, emotion calculation, and emotion expression can be defined (see Figure 1.1) [Pudane et al., 2019]. Emotion-aware systems vary in type and depending on the specific application they usually include (a subset of) these components. A system can still perform well if it has just a few functional blocks. For example, intelligent tutoring systems (i.e., learning environments supporting individual learning by adapting the learning process to the user) often adapt to a user's emotions by recognizing the emotions and expressing them [Pudāne and Lavendelis, 2017].

Emotion recognition. Emotion recognition is conducted by exploiting an emotional signal from different data modalities. Such modalities encompass sensors that do not require physical contact (e.g., video cameras) and sensors requiring physical contact with the human body (e.g., biosensors). In this thesis, we focus on this emotion recognition component as it is a crucial part of every emotion-aware system because such a system can never respond to the affective states of users without recognizing their affective states. The emotion recognition component is a prerequisite for the emotion calculation and emotion expression components. It is noteworthy that emotion recognition is a very challenging problem due to the fuzziness and subjectivity in the expression and experience of emotions [Calvo and D'Mello, 2010].

Emotion expression. The emotion expression component enables the system to express and visualize emotions. For example, such functionality can be achieved through affective conversational agents, widgets carrying emotional information, and changes in music, color, and lighting. Emotions can be expressed based on the recognized emotion of the user or after being processed by the rational reasoning component, i.e., taking into account the state of the system. To express the system's emotional state, the recognized emotion must first be processed by the emotion

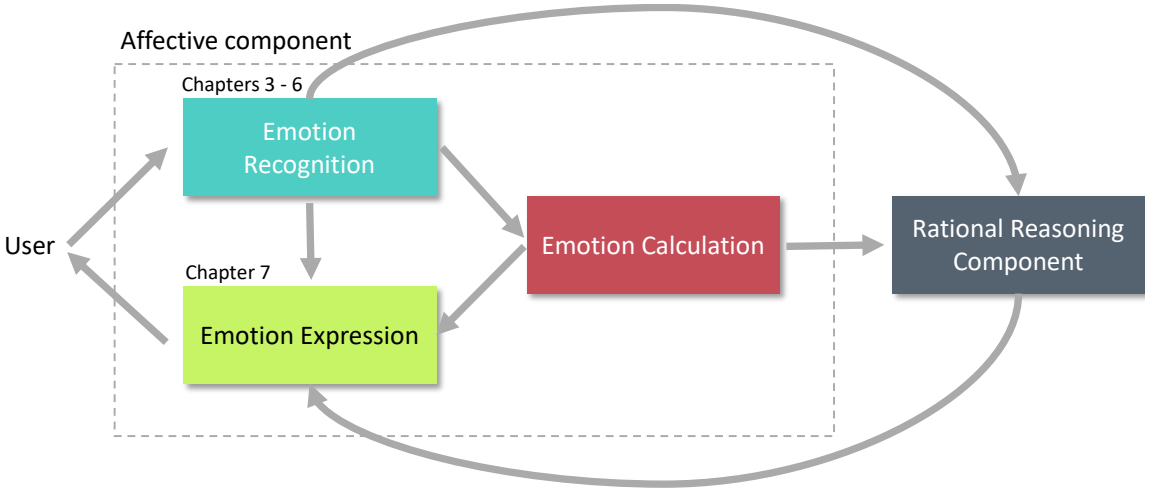


Figure 1.1: *Outline of an emotion-aware system (adapted from [Pudane et al., 2019]). The affective processes consist of emotion recognition, emotion expression (emotion feedback), and emotion calculation. The rational reasoning component adapts the state of the system based on the recognized emotion and the emotional state of the system.*

calculation component. In this thesis, we will touch on the visualization of emotions in Chapter 7 by presenting two intuitive widgets for visualizing emotions and communicating emotions to others.

Emotion calculation. In contrast to pure emotion recognition, emotion calculation enables the system to have and feel emotions by enriching the system with emotion mechanisms similar to those of humans. Emotions are generated by evaluating events in relation to the system's expectations, needs, and goals. This is consistent with the appraisal theory of emotions, which relates subjective emotional experience to the appraisal of stimuli (detailed in Section 1.1.2). For example, emotion calculation is an important part of social robots to enable the robot to have its own emotions [Pudane et al., 2019].

Rational reasoning component. This component can receive direct input from the emotion recognition and emotion calculation component. Its purpose is to adapt the state of the system (e.g., the user model) based on the recognized emotion of the user and the emotional state of the system (i.e., the emotion the system feels). The updated state can then be passed to the emotion expression component for visualization purposes.

1.1.1 Applications

The ability to predict affective states has a broad range of applications. In the following, we detail possible applications in mental health, awareness, and education that can benefit from affective state predictions.

Mental health. Emotions are closely related to physiological and mental health [Breazeal, 2011]. Thus, recognizing a person's emotion can help with the treatment of health problems by either calling a caretaker or by the intervention of the system with the user itself [Riva et al., 2015]. For example, Woebot [2021] is one of many therapeutic chatbots available for Android and iOS devices. Using methods from cognitive behavioral therapy, Woebot aims to increase the overall mood of users and has been shown to reduce symptoms of depression and anxiety [Fitzpatrick et al., 2017]. Woebot uses predefined questions to adequately adapt the conversation to the mood of the user, inferring the mood directly from the chat messages provided by the user. The bot can adapt the responses to the users' changing affective state. Similarly, tracking the intensity and duration of positive and negative valence in real-time using a heart rate monitor enables detecting depression at its early stages [Leon et al., 2011]. Other applications could also benefit from affective predictions, such as customer service applications (e.g., Zendesk [2021]).

Awareness. Knowledge about affective states can be leveraged to increase self-awareness and to convey awareness of affective states to others. Textual or graphical elements can be used to make users aware of their affective states. Such feedback can make users think about their affective state and encourage them to take regulatory actions (e.g., taking a break), which can have an impact on the user's well-being [Lane et al., 2012]. Furthermore, it may allow the user to foster self-regulation, detect potential stress causes, and adjust daily routines based on the extracted information. If the user agrees, these affective states can be communicated to others using status messages that are common on social networks and chat applications. Figure 1.2A provides an example of our visualization for valence, arousal, and dominance (detailed in Chapter 7). The circle is divided into three equal-sized segments, one for each dimension. The parts of the segments are filled according to the level of the corresponding dimension. Figure 1.2B shows how our visualization could be used as part of the header in a chat application.

Education. Affective states play an important role in the educational context and can directly influence a student's motivation, problem-solving ability, and learning gain [Baker et al., 2012; Csikszentmihalyi, 2008; Miserandino, 1996]. For example, learning outcomes have been found to decrease if frustration is persistent during problem-solving, whereas overcoming a state of frustration can have a positive effect on learning [Baker et al., 2012]. Teachers having access to the visualizations of the affective states of their students can provide feedback to the students based on the

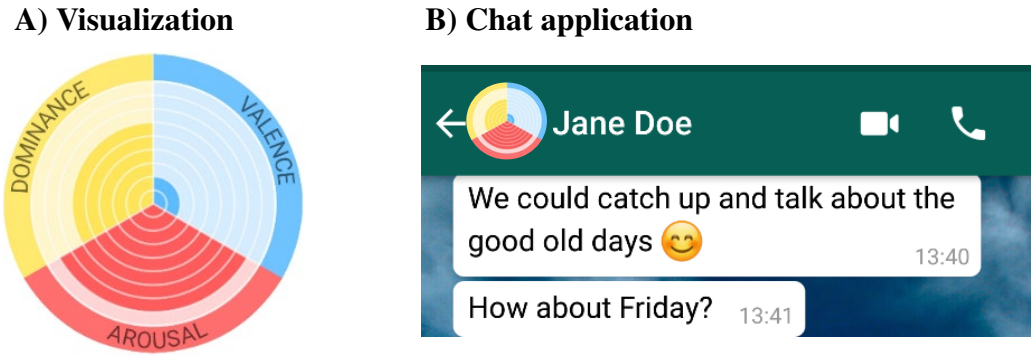


Figure 1.2: *A possible visualization of affective states. A) One segment for each affective dimension. The parts of the segments are filled according to the level of the corresponding dimension. B) Example of how the visualization can be used in a chat application.*

affective information, which can increase the learning experience [Grawemeyer et al., 2015]. A further application of affective state prediction is as part of intelligent tutoring systems. Based on the affective states of students, affect-driven feedback and instructional help messages can be provided. [Cabestrero et al., 2018]. For example, Santos et al. [2016] developed a learning environment for language learning which provides personalized feedback (e.g., playing songs) based on the detection of relaxed and nervous emotional states of students. Finally, prediction of affective states can also be useful to teach children to recognize and interpret their emotions, e.g., during taking photos and recording videos [Leijdekkers et al., 2013] or by engaging with virtual agents representing real people and their emotions [Bertacchini et al., 2013]. This can be especially useful for autistic children having difficulty recognizing emotions in others and sharing emotions with others.

1.1.2 Affect Modelling

In this section, we contextualize our engineering goals within different modeling approaches of affective states. Typically, emotions, and affects in general, are modeled using a categorical, dimensional, or appraisal approach [Politou et al., 2017].

Categorical models. Categorical models describe emotions as innate, discrete, and separately identifiable, and as universal to all humans (i.e., cross-cultural) [Colombetti, 2009]. Based on Tomkins model [Tomkins, 1962] stating that emotions are primarily facial behavior and body response is secondary, Ekman et al. [1987] proposed six basic universal emotions (i.e., sadness, happiness, anger, fear, disgust, and surprise). According to the Facial Action Coding System (FACS) [Ekman and Rosenberg, 1997], each basic emotion is associated with a unique facial expression.

For example, happiness is identified through the raising of the mouth corners and tightening of the eyelids, and anger is identified through eyebrows lowering, lips pressing and eyes bulging. The basic emotions can also be blended to form complex emotions [Shoumy et al., 2020]. For example, a feeling of happiness and sadness will result in melancholy.

Dimensional models. Dimensional models conceptualize emotions as a combination of several psychological dimensions. The circumplex model proposed by Russell [1980] is a two-dimensional model representing affective states in terms of two orthogonal and bipolar dimensions, i.e., valence and arousal. According to this model, each affective state is a linear combination of valence and arousal in different degrees. Valence measures the pleasantness of an affective state. For example, anger and fear are unpleasant emotions (negative valence), but happiness is a pleasant emotion (positive valence). Arousal refers to the perceived intensity of an event. For example, anger is typically an intense feeling (high arousal), whereas boredom has a low arousal value. This two-dimensional model was further extended by an additional dominance dimension [Mehrabian, 1996]. Dominance represents how controlling and dominant one feels about a situation. For example, anger is typically a moderately dominant emotion, whereas boredom is a non-dominant emotion.

Appraisal models. According to the appraisal theory, emotions are generated by the evaluation of the internal state of a person and the state of the outside world [Gunes and Pantic, 2010]. Stated differently, emotions are primarily caused by cognitive processes, specifically by appraisals of objects as relevant to one's well-being. In contrast to the categorical and dimensional models, the same situation or event can cause different emotions for different people. In particular, the emotions depend on how the people are appraising the situation taking into account their history, goals, needs, and expectations.

In this thesis, we use categorical models (i.e., basic emotions) and dimensional models (i.e., valence, arousal, and dominance) to represent and measure the emotional state of a person. Appraisal models have shown to be too complex for real-world applications [Calvo and D'Mello, 2010] and were, therefore, excluded in this thesis.

1.1.3 Data Modalities

The human body responds to emotions through various physical and physiological signals [Kanjo et al., 2015]. Physical signals encompass, for example, facial expressions, gestures, and movements. Physiological signals include, among others, skin conductance and pulse rate. Such signals can serve as input to machine learning models enabling the prediction of emotions. Two important factors affecting the value of each data modality as a viable affect detection channel are the reliability of the signals in real-world environments (e.g., in classrooms and at home) and the degree

of privacy protection of the signal [Calvo and D'Mello, 2010]. Models applicable in real-world environments are more useful for the users as they can support them in their daily routine, though such models and signals are often accompanied by a privacy trade-off. People are usually sensitive when personal data is used or personal information is revealed from signals or the combination of signals [Politou et al., 2017]. Moreover, not only the users themselves can be affected by privacy issues but also other people in the vicinity of the users (e.g., a camera recording other people in the field of view).

Privacy. Although an invasion of privacy can be desired and beneficial for the users (e.g., user-tailored services and personalization features), the collected data may also be used to identify users and to extract sensitive information (e.g., users habits and relations) which is not disclosed to the users [Christin et al., 2011]. Due to the lack of a common understanding of privacy [Newell, 1995], we introduce our definition of privacy tailored to our context. We link privacy to personal identifiable information (PII) which is information that can be used on its own or combined with other information to identify or trace an individual (e.g., name, social security number, biometric records, date, and place of birth, etc.) [McCallister et al., 2010]. PII exists also in the legislation of many countries (e.g., in the United States, the National Institute of Standards and Technology's guide to protecting the confidentiality of personally identifiable information [McCallister et al., 2010] and in the European Union, the General Data Protection Regulation [Voigt and Von dem Bussche, 2017]). Unauthorized access, use, or disclosure of PII can harm individuals in terms of identity theft, blackmail, and embarrassment. Based on the definition of PII, we establish the following three requirements which a data source must fulfill to protect privacy:

- **R1.** A user cannot be identified or traced from the data alone or in combination with other information [McCallister et al., 2010].
- **R2.** Other people cannot be identified or traced based on the collected data from the user [Christin et al., 2011; Politou et al., 2017]. This includes people in the vicinity of the user (e.g., sitting next to the user) but also people in contact with the user in another way (e.g., calling or messaging the user).
- **R3.** The user can trick the system to disguise the identity (e.g., changing the handwriting and typing differently) [Calvo et al., 2015]. This requirement only applies if R1 or R2 applies.

Another part of privacy is the requirement that a system should support users with usable and understandable mechanisms to provide the ability to control the release and the degree of granularity of data [Christin et al., 2011]. Potential privacy and ethical issues can also be related to the predicted emotions such as manipulating or influencing people's emotions [Kanjo et al., 2015; Politou et al., 2017].

Real-world usability. Data collection in laboratory environments is typically easier and provides less noisy signals than in real-world environments. On the other hand, to support users during their daily routines, a model for predicting affective states should be applicable in real-world scenarios. As such, we state three requirements for data sources to be applicable in real-world environments:

- **R4.** The data should be collected unobtrusively and energy-efficient (i.e., low power drain) so that the user is not disturbed in the daily routines and the device can be carried for long periods [Kanjo et al., 2015; Lane et al., 2010; Larradet et al., 2020; Macias et al., 2013]. Ideally, no additional hardware is needed to collect the data.
- **R5.** The signal of the data source should be robust under conditions observed in real-world environments (e.g., movement of users) to reduce corrupted and erroneous data being produced [Kanjo et al., 2015].
- **R6.** The costs of the data recording and hardware should be as low as possible [Politou et al., 2017]. Ideally, no additional costs arise (i.e., no additional hardware needed).

In this thesis, we investigate several data modalities to predict affective states. Similar to McCallister et al. [2010], we assign each requirement a score of 1 (requirement not fulfilled), 2 (requirement partially fulfilled), or 3 (requirement fulfilled). For each data source, we then sum up the scores of all requirements to obtain one score for privacy protection (a high score means high privacy protection) and one for usability in real-world environments (a high score means very usable). Table 1.1 lists the scores for all data modalities and all requirements as well as the total scores for privacy protection (R1–R3) and real-world usability (R4–R6). Figure 1.3 shows an overview of the data modalities in relation to their applicability in real-world environments and their degrees of privacy protection according to the total scores derived from the requirements. In the following, we detail the different data modalities and discuss each requirement.

Video data. Inspired by the fact that facial expressions are directly linked to basic emotions, we employ video data (i.e., facial expressions, eye gaze, eye blinks, body movement, and head movement) in Chapter 3 to predict affective states. Video data can be used to identify the user (R1 $=$ 1) and also people in the vicinity (R2 $=$ 1) [Christin et al., 2011]. Tricking the system to disguise the identity is only possible to a certain extent when wearing masks (R3 $=$ 1). Still, the people in the vicinity can be identified when only the user is wearing a mask. Video data can be collected unobtrusively using built-in webcams of tablets and smartphones. On the other hand, external cameras (e.g., GoPros) are visible to the user and thus more obtrusive (R4 $=$ 2). Video recordings are also draining the battery of smartphones and for external cameras, battery capacity is typically limited to a few hours. Video

Table 1.1: *The scores of the requirements for the protection of privacy (R1–R3) and usability in real-world environments (R4–R6) for each data modality. A higher score indicates higher privacy protection and higher usability in real-world environments.*

	Video	Biosensors	Stylus	Smartphone Touch	Smartphone Sensors
R1 (PII)	1	2	1	1	2
R2 (PII in vicinity)	1	3	3	1	3
R3 (Disguise)	1	2	2	2	3
R4 (Unobtrusiveness)	2	1	3	3	3
R5 (Robustness)	2	1	3	3	3
R6 (Costs)	2	2	2	3	3
Total R1–R3	3	7	6	4	8
Total R4–R6	6	4	8	9	9

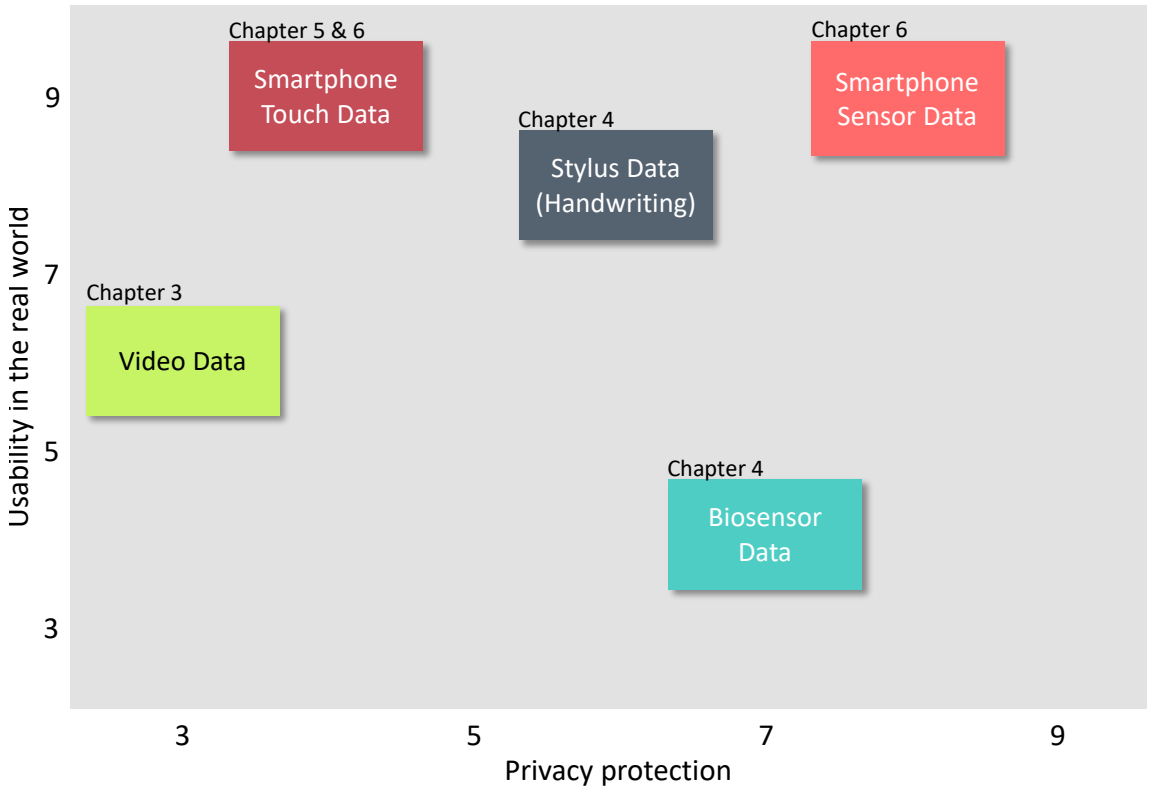


Figure 1.3: *Data Modalities in relation to privacy protection and usability in real-world environments. High scores indicate high usability in real-world environments and high privacy protection. Capturing video data does not protect the privacy and is more applicable in controlled environments whereas recording smartphone sensor data protects privacy more and is more applicable in real-world environments.*

recordings also depend on lighting conditions and view angle (R5 = 2). For example, often the face and body are not at all or only partially visible. In addition, facial expressions can be different from true facial expressions because people adjust their facial expressions due to social norms [Fridlund and Duchaine, 1996]. There is no additional cost for video recordings from smartphones or tablets but significant costs for external cameras (R6 = 2). Overall, this leads to a total score of 3 for privacy protection and a total score of 6 for usability in real-world environments. This is in line with Politou et al. [2017] who proposed low privacy protection for videos recorded from cameras.

Biosensor and stylus data. Based on the increase in the number and wearability of biosensor devices (e.g., integrated biosensors in watches) [Arroyo et al., 2009] and the fact that tablets bundled with a stylus are becoming increasingly available in households and classrooms and are inherently non-intrusive and mobile, we explore biosensors (i.e., heart rate, skin conductance, and skin temperature) and handwriting data recorded from a stylus for affective state prediction in Chapter 4. Users can be identified based on their handwriting and the handwriting of a person can potentially also be used to counterfeit documents, letters, and signatures (R1 = 1) [McCallister et al., 2010]. From the biosensor data, adversaries may identify health issues or diseases (R1 = 2) [Christin et al., 2011]. Neither from handwriting (e.g., solving math tasks using a stylus) nor from biosensor data other people in the vicinity of a user can be identified (R2 = 3). For handwriting, the system can be tricked by changing the style of writing although this can make it difficult to read the resulting text (R3 = 2). On the other hand, some biosensor signals can be consciously controlled by the user (e.g., heart rate [Pokrovskii and Polischuk, 2012] and skin temperature [Kozhevnikov et al., 2013]) while other biosensor signals cannot be controlled (e.g., skin conductance [Critchley, 2002]) (R3 = 2). While handwriting data can be collected unobtrusively during stylus usage (R4 = 3), biosensor data is typically collected using wrist bands, chest straps, and electrodes which can make longer-lasting recordings uncomfortable (R4 = 1). Further, handwriting recorded from a stylus provides usually a robust signal (R5 = 3) but biosensor signals can be degraded by motion artifacts and skin properties (R5 = 1) [Calvo and D'Mello, 2010]. Finally, tablets are often bundled with a stylus, otherwise, a stylus can be bought as a supplement to a smartphone or tablet (R6 = 2). Similarly, some smartwatches have already biosensor integrated but biosensor devices can also be bought in different price ranges (R6 = 2). In summary, the total score for privacy protection is 7 for biosensor devices and 6 for handwriting recorded from a stylus. According to our requirements, handwriting data is more usable in real-world environments (total score of 8) than biosensor data (total score of 4).

Smartphone touch and sensor data. Recent work has also suggested that being aware of one's current affective state can be particularly useful in the context of mobile devices as individuals become more dependent on smartphones for social

purposes [LiKamWa et al., 2013]. Smartphones are ubiquitous, unobtrusive, and provide a rich stream of continuous data. In this thesis, we exploit touch data and sensor data (i.e., gyroscope and accelerometer) recorded from smartphone devices (see Chapter 5 and 6). From touch data, a user can be uniquely identified by the touch characteristics of the typed text (R1 = 1) [Mahfouz et al., 2017]. Similarly, other people can potentially be identified by the typed text of the user, e.g., in a chat conversation (R2 = 1). Tricking the system is possible to some degree by changing the typing behavior or the typed text (R3 = 2). Information about the gait extracted from smartphone sensor data [Christin et al., 2011] and sensor fingerprinting [Hupperich et al., 2016] can be possible indicators of a user's identity (R1 = 2). Further, the vicinity of the user cannot be inferred from sensor data alone (R2 = 3). In addition, tricking the system is possible by rotating the smartphone, using the smartphone on a table, and trembling during usage (R3 = 3). Finally, both smartphone touch and sensor data can be collected unobtrusively with low battery drain in the background (R4 = 3), the signals are robust against noise (R5 = 3), and no additional costs are imposed due to the built-in sensors (R6 = 3). In summary, both touch and smartphone sensor data are usable in real-world environments (total score of 9) but collecting sensor data protects privacy more (total score of 8) than collecting touch data (total score of 4). This is in line with Politou et al. [2017] who declared high privacy protection for data recorded from accelerometers and gyroscopes.

1.2 Principal Contributions

In the following, we list the main technical contributions of the work presented in this thesis:

- **Camera setup and neural inpainting pipeline.** Most existing approaches for predicting affective states from camera recordings use external cameras (e.g., GoPro) or webcams, which are expensive, more difficult to handle, and are exposed to time synchronization problems. Therefore, we propose a cheap and unobtrusive camera setup for tablet computers and a deep learning-based image processing pipeline to reconstruct high-quality facial recordings. The setup requires a small mirror to be attached to the camera to improve the visibility of the face. Then, the image is reconstructed using a neural inpainting approach. We show that the mirror construction improves the visibility of the face in situations where external cameras struggle. With a qualitative and quantitative evaluation, we demonstrate that we can achieve results comparable to a GoPro camera. In particular, neural inpainting improves confidence in facial landmark detection by up to 88%.

- **Video-based features and affective state classification pipeline.** We present a vision-based model for predicting affective states. In our model, we fuse different existing approaches with novel features extracted from video recordings (i.e., head and body movement, eyes, and face). We evaluate our affective prediction model on data from a laboratory experiment with 88 participants. Participants were solving math tasks (active) and were exposed to emotional stimuli from pictures (passive). We show that our model accurately predicts two levels (low and high) of valence (up to 0.80 AUC) and arousal (up to 0.73 AUC) using data from the front camera.
- **Affective state classification pipeline based on biosensor and stylus data.** One of the main challenges of affective state prediction are privacy concerns related to the input modalities and the applicability in real-world settings. Therefore, we propose a system to detect affective states based on biosensor and handwriting data recorded from a stylus that is cheap and easy to operate, can be used outside a lab setting, is non-intrusive, and minimizes potential issues related to privacy. We evaluate our method by applying it to a math problem-solving scenario in which 88 participants provided answers in unstructured handwriting on a tablet device. We show that we can reach good classification accuracy (0.88 AUC) when using data from biosensors and handwriting in combination. We reach a comparable performance using only the data acquired by the stylus (0.84 AUC). These results suggest that a simple tablet with a stylus can be sufficient to reliably predict affective states, which leads to less intrusive and cheaper setups.
- **Generalized affective state prediction model.** We explore whether the affective state prediction model based on biosensor and handwriting data can be generalized over domains. For this purpose, we apply the model trained on math task solving to a passive setting with picture stimuli leading to a performance of 0.68 AUC.
- **Semi-supervised affective state classification pipeline based on smartphone touch heat maps.** We propose a non-invasive solution that can accurately predict affective states based on touch data from a mobile device. We generate two-dimensional heat maps of typing characteristics by considering only touch input (i.e., pressure and speed of typing) from the smartphone's on-screen keyboard. By using heat maps in contrast to raw data, we are also taking into account the spatial distribution of the data. To make use of the large amount of unlabeled data, we train a semi-supervised deep learning architecture on these heat maps to learn a low-dimensional feature embedding. We demonstrate the effectiveness of predicting the affective states in a data collection study with 70 participants engaged with a chat application. We show that our pipeline can accurately predict three classes

(low, medium, high) of valence (up to 0.84 AUC), arousal (up to 0.82 AUC), and dominance (up to 0.82 AUC). We also present results for the prediction of two levels (present vs. not present) of anger (0.84 AUC), happiness (0.88 AUC), sadness (0.87 AUC), surprise (0.76 AUC), and stress (0.80 AUC).

- **Affective state classification pipeline based on smartphone usage in the wild.** Based on the touch heat maps encoding pressure and speed of typing, we refine the heat maps by encoding key positions and using three channels to simultaneously leveraging multiple typing metrics. Our heat maps serve not only as input to our affective state prediction model but can also be used for visualization purposes to investigate typing behavior on smartphones. We evaluate our convolutional neural network model on data collected from 82 participants over 10 weeks in the wild. We show that we can accurately predict two levels (present vs. absent) of the basic emotions and stress (up to 0.86 AUC) and three levels (low, medium, high) of valence (up to 0.83 AUC), arousal (up to 0.85 AUC), and dominance (up to 0.84 AUC). In addition, we show that using two-dimensional heat maps created from smartphone sensor data (i.e., gyroscope and accelerometer), we can achieve similar performance for valence (up to 0.79 AUC), arousal (up to 0.83 AUC), and dominance (up to 0.81 AUC). Such a model leveraging smartphone sensor data only is less privacy-invasive, and thus has potential higher acceptance by users in real-world scenarios.

- **Emoji-based questionnaire for measuring affective states.** Existing questionnaires for measuring affective states typically have an old-fashioned layout and are not suitable for fast assessments of affective states on mobile devices (i.e., with space constraints). For example, the self-assessment manikin [Bradley and Lang, 1994] is a pictorial assessment used to quantify levels of valence, arousal, and dominance on a 9-point scale. Thus, we propose a simplified mobile-friendly version of the self-assessment manikin measuring valence, arousal, and dominance on five levels using an emoji-style pictorial assessment. Nowadays, emojis are common on smartphones for expressing emotions (e.g., in chat applications), thus an emoji-based encoding simplifies understanding for users and makes the assessment more motivating.

- **Widgets for visualizing affective states.** We present two application-dependent graphical user interface widgets that provide affective feedback for valence, arousal, and dominance. The widgets are designed to be compact and transparent such that they interfere as little as possible with other activities on the user's screen. The first widget focuses on an intuitive and fast assessment of the current affective state. The second widget concentrates on an exact, clear, and time-dependent visualization. We test the widgets on

intuitiveness and understandability and evaluate them in a user study with 644 participants.

1.3 Outline

This thesis is organized as follows.

- **Chapter 2** gives an overview of related work on input modalities, models and systems, and visualization of affective states.
- **Chapter 3** describes a pipeline for predicting affective states based on video data using a novel camera setup and a neural inpainting pipeline to improve the visibility of the face for front camera recordings. We evaluate our setting and pipeline using data from a study with 88 participants.
- **Chapter 4** presents an automated pipeline for predicting the affective states of participants solving tablet-based math tasks using signals from low-cost mobile biosensors and handwriting data recorded from a stylus. In contrast to video data, these data sources are less privacy-invasive and more applicable in real-world environments.
- **Chapter 5** describes a system that analyzes the user's text typing behavior on smartphones using a semi-supervised deep learning pipeline for predicting affective states. We evaluate the system using a data collection study with 70 participants on text conversations designed to trigger different affective responses.
- **Chapter 6** expands on the work in Chapter 5 by improving the modeling qualitatively and expanding the model by smartphone sensor data making our model less privacy-invasive. We evaluate our model using data collected in the wild from 82 participants over 10 weeks.
- **Chapter 7** describes two graphical user interface widgets that visualize the user's affective state, ensuring a compact and unobtrusive visualization. We evaluate the widgets in relation to a baseline widget and test the widgets on intuitiveness and understandability based on data from a user study with 644 participants.
- **Chapter 8** concludes this thesis with a discussion of the contributions, limitations, and an outlook to potential future work.

1.4 Publications

In the context of this thesis, the following peer-reviewed publications have been accepted:

> **R. WAMPFLER**, S. KLINGLER, B. SOLENTHALER, V. SCHINAZI and M. GROSS (2019). Affective State Prediction in a Mobile Setting using Wearable Biometric Sensors and Stylus. *Proceedings of the International Conference on Educational Data Mining (Montréal, Canada, July 2-5, 2019)*, pp. 224-233.
>
> **R. WAMPFLER**, S. KLINGLER, B. SOLENTHALER, V. SCHINAZI and M. GROSS (2020). Affective State Prediction Based on Semi-Supervised Learning from Smartphone Touch Data. *Proceedings of the Conference on Human Factors in Computing Systems (Virtual conference, April 25-30, 2020)*, pp. 1-13.
>
> **N. KOVAČEVIĆ**, R. WAMPFLER, B. SOLENTHALER, M. GROSS and T. GÜNTHER (2020). Glyph-Based Visualization of Affective States. *Eurographics/IEEE VGTC Symposium on Visualization (Virtual conference, May 25-29, 2020)*, pp. 121-125.
>
> **R. WAMPFLER**, A. EMCH, B. SOLENTHALER and M. GROSS (2020). Image Reconstruction of Tablet Front Camera Recordings in Educational Settings. *Proceedings of the International Conference on Educational Data Mining (Virtual conference, July 10-13, 2020)*, pp. 245-256.

This thesis is also based on the following planned publication:

> **R. WAMPFLER**, S. KLINGLER, B. SOLENTHALER, V. SCHINAZI, C. HOLZ, and M. GROSS. Affective State Prediction from Smartphone Keystroke and Sensor Data in the Wild. *Under review at the time of submission of this thesis*.

This thesis includes the contents of all the above papers as well as additional implementation and evaluation details not present in the papers.

CHAPTER 2

Related Work

This chapter provides an overview of previous research in the field of affective state prediction. Different data modalities were the major focus in research on affective state prediction models. Therefore, in Section 2.1 we review related work on different models and data modalities (i.e., video data, biosensor data, stylus data, and smartphone data) for predicting affective states. Then, in Section 2.2 recent work in the related field of stress prediction is discussed. Besides the prediction of affective states, the visualization of affective states is another important part of an affect-aware system. Thus, in Section 2.3 we cover related work on affective state visualization. More specific related work and background to methods presented in this thesis are discussed in the corresponding chapters.

2.1 Affective State Prediction

Different data modalities were used to predict affective states in different domains. In an educational setup, acoustic features from student voices during interaction with tutors were used to predict three levels of valence [Litman and Forbes-Riley, 2006]. Another line of research tried to predict affective states based on logged user interactions only, such as input and error behavior, timing, and help calls. Frustration, boredom, engaged concentration, and confusion were successfully predicted using interaction data for math tutoring systems [Kostyuk et al., 2018; Grawemeyer et al., 2016]. On the other hand, valence and arousal were predicted using mouse and keyboard interaction data from writing compositions in free text [Salmeron-Majadas et al., 2018]. Although large and powerful interaction data sets can be easily collected especially in online environments, the features are typically dependent

on the learning domain and the specific learning system. Generalized models were proposed, such as an engagement model for two different learning domains and tutors (spelling and math) [Käser et al., 2013], but these generalized methods typically have a lower accuracy as domain-specific features. Multimodal approaches fusing different data modalities were also introduced for the prediction of affective states. We refer to D'Mello et al. [2018] for a concise overview of such multimodal methods in educational settings. In addition, Kanjo et al. [2015] provides an overview of a variety of data sources that can be used for predicting affective states (e.g., physiological signals, facial expressions, speech, phone usage, social networks, and mobile network data). In the following, we will give an overview of existing work on video data, biosensor data, stylus data, and smartphone touch and sensor data. We will focus on applications in an educational context and on mobile devices such as smartphones.

2.1.1 Video Data

Prediction of affective states from video recordings is one of the most popular approaches nowadays as it allows different features to be exploited, such as body language and posture, head movement, eye gaze, and facial expressions [Zeng et al., 2008]. Bosch et al. [2015] calculated statistics (i.e., maximum, median, and standard deviation) of the frame-level likelihood values of 19 different action units (i.e., facial muscle movements identifying independent motions of the face), the head position, and gross body movement from webcam video recordings of students playing an educational physics game. They predicted two levels (present vs. absent) of boredom (0.61 AUC), confusion (0.65 AUC), delight (0.87 AUC), engagement (0.68 AUC) and frustration (0.63 AUC). Based on this work, Kai et al. [2015] found that an interaction-based model using timing and counting-based features performs worse than the video-based model. Similarly, using a math tutor, Arroyo et al. [2009] found facial expressions to be more predictive for confidence, frustration, excitement, and interest than conductance bracelets, pressure mice, and a posture analysis seat.

Also for other tasks, facial expressions were found to be a good predictor for affective states. In text comprehension tasks, two levels of confusion (0.64 AUC), engagement (0.55 AUC), and frustration (0.61 AUC) were successfully predicted using 20 different action units [Chen et al., 2015]. On the other hand, Grafsgaard et al. [2013] found upper face movements predictive for engagement, frustration, and learning in a setup consisting of a programming tutor and a webcam. Finally, based on eye gaze features (e.g., fixation and view angle) extracted from a specialized eye capturing device, boredom (69%) and curiosity (73%) were successfully predicted on two levels during interaction with an intelligent tutoring system [Jaques et al., 2014].

In contrast to facial expressions, body movement and posture are ordinarily unconscious and unintentional, and thus not susceptible to social editing (i.e., adapting

expressions due to social norms) [Calvo and D'Mello, 2010]. Moreover, gross body motions can be differentiated over long distances, whereas for facial expression analysis typically short distance and high-resolution recordings are necessary [Walk and Walters, 1988]. Thus, Sanghvi et al. [2011] employed posture and movement features to predict two levels of engagement of children playing chess with a robot with an accuracy of 82%. Posture and body movement can also encode other emotions such as anger (leg stepping back, elbows bent, and head bent forward, etc.), boredom (head bent backward and collapsed upper body, etc.), and fear (slow movement, knees bent apart, forearms raised, and head bent backward, etc.) [Shoumy et al., 2020]. A survey of different video-based approaches for predicting affective states is provided by Zeng et al. [2008].

In summary, a majority of the existing vision-based approaches use external devices, such as webcams, and rely on posed facial expressions to predict basic emotions. Therefore, we present in Chapter 3 a novel setup for reliably recording the face and body of users based on the front camera of tablet computers, and hence without the need for expensive devices or synchronization between the devices. We demonstrate the usefulness of our setup by predicting affective states in terms of valence and arousal using data from an experiment containing spontaneous (non-posed) facial expressions. Finally, for our vision-based model, we fuse different existing approaches with novel features.

2.1.2 Biometric Sensors

Biometric sensors provide an objective measure of the physiological reactivity of users engaging with a learning environment while minimizing interference with the actual task [Blanchard et al., 2014; Jraidi et al., 2014; Kim et al., 2004; Salmeron-Majadas et al., 2015]. Indeed, education research investigated the effectiveness of a variety of physiological signals used to infer affective states. Electrodermal activity (measuring electrical conductivity as a function of the activity of sweat glands on the skin surface), skin temperature, and heart rate were generally found to be good predictors of emotions [Jraidi et al., 2014; Kim et al., 2004; Salmeron-Majadas et al., 2015] and mind wandering [Blanchard et al., 2014] across different tasks including math learning [Jraidi et al., 2014; Salmeron-Majadas et al., 2015], scientific text reading [Blanchard et al., 2014], and audio, visual and cognitive stimuli in general [Kim et al., 2004]. Other used physiological signals consist of electroencephalogram (measuring brain activity), electromyography (measuring muscle activity), and breath rate [Santos, 2016].

Kim et al. [2004] used a complex setup consisting of audio, visual, and cognitive stimuli to elicit different emotions. Using data from 50 subjects, they predicted sadness, anger, stress, and surprise with an accuracy of up to 61.8% using a combination

of electrodermal activity, skin temperature, and heart rate measures. In contrast, Blanchard et al. [2014] were capable of predicting self-reported mind wandering on two levels (present vs. absent) with 60% accuracy using electrodermal activity and skin temperature while students answered questions of varying difficulty and point value after reading a scientific text. In the context of math learning, Jraidi et al. [2014] used Bayesian networks to analyze stress, confusion, boredom, and frustration from changes in skin conductance, heart rate, and electroencephalogram activity while participants solved math tasks. They reported an accuracy ranging from 81% to 90% for a three-level assessment of the emotions. Dzedzickis et al. [2020] provide a review of work leveraging physiological signals for predicting affective states also outside the educational domain.

These previous works mainly focused on expensive, high-quality sensors to provide medical-grade accuracy for the measurement of physiological signals. In contrast, in Chapter 4, we gather such data in a non-intrusive and easy-to-use way.

2.1.3 Stylus

Predicting affective states based on stylus data is still a relatively new research topic. Likforman-Sulem et al. [2017] used ductus (i.e., number of strokes) and timing features extracted from figure drawings and writing given words to predict anxiety (60%), depression (73%), and stress (60%) on two levels (present vs. absent) for 129 participants using a Support Vector Machine and a Random Forest classifier. A feature importance analysis revealed that both in-air and on-paper features were relevant for predicting emotional states. For the prediction of depression, figure drawings were most predictive, whereas for the prediction of anxiety and stress both figure drawings and writing were predictive. Fairhurst et al. [2015] conducted an experiment for predicting stress and happiness by letting participants writing down a given list of words and describing a visual scene in their own words. They extracted features related to velocity, acceleration, and pressure of writing and reported an accuracy on two classes of up to 70% for stress prediction and 80% for predicting happiness using a Support Vector Machine classifier. On the other hand, Zhou et al. [2014] used digital pen data (on digital paper) in a collaborative math solving setting (16 math problems with 4 difficulty levels) to predict the students' expertise (expert vs. non-expert) as well as identifying the dominant domain expert among the students with an accuracy of up to 83%. From the stylus data, they extracted features related to stroke distance, stroke duration, writing speed, and pressure.

The ground truth for all three presented works was gathered using a single questionnaire for each participant. Instead, probing the affective state of subjects in regular intervals, as done in this thesis, provides a more fine-grained view into the emotional regulation of people. Recently, handwriting was also considered for predicting per-

sonality traits (e.g., openness, conscientiousness, extraversion, agreeableness, and neuroticism) by extracting characteristics related to word slant, pressure of writing, and the space between lines, and the size of lines, words, and characters [Remaida et al., 2020].

2.1.4 Smartphone Data

We focus in this section on methods that base the prediction on typing characteristics on computer keyboards and the various data sources available on smartphones (e.g., touchscreen, gyroscope, and accelerometer). Typing characteristics on computer keyboards are related to smartphone touchscreen data in so far that typing patterns can resemble the typing patterns on smartphone keyboards.

Most available affect-aware smartphone systems are complex in terms of the amount and nature of the data modalities involved. Systems were built from smartphone sensor data (e.g., accelerometer, Bluetooth, microphone, and GPS) to grasp user movements and conversational cues [Rachuri et al., 2010]. Other systems included the context of the user data (e.g., location and weather) [Bogomolov et al., 2013; Lee et al., 2012], communication data (e.g., call and SMS logs) [Bogomolov et al., 2013; LiKamWa et al., 2013; Pielot et al., 2015], and interaction data (e.g., web browsing and application usage) [LiKamWa et al., 2013; Pielot et al., 2015]. Such systems provided decent performance with accuracies up to 71% for predicting various emotions (i.e., happy, sad, fear, anger, and neutral) [Rachuri et al., 2010] and 80% for predicting three levels of happiness (i.e., happy, neutral, and unhappy) [Bogomolov et al., 2013]. Nevertheless, such complex systems are often privacy-invasive and computationally demanding.

Other more lightweight approaches for predicting affective states exploited touch and typing behavior. Gao et al. [2012] predicted four states (i.e., excited, relaxed, bored, and frustrated), each with two levels, with an accuracy between 69% and 77% as well as two levels of arousal and valence with an accuracy of 89%. These researchers used touch pressure and speed of touch features recorded while users were playing a game. Previous work also employed touch data from chat conversations. Lee et al. [2012] predicted Ekman's six basic emotions and a neutral state with 67% accuracy using a Bayesian network classifier based on behavior data (i.e., typing speed and touch count) and context data (i.e., location and weather) collected while users used the Twitter application. Interestingly, they found that the speed of typing was the most predictive factor. On the other hand, Ghosh et al. [2017] predicted four states (i.e., happy, sad, stressed, and relaxed) with a performance of 0.84 AUC using touch statistics (inter-tap durations, number of special characters, and number of deletes). These researchers jointly modeled the typing characteristics and the persistence of emotions by adapting the reported emotions based on a Markov chain.

Other researchers [Huang et al., 2018] predicted depression and mania on a regression scale using a personalized deep learning model for bipolar subjects by leveraging temporal dynamics and fusing accelerometer and keyboard metadata (duration of a keypress, time since the last keypress, and distance to the last keypress). Interestingly, Leow et al. [2019] found a positive correlation between higher accelerometer displacements and depression as well as mania.

Keystroke dynamic features such as pressure, latency, and duration were also used on computer keyboards [Lv et al., 2008]. Using these features, Epp et al. [2011] predicted 15 emotional states on two levels (present vs. absent) with an accuracy between 77% and 88%. Kołakowska [2013] provides an overview of other work on predicting affective states based on computer keyboards.

In contrast to touchscreen data and sensor data such as GPS and Bluetooth, accelerometer and gyroscope sensors provide a less privacy-invasive way of predicting affective states. In addition and in comparison to GPS and Bluetooth, recording accelerometer and gyroscope data also drain the battery less [Kołakowska et al., 2020]. Olsen and Torresen [2016] predicted valence and arousal on three levels using data from the accelerometer during walking segments by designing sophisticated step features. Using data from a study with 10 participants they reported an accuracy of 50.9% for valence (multilayer perceptron) and 75% for arousal (Support Vector Machine). On the other hand, Garcia-Ceja et al. [2015] predicted three levels of stress in a working environment using data from the accelerometer. By building the model based on data from similar behaving users, they achieved an accuracy of 60%. By personalizing the model per user, the accuracy increased to 71%. Most other works used accelerometer and gyroscope measurement only together with touchscreen, GPS, or Bluetooth recordings. A detailed overview of other related work on predicting affective states based on smartphone sensors is provided by Kołakowska et al. [2020].

In Chapter 5, we are using a lightweight approach by considering pressure and speed characteristics of touch data and employing a semi-supervised pipeline on heat maps extracted from this data. Moreover, we are using a pressure-sensitive display instead of the contact area [Gao et al., 2012] to approximate pressure, and we are also considering dominance, which we believe might be necessary for finer-grained distinctions between affective states. In Chapter 6, we are then comparing the predictive power of models based on touchscreen data and sensor data (i.e., gyroscope and acceleration) collected in the wild.

2.2 Stress Prediction

Stress arises from the transition from a calm to an excited state (negative or positive) [Shoumy et al., 2020]. Nowadays stress is omnipresent in our society and has a negative influence on mental health, physical health, productivity, decision-making capabilities, and situational awareness [Politou et al., 2017]. The ubiquity of mobile devices led to a surge in research focusing on the prediction of stress based on smartphone usage [Zautra, 2006]. Researchers used different data modalities for predicting stress based on smartphone data. These include behavioral metrics such as call and text logs and location data stemming from GPS [Bauer and Lukowicz, 2012; Bogomolov et al., 2014], application usage patterns [Ferdous et al., 2015], voice recordings [Lu et al., 2012], and video recordings [Carneiro et al., 2012].

Apart from being invasive (e.g., sharing of text logs), relying on these data modalities for the prediction of affective states has also the disadvantage of draining the smartphone battery (e.g., the high power consumption of GPS sensors). As such, other work focused on using sensor-based smartphone data, including touchscreen data and accelerometer data [Carneiro et al., 2012; Garcia-Ceja et al., 2015]. Carneiro et al. [2012] used patterns, accuracy, intensity, and duration of touch events as well as hand gestures to predict stress in real-time while users played a mentally challenging mobile game. In addition, Hernandez et al. [2014] showed that typing pressure and the size of the contact area with the mouse tend to increase during stressful situations (i.e., expressive writing, text transcription, and mouse clicking). To measure pressure and contact area, this work relied on pressure-sensitive computer keyboards and capacitive mouses, respectively. Recently, Exposito et al. [2018] conducted a similar study on the smartphone and showed that typing pressure increases during stressful situations (i.e., expressive writing). In addition, Sarsenbayeva et al. [2019] showed that stress increases tapping frequency but decreases tapping accuracy. These researchers also found that text difficulty had a larger effect on typing performance (measured as the ratio between the number of errors and the number of entered characters) than the stress level. Finally, other researchers proposed a multimodal approach jointly measuring accelerometer data, microphone data, and social activity data from calls and text messages [Maxhuni et al., 2016].

Most existing approaches for stress prediction used two [Bogomolov et al., 2014; Lu et al., 2012; Bauer and Lukowicz, 2012] or three classes [Maxhuni et al., 2016], and the stress measurement tool of choice were self-reports [Hernandez et al., 2014; Maxhuni et al., 2016; Ferdous et al., 2015]. Achieved performance ranged from 83% to 100% for two classes (stressed vs. non stressed) [Hernandez et al., 2014] and 71% for three classes (low, medium, high) [Maxhuni et al., 2016].

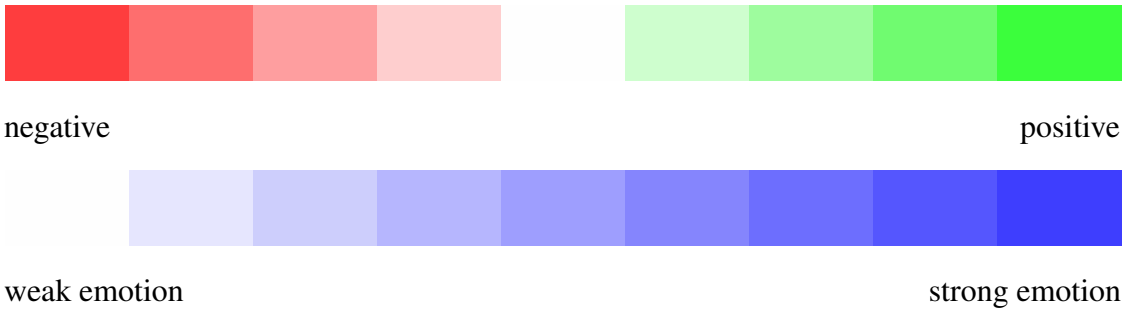


Figure 2.1: *Example visualization of valence (top) and arousal (bottom) based on work by Cernea et al. [2013]. Saturation and color of each bar denote the level of the corresponding dimension.*

2.3 Visualization of Affective States

To measure valence, arousal, and dominance, the Self-Assessment Manikin [Bradley and Lang, 1994] was proposed, which uses glyph-based visualizations [Borgo et al., 2013]. Each dimension is assessed on a 9-point scale where the different levels of each dimension are denoted by glyphs. This makes the approach widely applicable because it relies on a universal and language-independent representation.

Cernea et al. [2013] developed a visualization of valence and arousal for user interface components such as buttons. The affective states are visualized by displaying a color bar per dimension conveying different affective information by adjusting the color and saturation of the bar. Valence is displayed using a divergent color map, and arousal is displayed using a sequential color map (see Figure 2.1). The bars allow multiple states to be displayed at the same time. The bar is divided vertically into multiple parts, each of which corresponds to an affective state at a certain time, i.e., time is mapped to the horizontal axis. This visualization allows a comparison among user interface components because each component has multiple affective states assigned to it. However, it impedes an exact evaluation of the current level because levels are mapped to color saturation. Furthermore, the valence bar uses red and green, which is problematic for color vision deficiencies.

In a later work, Cernea et al. [2015] combined the visualizations of valence and arousal. They placed a closed curve around the user interface component under consideration. The outline exhibits waves or spikes depending on how high or low valence is. The higher the arousal, the higher the pulsation frequency of the curve. While this visualization impedes a visualization of affective states over time, it allows for a more intuitive and compact reading of the affective state. However, the visualization of valence may not be suitable if the visualization is too small, because differences between spikes and curves become hard to see.

CHAPTER 3

Affective State Prediction Using Video Data

Tablet computers have found quick application in education [Ditzler et al., 2016] as the technology offers new opportunities to students and teachers. It has been shown that tablets can influence learning pathways [Falloon, 2013] and improve digital skills [Reid and Ostashewski, 2011]. Moreover, tablets typically have built-in cameras, which can be used to unobtrusively record the student during the learning. Such data offers valuable clues to experts about the student's learning behavior and attention. Student observation has been implemented in studies with external camera setups [Zaletelj and Košir, 2017]. Such frontal-view camera data can also be used for predictions of the affective states of a student based on facial feature extraction [Pham and Wang, 2018], which works robustly even with low-resolution recordings [Nguyen et al., 2017].

Using external cameras for frontal view recordings of students provides an optimal viewing angle for robust facial feature extraction and affective state prediction. However, such setups require externally positioned cameras, which can be obtrusive and further depend on timestamp synchronization with the digital learning environment. Using tablet computers for learning circumvents these problems, as the built-in camera can be leveraged and timestamps are inherently in sync. Built-in cameras have, however, a sub-optimal viewing angle, leading to partially occluded and skewed faces in the recordings that makes it difficult to robustly extract facial features for affective state prediction.

In this chapter, we therefore propose a camera setup for tablet computers and a deep learning-based image processing pipeline to reconstruct high-quality facial recordings of students. The setup requires a small mirror to be attached to the camera to improve

the visibility of the face. Then, the image is reconstructed using a neural inpainting approach. We demonstrate the advantage of this setup and our reconstruction by an application for predicting affective states. The high quality of the reconstructed image enables facial feature extraction, such as head pose, eye gaze, and facial landmarks. We compare our method with an external camera setup (GoPro camera) and show that we can achieve a similar performance for predicting two levels (high and low) of valence and arousal for students performing active tasks, i.e., solving math tasks (up to 0.73 AUC) and students performing passive tasks, i.e., exposure to emotional stimuli from pictures (up to 0.80 AUC).

3.1 Background

Image inpainting is an image processing method to reconstruct missing or corrupted regions of an image. Common application areas include image restoration (e.g., removing scratches and text) [Liu et al., 2018], photo-editing (e.g., object removal) [Sarpate and Guru, 2014], and image coding and transmission (e.g., recovering the missing blocks) [Wang et al., 2006]. In Section 3.2, we focus on the specific task of face completion. Popular non-learning-based approaches applied to faces consist of patch-based methods, where image patches are copied to missing areas. Similar patches can be identified by using a face image dataset [Zhuang et al., 2009]. We refer to Guillemot and Le Meur [Guillemot and Le Meur, 2013] for a complete overview of non-learning-based models.

While non-learning-based methods can have difficulties ensuring consistent image structures [Iizuka et al., 2017; Pathak et al., 2016; Yeh et al., 2017], learning-based approaches typically generate smoother results. A popular line of learning-based methods uses generative adversarial networks (GAN) to inpaint missing regions of an image. GANs consist of a generative network to create a new image and a discriminator network to distinguish the new image from actual ground truth images. Using such a GAN approach, Malesevic et al. [2019] reported a peak signal-to-noise ratio (PSNR) of up to 20.57 for inpainting missing regions in faces. A similar performance of up to 20.2 PSNR and 0.84 structural similarity (SSIM) was achieved by Li et al. [2017] using an encoder-decoder network as the generator, a local and global loss function, and a semantic regularization term. On the other hand, Liao et al. [2018] used a collaborative model by training a GAN simultaneously on multiple tasks (i.e., face completion, landmark detection, and semantic segmentation). Using this knowledge-sharing approach, they reported a PSNR of up to 31.5 and an SSIM of 0.97 on face inpainting.

Convolutional neural networks (CNN) have been used for image inpainting as well. The encoder compresses the image with convolutional operations into a latent space, and the decoder reconstructs the image from the compressed representation. Guo et

al. [2019] proposed an encoder-decoder network using full-resolution residual blocks. For face inpainting, they reported a PSNR of 29 and an SSIM of 0.95. On the other hand, Liu et al. [2019] achieved a PSNR of 34.69 and an SSIM of 0.99 by adding a coherent semantic attention layer to the encoder. One disadvantage of this method is its long runtime of 0.82 seconds per image of size 256×256 rendering this method inapplicable for real-time video processing with more than one frame per second. Another problem with existing CNN-based methods is that the convolution operations are applied both to the valid and missing pixels at the same time, which can lead to visual artifacts (e.g., color discrepancy and blurriness). To overcome this issue, Liu et al. [2018] proposed partial convolutions, where the convolution operations are only applied to valid pixels by masking regions that need to be inpainted. The mask is updated during training of the network, including newly inpainted values. The authors demonstrated that the approach could produce semantically meaningful predictions also for inpainting regions with different shapes and sizes, achieving a PSNR of up to 34.34 and an SSIM of up to 0.95. We use this partial convolution approach to inpaint missing regions in images from front camera recordings. The dataset used for training the network is tailored to our use case.

3.2 Camera Setup

In this section, we present a low-cost hardware setup for recordings from the integrated front camera of a tablet computer, maximizing the visibility of the face of the users. Videos and images captured by the front camera are preprocessed, and missing parts are inpainted using a deep learning model to reconstruct the face of the users. Our approach is image-based and processes captured videos frame by frame.

3.2.1 Hardware Setup

While working on a tablet (e.g., writing with a stylus) it is convenient to have the device lying on the table (see Figure 3.1A). Due to the field of view of the front camera, only part of a users' face is visible. To adjust the field of view of the front camera, we attached a circular mirror (3 cm radius) to the tablet using a hinge (see Figure 3.1B). The hinge was fixed with glue so that the mirror would remain in a stable position. The mirror was mounted with an angle of 75 degrees relative to the tablet. This angle was chosen so that the visibility of the face was maximized. Due to the mirror setup, the upper part of the recordings is mirror-inverted (see Figure 3.1C). Depending on the conditions of the illumination of the recording environment, the exposure time of the camera of the recording device (e.g., tablet) needs to be adapted accordingly so that the camera focuses on the face instead of

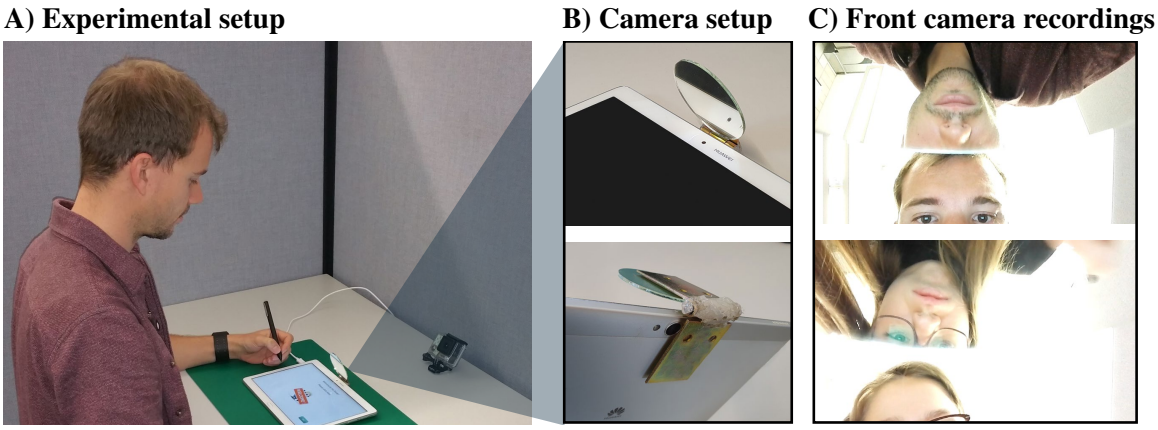


Figure 3.1: *The hardware setup. A user is working on the tablet (A). A mirror is attached to the tablet using a hinge (B). Due to the mirror reflections, the field of view of the front camera is changed so that the face of the participant is visible (C).*

the background. This adjustment of the exposure time can lead to an overexposed background (see Figure 3.1C).

3.2.2 Image Processing Pipeline

A raw image captured by the front camera is split by the mirror into two parts with the upper part of the image being mirror-inverted (see Figure 3.2A). To reconstruct the image, we propose a series of processing steps applied to the image (i.e., flattening the splitting boundary, face composition, image rotation, and extracting the face area). Image rotation and extraction of the face area are conducted as a preprocessing step for inpainting. Further, to train our inpainting model at a later stage, we assume that we have access to a dataset Ψ of square-shaped face images.

Splitting boundary. We apply a transformation to flatten the splitting boundary of the image (green line in Figure 3.2A), which simplifies image processing in the later stages and improves the final results qualitatively. We divide the image into 16 rectangles with equal width. An example of such a rectangle is shown in purple in Figure 3.2A. For each such rectangle, we transform the region defined by the vertices p_1, p_2, p_3, and p_4 into the region defined by the vertices p_1, p_2, p_5, and p_6 using a perspective transformation with linear interpolation. The location of these points can be calculated beforehand (or read from the image) because the mirror remains in a fixed position. The result of the transformation is shown in Figure 3.2B, where the splitting boundary (green) is a straight line.

Face composition. We rearrange the image by moving the part below the splitting boundary to the top and the flipped upper part to the bottom (see Figure 3.2C). The cut line defined by the mirror is shown in black. In addition, we adapt the height

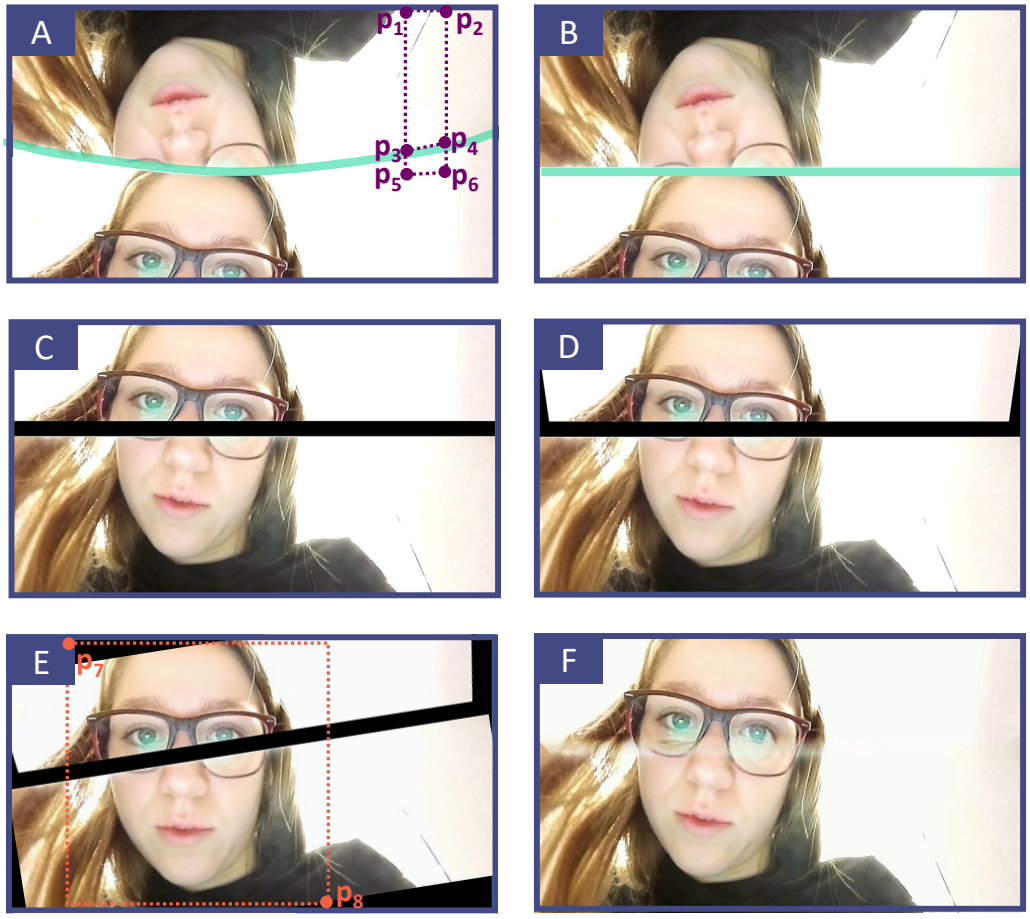


Figure 3.2: *The main inpainting steps. The splitting boundary of front camera recordings (A) is flattened using a perspective transformation (B). The face is reconstructed from the upper and lower parts (C) and warped so that the upper and lower part match (D). Finally, after horizontally aligning the eyes (E), the missing regions (black) are inpainted (F).*

of this cut line because depending on the distance of the face, the missing part is increasing (increasing distance) or decreasing (decreasing distance). As a next step, we push the bottom corner of the upper face towards the middle by applying a second perspective transformation to the image so that the upper and lower part of the face are matching (see Figure 3.2D).

Image rotation. We then rotate the front camera image so that the eyes are horizontally aligned (see Figure 3.2E). Using dlib [King, 2009], we extract the coordinates of the facial landmarks belonging to the left and right eye. From these landmarks, we calculate the position of the center of each eye and rotate the image around the midpoint between the eye centers so that the line connecting the center of the eyes is horizontally aligned.

Face area. We extract the face area by computing a square bounding box encompassing the face (see the orange box in Figure 3.2E). This bounding box is defined by the vertices $p_7 = (x_7, y_7)$ and $p_8 = (x_8, y_8)$ and is given by

$$x_7 = c_{x,I} - \frac{w_{I_\Psi}}{2} * \frac{\delta_I}{\delta_{I_\Psi}} \tag{3.1}$$

$$x_8 = c_{x,I} + \frac{w_{I_\Psi}}{2} * \frac{\delta_I}{\delta_{I_\Psi}} \tag{3.2}$$

$$y_7 = c_{y,I} - \frac{c_{y,I_\Psi}}{h_{I_\Psi}} * (x_8 - x_7) \tag{3.3}$$

$$y_8 = c_{y,I} + \frac{h_I - c_{y,I_\Psi}}{h_{I_\Psi}} * (x_8 - x_7), \tag{3.4}$$

where I and I_Ψ denote an image of the front camera and an image in the dataset Ψ, respectively. The width and height in pixels of an image are given by w and h. The x- and y-coordinate of the midpoint between the left and right eye are denoted by c_x and c_y, respectively, and δ is the distance between the eyes. Here, we assume that the origin is located at the top left of the image.

The part of the front camera image I outlined by the orange bounding box is then resized to the resolution $w_{I_\Psi} \times h_{I_\Psi}$ using bilinear interpolation. If the head of the user is close to the mirror, the face covers the full height of the image, and the bounding box might go over the upper and/or lower image borders. In such a case, we fill the parts overlapping the image with black pixels to get consistently sized bounding boxes (note that for visualization purposes only, the orange box in Figure 3.2E does not reflect this but instead is cut at the image border). We use the face detector of dlib [King, 2009] to test if a face and hence the landmarks of the eyes are identified in the image. In cases where the face cannot be detected, we use the landmarks of the eyes of the last image where the face could be identified (assuming that we have a video recording available, i.e., a series of images).

Inpainting missing area. As the last step in our image preprocessing pipeline, we inpaint the missing parts in the bounding box of the image (black region of the orange box in Figure 3.2E) with the neural inpainting approach of Liu et al. [2018] described in Section 3.2.3. We apply the neural inpainting only to the bounding box because it contains the important parts of the face (i.e., eyebrows, eyes, and mouth). We inpaint other parts of the image outside the bounding box using a simple Navier-Stokes based inpainting method provided by OpenCV [Bradski, 2000] which is based on a circular neighborhood of three pixels for each inpainted pixel. Finally, we rotate the image back to its original orientation. This then leads to the final reconstructed image shown in Figure 3.2F.

Figure 3.3: *Two example masks applied to images of the CelebA-HQ dataset [Karras et al., 2018].*

3.2.3 Neural Inpainting

For the neural inpainting approach, we use the dataset Ψ of square-shaped face images with customized missing regions tailored to our application of tablet front camera recordings and then train the network on this dataset.

Training dataset. The model is trained on a large corpus of images from the dataset Ψ together with a mask for each image indicating the missing parts (a mask is a matrix with the same size as the image having a $'1'$ entry for missing pixels and a $'0'$ entry otherwise). We create the corresponding mask randomly and similar in shape (rectangle) to the expected mask in our front camera recordings (see Figure 3.3 for an example of two such masks applied to two images from the CelebA-HQ dataset [Karras et al., 2018]). Note that the mask (missing image region) is not necessarily horizontal but rotates if a user is rotating the tablet or the head (vertical in the extreme).

Inpainting method. Liu et al. [2018] use a neural network that consists of an encoder E and a decoder D. The encoder network transforms the input image $\mathbf{I} \in \mathbb{R}^{M \times N}$ into a low-dimensional (latent) space $\mathbf{z} = E(\mathbf{I})$. The decoder then reconstructs the original image based on this low-dimensional representation $\hat{\mathbf{I}} = D(\mathbf{z})$. The encoder and decoder networks consist of $n = 8$ partial convolutional layers denoted as $E_1, \ldots, E_n$ and $D_1, \ldots, D_n$ for the encoder and decoder networks, respectively. Before each convolution operation, the image is constrained by the mask to condition the operation on only valid pixels. The mask is updated for the next layer removing masking for pixels where the convolutional operation operated on unmasked values. In addition,

each layer in the encoder network E_i is connected to the corresponding layer in the decoder network $D_i, \forall i \in \{1 \ldots, n\}$ using skip links. These skip links allow for copying unmasked pixels directly from the encoder to the decoder without passing the bottleneck (latent space). To direct the training of the network towards semantically meaningful inpaintings, a combination of four loss functions is used (i.e., per-pixel loss, perceptual loss, style loss, and total variation loss). Using these loss functions smooth transitions of the predicted masked values into their neighboring pixels are also taken into account. As activation functions Rectified Linear Unit (encoder) and a leaky version of a Rectified Linear Unit (decoder) are used.

3.3 Affective State Prediction

Our classification pipeline can be generally applied to any recordings captured with a tablet front camera or an external camera (such as a GoPro). Our method assumes that we have access to reports of affective states of users based on the circumplex model of affect [Russell, 1980]. The circumplex model defines affective states in a two-dimensional space spanned by valence and arousal (see Chapter 1.1.2). The classification task then amounts to preprocessing the camera recordings to adjust the brightness and the frame rate and predicting valence and arousal based on features extracted from the adjusted camera recordings. Affectiva [McDuff et al., 2016] provides out-of-the-box predictions of the basic emotions and valence based on images and video recordings. However, initial tests revealed that these predictions are not of sufficient quality when applied to our use case. Thus, we developed our own set of features incorporating some additional features not taken into account by Affectiva, such as movement and fidgeting. Moreover, by using our own extracted features, we can predict arousal in addition to valence.

3.3.1 Preprocessing

First, we resample the camera recordings using FFmpeg [Tomar, 2006] to a constant frame rate close to the mean frame rate. Depending on the recording device, the frame rate can vary (e.g., the frame rate can drop due to the higher load of the device). A constant frame rate facilitates the extraction of the features and the processing of the recordings in later stages. In addition, we adjust the brightness of the recordings based on the brightness estimation of Affectiva [McDuff et al., 2016] to improve the lighting of the face for the analysis. Depending on the conditions of illumination at recording time the face can be underexposed (too dark) or overexposed (too bright, e.g., when the camera is directed towards a lamp). This can hinder the accurate detection and extraction of facial features such as landmarks.

3.3.2 Feature Extraction

From the camera recordings, we extract several different feature types. We design all features such that they are independent of the frame rate (e.g., using percentages instead of absolute positions) to support cameras with different frame rates. To extract facial landmarks, eye gaze, and head position from the camera recordings, we rely on OpenFace [Baltrusaitis et al., 2018] using static extraction (i.e., per frame without calibrating to a person). OpenFace also provides a confidence value $c(i) \in [0, 1]$ for each frame i indicating the confidence in the landmark detection estimate. If $c(i) < 0.82$, we discard the frames $i - 5, \ldots, i + 5$ (i.e., 11 frames). The number of frames to discard (11) and the threshold (0.82) were heuristically determined. All features are computed over a window containing N frames. If, after considering the confidence value, less than 80% of the frames are remaining, we discard the window and the corresponding data point. Again, this threshold was determined heuristically. Where appropriate, we calculate for the different feature types basic statistics over the window (i.e., maximum, minimum, relative position of minimum and maximum, mean, standard deviation, and the slope of a fitted linear regression line), providing 282 features in total. In addition, to correct for differences between individuals related to facial expressions and posture, we normalize each feature according to a baseline by subtracting the feature calculated over a baseline period (e.g., watching a nature video putting the individuals in a relaxed state).

Action units. Facial action units (AUs) are based on the Facial Action Coding System (FACS) and identify independent motions of the face [Ekman and Friesen, 1978]. We extract basic statistics of the intensity (from 0 to 5) of 17 AUs covering motions in the eye, cheek, nose, mouth, and chin region. In addition, for each AU, we calculate the percentage of the presence (absent versus present) in the window. Moreover, the AUs can be directly mapped to the six basic emotions identified by Ekman [1999]. Thus, for each basic emotion, we also calculate the basic statistics of the corresponding added-up AUs.

Eye blinks. Researchers found a correlation between eye blink frequency and stressful situations in a car driving simulation [Haak et al., 2009]. Similarly, a correlation between eye blinks and affective states in learning environments was found [McDaniel et al., 2007]. Here, we base the eye blink detection on the signal from the AU that represents eye closure as a continuous signal (from 0 to 5) with peaks indicating potential eye blinks. We detect peaks belonging to an eye blink by thresholding the signal according to the ratio between the prominence (how much a peak stands out measured as the vertical distance between the peak and its lowest contour line) and the width of a peak. Heuristically, we found a threshold of 0.026 to provide the best results. We found that taking into account the width of the peaks is necessary to accurately detect peaks belonging to eye blinks because the

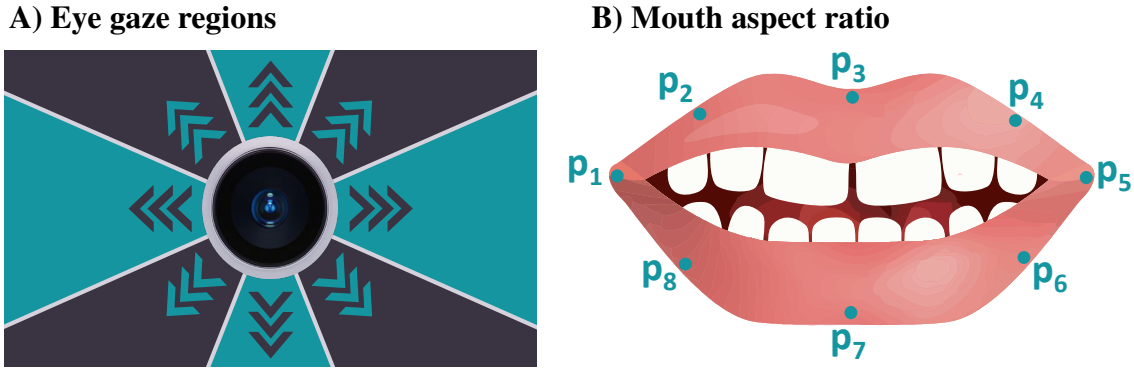


Figure 3.4: *Eye gaze regions and mouth aspect ratio (MAR). The gaze angle is discretized into nine different gaze regions, including the center (gazing towards the camera lens) (A). MAR is calculated based on the height and width of the mouth (B).*

prominence of the peaks differs among users and head pose. We extract the number of blinks and the basic statistics of the duration between blinks, the prominence, and the width of each blink. In addition, inspired by interbeat intervals (time intervals between individual heartbeats) and the calculation of heartbeats thereof, we linearly interpolate the duration between two consecutive peaks surviving the threshold (i.e., eye blinks) to infer a continuous signal. We then calculate the number of eye blinks for every frame by taking the inverse of this interpolated signal. Subsequently, we again calculate the basic statistics over the number of eye blinks.

Eye gaze. The intention behind features related to eye gaze is that individuals might look away when thinking while solving math tasks or when looking at emotionally disturbing pictures. Thus, we compute the basic statistics on the angle in the x-direction (looking left-right) and y-direction (looking up-down) of the eye gaze averaged for both eyes and measured in radians in world coordinates. In addition, we discretize the eye gaze angle by defining nine different gaze regions (see Figure 3.4A). The center corresponds to a line of gaze directed towards the camera lens. For each of the nine regions, we count the number of occurrences and normalize it over $s * \text{fps}$, where s is the window size and fps is the frame rate per second (so that it is independent of the used camera, i.e., the frame rate).

Mouth aspect ratio. Previously, the mouth aspect ratio (MAR) was used to detect driver drowsiness [Singh et al., 2018]. It is defined by the ratio between the height and the width of the mouth, which is increased when opening the mouth (see Figure 3.4B):

$$\text{MAR} = \frac{\|p_2 - p_8\| + \|p_3 - p_7\| + \|p_4 - p_6\|}{3 * \|p_5 - p_1\|}. \tag{3.5}$$

Each point $p_i, \forall i \in \{1, \ldots, 8\}$, is defined as the average of the inner and outer mouth landmarks. From the MAR, we calculate the basic statistics.

A) Original **B) Fidgeting**

Figure 3.5: *Fidgeting of a user. From the original image (A), the fidgeting image (B) is calculated by pixel-wise thresholding the difference of the current (A) to the past grayscale images.*

Head movement. From the longest head moving sequence of an individual in the window, we extract the position of the first frame of the sequence in relation to the beginning of the window, the duration of the movement, and the total distance of the movement. The position of the first frame and the duration are normalized by $s * \text{fps}$. We also sum up the total distance moved over the entire window to capture individuals continually moving back and forth. In addition, we calculate the basic statistics of the velocity and acceleration of the head movements in the window. All these features are extracted for the x-axis, y-axis, and z-axis separately. Finally, we also extract the basic statistics of the distance of the head to the camera in the three-dimensional space.

Fidgeting. Navarathna et al. [2014] introduced a fidgeting index for predicting movie ratings from audience behavior by calculating the total energy individuals are using for the movement. In contrast to features related to the head movement, fidgeting captures all the movement in the video (i.e., also body and face). First, we define the grayscale adaptive background b_{gray}, which is a weighted average of past frames. To calculate the energy E for a new frame f_{gray} (converted into grayscale), we subtract the adaptive background b_{gray} from f_{gray}, binarize the image by thresholding it, and then calculating the percentage of surviving pixels with respect to the camera resolution (see Figure 3.5B). We have chosen the threshold such that noise from the background is minimized, and the visibility of movements is maximized. Finally, the adaptive background is updated using

$$b_{\text{gray}} = (1-a) * b_{\text{gray}} + a * f_{\text{gray}}, \tag{3.6}$$

where a is a weight term (we found $a = 0.2$ to provide the qualitatively best results). From the energy E of each frame in the window, we calculate basic statistics, sum up the energies over all frames and use the position of the frame with minimum and maximum energy normalized by $s * \text{fps}$.

3.3.3 Classification

We build the ground truth for our classifiers by splitting valence and arousal into two levels (high and low). We then use classifiers to predict these levels based on the features extracted from the camera recordings. In addition, we remove features having a correlation greater than a threshold, select features based on the ANOVA F-value between the class labels and the features, and standardize the features to have zero mean and unit variance. We use four different classifiers (i.e., Random Forest, Support Vector Machine, k-Nearest Neighbors, and Gaussian Naive Bayes) because these classifiers have been most promising in initial tests and they have shown to provide good results for predicting affective states from video data in other works [Bosch et al., 2015; Calvo and D'Mello, 2010; Jaques et al., 2014]. We use leave-one-user-out cross-validation to evaluate our models, which ensures that data of a participant is not used for training and testing at the same time. Finally, we optimize the hyperparameters (i.e., number of selected features, the threshold for removing correlated features, and parameters of the model) using random search with nested cross-validation.

3.4 Experiment

We conducted a controlled lab experiment with 88 participants to test our pipeline. The experiment was approved by the ethics board of ETH Zurich. In the experiment, participants solved approximately 40 math tasks chosen to trigger different affective states. The math tasks were chosen because they are an integral part of the educational curriculum. However, instead of relying on a math-based intelligent tutoring system, we designed specific math tasks to increase the probability of evoking a wider range of affective states.

3.4.1 Experimental Setup

Participants. We recruited 88 participants (45 female) between ages of 18 and 29 (mean $=$ 22.1 years, standard deviation SD $=$ 2.0 years) from ten different engineering and natural science departments of the second and third year of the Bachelor program of ETH Zurich. We excluded participants suffering from cardiovascular pathologies, smokers, and participants suffering from evident mental pathologies (score $>$ 4 in the Patient Health Questionnaire [Kroenke et al., 2001]). In order to control for external factors, we kept the humidity and room temperature at an average of 21.7 °C (SD $=$ 0.59 °C) and 32.6% (SD $=$ 5.3%), respectively. Figure 3.6 presents the experimental setup.

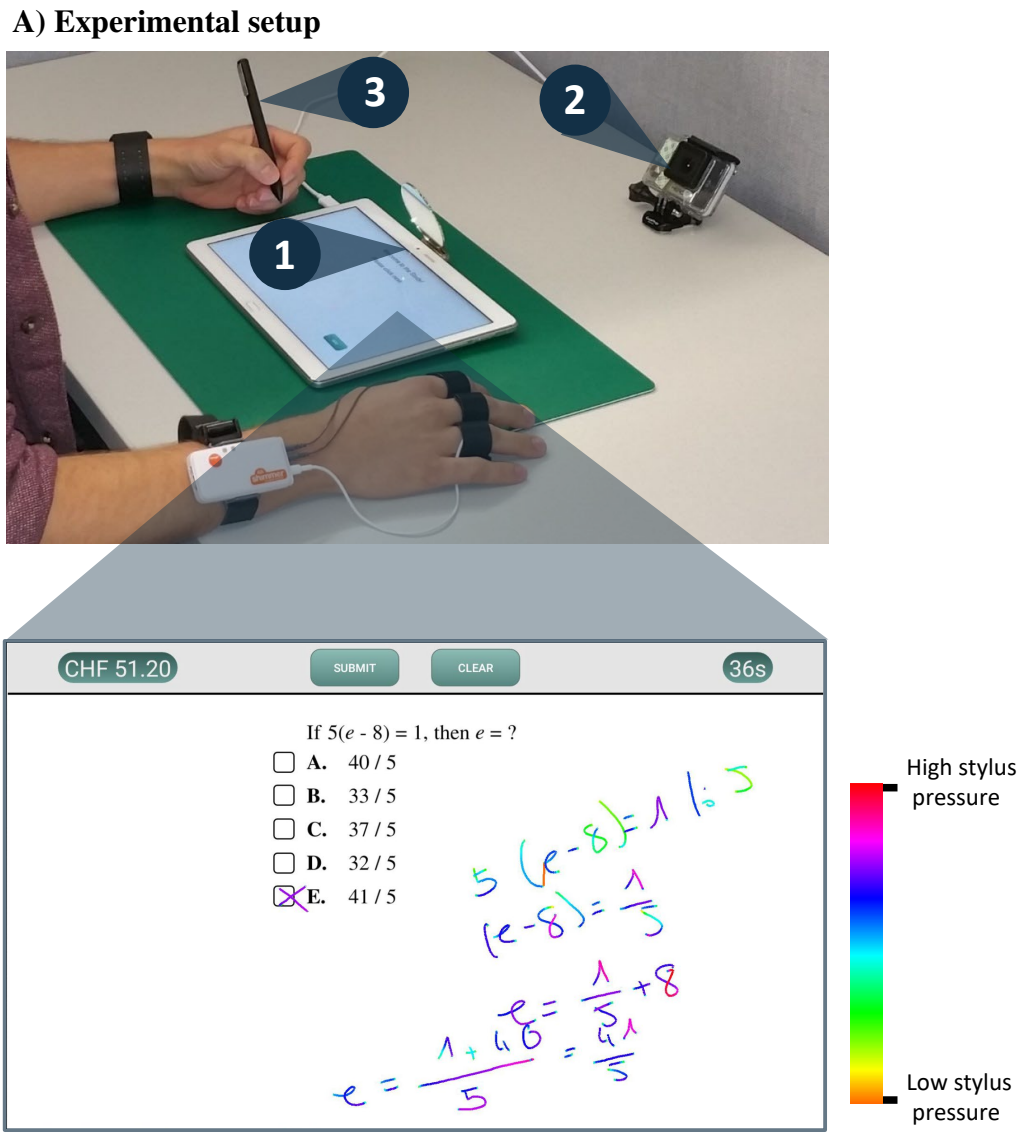


Figure 3.6: *A participant completing the math tasks. A) Participant were recorded by (1) the tablet front camera and (2) a GoPro HERO3. All interactions with the tablet were conducted with a stylus (3). B) The task interface allows participants to write solution paths directly onto the screen (the stylus pressure is color-coded for visualization purposes only).*

Devices. During the experiment, participants interacted with a Huawei MediaPad M2 10.0 running Android 5.1 to solve the different math tasks. All interactions with the tablet were conducted with a Wacom Bamboo Ink stylus. Participants were recorded by the front camera (resolution of 1280×720 pixels) using our proposed mirror construction setup and a GoPro HERO3 camera (frame rate per second fps of 59.94 and a resolution of 1920×1080 pixels) (see the setup in Figure 3.6A). Due to the varying load of the tablet during the experiment, the frame rate per second was variable (mean $= 20.2$ fps, SD $= 1.92$ fps). We resampled the recordings from the tablet and the GoPro to a frame rate per second of 25 and 60, respectively. To synchronize the timestamps between the GoPro and the tablet, a beep signal was played on the tablet before the start of each session. In addition, we also recorded signals from biosensor devices which we used for our model presented in Chapter 4.

3.4.2 Experimental Procedure

We used the self-assessment manikin (SAM) [Bradley and Lang, 1994] to measure valence and arousal on a scale from one (most negative, lowest arousal) to nine (most positive, highest arousal). For triggering the affective states we used math tasks and pictures from the International Affective Picture System (IAPS) [Lang et al., 2008]. The IAPS is a database of 1182 pictures typically used in emotion research and has been standardized in terms of valence and arousal based on SAM ratings. The set of IAPS pictures presented to the participants was sampled to cover similar affective responses as those expected to be evoked by the different math tasks.

An overview of the study procedure is presented in Figure 3.7A. The experiment lasted an average of 90 minutes for each participant. Upon arriving at the lab, participants completed a demographics questionnaire and were given an oral overview of the procedure. This included an explanation of the SAM questionnaire based on four example pictures from the IAPS presented on paper. Next, participants started working independently on the tablet by first watching a 7 minutes nature video (biosensor baseline), followed by the stylus baseline that consisted of writing an English sentence with the stylus (the biosensor and stylus baseline are used in Chapter 4). Participants were then presented with 40 pictures from the IAPS in random order. Each picture was shown for 10 seconds and was directly followed by the SAM rating (valence and arousal) and a 10 seconds fixation cross. In total, we collected 3400 ratings from all participants. After rating the IAPS pictures, participants were asked to watch the nature video one more time before completing the math tasks. Before finishing the experiment, participants completed a paper questionnaire about their overall mood, comfort level while wearing the sensors, nervousness, and sweating level.

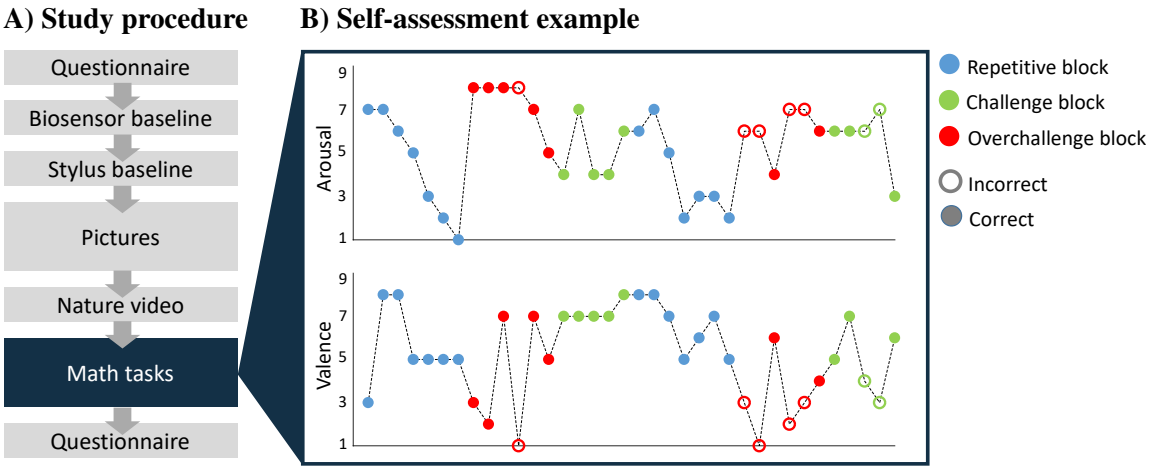


Figure 3.7: *Overview over the different parts of the study. A) Overall experimental procedure. B) Changes in valence and arousal for one participant in relation to task type and answer.*

3.4.3 Experimental Tasks

To trigger different affective states, we created three different math task conditions by varying the difficulty level, available time for completion, and monetary reward of the task. These types of manipulations were shown to be effective at eliciting different affective states in reading comprehension [Blanchard et al., 2014] and math tasks [Saneiro et al., 2014].

Task design. The math tasks were taken from an ACT data set [ACT, 2017] that provided difficulty ratings from 0.12 (most difficult) to 0.96 (simplest). We conducted a pilot study (same conditions, 11 participants) to get an indication of the time needed to solve the different tasks. Based on this timing information and the tasks from the ACT data set we generated the following three conditions.

1) *Repetitive condition.* For the *repetitive condition* we created random variants (by substituting the numerical values in the task) of two easy tasks from the ACT data set (difficulty of 0.76 and 0.83). The time available to solve each task was set between 60 and 75 seconds at random. This provided participants with more than sufficient time to come up with a solution for each task. Correctly solving a task in the *repetitive condition* granted only a minor monetary reward (+CHF 0.2) and a minor penalty (−CHF 0.2) for incorrect solutions. The *repetitive condition* was designed to trigger emotions such as boredom and fatigue.

2) *Challenge condition.* For the *challenge condition* we selected math tasks from the ACT data set with medium difficulty (difficulty $\in [0.58, 0.69]$) and provided participants with a larger monetary reward (+CHF 2) for correct solutions and the

same small penalty as the *repetitive condition* (−CHF 0.2) for incorrect solutions. Participants were provided with sufficient time to solve the tasks based on data from the pilot study min = 53 s, max = 93 s). The *challenge condition* was designed to provide diversified tasks for a more engaging and interesting experience, while the larger monetary reward provided a bigger incentive (higher-stakes) for participants to perform well with a relatively small penalty in case of mistakes.

3) *Overchallenge condition.* For the *overchallenge condition*, we selected the math tasks with high difficulty in the ACT data set (difficulty $\in [0.25, 0.53]$). Participants received small monetary rewards for correct solutions (+CHF 0.2) and a large penalty (−CHF 2) for incorrect solutions. The time to solve each task was set to be insufficient for most participants based on data from the pilot study (min = 25 s, max = 51 s). The *overchallenge condition* was designed to provide a frustrating and annoying experience to participants.

The math tasks were presented in six blocks (two in each condition) each containing a different number of tasks (*repetitive condition* 13 tasks, *challenge condition* 5 tasks, *overchallenge condition* 6 tasks). A similar block design for math tasks was already applied in previous work [Saneiro et al., 2014]. Moreover, we believe that a sequence of tasks is necessary to trigger an affective state. The first three blocks presented were randomly sampled. However, the succeeding three blocks were fixed to the same order as the first three blocks (but contained different tasks). In addition, the maximum time for each block was limited to 5 minutes to ensure that the math part of the experiment does not go over 30 minutes. After each block, a fixation cross was shown for 30 seconds to reduce potential carry-over effects of affective states. At the end of each math task, participants were asked to fill in the 9-point SAM scale to report their current valence and arousal level (in total, we collected 3026 ratings from the participants). Figure 3.7B depicts the changes in the valence and arousal ratings for one participant in relation to the block type and task answer (correct vs. incorrect). We see that for the repetitive tasks, valence and arousal are decreasing over time leading to a shift towards boredom. Additionally, for incorrectly solved tasks, valence drops and arousal tends to increase. After the repetitive blocks, we see a decrease in valence and an immediate steep increase in arousal that may be attributed to the increase in difficulty from the repetitive block to the overchallenge block. On average participants finished with CHF 44.3 (min = CHF 22.2, max = CHF 62.8). At the end of the experiment, each participant was compensated with a minimum of CHF 40.

Math task interface. Participants were asked to provide a solution path for every task anywhere on the screen and then to select their answers from five multiple-choice alternatives (see Figure 3.6B). Participants received immediate feedback on whether their answer was correct. A timer located on the top right corner of the interface informed participants about the time left to respond and started to blink when less

than 10 seconds remained. When the time was up and the participant did not submit a solution, the answer was considered wrong. The cumulative amount of money earned was displayed on the top left of the interface.

3.5 Results

We conducted a qualitative and quantitative evaluation of our mirror setup and image processing pipeline with neural inpainting and investigated the applicability of our setup to predict affective states during math-solving tasks (active) and exposure to emotional stimuli from images (passive). For training the neural inpainting model, we used the celebA-HQ dataset [Karras et al., 2018] consisting of 30000 face-aligned colored images from celebrities with a resolution of 1024×1024 pixels (we downsampled the images to 512×512 pixels). We split the dataset into a training set of 25000 images, a test set of 2500 images, and a validation set of 2500 images. We set the parameters for the network in the same way as proposed by Liu et al. [2018]. The results of the affective state prediction are based on a Random Forest classifier since this was the best performing model. Hyperparameters were optimized using random search with 50 iterations. Finally, for measuring the performance of our model, we used the area under curve (AUC) of the receiver operating characteristic curve and accuracy (chance level is 0.5).

3.5.1 Study Validation

Our study was designed to trigger affective states across the entire valence-arousal space. As a first step, we investigate if our study design worked by examining if the different parameters acted as intended. In our task design, we varied task difficulty, monetary reward, and the available time for task completion. We performed a per-task Kendall's tau correlation analysis between these three parameters and the arousal and valence ratings of the participants. For the task difficulty and the percentage of remaining time, we found high correlations for both valence (-0.2; $p < 10^{-59}$ and 0.22; $p < 10^{-80}$) and arousal (0.27; $p < 10^{-102}$ and -0.27; $p < 10^{-117}$). Participants shifted towards frustration (decreasing valence and increasing arousal) with increasing task difficulty or with a reduction in the time remaining to complete the task. Interestingly, the effect size on valence and arousal is almost identical. In contrast, monetary reward appears to have a much larger effect on valence (0.47; $p < 10^{-295}$) than on arousal (-0.06; $p < 10^{-4}$). Accounting for potential superficial correlations (e.g., task duration) is an important part of our study design. We found a significant Kendall's tau correlation between the task duration and the user ratings of 0.17 ($p < 10^{-48}$) and -0.11 ($p < 10^{-22}$) for arousal and valence, respectively.

Table 3.1: *Means of framewise confidence in landmark detection for different camera sources, tasks (math and IAPS) and the full recordings. Confidence values range from 0 (not confident) to 1 (fully confident). Standard deviations are given in brackets.*

Source	IAPS	Math	Complete
Front (no inpainting)	0.79 (0.36)	0.48 (0.45)	0.68 (0.42)
Front (inpainting)	0.94 (0.14)	0.90 (0.22)	0.93 (0.18)
GoPro	0.97 (0.08)	0.93 (0.17)	0.95 (0.12)

The participants had the most remaining time available for the tasks in the boredom block (mean = 42 s, SD = 6 s), followed by the engagement block (mean = 23 s, SD = 8 s) and the frustration block (mean = 5 s, SD = 2 s). The tasks in the boredom block were solved correctly by most participants (mean = 97%, SD = 4%), whereas the participants performed poorer for the tasks in the engagement block (mean = 70%, SD = 2%) and frustration block (mean = 41%, SD = 2%). Altogether, it appears that our tasks worked as intended.

3.5.2 Face Recognition

We provide qualitative and quantitative results of our setup using neural inpainting. In particular, we compare our results to recordings taken by the GoPro camera.

Qualitative evaluation. Figure 3.8 shows the facial landmarks detected by OpenFace for three participants from the front camera without inpainting, using neural inpainting, and from the GoPro. The positions of the detected landmarks without inpainting are inferior compared to neural inpainting. For participant 3, the landmarks at the upper face (eyebrows, eyes, and nose) are misaligned without inpainting. Often no facial landmarks could be detected (see Figure 3.8 participants 1A and 2A). With our neural inpainting approach, we achieved a qualitatively good recovered image independent of the position of the missing region (e.g., eyes and mouth). It is noteworthy that the inpainting and facial landmark detection also worked for participants wearing glasses. The detected landmarks after neural inpainting are similar to the landmarks detected from the GoPro recordings (see Figure 3.8C). Depending on the position of the head, the landmarks of the eyes and the mouth can become locally condensed in the GoPro recordings, and it might be hard to distinguish slight facial movements. On the other hand, from the front camera, the recordings are frontal, and the variations of facial parts (e.g., eye and mouth) are better visible.

Quantitative evaluation. Table 3.1 presents the average confidence in landmark detection of OpenFace over all frames for the IAPS and math-solving tasks and

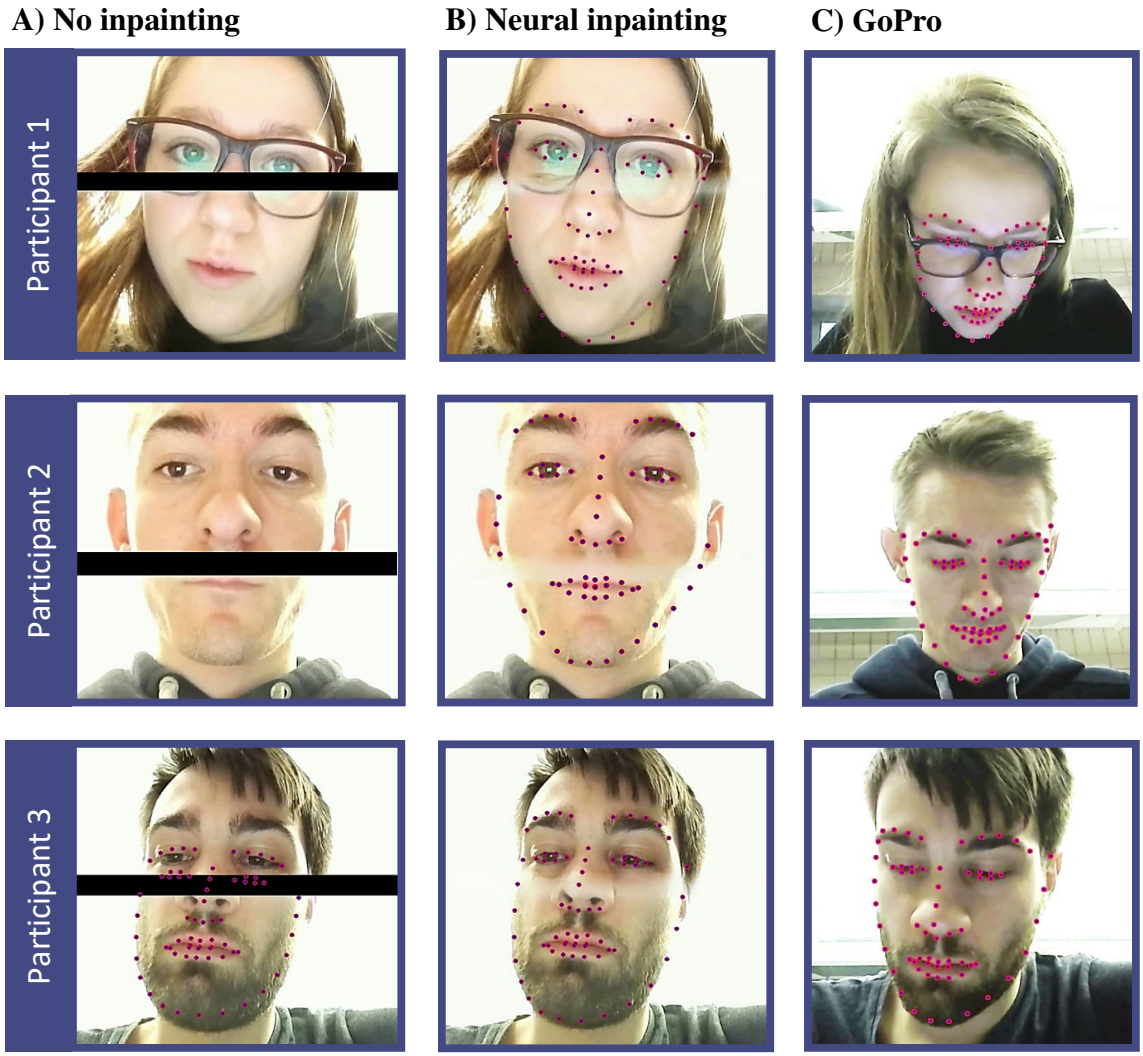


Figure 3.8: *Recordings of three participants. The facial landmarks were detected from the front camera recordings without inpainting (A) and with neural inpainting (B) and from the external GoPro camera (C). If no landmarks are visible, no landmarks were detected by OpenFace.*

the full recordings (including also parts not belonging to the IAPS and math tasks). Reported confidence values by OpenFace are between 0 (not confident) and 1 (fully confident). Without inpainting, the confidence values are low, and standard deviations are high due to the imperfect recognition of landmarks. Without inpainting, landmarks were often only detected correctly when the missing regions were situated above the eyebrows (i.e., no landmarks were affected). After applying neural inpainting, the confidence values increased by 19% and 88% during IAPS and math sequences, respectively. When considering the full video recordings, the increase amounts to 37%. In addition, the standard deviation decreased substantially. This increase of

Table 3.2: *Performance of Random Forest on the math and IAPS data from two levels (low and high) of valence and arousal based on the front camera recordings with neural inpainting and the GoPro recordings. The chance level for accuracy and AUC is 0.5.*

Source	Data	AUC	Accuracy
Front camera	Math (valence)	0.73	68%
	Math (arousal)	0.54	57%
	IAPS (valence)	0.80	73%
	IAPS (arousal)	0.70	66%
GoPro	Math (valence)	0.76	72%
	Math (arousal)	0.58	62%
	IAPS (valence)	0.78	72%
	IAPS (arousal)	0.73	67%

confidence leads to an increase in the number of samples (if a window used during feature extraction contained less than 80% frames with a confidence value above 0.82 we discarded the corresponding data point). For IAPS, this leads to 348 and 383 additional samples for valence and arousal, respectively. For the math tasks, this amounted to 1233 and 1179 additional samples for valence and arousal, respectively. Finally, the confidence in landmark detection of the GoPro recordings is comparable to the front camera recordings with neural inpainting. In general, for recordings taken during exposure to a stimulus set of images the mean confidence is higher than during math tasks. This can be attributed to the fact that while solving math tasks, participants were moving more, which leads more often to suboptimal head positions for landmark detection. This finding is also reflected in the higher standard deviations of the confidence values for math tasks.

3.5.3 Classification Performance

Before predicting the affective states, the reconstructed front camera recordings and the GoPro recordings were preprocessed (see Section 3.3.1). Features were extracted using a 10 seconds window encompassing the on-screen time of each picture and the last 10 seconds of each math task because each picture was presented for 10 seconds and the minimum task duration was 10 seconds. Table 3.2 presents the performance of our model for predicting two levels (low and high) of valence and arousal. Based on the findings that the confidence in landmark detection increased up to 88% with neural inpainting, we used only the front camera recordings with neural inpainting. Using these recordings, our model achieved a performance of 0.73 AUC and 0.80 AUC for predicting valence on math tasks and IAPS, respectively. For predicting

Table 3.3: *Number of occurrences of each feature type in the ten most predictive features. The numbers are provided for each of the four models (MV = math valence, MA = math arousal, IV = IAPS valence, IA = IAPS arousal).*

Feature Type	MV	MA	IV	IA
Action units	0	2	2	3
Eye blinks	1	4	0	1
Eye gaze	1	2	2	0
Mouth aspect ratio	0	0	0	0
Head Movement	5	2	5	6
Fidgeting	3	0	1	0

arousal, the performance drops and is only at random level for math tasks (0.54 AUC), while for IAPS it is above random (0.70 AUC). A similar pattern is visible for the GoPro recordings. While for predicting arousal based on the math tasks, the performance is close to random (0.58 AUC), all other predictions are above random. In summary, the predictions using the front camera are comparable to using the GoPro recordings with a maximum difference of 0.04 AUC. For predicting valence based on IAPS, the performance from the front camera recordings (0.80 AUC) exceeds the performance achieved by using the GoPro (0.78 AUC).

Feature importance. Table 3.3 presents the number of occurrences of each feature type in the 10 most important features for each of the 4 models. We analyzed the feature importance using the Gini importance measure provided by the Random Forest classifier. Features related to head movement contributed the most to predicting valence based on math tasks (five features) and valence and arousal based on IAPS (five and six features). For predicting arousal based on math tasks, eye blinks provided 4 out of the 10 most important features. There were no MAR features among the top 10 features for any model. However, all feature types appeared in the top 30 ranked features of each model. For the model based on the math tasks, the maximum moved distance in the x-direction and the number of eye blinks were the highest scoring features for predicting valence and arousal, respectively. For the model based on IAPS, the mean acceleration in the x-direction and mean velocity in the x-direction were most important for predicting valence and arousal, respectively. Interestingly, head movement along the x-axis (left and right) was more informative than along the z-axis (forward and backward).

3.5.4 Runtime

We conducted a runtime analysis of the different parts of our inpainting pipeline and affective state prediction model. Our computing environment consisted of an

Intel® Core™ CPU i9-9900K @ 3.60GHz and an NVIDIA GeForce® RTX 2080 Ti. Processing one frame consisted of flattening the splitting boundary, face composition, image rotation and extracting the face area (mean = 17.07 ms, SD = 4.74 ms), detecting the position of the eyes using dlib (mean = 74.66 ms, SD = 6.43 ms), using the deep learning model to inpaint missing regions in the face (mean = 76.25 ms, SD = 13.81 ms) and inpainting the background of the image (mean = 47.01 ms, SD = 11.87 ms). Summing up these values leads to a processing time for one frame of 214.99 milliseconds. Prediction of a new data point consisted of feature extraction (mean = 16.37 ms, SD = 2.18 ms) and using the Random Forest classifier for predicting valence and arousal (mean = 6.43 ms, SD = 10.52 ms), leading to a total prediction time of 22.8 milliseconds.

3.6 Discussion

Our findings show that it is possible to use our tablet-based front camera setup and processing pipeline to accurately capture users for extracting features such as facial landmarks and movement of the head and body. Our neural inpainting pipeline provides a qualitatively accurate restoration of missing regions caused by our mirror construction setup and increases the confidence in landmark detection by up to 88%. Compared to recordings from a GoPro camera, our setup provides better results in terms of face visibility (frontal view). Thus, it potentially facilitates the recognition of minor facial movements (e.g., mouth and eyes). In particular, for solving math tasks we found the recording conditions of the GoPro more challenging due to the viewing angle (participants were bending over the tablet). This resulted in lower confidence in landmark detection (0.93 for math tasks versus 0.97 for IAPS). Similarly, the front camera recordings with neural inpainting showed higher confidence in landmark detection during exposure to pictures from the IAPS (0.94) compared to solving math tasks (0.90). During the exposure to a stimulus set of images from the IAPS dataset, participants were sitting straight, implicating that the splitting boundary was located at the forehead, which made inpainting easier. In contrast, during solving math tasks, the splitting boundary was often located in the middle (eye) or lower part of the face (mouth), creating a more challenging situation for our neural inpainting model.

We showed the applicability of our setup for predicting affective states during active (math-solving) and passive (exposure to pictures) tasks based on the recordings from the front camera. Our model achieved better performance on IAPS (up to 0.80 AUC) than on the math tasks (up to 0.73 AUC). Due to the active involvement of the participants while solving math tasks, participants were moving more, which made accurate tracking of facial landmarks, AUs, and eye gaze more demanding. In addition, our model performed better for predicting valence (0.73 AUC and 0.80 AUC) than arousal (0.54 AUC and 0.70 AUC). Although affective states are universal,

they also have components that are individual to a person [Elfenbein and Ambady, 2003]. This makes it harder to predict an affective state of a person without having training data available of that person. Comparing the performance of our affective prediction pipeline to other research is difficult because most existing work [Calvo and D'Mello, 2010; Zeng et al., 2008] predicted basic emotions and used other settings.

Our analysis of the feature importance showed that head movement is a predictive feature in contrast to MAR. Some AUs capture movements of the mouth. Thus, we analyzed the correlation between MAR and AUs specific to the mouth region. The correlations between the MAR feature and the AUs specifying lip corner puller (-0.15, $p = 0.15$), opening the mouth (0.25, $p = 0.13$) and jaw drop (0.045, $p = 0.26$) have all been low and not significant.

In comparison to recordings from the GoPro, our model based on front camera recordings performed equally well and even better for predicting valence on IAPS (0.80 AUC versus 0.78 AUC). This renders our setup a viable alternative to more expensive equipment such as a GoPro. Our setup comes at low costs (CHF 5), is unobtrusive, can easily be mounted, is flexible in the application (e.g., in classrooms or at home), and eliminates the need for synchronizing different devices. In contrast to external cameras, the camera (i.e., the lens) in our setup is small and unobtrusive. Some participants reported after the experiment that they got slightly distracted by the GoPro but not by our mirror setup. Similarly, in the video recordings, we recognized that participants were sometimes glancing at the GoPro. Finally, with a processing time of 214.99 milliseconds per frame, our pipeline can handle four frames per second. Our affective prediction pipeline is capable of making 43 new predictions every second.

We acknowledge potential limitations to our approach presented in this chapter. Our setup is constrained by the lighting conditions, head pose, and occlusions from hand movement. We believe that other camera setups suffer from the same constraints. Further, our mirror construction is a prototype and not yet ready for production. Although during the experiment the construction proved to be stable, it can be improved in terms of stability and flexibility. Neural inpainting provided qualitatively satisfactory results for most facial parts. However, if the splitting boundary is covering the eyes (i.e., both eyes are occluded), it is hard for the inpainting model to reconstruct the eyes at a qualitatively high level. Consequently, the landmark detection cannot recover eye gaze and eye blinks, but still detects other facial features. In addition, although the CelebA-HQ dataset consists of facial images from celebrities with diverse ethnicity, age, and facial characteristics (e.g., glasses and facial hair), our inpainting method might be less appropriate for users who are underrepresented in the CelebA-HQ dataset. We further acknowledge that our experiment is restricted to math tasks and exposure to emotional stimuli from pictures in a lab environment

with bachelor students. We are optimistic that our approach generalizes to a broader population and to other tasks given that we used active (math-solving) and passive (exposure to pictures) tasks and assuming a proper baseline normalization of the features. In addition, participants reported that the setup was comfortable and that they could act in a natural way. Finally, we predicted valence and arousal on two levels omitting data points in the medium range (4 to 6). We mainly built our affective prediction model for investigating the applicability of video-based features. In the next chapter, we will consider affective regions in the valence-arousal space. In addition, by taking into account other data modalities (i.e., biosensors and handwriting) we will try to overcome limitations inherent to the approach presented in this chapter (e.g., deteriorated performance due to camera occlusions, lighting conditions, and head pose).

3.7 Conclusion

In this chapter, we presented a hardware setup consisting of a cheap and unobtrusive mirror construction to improve the visibility of the face in tablet-based front camera recordings. Recordings were processed using an inpainting pipeline consisting of a neural network for reconstructing missing data in the recordings. We showed that the mirror construction improved the visibility of the face in situations where external cameras (e.g., GoPro) struggle. With a qualitative and quantitative evaluation, we demonstrated that we could achieve results comparable to a GoPro camera. In particular, neural inpainting improved confidence in facial landmark detection by up to 88%. We showed the applicability of our setup and processing pipeline on affective state prediction based on front camera recordings. Our model consisted of features capturing information from movement, eyes, and face. We evaluated our affective prediction model on data from a lab experiment with 88 participants using leave-one-user-out cross-validation. Participants were solving math tasks (active) and were exposed to emotional stimuli from pictures (passive). Our model accurately predicted two levels (low and high) of valence (up to 0.80 AUC) and arousal (up to 0.73 AUC) using data from the front camera. These results were comparable to results obtained using recordings from a GoPro camera (up to 0.78 AUC for valence and up to 0.73 AUC for arousal). Our setup is cheap (CHF 5), easy to mount, and can be used in classrooms or at home. Besides affective state prediction, it can be used to monitor students or analyzing attention. Most existing approaches use external cameras such as GoPros or webcams, which are more expensive, more difficult to handle, and are exposed to time synchronization problems. In our setup, the camera data is recorded on the same device as the task is conducted, and thus we circumvent such time synchronization issues in an elegant way.

CHAPTER 4

Affective State Prediction Using Biometric Sensors and Stylus

Previously, a wide range of data sources has been used to predict affective states including audio [Saneiro et al., 2014] and interaction data [Grawemeyer et al., 2016; Kostyuk et al., 2018]. Data from biosensors (e.g., measuring muscle activity [Conati and Maclaren, 2009] and heart rate [Blanchard et al., 2014]) have also been used to predict emotions. However, most of these devices are typically restricted to lab settings, expensive and difficult to operate, and somewhat intrusive. Recently, a variety of portable and low-cost biosensor devices became available (e.g., Shimmer GSR+, Polar H10, and Empatica E4). These devices have the potential to transform affective research because they can be used to monitor a user's physiological state at home or in a classroom.

In this chapter, we explore a low-cost mobile setup to detect the affective state based on biosensor and handwriting data. Our goal is a system to detect affective states that is cheap and easy to operate, can be used outside a lab setting, is non-intrusive, and minimizes potential issues related to privacy. We consider biosensor data from skin conductance, heart measures, and skin temperature. In addition, we also evaluate handwriting data recorded by a stylus to predict the affective state. Here, we use the fact that tablets bundled with a stylus are becoming increasingly available in households and classrooms and are inherently non-intrusive and mobile.

We propose a generic pipeline in which we process the data from the biosensors and stylus to extract a set of features for each of the sensors. We then use a classification model to predict the current affective region in the valence-arousal space of emotions. Our method allows researchers to define arbitrary areas of interest in the valence-

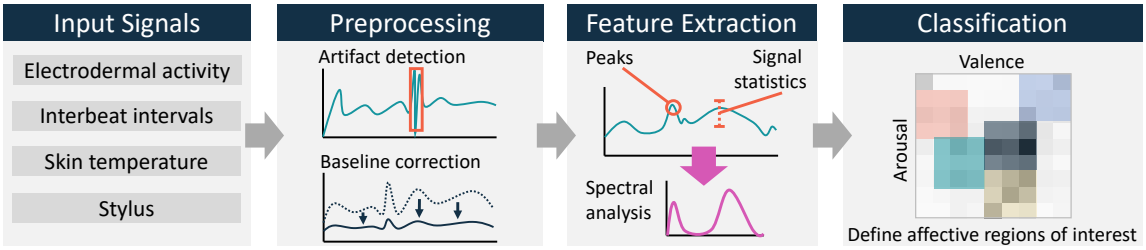


Figure 4.1: *The classification pipeline. Stylus and biosensor data are gathered during task solving processes. After preprocessing the signals, features are extracted and used to classify the affective regions of interest.*

arousal space and can be applied to a wide range of applications and questions of interest.

We evaluate our method by applying it to the math problem-solving scenario presented in Chapter 3 in which participants provided answers in unstructured handwriting on a tablet device. Best performance is reached when data from all sensors is used for the prediction (0.88 AUC). Interestingly, we reach a comparable performance using only the data acquired by the stylus (0.84 AUC). These results suggest that a simple tablet with a stylus can be sufficient to reliably predict a student's affective state. Finally, we also explore whether the affective state model could be generalized over domains. For this purpose, we apply the trained model to a passive setting with picture stimuli leading to a performance of 0.68 AUC.

4.1 Method

We present a classification pipeline that automatically predicts affective states based on low-cost and mobile biosensor devices and stylus pens. Our pipeline assumes that we have access to reports on affective states of users based on the circumplex model of affect [Russell, 1980]. The classification task then amounts to classifying regions within this space using a combination of signals from biosensor and stylus devices. For this purpose, we build a generic affective predictor (see Figure 4.1). Recorded stylus and biosensor data are preprocessed and the relevant features are extracted to train a classification model for the specific affective regions.

4.1.1 Input Signals

During the task solving process biosensor and stylus data are recorded. Physiological responses are modulated by the autonomic nervous system (ANS) which, in turn, reacts to affective states [Andreassi, 2010]. The ANS controls the function of our

organs and glands. The ANS consists of the sympathetic (mobilize the body's fight-or-flight response) and parasympathetic (controls the body's rest-and-digest response) branches.

Electrodermal activity (EDA). EDA is an indicator of the emotional state of a person reflected by the variation in the electrical characteristics of the skin as a result of sweating [Benedek and Kaernbach, 2010]. EDA is affected by the sympathetic nervous system and is quantified by measuring the amount of current flowing between electrodes attached to the skin. Changes in affective states can lead to subtle variations in the level of sweat that can be detected as the changes in the current. Typically, the EDA signal is decomposed into tonic (low frequency) and phasic (high frequency) components. The tonic signal varies in terms of tens of seconds while the phasic signal reacts within seconds after an external stimuli [Fritz et al., 2014].

Interbeat intervals (IBIs). IBIs are the time intervals between consecutive heartbeats in normal heart function. This natural variation is also known as heart rate variability (HRV). HRV reacts within a few seconds to changes in sympathetic and parasympathetic activation [Malik et al., 1996]. The heart rate (HR) can be computed as the inverse of the IBI averaged over a certain time window.

Skin temperature (ST). Skin temperature measures the thermal response of human skin. Vasoconstriction (e.g., provoked by an affective state) can increase blood flow, and consequently, skin temperature [Kim et al., 2004]. Skin temperature is modulated by both the sympathetic and parasympathetic nervous system [Calvo et al., 2015].

Stylus. Tablet devices often come equipped with stylus pens as accessories that can provide precise and pressure-sensitive input. Stylus data consists of the applied pressure during writing and the pixel positions of the written text. From these measurements, handwriting characteristics related to time and ductus can be calculated. Handwriting characteristics can be affected by cognitive processes and are indirectly connected to the ANS [Smith and Smith, 1991]. For example, increased muscle contraction due to an increase in the sympathetic nervous system can lead to increased pressure applied to the stylus.

4.1.2 Preprocessing of Signals

During preprocessing, the raw input signals are filtered to detect artifacts from movement and muscle contraction. The signals are also corrected for differences between individuals using baseline recordings for each individual.

Artifact detection. We follow the procedure outlined by Greco et al. [2016] to decompose the EDA into tonic, phasic, and an additive white Gaussian noise component with a convex optimization approach that accounts for signal filtering and detrending.

For IBIs, detrending is not necessary for the preprocessing [Yoo and Yi, 2004], and we use the criterion beat difference for artifact detection [Hovsepian et al., 2015].

Baseline correction. Similar to previous work [Jraidi et al., 2014; Salmeron-Majadas et al., 2015], we collect baseline data for all sensors to account for individual differences in stylus and biosensor signals related to writing habit, ambient temperature, and dryness of the skin. Baseline data is collected while individuals remain in a relaxed state (e.g., watching a nature video). We search for the minimum value of each biosensor signal during the relaxation phase over a 10 seconds window using a sliding window approach to be robust against outliers. Due to possible signal lags, we search the minimum for each signal separately. We then normalize the biosensor data by subtracting the feature values calculated over the corresponding 10 seconds interval of the baseline from the actual feature values computed during task solving. Stylus data is normalized by subtracting a baseline for all features computed over the handwriting of an English sentence.

4.1.3 Feature Extraction

In the proposed pipeline, we extract several different feature types from the stylus and biosensor signals. Where appropriate, we compute basic statistics for these features types including the mean, standard deviation (SD), minimum and maximum, and the linear trend (slope of a fitted linear regression line). A summary of all extracted features is presented in Table 4.1. Because we extracted the stylus features over the whole task, we excluded all features having a significant Spearman correlation to the task duration (features greyed out in Table 4.1).

Electrodermal activity. For EDA, we decompose the signal into phasic and tonic components and calculate standard statistics (i.e., mean, SD, min, max, slope). For the phasic component, we also calculate the area under the curve (AUC) [Betella et al., 2014] and the number of peaks using zero-crossings of the smoothed gradients of the signal [Kim et al., 2004]. Based on the extracted peaks, we further compute amplitude statistics (i.e., mean, min, max) [Züger and Fritz, 2015].

Interbeat intervals. From the IBI recordings, we extract temporal and frequency features. In the temporal domain, we calculate the percentage of successive IBIs that differ by more than 50 milliseconds (pNN50) and 20 milliseconds (pNN20) as well as the standard deviation and root mean square of successive differences between adjacent IBIs (SDSD and RMSSD) [Malik et al., 1996; Shaffer and Ginsberg, 2017]. For the frequency domain, it is well known that the distribution of spectral power gives an indication of physiological activation [Betella et al., 2014]. Therefore, we extract a feature related to the high frequency (HF) band of 0.15 Hz–0.40 Hz by a Fast Fourier transform of the cubic spline interpolated signal [Malik et al., 1996;

Table 4.1: *Extracted biosensor and stylus features. For each signal, the features are sorted according to their importance (based on our experiments). The 10 most predictive features are highlighted in bold. SD refers to the standard deviation.*

Signals	Features
EDA	**Phasic AUC**, **Phasic Mean**, **Tonic SD**, Tonic Max, Tonic Mean, Tonic Min, Phasic SD, # Phasic Peaks, Tonic Slope, Max Phasic Peak Amplitude, Min Phasic Peak Amplitude, Phasic Slope, Mean Phasic Peak Amplitude
Heart	**IBI SDSD**, **IBI RMSSD**, **IBI SD**, IBI pNN20, HR Mean, IBI High Frequency, IBI pNN50, IBI Mean, HR Min, HR Max, HR SD, HR Slope
Temperature	Max, Mean, Min, Slope, SD
Stylus	**#Strokes/Mean Speed**, **Mean Distance between Strokes**, **Max Distance between Strokes**, **SD Distance between Strokes**, Mean Pressure, Max Pressure, Mean Stroke Acceleration, Max Stroke Acceleration, Max Stroke Speed, Max Speed between Strokes, Mean Speed between Strokes, SD Speed between Strokes, SD Stroke Speed, SD Stroke Acceleration Excluded[1]: %Writing, {SD, Slope, Skewness} Pressure, {Mean, Min, Slope} Stroke Speed, {Min, Slope} Stroke Acceleration, Min Speed between Strokes, Min Distance between Strokes, #Strokes/Minute

[1] Excluded due to our experimental setup (features having a significant Spearman correlation to the task duration)

Shaffer and Ginsberg, 2017]. Based on the IBIs, we compute the heart rate for which we extract several standard statistics (i.e., mean, SD, min, max, slope).

Skin temperature. We extract several statistics (i.e., mean, SD, min, max, slope) from the temperature signal [Shi et al., 2010; Züger and Fritz, 2015].

Stylus. From the stylus data, we derive features related to the pressure applied by the pen as well as timing and location information. Previous research has successfully employed these features to predict affective states [Fairhurst et al., 2015; Likforman-Sulem et al., 2017]. From the pressure data, we compute standard statistics (mean, SD, max, min) per stroke and average these over an entire task. Additionally, over each task, we compute the slope of a linear regression fit to the pressure values and the statistical skewness of the pressure distribution. We also compute standard statistics (i.e., mean, SD, max, min, slope) of the speed and acceleration of the strokes. For the handwriting data, we discriminate between the actual writing process and the think time while completing the task [Likforman-Sulem et al., 2017]. During writing, there are always small time gaps between strokes that cannot be attributed to thinking but belong to the writing process itself. Because writing patterns are different for every user, we infer an individual threshold for each user to distinguish if the time

between two strokes belongs to thinking or to the actual writing process. We chose this threshold as the 80% cut-off value of the distribution of the time between the strokes over the stylus baseline (cropping the right tail of the distribution). Based on this threshold, we derive a feature measuring the percentage of writing (i.e., the time spent in the writing process). Additionally, we compute the statistics (i.e., mean, SD, max, min) on the speed between consecutive strokes having time differences below the threshold (writing process) and on the distance between strokes having time differences above the threshold (thinking).

4.1.4 Classification

To train our classification algorithms ground truth is built by defining arbitrary non-overlapping regions of interest in the two-dimensional valence and arousal space based on the affective labels which can be gathered, for example, through self-reports or expert labelers. We then use a classification model to predict the affective region an individual is likely to be in during task solving based on the recorded biosensor and stylus data. Before applying the classification algorithm, we standardize all features to have zero mean and unit variance. We propose the usage of four different classifiers (i.e., Random Forest, Support Vector Machine, k-Nearest Neighbors, and Gaussian Naive Bayes). We select these classifiers because they are among the most widely used in machine learning and have shown to provide good results on biosensor and stylus data [Fritz et al., 2014; Likforman-Sulem et al., 2017; Zhou et al., 2014]. All models are evaluated using leave-one-user-out cross-validation which ensures that data from the same user is not in the testing and training set at the same time. Hyperparameter optimization is performed using nested cross-validation and randomized search.

4.2 Results

We compared different versions of our classification pipeline using only a subset of the sensors with a focus on the difference between stylus and biosensors. All results are based on Random Forest (using 500 trees, balanced class weights, and hyperparameter optimization using randomized search with 100 iterations) given that this was the best performing classifier. To measure the performance of our classifiers, we used accuracy (chance level $=$ 1/# classes) and micro-averaged area under curve (AUC) of the receiver operating characteristic (ROC) curve (chance level $= 0.5$), which aggregates the contributions of all classes to compute the average metric. Because both metrics are affected by class imbalance, we also considered the macro-averaged AUC (chance level $= 0.5$) which is the average of the class-wise

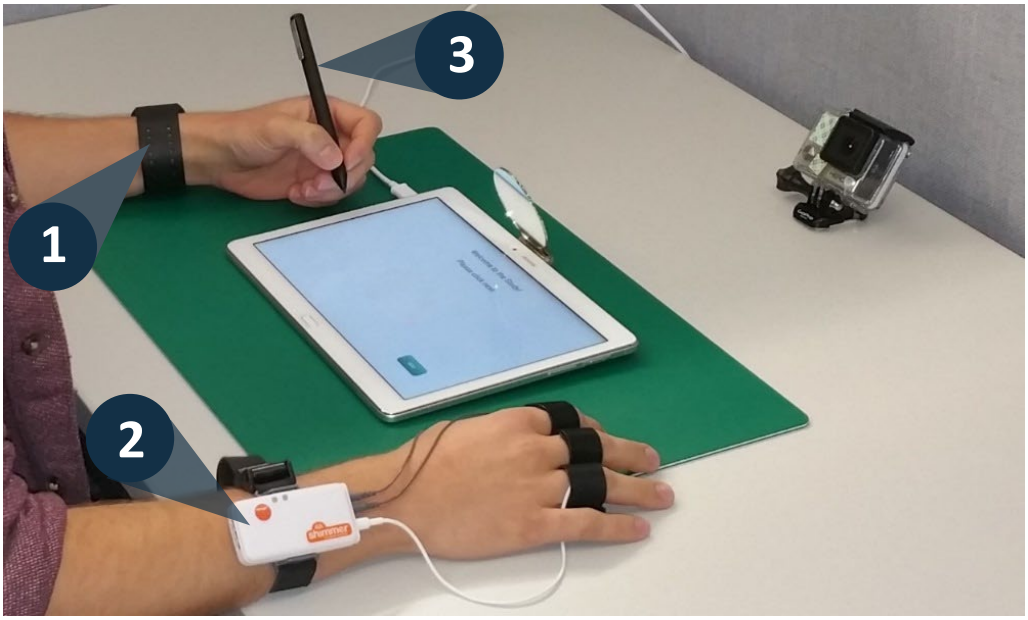


Figure 4.2: *Experimental setup. During each session data is recorded from different devices. (1) An Empatica E4 recording skin temperature on the dominant hand. (2) A Shimmer GSR+ measuring skin conductance and wrist acceleration on the non-dominant hand. All interactions with the tablet were conducted with a stylus (3). Participants also wore a Polar H10 chest belt (not visible in the image) for recording heart activity.*

AUCs giving each class the same weight. To derive the standard deviation for each metric, we employed an additional 10-fold cross-validation.

4.2.1 Experiment

We reused the dataset that we collected in a laboratory experiment described in Chapter 3.4). Figure 4.2 shows again the experimental setup. All interactions with the tablet were conducted with a Wacom Bamboo Ink stylus at an average sampling rate of 250 Hz (SD $=$ 25 Hz) and with 2048 levels of pressure sensitivity. We measured skin conductance and wrist acceleration of the participants using a Shimmer GSR+ device. To test the accuracy of the device, we compared its measurements with a state-of-the-art ADInstruments PowerLab 8/35 device (connected through the ADInstruments FE116 GSRAmp signal amplifier) over a 23 minutes recording of a user watching a nature video and picture stimuli. Results revealed a strong and significant cross-correlation value of 0.96 ($p < 10^{-100}$) between the two signals. These results suggest that the smaller, mobile and more affordable Shimmer GSR+ device may be sufficient to detect changes in affective states. During the experiment,

the Shimmer GSR+ device was worn on the non-dominant hand with the electrodes placed at the proximal phalanx of the index and middle finger [Calvo et al., 2015]. Data was recorded at a sampling rate of 100 Hz. As part of the Shimmer GSR+ setup, we also attached an optical pulse sensor providing a photoplethysmogram signal on the ring finger. However, photoplethysmogram data was of poor quality and consequently discarded from the analysis. Prior to electrode attachment, we asked participants to wash their hands with lukewarm water [Boucsein et al., 2012].

We measured heart activity of the participants using a Polar H10 chest belt. The Polar H10 belt provides IBIs and post-processed heart rate data by monitoring electrical changes on the surface of the skin. A predecessor of this device (Polar H7) was shown to provide accurate data when compared to an expensive lab device (Cosmed Quark T12x system) [Plews et al., 2017].

We recorded the skin temperature using the infrared thermopile sensor of the Empatica E4 device (sampling rate $=$ 4 Hz; resolution $=$ 0.02 °C). Since the sensor was attached to the dominant hand (used for writing during the tasks), other signals that the wristband can provide (EDA and blood volume pulse) were affected by motion artifacts and discarded from the analyses.

According to a questionnaire we asked the participants to fill out at the end of the experiment, the Empatica E4 was the most comfortable device (78% very, 17% medium, 5% little), followed by the Polar H10 chest belt (55%, 43%, 2%) and the Shimmer GSR+ finger electrodes (28%, 49%, 23%). The signals from the biosensor devices were streamed to the tablet using the Bluetooth Low Energy protocol.

4.2.2 Data Analysis

Input signals. Given that we detected a very low amount of artifacts across participants (EDA $=$ 0.015% and IBI $=$ 0.71%), we refrained from removing them from the analysis. Visual inspection of the skin temperature recordings revealed a slow linear increase of the temperature over the course of a participant's session. This change in temperature may be due to the skin warming up under the wristband and independent of the affective state of the participants. We removed this linear trend from all measurements by subtracting the result of a linear least-squares fit to the signal. We did not observe any other artifacts for skin temperature. The biosensor features listed in Table 4.1 were computed using a window of 10 seconds since the minimum task duration was 10 seconds. For the stylus features, we used an implicit window over the entire task. In addition, we excluded all data points having at least one missing value.

Clustering of ratings. Figure 4.3 presents the distribution of the participants' ratings

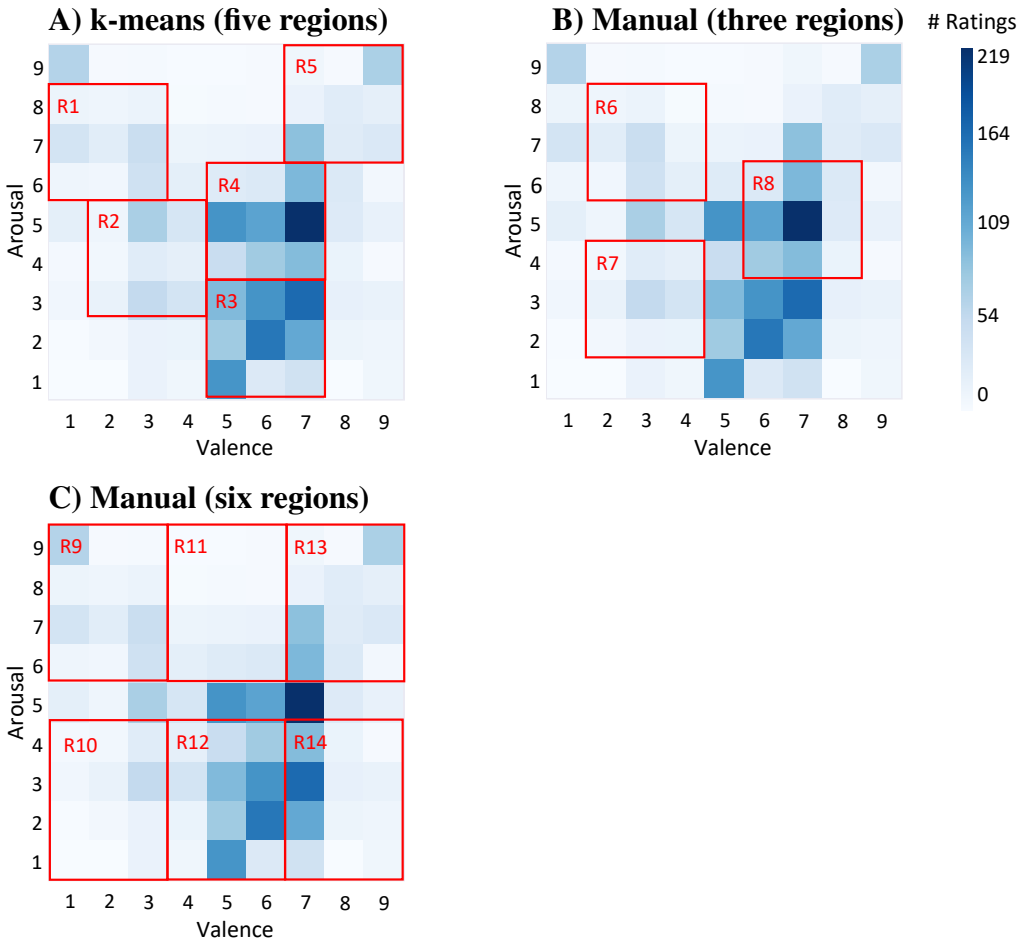


Figure 4.3: *Heat maps showing the distribution of the participants' ratings on the math tasks. The red rectangles represent the different regions. A) five regions automatically chosen using k-means clustering. B) Three regions manually selected. C) Six regions manually selected.*

in the valence-arousal space (dark and light blue refers to a high and a low number of data points, respectively). A v-shape is visible with most ratings being made at a valence and arousal level of seven and five, corresponding to a positive medium intense state (e.g., interest). Several ratings were made at the extremes (top left and top right) of the valence-arousal space corresponding to states of distress and excitement that are associated with very good and very poor performance. To uncover the underlying clusters in the data, we applied k-means clustering in this two-dimensional valence and arousal space. Using the Bayesian information criterion, we found an optimal number of five clusters. We defined region boundaries (shown by the red rectangles in Figure 4.3A) as the arithmetically rounded value of the centroid of each cluster plus and minus the standard deviation of the participants' ratings in the corresponding cluster. We observed that the regions are all of equal

Table 4.2: *Performance of Random Forest on the math data for different signals and regions. AUC_{micro} and AUC_{macro} represent micro-averaged and macro-averaged AUC, respectively. The chance level for accuracy is 1/# regions and for AUC it is 0.5. The standard deviations are given in brackets.*

Regions	Signals	AUC_{micro}	AUC_{macro}	Accuracy
k-means	EDA	0.80 (0.02)	0.75 (0.03)	50% (4%)
(5 Regions)	Heart	0.81 (0.01)	0.73 (0.01)	52% (2%)
	Temperature	0.69 (0.03)	0.59 (0.03)	37% (4%)
	Stylus	0.84 (0.01)	0.76 (0.02)	59% (2%)
	Biosensors	0.86 (0.01)	0.81 (0.02)	60% (2%)
	Biosensors & Stylus	0.88 (0.01)	0.83 (0.02)	64% (2%)
Manual	EDA	0.81 (0.02)	0.69 (0.04)	66% (2%)
(3 Regions)	Heart	0.79 (0.02)	0.66 (0.03)	62% (3%)
	Temperature	0.76 (0.01)	0.60 (0.04)	60% (3%)
	Stylus	0.83 (0.02)	0.72 (0.02)	67% (3%)
	Biosensors	0.84 (0.01)	0.76 (0.03)	67% (1%)
	Bisensors & Stylus	0.87 (0.01)	0.80 (0.02)	67% (2%)
Manual	EDA	0.80 (0.02)	0.72 (0.03)	46% (3%)
(6 Regions)	Heart	0.78 (0.01)	0.72 (0.02)	44% (2%)
	Temperature	0.70 (0.02)	0.61 (0.02)	35% (3%)
	Stylus	0.81 (0.01)	0.75 (0.02)	48% (2%)
	Biosensors	0.85 (0.02)	0.80 (0.02)	57% (4%)
	Bio-sensors & Stylus	0.87 (0.02)	0.83 (0.03)	61% (3%)

size and cover the area of the v-shape. Based on the categorization of Russell [1980] and Scherer [2005] we identify the following regions, their sizes and corresponding affective states: Region R1 (213 data points; frustrated, annoyed), region R2 (284; bored, taken aback), region R3 (965; attentive, serious), region R4 (861; expectant, confident), region R5 (295; excited, triumphant). Together, it appears that the math task covered a broad range of affective states relevant for learning and that positive states (R3, R4, R5) dominate.

4.2.3 Classification Performance

Table 4.2 and Figure 4.4A present the predictive performance of the model based on the five defined regions. Using all sensors, the model achieved an accuracy of 65% (chance level = 20%). Here, the slightly lower value for the macro-averaged AUC (0.83) compared to the micro-averaged AUC (0.88) may be related to class imbalance.

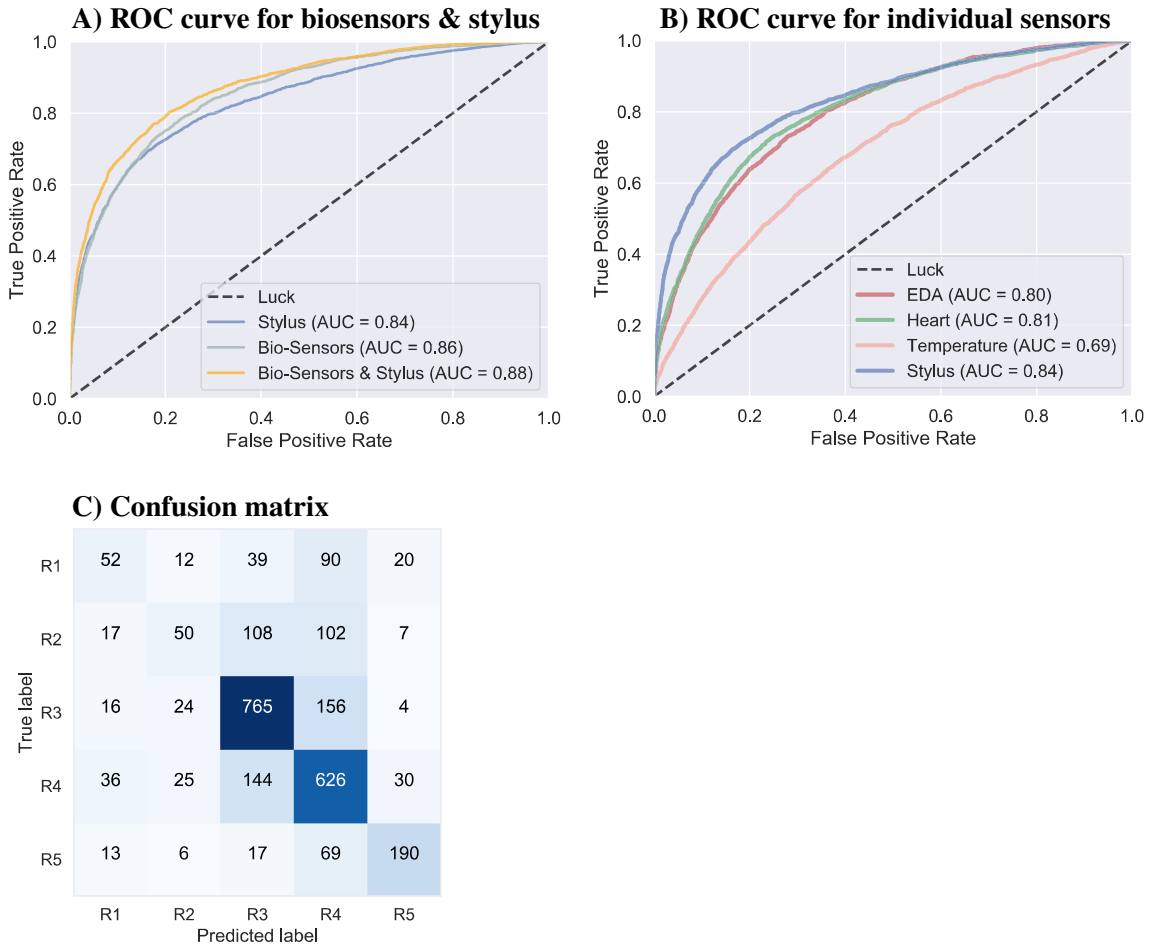


Figure 4.4: *ROC curves and micro-averaged AUC scores for five regions chosen by k-means clustering for (A) the biosensors, stylus and the combination of biosensors and stylus and (B) the individual biosensors & stylus. (C) The confusion matrix is computed by using the combination of biosensors and stylus.*

Figure 4.4C depicts the confusion matrix based on all sensors. The matrix shows that regions R1 and R2 are more difficult to predict than the other regions. This may be due to the lower number of data points collected for these regions. As expected, the larger the distance between the regions, the easier it is for the model to discriminate between them.

Feature importance. Table 4.1 presents the 10 most important features (in bold). The features are sorted according to their relative importance which we computed using permutation feature importance (permuting each feature 100 times and measuring the mean decrease in micro-averaged AUC). We obtained the same relative feature importance ordering using the Gini importance measure. EDA and heart measures provided 3 out of the 10 most important features and stylus features contributed with

4 of the most important features. There were no skin temperature features among the top 10 features. Regarding the heart measures, the features related to IBIs were more important than HR features. An interesting observation can be made for the stylus features. Features related to the distance between strokes appear to be more important than speed between stroke features indicating that the spread of writing attributed to thinking (i.e., how the writing space is covered) provides more information than the actual writing behavior.

4.2.4 Sensor Comparison

Biosensors. If we consider the individual sensors (Figure 4.4B), skin temperature performs substantially worse (-0.11 AUC) compared to EDA (0.80 AUC) and heart rate measures (0.81 AUC). The combination of all the biosensors (Figure 4.4A) provides only marginal performance improvements ($+0.05$ AUC) compared to the individual sensors.

Stylus. Our most important finding is that the stylus performs equally well as the biosensors (Figure 4.4B), rendering the data from the biosensors redundant and unnecessary for the prediction of affective states. The performance of the stylus is only marginally inferior (-0.02 AUC) when compared to the combination of all biosensors. In contrast, the combination of the biosensors and the stylus achieves a slightly higher performance ($+0.02$ AUC) compared with the biosensors and stylus alone (Figure 4.4A). This might be an indication that they may contain complementary information, although the difference appears to be small.

4.2.5 Affective Region Analysis

To investigate the ability of our pipeline to predict different affective regions based on the recorded biosensor and stylus data, we defined two additional coverings of the valence and arousal space (Figure 4.3B and Figure 4.3C). Based on Russell [1980] and Scherer [2005] we manually defined specific regions associated with frustration (annoying; region R6, 185 data points), boredom (taken aback; region R7, 199) and interest (engaged concentration, flow; region R8, 720) as shown in Figure 4.3B. For example, in education, it is important to distinguish these three regions due to their impact on learning gain [Baker et al., 2012; Csikszentmihalyi, 2008; Miserandino, 1996]. To cover the valence and arousal space evenly, we manually defined the six regions shown in Figure 4.3C, dividing arousal in two and valence into three components (the number of data points from region R9 to R14 are 287, 154, 134, 852, 432 and 506). The results for both space partitionings are listed in Table 4.2 (note that chance level for the accuracy is 33% for three regions and 16.66% for six regions). The performance of the classification of three regions outperforms the

one for five and six regions in terms of accuracy. On the other hand, when taking into account the AUC, there is no substantial difference in performance between the different coverings. This difference between accuracy and AUC stems from the fact that predicting only three regions is a much easier task than predicting five or six regions. This is in line with the finding that the accuracy for predicting five regions is slightly higher than for six regions. Nevertheless, we can conclude that we saw that our approach can provide good results for three different coverings. Thus, we come to the conclusion that our pipeline is rather flexible being able to handle different regions in the valence-arousal space. Compared to previous work relying on fixed affective states, our approach has the advantage that the regions do not have to be pre-defined allowing for much more flexible use.

4.2.6 Model Transfer

In addition to the math tasks, we also gathered biosensor data as well as valence and arousal ratings from the participants while they observed pictures from the IAPS. We used this data to investigate our model's capacity to generalize to more passive tasks, such as looking at pictures. To predict the affective regions of interest, we applied our model trained on the biosensor data recorded during math task solving to data collected while participants viewed and rated the set of IAPS pictures. When we consider the 5 different regions (Figure 4.3A), the model's accuracy reaches 39% (chance level $= 20\%$, $\text{AUC}_{\text{micro}} = 0.68$, $\text{AUC}_{\text{macro}} = 0.64$). When we train and evaluate a model directly on the picture data, we achieve a slightly better classification performance (accuracy $= 42\%$, $\text{AUC}_{\text{micro}} = 0.74$, $\text{AUC}_{\text{macro}} = 0.66$). There may be several reasons behind the suboptimal performance when predicting affective states during the picture task. These include sociocultural aspects when rating emotions based on pictures (e.g., rating how it is expected), old and low-resolution pictures from the IAPS data set, media influence desensitizing participants to the content of the IAPS, and the fact that the math and picture domains are very different. Together these initial results indicate that building a general predictor of affective states might be possible, but further experiments are necessary.

4.3 Discussion

In this chapter, we presented a generic pipeline for predicting affective regions of interest using biosensor and stylus data. We validated our pipeline for the case of math solving tasks and demonstrated that our pipeline can accurately predict various regions in the valence-arousal space (up to 0.88 AUC). In addition, we compared different input signals with each other. The performance of the Shimmer GSR+ (measuring skin conductance) and Polar H10 (measuring heart activity) were on the

same level (up to 0.81 AUC). Due to the higher cost of the Shimmer GSR+, we recommend the usage of the Polar H10. Besides the cost factor, the Polar H10 is also more robust against movement artifacts and more comfortable to wear.

Moreover, we found that the classification performance using only stylus data is comparable to the classification performance based on the biosensors. Taking into account the emerging digitization of education and the spread of tablets in schools and private households, these results make the stylus a preferred alternative to biosensors for measuring affective states in classrooms. Using biosensors in classroom settings can be cumbersome and costly as it requires the purchase and synchronization of several devices. In contrast, systems that depend on a stylus only are cheaper than systems relying on biosensor devices, and styluses often come bundled with mobile devices, such as tablets or smartphones. In addition to being cheaper and more ubiquitous, styluses are easier to setup (e.g., no attachment of electrodes, no motion artifacts) and less intrusive. Furthermore, stylus data is not only restricted to digital devices but can also be recorded using digital pens. Finally, we demonstrated the possibility of a generalized model for predicting affective states by applying the model trained on the data from the math tasks (active part) to pictures from the IAPS (passive part) reaching a performance of 0.68 AUC.

There are some potential limitations to our approach presented in this chapter. First of all, the experimental setup was restricted to a lab environment and the population of bachelor students may limit generalization to students at other levels. We assume that given a proper baseline correction the signals are also predictive for a heterogeneous group of people. Another limitation is the restriction to math tasks. Similar to biosensor data, we believe that handwriting data carries affective information independent of the task. Thus, we expect our approach to work also in other domains involving handwriting, such as solving exercises for different school subjects and writing essays.

CHAPTER 5

Affective State Prediction Using Smartphones in the Lab

Recent work has suggested that being aware of one's current affective state can be particularly useful in the context of mobile devices as individuals become more dependent on smartphones for social purposes [LiKamWa et al., 2013]. Here, chat applications are especially relevant as they currently rank as the most used applications on smartphones [Androidrank, 2021].

The majority of methods to detect affective states rely on biosensor data (see Chapter 4) or camera data to infer emotions from facial expressions (see Chapter 3). However, most of these setups are privacy-invasive and potentially costly, which can limit their applicability in real-world environments. As such, researchers have explored different methods to infer affective states directly from smartphone data, including sensor inputs (e.g., accelerometer and gyroscope) [Lane et al., 2012], application usage patterns [Bachmann et al., 2015; LiKamWa et al., 2013], and typing speed [Gao et al., 2012].

In this chapter, we propose a non-invasive solution that can accurately predict affective states based on sensor data from a mobile device (see Figure 5.1). We achieve this by considering only touch input from the smartphone's on-screen keyboard to generate two-dimensional heat maps of typing characteristics. We train our semi-supervised deep learning architecture on these heat maps to learn a low-dimensional feature embedding. The subsequent classification can predict valence (up to 0.84 AUC), arousal (up to 0.82 AUC), and dominance (up to 0.82 AUC) on three levels each (low, medium, high). We demonstrate the effectiveness of predicting the affective states based on the touch characteristics of smartphone users in a data collection study with 70 participants engaged with a chat application.

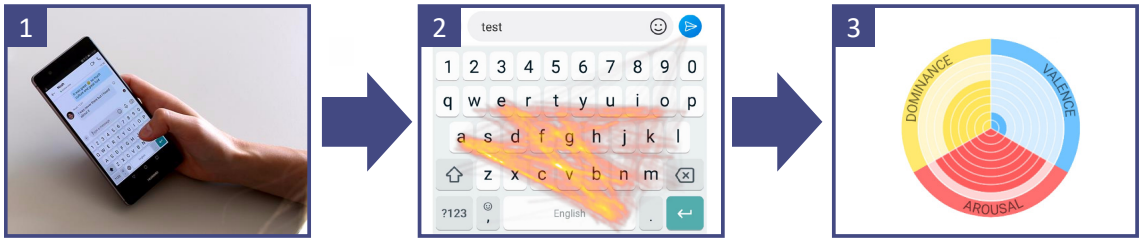


Figure 5.1: *Our system extracts touch input characteristics of users while typing on smartphones (1) and aggregates these metrics into two-dimensional heat maps (2). A semi-supervised classification pipeline dynamically predicts affective states (valence, arousal, and dominance) of the user (3).*

5.1 Method

We present a semi-supervised classification pipeline for predicting affective states based on touch data collected during typing on smartphones. While touch data is continuously available, ground truth is typically only available in certain intervals (e.g., from self-reports). To make use of the large amount of unlabeled data, we employ variational autoencoders to infer meaningful low-dimensional embeddings from two-dimensional heat maps (see Figure 5.2A). In a second step, we add a fully connected classification layer to the learned data encoder and fine-tune the entire network for the classification of affective states (see Figure 5.2B). In the following, we provide details on every part of our method.

5.1.1 Heat Maps

Modern smartphones allow for the collection of accurate information about the user's screen inputs. An input $e_i = (x, y, t)$ is defined by the coordinates (x, y) on the screen and the timestamp t in milliseconds. A single touch event $E = [e_1, \ldots, e_n]$ can consist of n touch inputs from the time the user initially touched the screen (e_1, touch down) until he or she releases the screen (e_n, touch up). Based on the raw input data, we can extract several touch event metrics: Down-down speed provides information about the typing speed and is equal to the time difference between two consecutive touch downs normalized by the distance. Up-down speed is equal to the time between a touch up and the subsequent touch down normalized by the distance. Up-down speed provides information about the speed between touch events. In contrast to previous research [Araújo et al., 2005; De Luca et al., 2012; Monrose et al., 2002], we do not account for touch duration (i.e., down-up speed) since touch events often consist of a single input $E = [e_1]$ for which no duration can be computed. All metrics are standardized based on the mean and standard deviation during a baseline typing period.

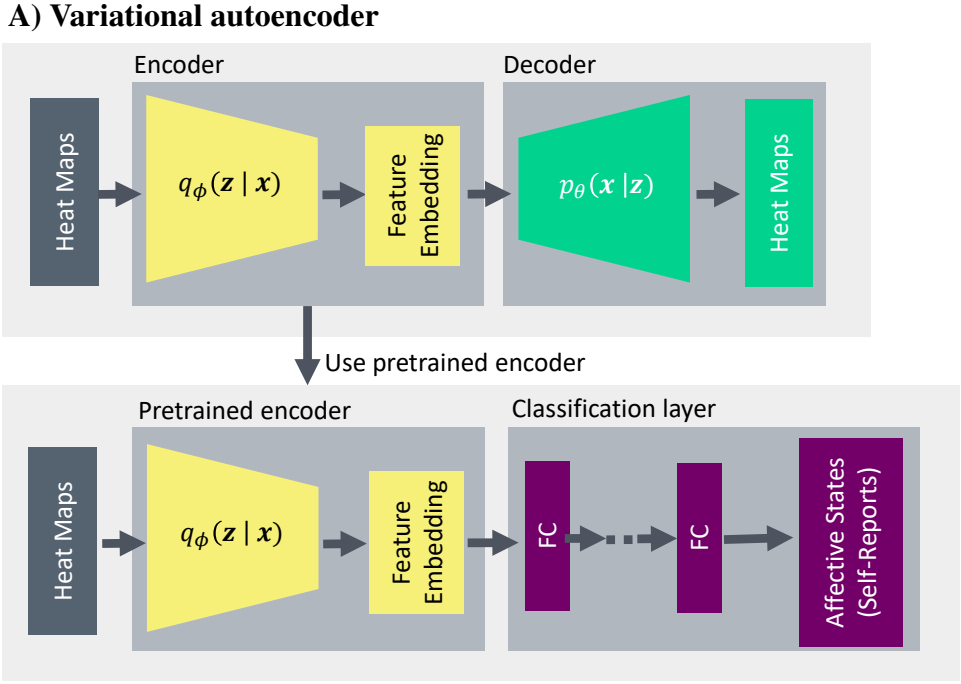


Figure 5.2: *Overview of the main steps of our model. A) A variational autoencoder is trained on heat maps created from smartphone touch data to learn an efficient low-dimensional feature embedding. B) For classification, the low-dimensional embedding is used as input to fully connected layers.*

Since touch inputs are inherently spatial, we aggregate the touch event metrics into two-dimensional heat maps. These heat maps cover the keyboard region and the send button (see the red dashed line in Figure 5.4B) as we only include keyboard inputs in this chapter. We use a sliding window with a window size of 180 seconds shifted by five seconds to extract a sequence of heat maps for each user. Since the down-down speed and up-down speed metrics always correspond to two consecutive touch events E_i and E_{i+1}, we assign their value to every pixel on a straight line between the events (see Figure 5.3B and 5.3C).

Finally, we apply Gaussian smoothing to the heat maps to reduce high-frequency noise. We use a kernel of size $k = 31 \times 31$ pixels, which is twice the typical key distance in pixels, and prevents smearing into neighboring keys while keeping inter-key resolution high. In addition, we use $\sigma = 5$ provided by OpenCV [Bradski, 2000].

Figure 5.3 shows examples of extracted heat maps for pressure, down-down speed, and up-down speed. The colors in the heat maps are for visualization purposes only. In our pipeline, we only use one value per pixel.

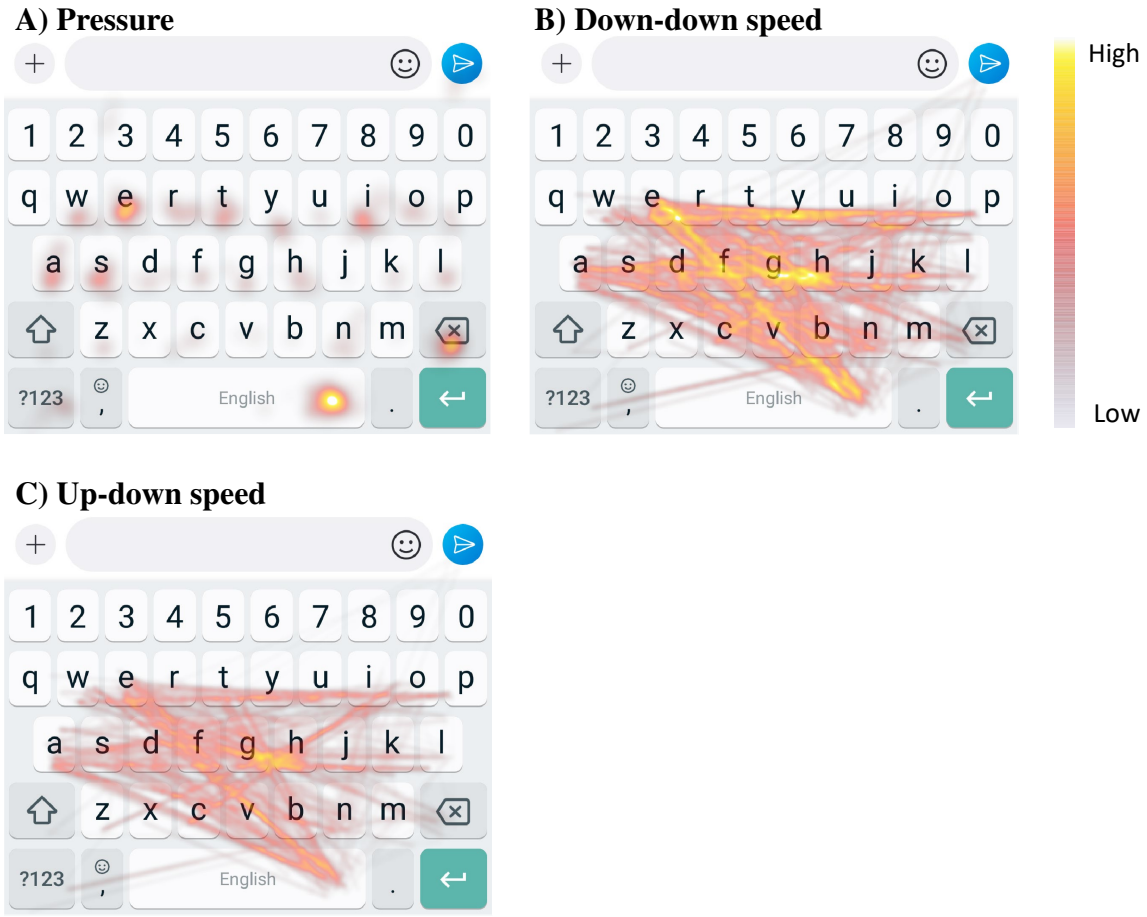


Figure 5.3: *Examples of heat maps extracted from the touch events of a user. A) The color indicates the average pressure applied. B) and C) Consecutive touch events are connected by a line, and the color indicates the down-down and up-down speed between these two events, respectively. The colors are for visualization purposes only.*

5.1.2 Variational Autoencoder

While touch data is available continuously, labels are sparse. We make use of the unlabeled data by learning a low-dimensional representation of the heat maps that capture as much information from the original heat maps as possible. To extract such a low-dimensional representation (also called latent space or embedding), we employ a particular type of neural network called variational autoencoder [Kingma et al., 2014] (see Figure 5.2A). Variational autoencoders have the advantage of providing representations with disentangled factors and allow control over modeling the latent distribution (in our case, multivariate Gaussian) [Higgins et al., 2016; Kingma and Welling, 2013]. Previous research has shown that variational autoencoders are capable

of automatically learning meaningful low-dimensional representations in different domains [Aksan et al., 2018; Klingler et al., 2017].

A variational autoencoder consists of an encoder and decoder. The encoder network $q_\phi(\mathbf{z}|\mathbf{x})$ learns an efficient compression of the input data $\mathbf{x}$ (heat map) into a low-dimensional space $\mathbf{z}$ using a deep neural network parameterized by ϕ. The decoder network $p_\theta(\mathbf{x}|\mathbf{z})$ reconstructs the input based on sampling from the distribution of the latent space. Here, θ are the parameters of the decoder network. We train the variational autoencoder using the loss function

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \beta\,\mathrm{KL}\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right],$$

where KL denotes the Kullback-Leibler divergence. The left term measures the reconstruction quality, and the right term regularizes the latent space towards the prior $p(\mathbf{z})$. By using the Lagrangian multiplier β, we introduce a trade-off between reconstruction quality and disentanglement of the latent factors fostering a more efficient encoding. This modification of the loss function has been successfully used for training variational autoencoders [Higgins et al., 2017].

For the autoencoder, we use two-dimensional convolutions with symmetric encoder and decoder. Depending on the resolution of the input heat maps, it is necessary to down-sample the heat maps to reduce training time. Input data is commonly scaled before training. We use Min-Max scaling of the heat maps per user.

5.1.3 Classification

We take advantage of the learned low-dimensional representation by adding a classification network to the pre-trained encoder network (see Figure 5.2B). The classification network consists of fully connected layers with rectified linear unit activations except for the last layer, where we use softmax activation for the classification output. The different heat maps are aggregated by stacking the latent space of the individual heat maps.

The fully connected network is trained on the labeled data (heat maps and corresponding affective states) using backpropagation that minimize the cross-entropy loss. Fine-tuning the classification network has shown good performance in other domains [Sun et al., 2016].

5.2 Experiment

We conducted a controlled lab experiment to validate our pipeline for the prediction of affective states based on smartphone touch data. The experiment was approved by the ethics board of ETH Zurich. During the experiment, we collected smartphone touch data while participants interacted with a chat application (i.e., Skype) for approximately 70 minutes. We used text-based chat conversations because they are widely used [Androidrank, 2021] and would be familiar to the participants in the study. In addition, these applications require interaction with the smartphone and can provide the data necessary for testing our prediction model.

5.2.1 Participants

We recruited 70 participants (35 female) between the ages of 18 and 31 (mean = 23.0 years, standard deviation SD = 2.7 years) from 20 different departments at the master and bachelor level of ETH Zurich and the University of Zurich. We only considered participants that were fluent in English[1] and used smartphone-based chat applications on a daily basis. We excluded participants taking any type of medication, tranquilizers, or psychotropic drugs (e.g., anti-depressants) as well as participants affected by any type of the autism spectrum disorders. To control for external environmental factors, we kept the room temperature and the humidity at an average of 23.9° (SD = 0.24°) and 30.1% (SD =3.6%), respectively. All participants provided written informed consent before the start of the experiment and were rewarded with CHF 45 for their participation. Participants were rewarded with an additional CHF 5 if they missed only one response window when completing the self-report measures.

5.2.2 Apparatus

Participants interacted with five contacts within the Skype application on a Huawei P9 Plus smartphone running Android 7.0. This smartphone provides over 17000 levels of touch pressure sensitivity. The software keyboard used was Gboard with auto-correction and spell-checker features disabled. Throughout the experiment, we recorded their interaction with the device, including sensor (acceleration and orientation) and touch (pressure and position) data. In addition, participants used a Huawei MediaPad M2 tablet to report their emotional state at regular intervals during the experiment. Figure 5.4A presents the experimental setup.

[1] A post-experiment questionnaire revealed that 94% of the participants judged their English level to be "proficient" (C2) or "advanced" (C1) according to the Common European Framework of Reference for Languages.

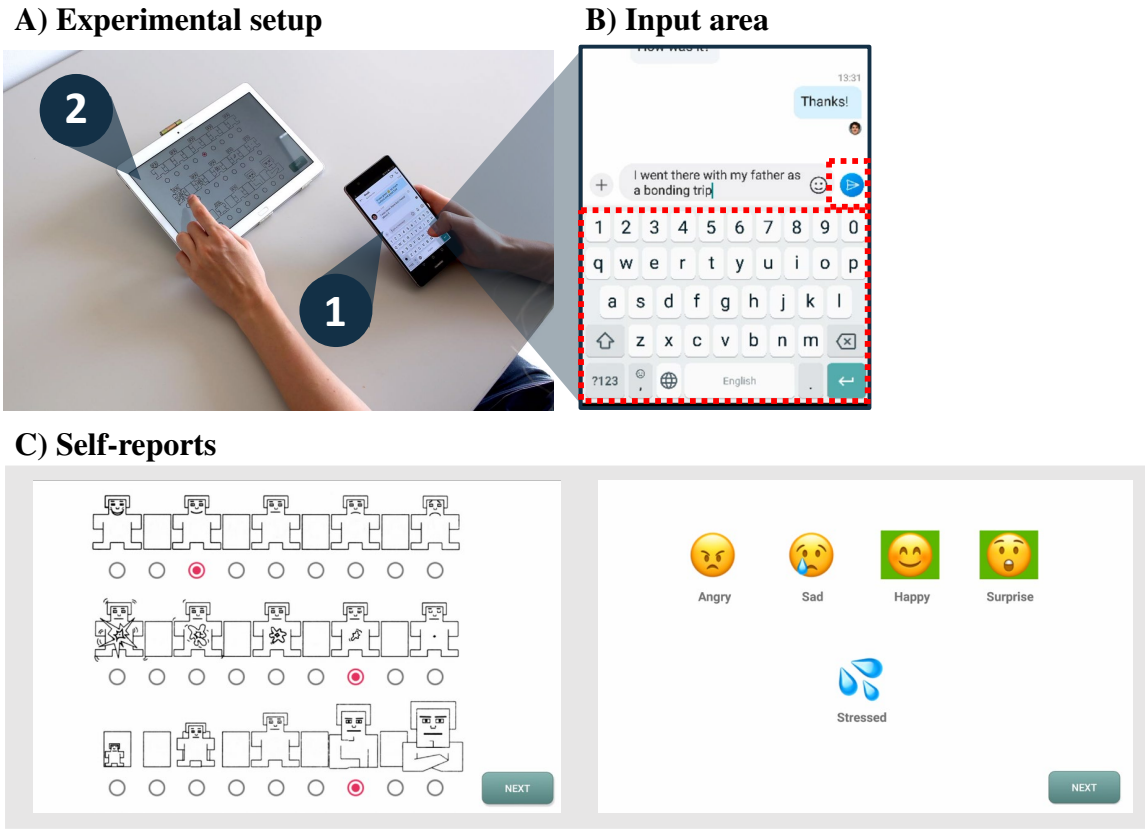


Figure 5.4: *Experimental setup. A) During each session, participants engaged in chat conversations using Skype on a smartphone (1). At regular intervals, participants were asked to complete self-reports on a tablet (2). B) Chat interface and the region that was considered in the prediction model (red-dashed area). C) Self-reports for capturing valence, arousal and dominance (left), basic emotions, and stress level (right).*

5.2.3 Self-Reports

To gather ground truth data for our model (valence, arousal, and dominance on a 9-point scale), we asked participants to complete the Self-Assessment Manikin (SAM) [Bradley and Lang, 1994] at regular intervals during the experiment. Participants were also asked to select from a series of basic emotions (i.e., anger, sadness, happiness, and surprise) represented by different emojis. To ensure that choices were independent and to enable blending of basic emotions to form complex emotions, participants were allowed to simultaneously select more than one emoji at a time (e.g., anger and surprise). The basic emotions did not include fear and disgust after a pilot study ($n = 8$) revealed that participants did not experience these emotions during the chat conversations. However, participants had the choice of selecting a

'stress' emoji when reporting their emotions. Participants were allowed to select all possible combinations of the basic emotions and stress without any restrictions. Figure 5.4C shows an illustration of the self-reports.

5.2.4 Procedure

Before the day of the experiment, participants were asked to complete the Patient Health Questionnaire [Kroenke et al., 2001] and the Big Five Inventory [John et al., 1991; John et al., 2008] as measures of mental health and personality traits, respectively. On the day of the experiment, the participants were given an oral overview of the procedure, including an introduction to the self-report questionnaires and an explanation regarding the use of the smartphone. The experimenter then exited the room and used one of the Skype contacts (guiding contact) to start a conversation (five minutes) with the participants. During this conversation, the experimenter asked six predefined questions about well-being, age, living place, work, hobbies, and family. These questions were used to make the participants comfortable with the keyboard and the handling of the smartphone. Next, participants were instructed to watch a nature video for five minutes on the smartphone that was used as relaxation and allowed them to acclimate to the room environment. At the end of the nature video, participants were asked to type two well-known pangrams (149 characters) that served as a baseline for touch input during the modeling stage. During the main phase of the experiment, participants chatted with four different Skype users. These Skype users were fake accounts created and controlled by the experimenter sitting in an adjacent room. After finishing all four chat conversations, participants were asked to type once again the two pangrams. Finally, participants completed an exit questionnaire on smartphone use, demographics, and overall mood. Figure 5.5A provides an overview of the procedure used in the experiment.

At the beginning of the experiment, participants were given an oral explanation regarding the procedure for answering the self-reports and had a chance to practice with four examples. We collected a total of 1893 self-reports covering a large range of the SAM response space.

During the experiment, participants were alerted with an audio notification when it was time to complete the self-report. At this time, the SAM and emojis appeared on the tablet, and participants had 20 seconds to start the self-report. This time buffer, allowed participants to finish the current sentence in the Skype conversation without having to rush to complete the self-report. If participants were slow to respond, the tablet started to vibrate as a final reminder. After completing the self-reports, a delay of 90 seconds was introduced until the next self-report was presented. This time interval was decided based on feedback from participants in the pilot experiment as it

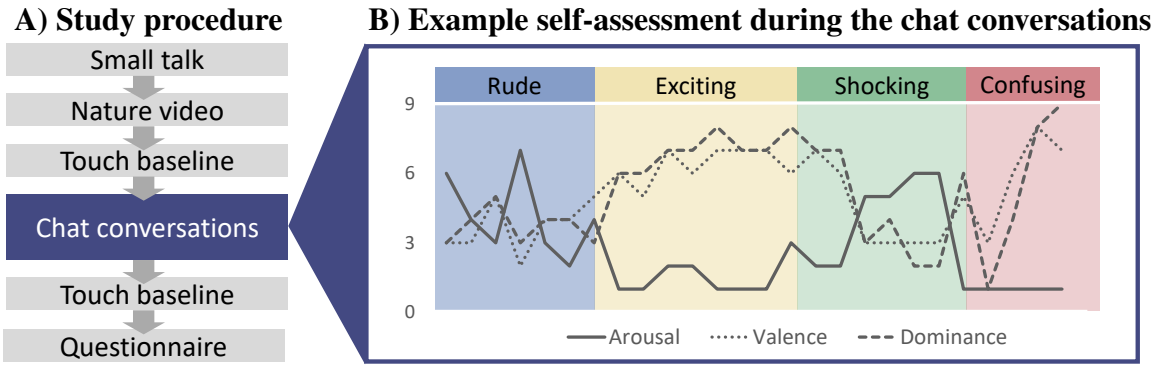


Figure 5.5: *Overview of the different parts of the experiment. A) Overall experimental procedure. B) Changes in valence, arousal, and dominance for one participant during the four chat conversations.*

provided the best trade-off between the amount of data collected and the number of interruptions.

There was no restriction for participants in using one or two hands to type. Nevertheless, in a post-experiment questionnaire, we found that only eight participants used the right hand, and one participant used the left hand while typing. All other remaining 61 participants used both hands.

The device was locked in portrait mode because this is the usual orientation for chatting (no participant complained). Swiping gestures, auto-completion, and auto-correction were disabled to make it consistent between participants. On the other hand, participants could use the symbolic keyboard, and they could chat during emotional surveys. Restricting chatting during emotional surveys is difficult because the survey was filled in on an external tablet. Apart from the 20 seconds time limit to start the rating, we did not put a time limit to fill in the survey to not put additional pressure on the rating, which could introduce a negative bias.

53% of the user were iOS users. According to the post-questionnaire, 49% of the participants felt very comfortable with the smartphone, 44% medium comfortable, and only 5% little comfortable (with the lowest rating being not at all comfortable). Thus, we conclude that although the keyboard differs between smartphones, participants were not affected too much by readjusting to another keyboard.

5.2.5 Tasks

To trigger different affective states, we created four different types of chat conversations (i.e., exciting, shocking, rude, and confusing) by varying the content and context of the text messages sent to the participants. Participants saw a list of five contacts in

the Skype application on the smartphone that was provided to them. Each contact was associated with one of the conversation types. A fifth contact representing the experimenter was created to guide participants through the experiment and to provide help in case of questions.

To make the chat task more credible, we employed NVIDIA's face generator [Karras et al., 2019] to create fake profile pictures for each of the four contacts. We used the image of the experimenter for the fifth contact. In addition, participants were told that the four contacts were real people sitting next door. All conversations followed a predetermined script to keep them consistent across participants. The paragraphs below describe in more detail each of these conversations.

Exciting conversation. During this conversation, the participants were chatting about their most beautiful holiday experience. This conversation was designed to make participants remember and reminisce, leading to positive feelings (e.g., enjoyment).

Shocking conversation. This conversation focused on the topic of the Rohingya refugee crisis, which is an ongoing persecution of Muslim Rohingya people in Myanmar by the government. This conversation was intended to sadden the participants leading to negative feelings (e.g., anger).

Rude conversation. In this conversation, we asked participants to provide help with a malfunctioning smartphone. Independent of the help participants provided, they could not resolve the issue at any point during the conversation. Here, the Skype contact chatting with participants became increasingly rude and was intended to trigger negative feelings (e.g., anger) and surprise.

Confusing conversation. For this conversation, we used Cleverbot [Carpenter, 2011]. Cleverbot is a well-known chatbot that learns from past conversations. We found this chatbot to be a good way to trigger confusion, anger, and surprise. We reset the chatbot engine for every participant to avoid introducing potential bias from conversations with previous participants. A post-experiment questionnaire revealed that 63% of the participants did not recognize that this conversation was with a chatbot.

The order of conversations was randomized across participants with the exception that the confusing conversation was always last to prevent participants from behaving differently should they recognize that they were chatting with a chatbot [Hill et al., 2015]. This lead to the counterbalancing of three conditions and a total of six orders. With our randomization approach, we achieved an almost complete counter balanced distribution (12, 11, 11, 14, 10, 12). In general, the average duration of the rude and confusing conversations (836 seconds and 650 seconds) was shorter compared to the exciting and shocking conversations (1212 seconds and 1272 seconds). These shorter

durations may be related to the fact that participants became tired of engaging in the conversations.

Figure 5.5B depicts the changes in valence, arousal, and dominance during the four chat conversations for one participant. The figure shows that valence increases during the exciting and confusing conversations and decreases for the other two conversations (we see the opposite pattern for arousal). The rude and shocking conversations seemed to be more intense than the exciting and confusing conversations. We also see that dominance is following a similar pattern than valence with the participant feeling more in control during the exciting and confusing conversations.

5.3 Results

We evaluated our classification pipeline based on the data we collected during the experiment. We collected 1893 self-reports on the affective and emotional state of participants that were used as the ground truth to our model. Because the SAM is scored on a 9-point scale, we evaluated the performance of the classifier for three classes (low, medium, high) of valence, arousal, and dominance. We also recorded 3720 minutes of touch data from which we extracted 44625 heat maps for each of the three types of heat maps (i.e., pressure, down-down speed, and up-down speed). We also reveal the runtime of our method to analyze the real-time applicability of our approach. To measure the performance of our model, we calculated the accuracy (chance level is 0.33 for three classes and 0.5 for two classes) and the micro-averaged area under curve (AUC) of the receiver operating characteristic (ROC) curve (chance level is 0.5). The micro-averaged AUC aggregates the contributions of all classes by considering each element of the label indicator matrix as a label. Because these two metrics are both affected by class imbalance, we also calculated the macro-averaged AUC (chance level is 0.5) by taking the mean of the class-wise AUCs. We evaluated our model using leave-one-user-out cross-validation to ensure that data of an user is not used for training and testing at the same time.

5.3.1 Network Parameters

Variational autoencoder. For each of the three types of heat maps, we trained a variational autoencoder to learn a low-dimensional representation. We used a resolution of the heat maps of 80×64 pixels. To find the network parameters, we employed the approach described by Bengio [2012]. Specifically, we increased the number of layers, and the number of features maps per layer until a good fit of the data was achieved (i.e., the loss was minimal). For the pressure heat maps, this resulted in a variational autoencoder consisting of two layers (32 and 64 feature maps) for the encoder and decoder, a kernel size of 4×4, and a latent space with ten

dimensions. For the down-down speed and up-down speed heat maps, this resulted in a variational autoencoder with four layers (32, 64, 128, and 256 feature maps) for the encoder and decoder, a kernel size of 3×3 and a latent space with 20 dimensions. In comparison to the network for the pressure heat maps, the down-down speed and up-down speed network was deeper and with a dimensionality of the latent space twice as high due to the higher complexity of the heat maps. For both networks, we used a stride of 2×2 for each convolution. We chose a relatively small $\beta = 0.00001$ (compared to [Higgins et al., 2017]) because of the difference in magnitude between reconstruction loss and the Kullback Leibler divergence. We trained the variational autoencoders for 200 epochs with a batch size of 64 on 40162 heat maps and used 4463 heat maps as the validation set.

Fully connected network. The network parameters for the fully connected network used for classification were defined using a randomized search with 50 iterations. We trained the network using nested leave-one-user-out cross-validation for 100 epochs with a batch size of 8. All networks were implemented using the Keras framework with TensorFlow™ back-end and optimized using Adam optimization with standard parameters [Kingma and Ba, 2015].

5.3.2 Experimental Validation

We conducted three Kruskal-Wallis tests to investigate whether the four text conversations elicited different levels of valence, arousal, and dominance. Results revealed significant differences in terms of valence ($H = 144.431$, 3 d.f., $p < 0.001$), arousal ($H = 19.461$, 3 d.f., $p < 0.001$) and dominance ($H = 39.982$, 3 d.f., $p < 0.001$). We performed five additional ANOVAs to investigate whether there were significant differences in terms of the basic emotions and stress reported by participants during the four conversations. For the ANOVAs, we added the times that participants reported a specific basic emotion or stress during each of the conversations. Here again, we found significant differences in terms of anger ($F(3,233) = 21.768$, $p < 0.001$), happiness ($F(3,233) = 238.068$, $p < 0.001$), sadness ($F(3,233) = 79.389$, $p < 0.001$), surprise ($F(3,233) = 6.158$, $p < 0.001$) and stress ($F(3,233) = 5.525$, $p = 0.001$). All tests are significant after Bonferroni correction. Table 5.1 presents the means and standard deviation for each of these variables (see Appendix A.1 for additional statistics). We also performed a series of correlations to investigate the relationship between the SAM ratings for valence, arousal, and dominance and the four basic emotions and stress. Table 5.2 presents the results for each of these correlations. Notably, these results suggest a close match between the SAM ratings and the four basic emotions and stress.

Table 5.1: *Means and standard deviations (in brackets) for the self-reported SAM, four basic emotions, and stress during the four conversations. Percentages for the four basic emotions and stress do not add to* 100% *since participants could either simultaneously pick more than one emotion or not pick an emotion at all.*

	Exciting	Shocking	Rude	Confusing	Total
Valence	7.3 (1.5)	3.3 (1.6)	4.8 (2.1)	5.2 (1.6)	5.2 (2.3)
Arousal	4.3 (2.1)	5.0 (2.2)	4.4 (2.2)	3.4 (1.9)	4.4 (2.2)
Dominance	6.3 (1.7)	4.8 (2.1)	5.3 (2.2)	5.1 (2.2)	5.4 (2.1)
Anger	0.7%	28.6%	25.4%	10.3%	15.7%
Happiness	77.5%	2.6%	16.7%	21.8%	33.1%
Sadness	2.9%	52.7%	10.5%	2.5%	21.0%
Surprise	7.6%	12.5%	18.7%	37.4%	16.0%
Stress	2.1%	8.0%	20.5%	15.6%	9.3%

Table 5.2: *Effect sizes of the Pearson correlations between valence, arousal, and dominance (from the SAM) and the four basic emotions and stress. Asterisks denote correlations that survived Bonferroni correction ($p = 0.003$).*

	Anger	Happiness	Sadness	Surprise	Stress
Valence	-0.55^*	$+0.79^*$	-0.62^*	-0.14	-0.30^*
Arousal	$+0.41^*$	$+0.07$	$+0.37^*$	$+0.03$	$+0.18$
Dominance	-0.19^*	$+0.43^*$	-0.24^*	-0.10	-0.33^*

5.3.3 Affective State Prediction

The performance of our model was evaluated with regards to the prediction of three classes (low $\in [1,3]$, medium $\in [4,6]$, high $\in [7,9]$) of valence (523, 712 and 660 data points), arousal (786, 758, 349) and dominance (375, 886, 632). We chose these three classes to cover the entire space considering all available ratings. Figure 5.6 and Table 5.3 present the performance of our model (ROC curves were calculated using the micro-averaging approach). See Table A.1 in the appendix for additional metrics.

Classification performance. Using all heat maps in combination, our model achieves an accuracy of 67% for valence, 63% for arousal, and 65% for dominance (chance level is 33%). Here, the slightly lower values for the macro-averaged AUC (0.83, 0.80, 0.80) compared to the micro-averaged AUC (0.84, 0.82, 0.82) may be attributed to class imbalance. If we consider the percentage of the most frequent class as baseline (valence $= 38\%$, arousal $= 42\%$, dominance $= 47\%$), the predic-

Table 5.3: *Performance for the prediction of three classes (low, medium, high) of valence, arousal, and dominance. AUC_{micro} and AUC_{macro} represent micro-averaged AUC and macro-averaged AUC, respectively. The chance level of accuracy and AUC is 0.33 and 0.5, respectively.*

Dimension	Heat Map	AUC_{micro}	AUC_{macro}	Accuracy
Valence	Pressure	0.75	0.74	56%
	Down-down	0.81	0.81	64%
	Up-down	0.79	0.79	61%
	Combination	0.84	0.83	67%
Arousal	Pressure	0.80	0.78	62%
	Down-down	0.75	0.73	55%
	Up-down	0.73	0.70	53%
	Combination	0.82	0.80	63%
Dominance	Pressure	0.79	0.77	63%
	Down-down	0.80	0.78	63%
	Up-down	0.78	0.76	61%
	Combination	0.82	0.80	65%

tions of our model are also above this baseline for all three dimensions. Figure 5.7 presents the confusion matrices for valence, arousal, and dominance based on the combination of all heat maps. The confusion matrices are calculated by predicting self-reports across all chat conversations. The matrices show that for valence, arousal, and dominance the low and high classes were often wrongly predicted as the medium class. As expected, the larger the distance between the classes, the easier it is to differentiate them for our model (i.e., the low class was only rarely confused with the high class and vice versa). Interestingly, for arousal, the medium class was most often wrongly predicted as the low class (Figure 5.7B), but medium dominance was more often confused with high dominance (Figure 5.7C).

Heat map comparison. Pressure is the best predictor of arousal (+0.05 AUC), while down-down speed and up-down speed are the best predictors for valence (+0.06 AUC). In terms of dominance, all three heat maps perform similarly (up to 0.80 AUC). Overall, the combination of all heat maps provides only marginal improvements compared to the individual heat maps (up to 0.03 AUC).

5.3.4 Affective Sequence Analysis

Affective states can change over time, and this may be characterized either by smooth transitions or abrupt changes (e.g., from low to high states). We hypothesize that the

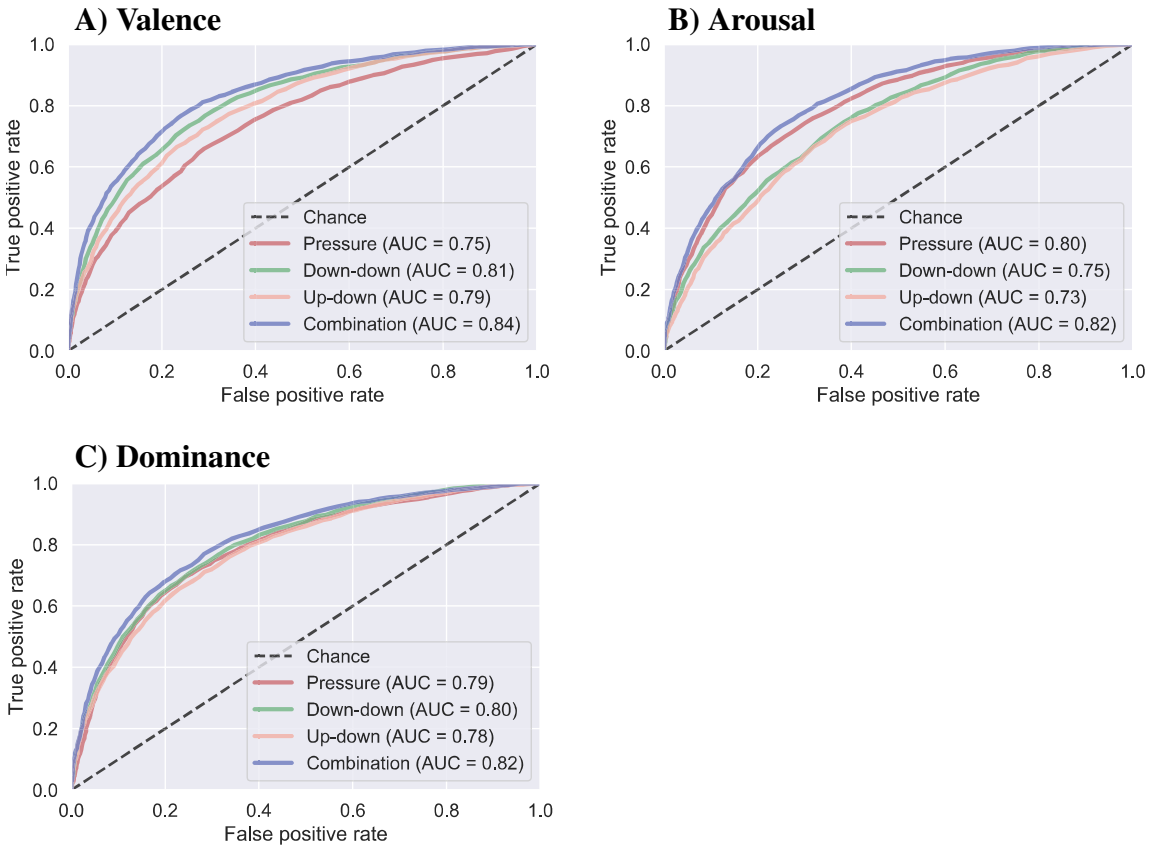


Figure 5.6: *ROC curves and micro-averaged AUC scores for classification of three levels (low, medium, high) of A) valence, B) arousal, and C) dominance.*

performance of our classifier can be affected by the period over which affective states are constant. For example, if affective states are alternating in short time, it can be much harder to make an accurate prediction compared to when affective states are constant over a longer period. This potential fluctuation in affective states, cannot be taken into account if we consider all labeled data from the conversations. As such, we recalculated the accuracy measure by considering only the data points for which the affective state was constant over a certain period (i.e., a specific number of preceding data points with the same class). Figure 5.8 shows the result of this accuracy measure for valence, arousal, and dominance. Here, a sequence length of zero corresponds to considering all data while sequence lengths of one, two, and three imply that we only considered data points having at least one, two, and three preceding data points with the same label. By excluding only immediate jumps (sequence length of one), we observe a steep increase in accuracy, reaching 78%, 75%, and 77% for valence, arousal, and dominance. In contrast, increasing the sequence length to two or three preceding data points provides only marginal improvements.

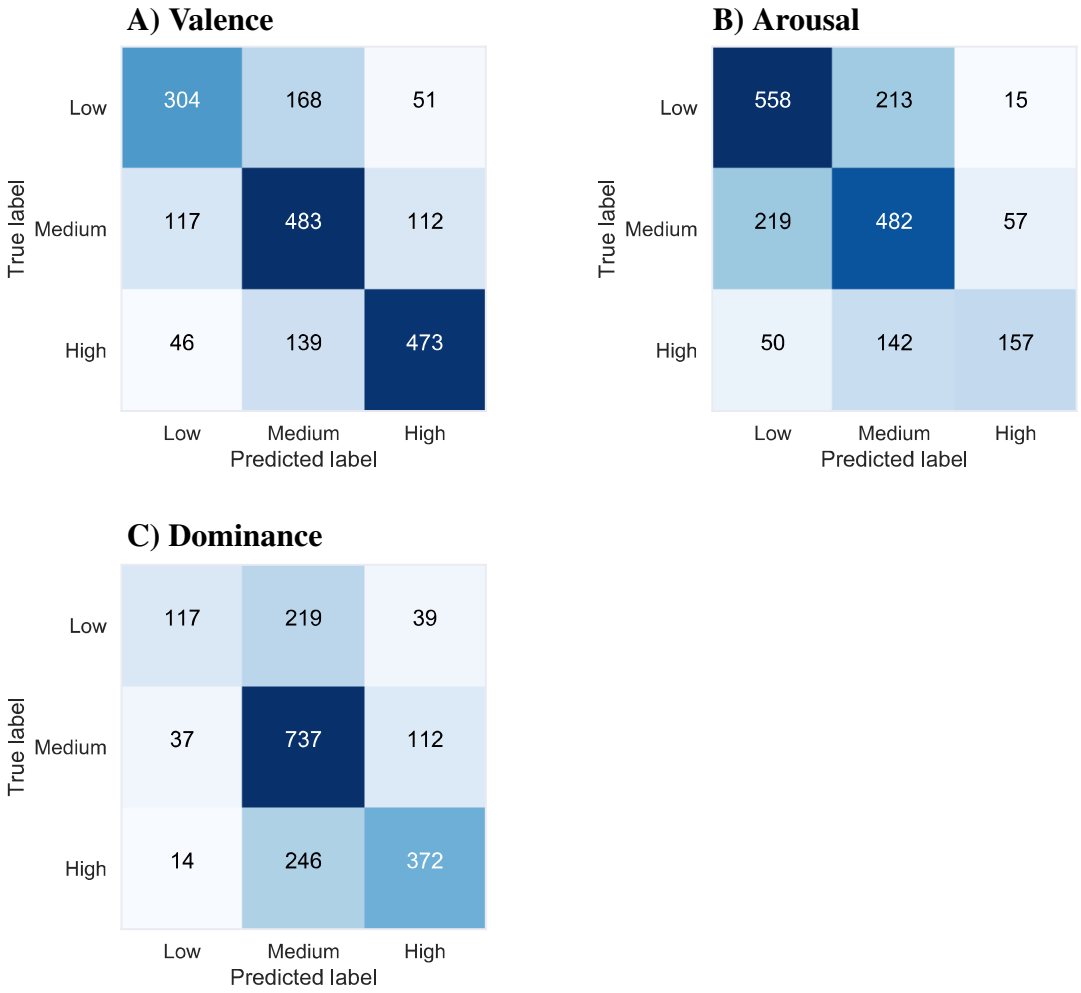


Figure 5.7: *Confusion matrices for classification of three levels (low, medium, high) of A) valence, B) arousal, and C) dominance. The confusion matrices are calculated by predicting self-reports across all chat conversations.*

5.3.5 Basic Emotion and Stress Prediction

With regard to the four basic emotions and stress, our classifier achieved a predictive performance of 87% (0.84 AUC) for anger, 81% (0.88 AUC) for happiness, 84% (0.87 AUC) for sadness, 84% (0.76 AUC) for surprise and 92% (0.80 AUC) for stress. The large differences between accuracy and AUC can be attributed to class imbalance (e.g., 164 vs. 1729 labels for stress). Altogether, these results reveal that our model is not only able to predict affective states measured in terms of valence, arousal, and dominance but is also predictive for a subset of the basic emotions and stress.

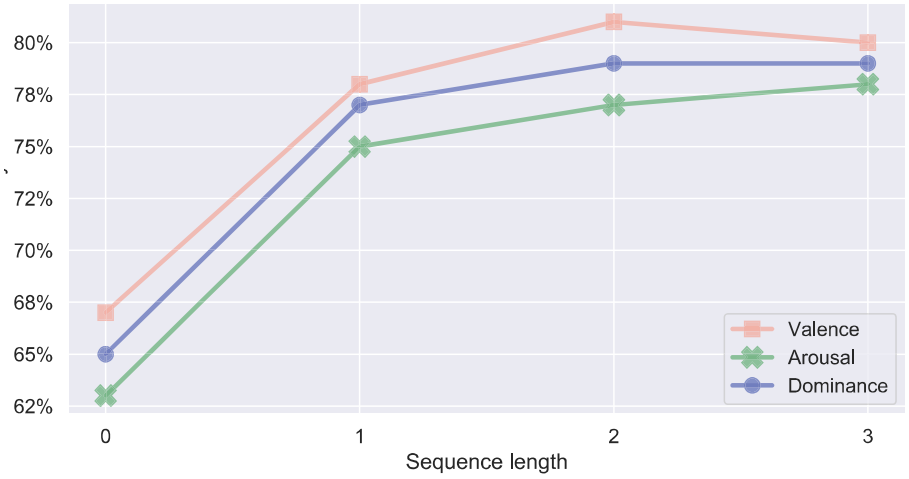


Figure 5.8: *Accuracy only considering data points with a specific number of preceding data points with the same class label.*

5.3.6 Runtime Analysis

For evaluating the applicability of our method for realtime predictions, we conducted a runtime analysis of the different parts of our model. Our computing environment consisted of an Intel® Xeon® CPU E5-2698 v4 @ 2.20GHz and an NVIDIA GeForce® GTX 1080 Ti. The training time of the variational autoencoder amounted to 182 minutes (*pressure* heat maps) and 611 minutes (*down-down speed* and *up-down speed* heat maps). For real-time applicability the training time of the networks does not matter because the networks can be trained beforehand on the existing data set and then used for the prediction of new data. Prediction of a new data point consisted of extracting heat maps (mean $= 0.38$ s, SD $= 0.09$ s), followed by extracting the low-dimensional embedding of the heat maps using the encoder (mean $= 0.065$ s, SD $= 0.0089$ s) and using the fully connected network for prediction (mean $= 0.002$ s, SD $= 0.003$ s). Summing up these values leads to a prediction time of 0.447 seconds. In other words, the system is capable of making two new predictions every second.

5.4 Discussion

In this chapter, we presented a complete classification pipeline that is capable of accurately predicting three classes (low, medium, high) of valence (up to 0.84 AUC), arousal (up to 0.82 AUC) and dominance (up to 0.82 AUC). In addition, we also showed that we could accurately predict two levels (present vs. not present) of stress

(0.80 AUC) and the basic emotions of anger (0.84 AUC), happiness (0.88 AUC), sadness (0.87 AUC) and surprise (0.76 AUC).

These predictions were based on heat maps generated from pressure and touch speed (i.e., down-down and up-down) collected during text conversations. We found that all three types of heat maps can predict valence, arousal, and dominance. Interestingly, while down-down speed showed the best performance for valence (0.81 AUC) and dominance (0.80 AUC), pressure was most predictive for arousal (0.80 AUC). These results may be related with the findings reported by Hernandez et al. [2014], suggesting that people apply more pressure on keyboards under stressful conditions. Moreover, affective states characterized by higher valence (e.g., excitement) can lead to higher typing speed, increasing the down-down speed and up-down speed, which has also been reported in previous work (e.g., Lee et al. [2015]).

The performance of our model cannot be directly compared with previous work due to differences in experimental setup. For example, Gao et al. [2012] used a game-based setting and different measures of emotional states while Huang et al. [2018] predicted mood on a regression scale. Our work in this chapter did not focus on the comparison of performance but instead on automatic feature extraction in a different setting. Our use of heat maps also allowed us to investigate the distributions of keystrokes as a measure of affective states (e.g., use of more backspaces when experiencing negative emotions). Interestingly, running our model using only spatial heat maps, we achieved a performance of only up to 0.60 AUC. Thus, we conclude that the distribution of keystrokes alone has only little predictive power.

We also showed that accuracy depends on the sequence of previous affective states and that accuracy tends to drop if affective states alternate. The reason for this is that when there is a preceding state belonging to a different class (e.g., low), noise is added to the window used for calculating the heat maps because this window contains touch data from both states whereby 1) the touch data is very different (e.g., low and high classes) or 2) the touch data is similar, but the class is different (e.g., low and medium classes).

Another noteworthy property of our model is its efficiency, which is particularly relevant for interactive applications in real-world environments. The computation of the heat maps, embedding, and prediction takes 0.447 seconds in total, meaning that the system can provide feedback on the user's emotional state in less than a second.

We acknowledge some limitations of the approach presented in this chapter. First, the experiment was restricted to a controlled lab environment and a population consisting of bachelor and master students. In addition, by querying emotions every 90 seconds, we might miss finer changes in emotions. A remedy would be to allow users to manually fill in self-reports when they face changes in emotions or to allow retrospective ratings. Finally, we acknowledge that using the pressure signal is limited

to devices supporting pressure measurement. Pressure can also be measured using the contact area of the fingertips, which is a supported measure by many smartphones nowadays, but this might negatively affect the performance of the prediction because the contact area can only approximate real pressure. In Chapter 6 we overcome some of these limitations by evaluating a refined system in real-world settings on different devices and by providing the users more freedom in choosing the moment for filling in the self-reports.

5.5 Conclusion

In this chapter, we presented a semi-supervised pipeline for predicting affective states and emotions based on heat maps generated from smartphone touch data. We validated our pipeline on touch data collected from text conversations in a lab experiment with 70 participants. We conducted the evaluation using a leave-one-user-out cross-validation, which ensures that our results generalize among users, and similar results can be expected when applying our pipeline to data from new users. We demonstrated that our pipeline could accurately predict three classes (low, medium, high) of valence (up to 0.84 AUC), arousal (up to 0.82 AUC) and dominance (up to 0.82 AUC). We also presented results for the prediction of two levels (present vs. not present) of anger (0.84 AUC), happiness (0.88 AUC), sadness (0.87 AUC), surprise (0.76 AUC), and stress (0.80 AUC). Considering the real-time applicability of our method (predictions are possible in less than 1 second), our pipeline can be useful in combination with our proposed visualization of affective states in Chapter 7. Our model provides an elegant way to combine features (i.e., the features are learned automatically by the encoder as part of the low-dimensional embedding) without explicit feature engineering. By using heat maps in contrast to raw data, we are also taking into account the spatial distribution of the data. In contrast to our other work using biosensor data (see Chapter 4) and video data (see Chapter 3), our approach is lightweight, less invasive, and can be used on different types of mobile devices. The findings of this chapter show a promising possibility of leveraging touch data to create emotion-aware chat conversations.

CHAPTER 6

Affective State Prediction Using Smartphones in the Wild

The ubiquitous use of smartphones for social interactions (e.g., chat applications and social networks), entertainment (e.g., music and video platforms), and news consumption provides a distinct opportunity for collecting information to recognize the affective states of users. In addition, smartphone use is also highly diverse in context and location (e.g., at home, on the train, or in school), which enables capturing the variability in affective states that may be used for prediction models in real-world environments. In Chapter 5, we leveraged data from a fixed setting and context (i.e., using a chat application while sitting at a table). Instead, in this chapter, we rely on data captured in real-world environments without any restrictions in smartphone usage and context.

In this chapter, we propose a system that can accurately predict affective states in real-world environments. We focus on typing-based applications (e.g., chat and browsing applications) as these are the most used applications [Androidrank, 2021] as well as smartphone sensor data (i.e., gyroscope and accelerometer sensors).

Using data from our in-the-wild user study with 82 participants, we show that we can accurately predict three levels of valence, arousal, and dominance. We also demonstrate a similar performance when using two-dimensional heat maps of just the inertial sensor data. We conclude that sensor data is a viable alternative to keyboard data for the prediction of affective states due to the continuous availability of data and the increased protection of privacy. The use of sensor data provides also greater flexibility in the wild and can increase the acceptance of users.

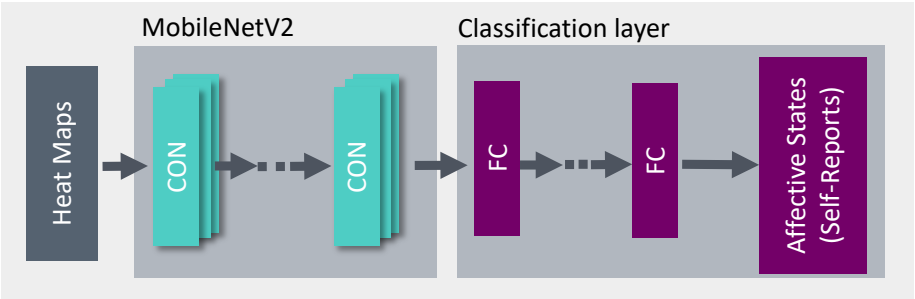


Figure 6.1: *Overview of the main steps of our model. A convolutional neural network (MobileNetV2 [Sandler et al., 2018]) is trained on heat maps created from smartphone keystroke and inertial sensor data. For classification of the affective states, the features learned by MobileNetV2 are used as input to fully connected layers.*

6.1 Method

Our model predicts affective states based on keyboard and inertial sensor data collected during smartphone usage (see Figure 6.1). We encode these data in two-dimensional heat maps and train convolutional neural networks to automatically extract meaningful features from the heat maps. For the classification of affective states, we then add a fully connected classification layer. The paragraphs below provide details on every step of our model.

6.1.1 Heat Maps

From the smartphone data collected in the wild, we generate two types of two-dimensional heat maps. First we create keystroke heat maps that encode typing characteristics of bigrams (i.e., key combinations) of consecutive keystrokes on the smartphone keyboard. Second, we create sensor heat maps encoding the distribution of the gyroscope and linear acceleration measurements.

Keystroke heat maps. A keystroke $k_i = (x, y, t_{\text{down}}, t_{\text{up}})$ is defined by the coordinates (x, y) on the screen as well as t_{down} and t_{up} providing the timestamp in milliseconds of pressing (touch down) and releasing (touch up) the key, respectively. A text $K = [k_1, \ldots, k_n]$ consists of n keystrokes. Based on the raw input data, we extract three keystroke metrics. First, "up-down" measures the time of moving from one key to the next key (up-down $= t_{i+1,\text{down}} - t_{i,\text{up}}$). Second, "down-down" measures the time of moving between keys as well as the hold time of the first keystroke (down-down $= t_{i+1,\text{down}} - t_{i,\text{down}}$). Third, "down-up" considers the time of moving between keys and the hold time of the first and second keystroke

(down-up $= t_{i+1,\text{up}} - t_{i,\text{down}}$). All three keystroke metrics are normalized by the distance between the keys.

Using a window of 80 keystrokes before the self-reports, we aggregate the keystroke metrics into two-dimensional heat maps covering all possible bigrams of characters (a–z including umlauts ä, ö, and ü) and special keys (i.e., delete, space, symbol, shift, return, period, comma, question mark, and exclamation point). In total, we consider 38 keys. We encode all possible key combinations in a 38×38 heat map H. The rows and columns encode all 38 keys taken into consideration using a centralized alignment of the keys. More frequently used keys in the English language [Solso and King, 1976] and German language [Best, 2005; Beutelspacher, 1996] are placed in the middle of the heat map (the space bar is considered to be the most frequent key and the exclamation point and q are the least frequent keys). The first and second keystroke in a bigram is encoded in the row and column, respectively. For example, $H(a, p)$ contains the keystroke metric (i.e., up-down speed, down-down speed, or down-up speed) calculated from the keys a (row) and p (column) of the bigram ap. We average all the values for each cell in the heat map H. In addition, all heat maps are standardized based on the mean heat map during a baseline typing period.

Figure 6.2 shows examples of extracted heat maps for up-down speed, down-down speed, and down-up speed. The colors in the heat maps are for visualization purposes only. In our model, we are using only one value per pixel. Up-down speed is larger than down-down speed and down-down speed is larger than down-up speed. From the heat maps, it is also visible that the highest up-down speed (see Figure 6.2A) is concentrated in the bigrams (*space bar, D*), (*E*, *N*), and (*space bar*, *shift key*). The bigram (*space bar*, *D*) has the largest down-down speed (see Figure 6.2B). The highest value for down-up speed (see Figure 6.2C) is associated with the bigram (*N*, *symbol key*).

Sensor heat maps. For creating the sensor heat maps, we extract the rate of rotation and linear acceleration of smartphones from the inertial sensors. Linear acceleration reports the gravity-subtracted acceleration of a smartphone in SI units (m/s^2) along three axes (i.e., x, y, and z). The linear acceleration sensor typically uses the gyroscope and accelerometer (providing acceleration including gravity) as input. On the other hand, a gyroscope is measuring the rate of rotation in radians per second around three-axis (i.e., x, y, and z). Linear acceleration and gyroscope measurements are relative to the smartphone's local coordinate system (the x-axis is parallel to the smaller screen side, the y-axis is parallel to the larger screen side, and the z-axis is normal to the screen). As a preprocessing step, we temporarily align the signals and convert the sampling rate to 100 Hz (i.e., downsampling and upsampling), which provides a noise reduction as a positive side effect.

We encode the three-axis combinations into separate heat maps: linear acceleration along the x-axis & rate of rotation around the z-axis, linear acceleration along the

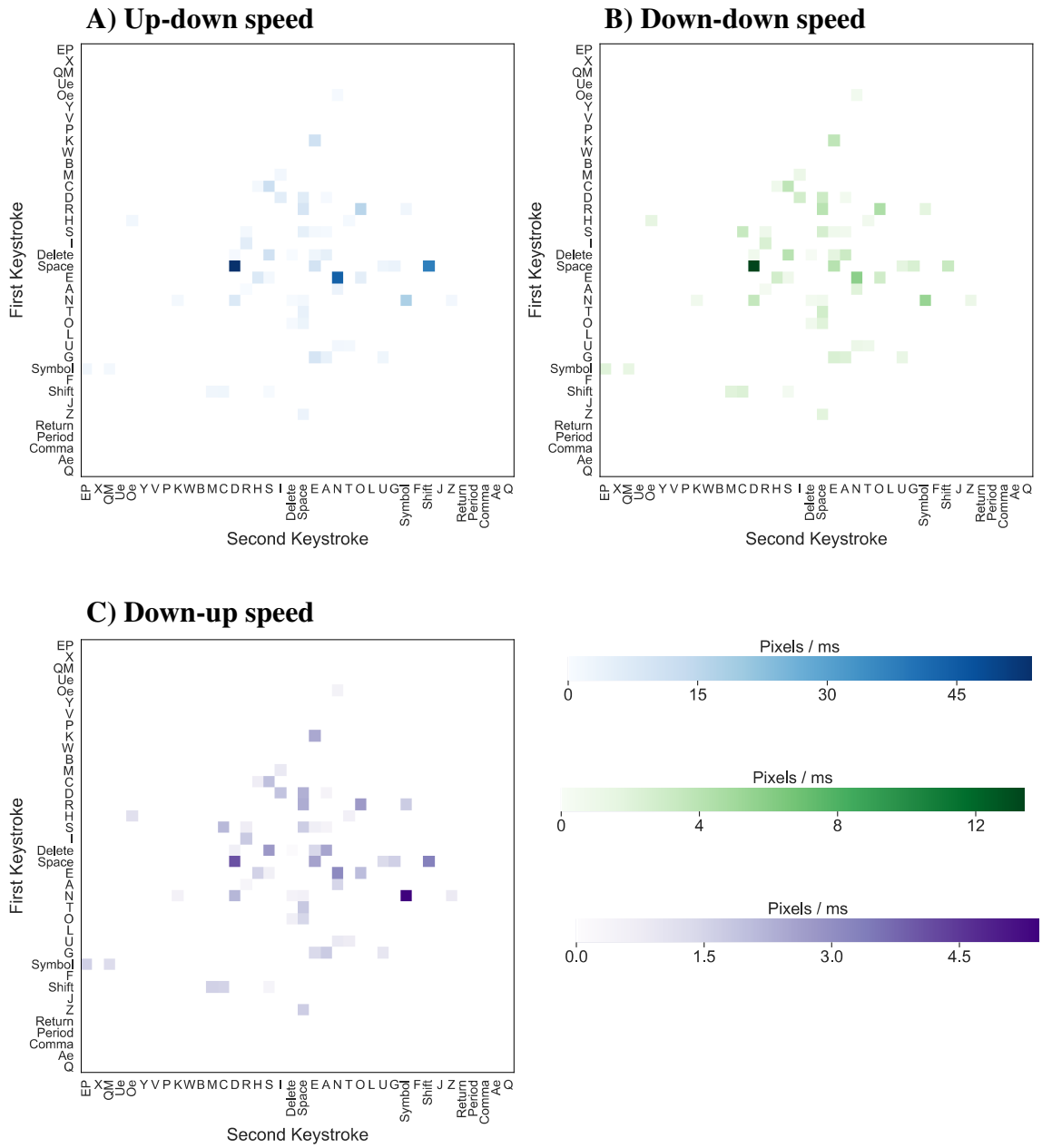


Figure 6.2: *Examples of keystroke heat maps extracted from 80 keystrokes of a selected participant. Abbreviations: exclamation point (EP), question mark (QM), ü (Ue), ö (Oe), and ä (Ae). Color saturation indicates the average up-down speed (A), down-down speed (B), and down-up speed (C) between consecutive keystrokes. The colors are for visualization purposes only.*

y-axis & rate of rotation around the x-axis, and linear acceleration along the z-axis & rate of rotation around the y-axis. We choose these axis combinations, because they reflect typical motion sequences. Using a window of 30 seconds, we bin the absolute sensor values into logarithmically spaced bins and count the number of values in each bin. We chose a logarithmic scale because the absolute sensor measurements are exponentially distributed. Thus, when taking the logarithm, the measurements become approximately normal distributed. Moreover, we believe that the distinction of smaller values is more important than larger values so that also micromotions can be adequately exploited [Ma et al., 2012]. For the heat maps, we use a resolution of 96×96 (i.e., 96 bins in each dimension) because it is divisible by a multiple of two, which is advantageous for the spatial downsampling in a convolutional neural network and it provides a sufficiently high resolution. We standardize all heat maps based on the mean heat map during a baseline period.

Figure 6.3 shows examples of extracted heat maps for the three-axis combinations (the colors are for visualization purposes only). The linear acceleration in the direction of the x-axis and the y-axis shows a large spread. The largest linear acceleration is contributed to the z-axis (i.e., moving the smartphone forth and back). The rotation around the x-axis and y-axis was slightly larger than around the z-axis.

6.1.2 Convolutional Neural Network

For the keystroke and sensor heat maps, we stack the three types of heat maps into three channels. To extract meaningful features from the keystroke ($38 \times 38 \times 3$) and sensor heat maps ($96 \times 96 \times 3$), we employ a particular type of convolutional neural network called MobileNetV2 [Sandler et al., 2018]. Affective labels are typically sparse and labeled datasets are relatively small for training a network for predicting affective states, making it prone to overfitting. Using a smaller but expressive network such as MobileNetV2 counters this effect. MobileNetV2 is a network optimized for low memory consumption and high execution speed and is parameterized to meet the resource constraints of mobile devices [Sahlol et al., 2020; Xiang et al., 2019].

The basic building block of MobileNetV2 is the bottleneck depth-separable convolution with residuals which consist of three operations [Sandler et al., 2018]. First, a 1×1 convolution layer expands the number of feature maps. Second, the depthwise convolution applies a single filter to each feature map. Finally, a pointwise convolution with a kernel size of 1×1 is used to combine the outputs of the depthwise convolutions (i.e., linear combinations of the feature maps) reducing the number of feature maps, and thus the amount of data flowing through the network. The factorization of the convolution into depthwise and pointwise convolutions reduces the computational cost and model size. In addition, the input and output to the basic

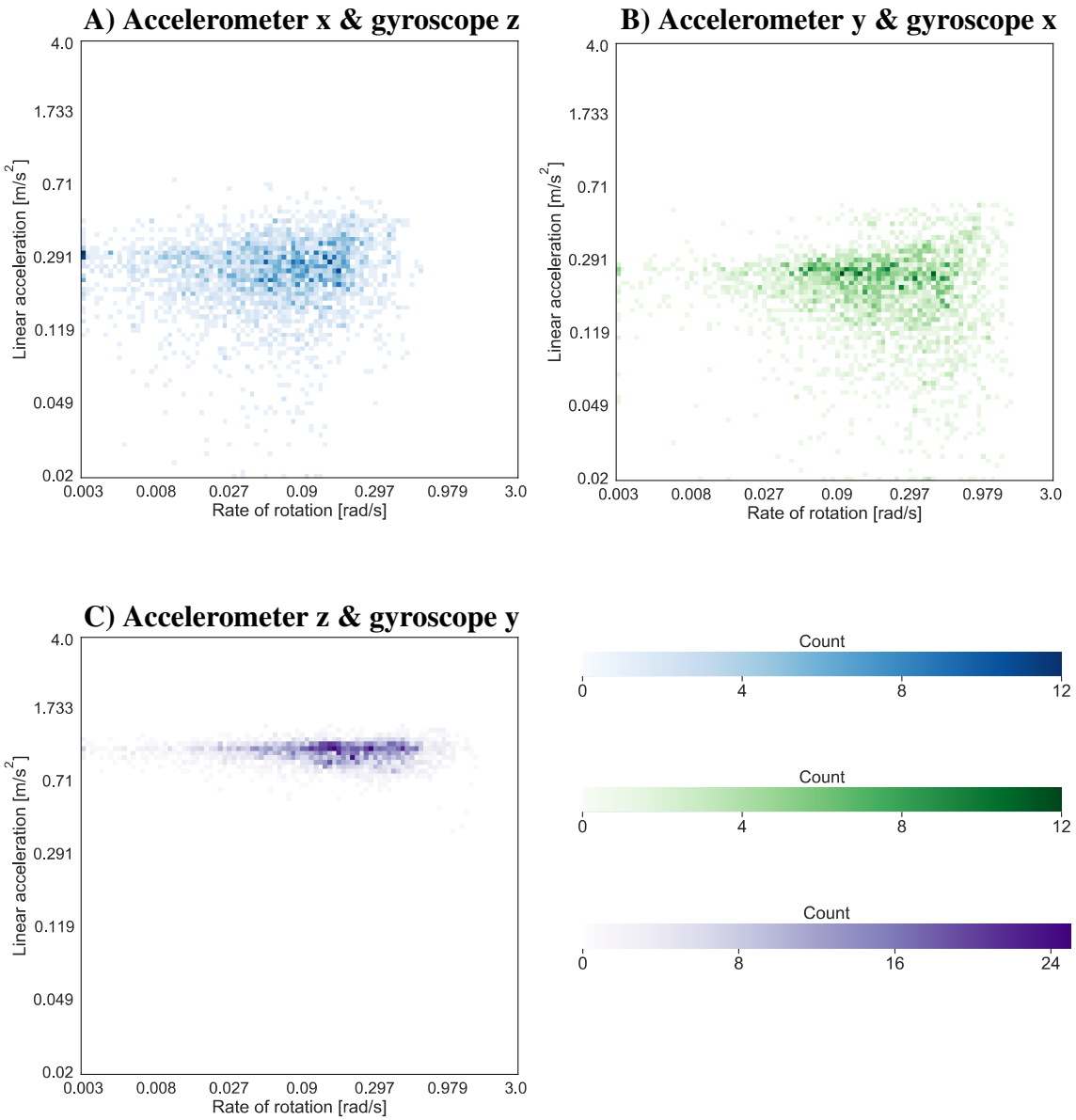


Figure 6.3: *Examples of sensor heat maps extracted from 30 seconds of the gyroscope and linear acceleration measurements of a selected participant. The color saturation indicates the number of sensor measurements for the combinations of the linear acceleration along the x-axis & the rate of rotation around the z-axis (A), the linear acceleration along the y-axis & the rate of rotation around the x-axis (B), and the linear acceleration along the z-axis & the rate of rotation around the y-axis (C). The colors are for visualization purposes only.*

building block are connected with a residual connection which enables faster training and better accuracy [Sandler et al., 2018].

MobileNetV2 was developed for images with a resolution of $224 \times 224 \times 3$ and consists of five downsampling layers (i.e., a stride of two). Thus, for the keystroke heat maps ($38 \times 38 \times 3$), we disable the first three downsampling layers (i.e., setting the stride to one). For the sensor heat maps ($96 \times 96 \times 3$), we disable the first downsampling layer. This modification of the network was successfully used on the CIFAR10 dataset (containing images with a resolution of $32 \times 32 \times 3$) [Ayi and El-Sharkawy, 2020]. Input data is commonly scaled before training. We use Min-Max scaling of the heat maps to the range $[-1, 1]$.

6.1.3 Classification

We take advantage of the learned features from the convolutional neural network by adding a classification network. The final output of the convolutional neural network is passed through a global average pooling layer and a fully connected layer with softmax activation. We aggregate the keystroke and sensor heat maps by stacking the output of the global average pooling layer of the pre-trained networks of the individual heat maps. For the combination of the sensor and keystroke heat maps, we use fully connected layers between the global average pooling and the softmax layer to foster the learning of mixtures of the extracted features from the heat maps. We train the whole network on the labeled data (heat maps and corresponding affective states) using backpropagation minimizing the cross-entropy loss.

6.2 Experiment

We conducted an experiment in the wild to validate our pipeline for the prediction of affective states based on smartphone keystroke and sensor data. The ethics board of ETH Zurich approved the experiment. During the experiment, we collected keyboard data, sensor data, and context data (e.g., foreground application) while participants used their smartphones in everyday life for approximately 70 days.

6.2.1 Participants

We recruited 82 participants (43 female, 39 male) between the ages of 18 and 43 (mean $=$ 23.0 years, standard deviation SD $=$ 3.64 years). Eleven participants were left-handed and seventy-one participants were right-handed. The majority of participants were students at the bachelor (61 participants), master (13 participants), and Ph.D. (3 participants) levels from ETH Zurich and the University of Zurich. The

remaining participants were an IT consultant, a nurse, a trainee, and two software engineers. We only considered participants that were German native speakers (due to the keyboard layout) and used typing-based applications (e.g., browsers and chat applications) daily on the smartphone. We excluded participants taking any type of medication, tranquilizers, or psychotropic drugs (e.g., anti-depressants) as well as participants affected by any type of autism spectrum disorders. We recruited only participants using Android devices (Android 7 to 10). The participants used a variety of devices from different manufacturers, i.e., Samsung (31 participants), Huawei (21), Xiaomi (7), OnePlus (7), Sony (3), LG (3), Google (2), Nokia (2), Blackberry (1), Fairphone (1), HTC (1), Lenovo (1), Oppo (1), and Wiko (1). The participants actively engaged for an average of 72 days (SD $=$ 2 days) in our experiment.

Compensation. Participants were rewarded for their participation depending on their level of contribution and received between CHF 60 and CHF 120 for submitting an average of 3 and 6 self-reports per day, respectively. One participant was awarded an additional CHF 1000 from a lottery draw. Depending on the number of submitted self-reports participants could reach three different levels providing a different number of tickets for the lottery: gold level (420 self-reports, 10 tickets), silver level (320 self-reports, 5 tickets), and bronze level (210 self-reports, 1 ticket). In addition, the participants with the highest number of average self-reports per day (after 70 days of participation) received additional tickets for the lottery (5 tickets for rank 1, 2 tickets for rank 2, and 1 ticket for rank 3). Such an incremental reward system and the chance to win an additional price via a lottery was already successfully employed in other works [Healey et al., 2010; Stieger et al., 2018; Wang et al., 2014].

6.2.2 Apparatus

To collect a large-scale dataset in the wild to validate our pipeline, we developed an Android application consisting of four main components: 1) a graphical user interface (GUI) providing the participant information, control, and statistics of the experiment, 2) a data logging component for collecting sensor data, context data, and usage logs in the background, and 3) a keyboard the participants had to use during the experiment. In the following, we detail the three components.

Graphical user interface. The main page of the app (see Figure 6.4A) provided information about the number of remaining days of participation and the number of self-reports until the next level is reached. Participants could manually start and pause the data recording. This mechanism enabled privacy when they did not want their data to be recorded.

Participants were required to have recording enabled for at least 90% of the time to be eligible for compensation. Furthermore, the experimenter could send messages in the form of notifications to specific or all participants (e.g., information about the

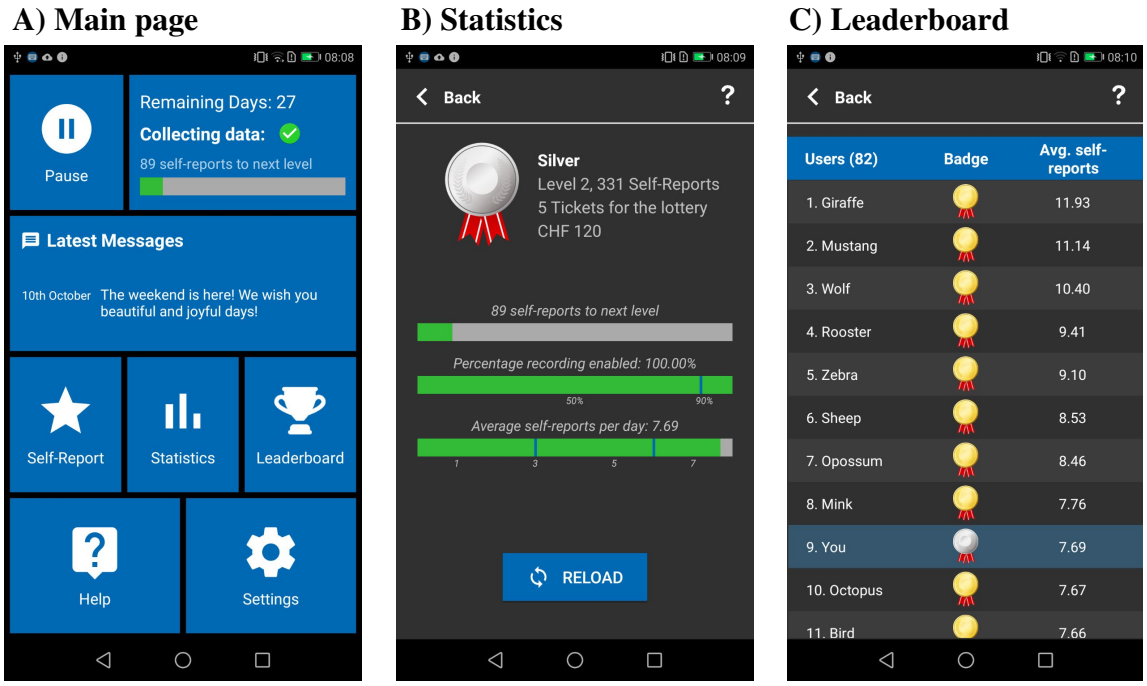


Figure 6.4: *Graphical user interface of the Android application. A) Main page of the application. B) Statistics providing information about self-reports and compensation. C) Leaderboard showing badges (level), average number of self-reports per day, and the rank. Users were assigned animal names to preserve anonymity.*

experiment or motivating messages). Participants could also access help information (i.e., help text, tutorial videos, and the information sheet) and change the settings (e.g., the storage location of the recorded data, finishing participation, and manual triggering synchronization with the server).

The statistics page (see Figure 6.4B) provided information about the number of self-reports, average number of self-reports per day, percentage of enabled recording, and information about compensation and lottery tickets. Finally, in the leaderboard (see Figure 6.4C), participants could track their rank in relation to the other participants in terms of the average number of self-reports per day. To maintain privacy of the participants, we assigned an animal name to each participant.

Data recording. When the phone was unlocked, the Android application logged the following data in the background: sensor data (i.e., accelerometer, gyroscope, magnetometer, proximity sensor, light sensor, and step counter), device usage logs (e.g., foreground application, charging state, screen orientation, ringer mode, timezone changes, and audio mode), and activity predicted by the activity recognition API of Google (i.e., still, in-vehicle, on a bicycle, running, on foot, tilting, and walking). We

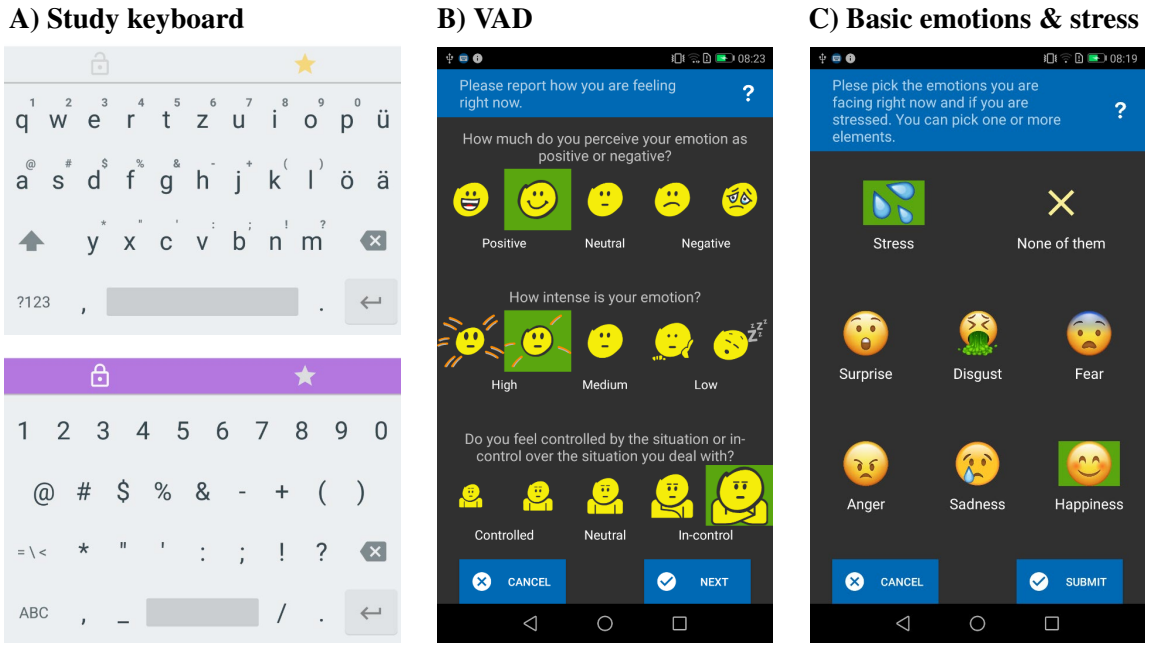


Figure 6.5: *The keyboard included in our application and the self-reports the participants had to fill in. A) Two additional buttons in the top bar for enabling private mode (left button) and starting a self-report (right button). The upper keyboard has private mode disabled and a self-report available (yellow star) and the lower keyboard has private mode enabled (purple top bar) and no self-report available. Self-reports captured valence, arousal, and dominance (B) and the basic emotions and stress (C). Selected items are highlighted with a green background.*

did not use all sources of data in this work. For example, we discarded activity from our analysis because on some devices there was a substantial lag in recognition of activities.

The data was uploaded in the background to a server several times during the day. Communication between the application and the server was encrypted and upload only took place when the smartphone was connected to Wi-Fi.

Keyboard. Our application included a keyboard with a layout similar to the default German Android keyboard (see Figure 6.5A). Participants were required to use our keyboard during the study. From the keyboard, we recorded touch-related data (i.e., position and timestamps). We did not record the pressed keys. In the modeling stage, we then mapped the touch positions to the keys. No data was recorded when participants typed passwords, phone numbers, names, postal addresses, and e-mail addresses.

The keyboard did not support auto-correction, auto-completion, and swiping. A pre-

experiment questionnaire revealed that before the experiment, 79% of the participants had never used swiping, 71% had never used auto-correction, and 75% had never or only rarely used auto-completion.

We extended the keyboard layout by two additional buttons at the top. The private mode button (left button in the top bar in Figure 6.5A) allowed participants to pause the recording of data directly on the keyboard. By pressing the star button for 2 seconds (right button in the top bar in Figure 6.5A), participants could fill in a self-report. Participants could start a self-report using the self-report button on the main page of the app (Figure 6.4A). Ninety-three percent of the submitted self-reports were started using the star button on the keyboard.

6.2.3 Self-Reports

To gather labeled data for our model, we asked participants to complete self-reports at regular intervals while using their smartphones. To quantify valence, arousal, and dominance, we adapted the Self-Assessment Manikin (SAM) [Bradley and Lang, 1994] in terms of the dimensions it represents and the number of levels. The SAM is not applicable on smartphones due to its old-fashioned style and the space constraints of smartphone screens. Based on the work by Hayashi et al. [2016] and feedback from participants in a pilot study ($n = 17$), we substituted the figures from the SAM with emojis and reduced the scale to five items (i.e., very low, low, neutral, high, very high). Emojis are commonly used in social networks and other communication applications. This familiarity made the self-reports more appealing and fostered a fast and accurate understanding of the experimental procedure by the participants.

Figure 6.5B shows an illustration of the emoji-based self-reports. For the valence dimension, we varied the emojis from a happy face (most positive) to a sad face (most negative). In the arousal dimension, the emojis varied from a awake emoji with large eyes (highest arousal) to a sleepy emoji (lowest arousal). Finally, for the dominance dimension, we increased the size of the emoji to portray control, similar to the SAM. Participants were also asked to select from a series of basic emotions (i.e., happiness, anger, sadness, surprise, disgust, and fear) and stress represented by different emojis (see Figure 6.5C). To track complex emotions, participants were allowed to select all possible combinations of the basic emotions and stress. Participants could also select *None of them* if none of the provided items applied (in that case no other items could be selected).

Following the guidelines by Schmidt et al. [2018] and Ghosh et al. [2019a], we used a combination of time-based and event-based schedules to trigger self-reports. A self-report became available (i.e., the star button on the keyboard turned yellow and started blinking) when four conditions were fulfilled. First, the participant typed at least 80 characters on the keyboard in the current session (we define a session as the period

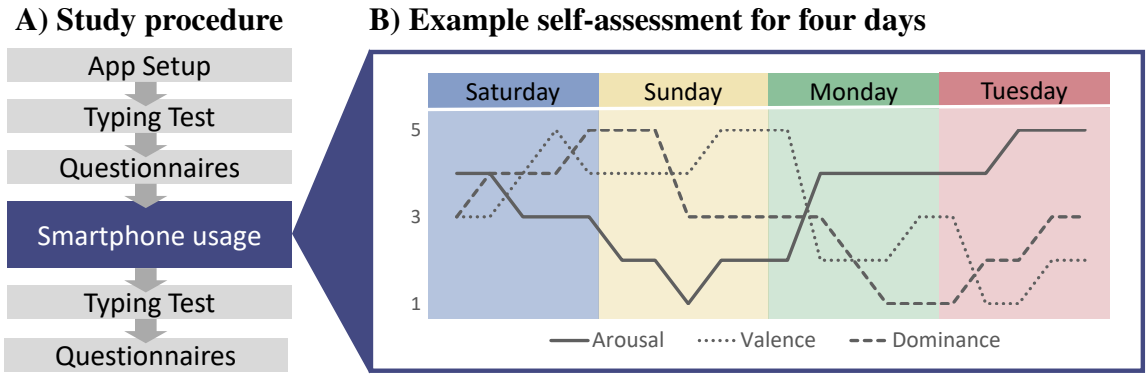


Figure 6.6: *Overview of the different parts of the experiment. A) Overall experimental procedure. B) Changes in valence, arousal, and dominance of a selected participant during four consecutive days.*

from unlocking the smartphone until it is locked again). Second, the smartphone was unlocked for at least 30 seconds in the current session. Third, between 30 minutes and 60 minutes elapsed since the last self-report was filled in. Fourth, data recording was enabled (i.e., private mode on the keyboard was disabled). Depending on the number of average self-reports per day, the minimum amount of time between self-reports (condition 3) was set to 30 minutes, 45 minutes, or 60 minutes. This helped to balance the number of self-reports per day and prevented participants from exaggerating submissions of self-reports. Once a self-report became available, participants could start the self-report until the smartphone was locked again (an additional margin of 10 seconds was provided in case participants accidentally locked the phone). We did not enforce a time limit for filling in the self-reports to not put additional pressure on the rating, which could introduce a negative bias. All these parameters were decided based on results from a pilot study ($n = 17$).

6.2.4 Procedure

Figure 6.6A provides an overview of the procedure used in the experiment. Before installing the application, we asked participants to read the information sheet and watch two YouTube tutorial videos explaining the application and the self-reports including four examples. The participants then installed the application from the Google Play Store. After opening the application for the first time, participants logged in with a username (i.e., an animal name) and password provided by the experimenter. After participants provided informed consent by selecting a checkbox, they were given a second chance to watch the tutorial videos. Next, the application requested participants to grant various device permissions. After setting up the application (i.e., watching two tutorial videos and granting device permissions), participants

conducted a typing test on their default keyboard used before the experiment and on the application keyboard before setting the application keyboard as the new default keyboard. The typing test consisted of six sentences in random order including two well-known pangrams (27, 30, 56, 37, 44, and 46 characters) [Dhakal et al., 2018; Palin et al., 2019].

After the setup was completed, participants used their smartphones for 10 weeks in everyday life, filling in self-reports in regular intervals. We collected a total of 30083 self-reports covering a large range of the valence-arousal-dominance space. Within the first week, we asked participants to fill in an online questionnaire on demographics and smartphone usage as well as the Patient Health Questionnaire [Kroenke et al., 2009] and the Big Five Inventory 2 [Danner et al., 2016] as measures of mental health and personality traits, respectively. At the end of the experiment, participants again typed the six sentences in random order on our keyboard and their default keyboard used before the experiment. Finally, participants completed an exit questionnaire on the self-reports (understandability, truthfulness, and frequency). The exit questionnaire also probed their perception of the application's keyboard and smartphone usage. Finally, participants were asked to fill in the Patient Health Questionnaire and the Big Five Inventory for a second time.

Figure 6.6B depicts the changes in valence, arousal, and dominance during four days of one selected participant. The figure shows that valence and dominance were highest on Saturday and Sunday and decreased on Monday and Tuesday. Arousal showed an opposite pattern with lower values on Saturday and Sunday and higher values on Monday and Tuesday.

6.3 Results

We used the data we collected during the study to evaluate our model. The 30083 self-reports served as labels to our model. We evaluated the performance of our pipeline in terms of accuracy (chance level is 0.33 for three classes and 0.5 for two classes), micro-averaged AUC (chance level is 0.5), and macro-averaged AUC (chance level is 0.5). We evaluated our model using leave-one-user-out cross-validation to ensure that data of an user in the test set is not used for training.

6.3.1 Model Parameters

Heat maps. To extract the keystroke and sensor heat maps, we used 80 keystrokes and 30 seconds before the filled in self-reports, respectively. We set these thresholds due to the minimum number of keystrokes (i.e., 80 keystrokes) and minimum time passed (i.e., 30 seconds) since the start of the session until a self-report could be filled

in. On the training data, we used all available heat maps to compute the baseline heat map. For heat map n of an user in the test set, we used the $n-1$ heat maps to compute the baseline heat map (i.e., the baseline gradually improves the more the user types). For creating the sensor heat maps we clipped linear acceleration to $4\,\mathrm{g} = 39.2\,\mathrm{m\,s^{-2}}$ as 98% of the sensor data were below 4 g. We clipped the gyroscope measurements at $5\,\mathrm{rad\,s^{-1}}$ because 99% of the measurements were below this threshold. In addition, we only considered linear acceleration and gyroscope measurements greater than $0.02\,\mathrm{m\,s^{-2}}$ and greater than $0.003\,\mathrm{rad\,s^{-1}}$, respectively.

We chose these thresholds because in a pilot study 95% of the sensor measurements were below these thresholds when the smartphones were lying flat on a table. Thus, we excluded noise inherent to the sensors. To exclude breaks during typing, we chose a threshold of 1 second between keystrokes for the keystroke heat maps. We motivate this threshold by the longest median time per character (400 milliseconds) [Buschek et al., 2018] and the fact that median $+ 3 *$ median absolute deviation $= 0.9$ s [Leys et al., 2013]. By choosing a conservative threshold of 1 second, we retain delays that are part of natural typing behavior.

Classification pipeline. For the sensor and keystroke heat maps, we trained the MobileNetV2 architecture and used a fully connected network for the classification of affective states. For the combination of the sensor and keystroke heat maps, we used a fully connected layer with 2048 units between the global average pooling and the softmax layer to foster the learning of mixtures of the extracted features from the heat maps. Due to the prevalent class imbalance, we used balanced class weights to give smaller classes more weight. To train the networks we minimized cross-entropy loss using 80 epochs and a batch size of 128. We optimized the networks using stochastic gradient descent with a momentum of 0.9 and a cyclical learning rate using an exponential decay ($\gamma = 0.99994$) with a minimum and maximum learning rate of 10^{-5} and 10^{-2}, respectively [Smith, 2017]. We implemented all networks using the Keras framework with TensorFlow™ back-end.

6.3.2 Experimental Validation

Smartphone usage. Figure 6.7A shows the smartphone usage over the hours of the day and each day of the week aggregated over all participants. Smartphone usage was lowest during the night (1 a.m. to 6 a.m.). During the day, smartphone usage was stable with a peak around 10 p.m. On Saturdays, participants used their smartphones least, whereas on Sundays usage was high throughout the day with peaks in the late afternoon and evening.

We collected data from 13071 hours of smartphone usage. On average we recorded 3533 sessions per user (SD = 1624 sessions, max = 8861 sessions, min = 1060 sessions). On average a session lasted for 183 s (SD = 532 s) and we recorded

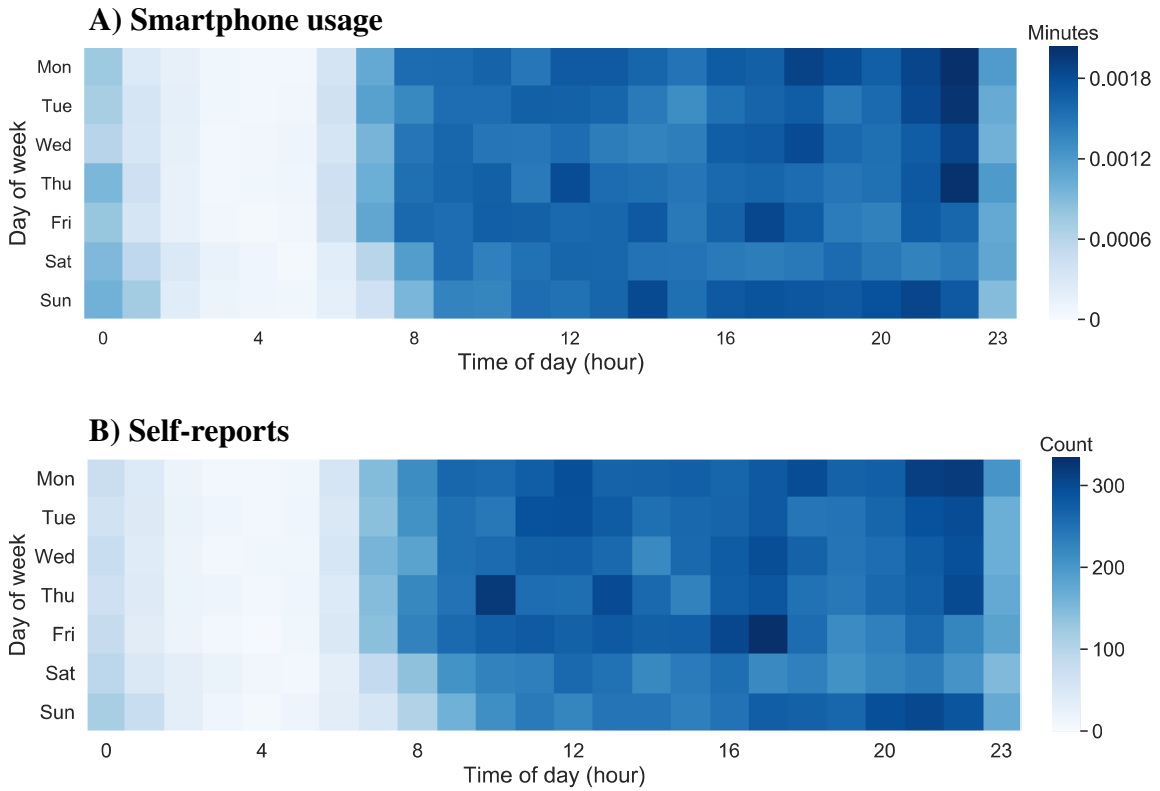


Figure 6.7: *The distribution of average smartphone usage (A) and self-reports (B) for the days of the week and the times of the day aggregated over all participants.*

an average of 47 keystrokes per session (SD = 174 keystrokes). The mean break between sessions was 11 min (SD = 72 min). Participants used the smartphone in landscape mode in 0.96% of the sessions. To simplify the analyses, we excluded all sessions where the smartphone was used in landscape mode. Participants could pause the recording by enabling the private mode. Participants used private mode only 0.026% of the time.

Self-reports. We collected a total of 30083 self-reports for valence (669 very low, 2767 low, 8071 neutral, 14642 high, 3934 very high), arousal (1643 very low, 5260 low, 12572 medium, 7591 high, 3017 very high), dominance (1866 very controlled, 3256 controlled, 12823 neutral, 8089 in-control, 4049 very in-control) and the basic emotions of anger (selected 1208 times), happiness (16425), sadness (1918), surprise (786), fear (1628), disgust (515), and stress (4795). On average, a participant submitted 402 self-reports (SD = 154 self-reports, min = 44, max = 835) totalling 5.64 self-reports per day on average (SD = 2.12 self-reports). Participants also spent an average of 6.76 s (SD = 27.27 s) filling in the self-reports. In addition, an average of 154 keystrokes (SD = 177 keystrokes) and 40 s (SD = 266 s) passed since the start of the session until a self-report was triggered.

Table 6.1: *Effect sizes of the Pearson correlations between valence, arousal, and dominance and the basic emotions and stress. Asterisks denote correlations that survived Bonferroni correction (p = 0.0024).*

	Anger	Happiness	Sadness	Surprise	Fear	Disgust	Stress
Valence	-0.30^{*}	$+0.55^{*}$	-0.37^{*}	-0.006	-0.22^{*}	-0.14^{*}	-0.24^{*}
Arousal	$+0.05^{*}$	$+0.27^{*}$	$+0.01$	$+0.04^{*}$	$+0.02^{*}$	-0.008	$+0.004$
Dominance	-0.16^{*}	$+0.17^{*}$	-0.18^{*}	-0.04^{*}	-0.17^{*}	-0.10^{*}	-0.20^{*}

Figure 6.7B shows the number of self-reports for the days of the week and the times of the day. As expected, a close match to smartphone usage is visible. Most self-reports were filled in between 7 a.m. and 11 p.m. Peaks are located on Mondays at 9 p.m. (318 self-reports) and 10 p.m. (323 self-reports), Thursdays at 10 a.m. (324 self-reports), and Fridays at 5 p.m. (338 self-reports).

We also performed a series of correlations to investigate the relationship between the valence, arousal, and dominance ratings, the basic emotions and stress. Table 6.1 presents the results for each of these correlations. The effect sizes are largest for valence and smallest for arousal. Notably, these results are a close match to the correlations found for the data collected in our laboratory experiment (see Table 5.2). We found the same direction for the correlations but smaller effects sizes.

Russell and Mehrabian [1977] provide a correspondence between valence, arousal, and dominance and the basic emotions based on laboratory experiments. In Table 6.2, we compare these values to the mean values obtained from the self-reports collected in our experiment. In Russel and Mehrabian's model, the affective dimensions (i.e., valence, arousal, and dominance) spanned the interval $[-1, 1]$. Thus, we mapped the self-reports collected in our experiment to the same interval to obtain a proper measure for comparison. The self-reports collected in our experiment closely match the correspondences found by Russel and Mehrabian. In contrast to Russel and Mehrabian, the mean values for valence, arousal, and dominance are smaller in our data. These differences may be related to the fact that we performed the experiment in the wild without using emotion-eliciting situations as stimuli. Notably, for anger, surprise, and disgust, the mean dominance value shows a reversed sign compared to Russel and Mehrabian's model. For stress, the mean values of all three dimensions (i.e., valence, arousal, and dominance) are around zero. It is known that stress can be positive and negative with different intensity levels [Folkman and Moskowitz, 2000; Folkman, 2008], thus potentially, positive and negative ratings cancel each other out leading to a mean close to zero.

A post-experiment questionnaire revealed that most participants completely (87%) or mostly (10%) understood the self-reports. All participants reported that they always

Table 6.2: *Mean values for valence, arousal, and dominance for the six basic emotions and stress. Results from our study are compared to the correspondences derived by Russell and Mehrabian [1977]. All measurements are mapped to the interval* $[-1, 1]$. *Values in brackets denote standard deviation.*

	Valence		Arousal		Dominance	
	Russel	Ours	Russel	Ours	Russel	Ours
Anger	-0.43	-0.36 (0.44)	0.67	0.19 (0.52)	0.34	-0.24 (0.56)
Happiness	0.76	0.54 (0.33)	0.48	0.20 (0.50)	0.35	0.24 (0.51)
Sadness	-0.63	-0.34 (0.50)	0.27	0.10 (0.54)	-0.33	-0.21 (0.51)
Surprise	0.40	0.29 (0.51)	0.67	0.20 (0.54)	-0.13	0.04 (0.52)
Fear	-0.64	-0.12 (0.51)	0.60	0.12 (0.53)	-0.43	-0.21 (0.53)
Disgust	-0.60	-0.18 (0.52)	0.35	0.04 (0.55)	0.11	-0.23 (0.57)
Stress	–	0.05 (0.49)	–	0.08 (0.51)	–	-0.08 (0.51)

(81%) or often (19%) filled in the self-reports truthfully. Most participants (59%) also felt that the self-reports had the right frequency. Only 11% and 30% of the participants reported that they found the self-reports being either too seldom or too often available, respectively.

Keyboard. We recorded an average of 7669 keyboard sessions per user (SD = 3341 sessions, min = 2255, max = 18445). We define a keyboard session as the time from opening to closing the keyboard. The ten most used keys were the space bar (13.5%), delete key (11.6%), E (8.3%), I (5.5%), A (5.1%), S (4.9%), N (4.7%), H (4.4%), T (4.0%), and R (3.8%).

In the modeling stage we generated the heat maps from key pairs. We hypothesize that the typing speed of a key pair depends on the involved thumbs (e.g., key combinations typed with the left and right thumb might be faster). Key pairs can be located in the left or right part of the keyboard (left-left and right-right) or belong to both parts of the keyboard (left-right and right-left). We define the left (right) part of the keyboard as all keys to the left (right) of the vertical line going through the keys z, h, and v (the space bar is split in the middle). Here, we assume that the left and right thumb are used for the left and right part of the keyboard, respectively. In a pre-experiment questionnaire, 84% of the participants reported to type with both hands (i.e., left and right thumb). An ANOVA revealed that there were significant differences in terms of the average typing speed for these four different key pair locations ($F(3, 336) = 110.395$, $p < 0.001$). Post hoc comparisons using the Tukey HSD test indicated that the mean typing speed for pairings of left-left (mean = 132 pixels per second (pps), SD = 61 pps), right-right (mean = 73 pps, SD = 35 pps), left-right (mean = 384 pps, SD = 211 pps) and right-left (mean = 505 pps, SD = 282 pps)

were significantly different (all $p < 0.001$). Only the typing speed between left-left and right-right was not significantly different ($p = 0.135$). By subtracting the baseline heat map per participant, we correct for these differences in typing speed. In addition, the convolutional neural network is capable of learning potential keyboard layout-based bias in typing speed.

At the beginning and at the end of the experiment participants typed six sentences in random order on their default keyboard and on our keyboard. A 2 (time) $\times$ 2 (keyboard) Aligned Ranked Transform (ART) ANOVA revealed that the typing speed was significantly higher ($F(1,216) = 53.460$, $p < 0.001$) at the end of the experiment (mean $=$ 1.380 characters per second (cps), SD $=$ 0.255 cps) than at the beginning of the experiment (mean $=$ 1.235 cps, SD $=$ 0.240 cps). We also found that typing speed was significantly higher ($F(1,216) = 5.571$, $p = 0.019$) for the default keyboard (mean $=$ 1.325 cps, SD $=$ 0.260 cps) than our keyboard (mean $=$ 1.290 cps, SD $=$ 0.235 cps). There was no interaction between keyboard and time ($F(1,216) = 1.909$, $p = 0.169$). The higher typing speed at the end of the experiment might be because participants already saw the sentences at the beginning of the experiment. We conclude that although participants used diverse keyboards before the experiment, they quickly became familiar with our keyboard.

Sensor data. In addition to the data recorded from the keyboard, we also recorded sensor data (i.e., linear acceleration, rate of rotation, and light intensity). In Figure 6.8 we show the valence, arousal, and dominance ratings in relation to these sensor measurements (we used moving averages to smooth the curves). For linear acceleration and the rate of rotation we calculated the magnitude $\sqrt{x^2 + y^2 + z^2}$. It is visible that valence and arousal increased with increasing linear acceleration (see Figure 6.8A), rate of rotation (see Figure 6.8B), and light intensity (see Figure 6.8C). Interestingly, dominance had only a small effect on the sensor measurements. Figure B.4 and Figure B.5 in the appendix reveal that happiness was higher the larger the linear acceleration and the rate of rotation were. On the other hand, sadness was decreasing with increasing linear acceleration and rate of rotation. Thus, we conclude that when participants were happy, they were more engaged in smartphone usage (e.g., more jittery and moving more) and remained calmer during sad episodes. Similarly, with increasing ambient light intensity, participants tended to be happier and less sad (see Figure B.6). Noteworthy, the stress level is reduced when ambient light intensity is increasing. Our final model did not employ data from the light sensor because it did not provide significant performance improvements.

Context data. The five most-used application were YouTube (14.2% of the total usage time), WhatsApp Messenger (13.1%), Instagram (8.7%), Google Chrome (8.2%), Netflix (2.6%), and Snapchat (1.9%). Accordingly, the top five application categories where participants spent most time were communication (28.7% of the total usage time), social networks (17.2%), video players (14.4%), games (8.4%), and entertain-

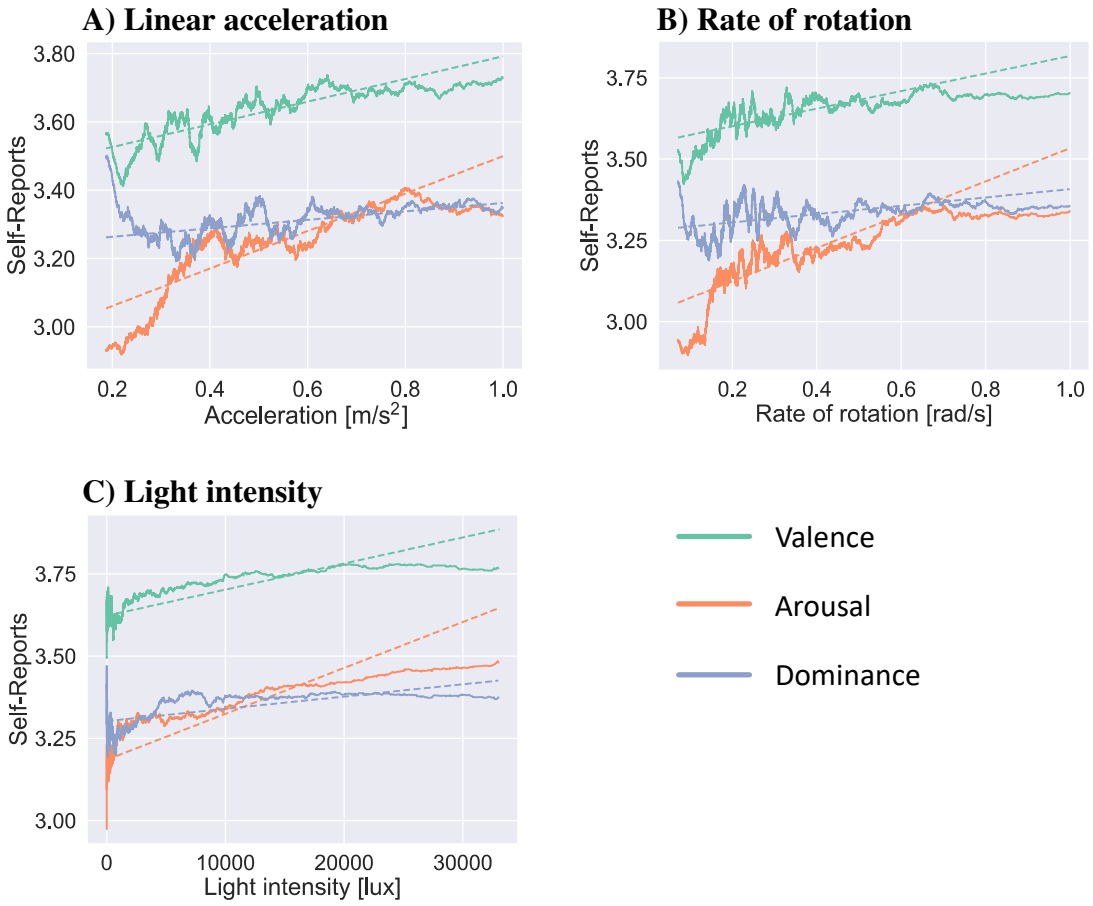


Figure 6.8: *Valence, arousal, and dominance in relation to the magnitude of linear acceleration (A), the magnitude of the rate of rotation (B), and the light intensity (C). The affective dimensions were encoded in the interval* $[1, 5]$*. The dashed regression lines show the linear trends in the data.*

ment (4.1%). Figure 6.9 shows the average valence, arousal, and dominance ratings for each application category. For each category, we considered all self-reports for which a corresponding application was used in a 30 seconds window before the self-report. The category 'events' is characterized by low valence and low dominance and high arousal. On the other hand, for the category 'sports' participants faced high valence and high dominance and low arousal. Figure B.3 in the appendix provides the same statistics for the basic emotions and stress level. Interestingly, applications belonging to the category 'music & audio' and the category 'lifestyle' provoked the highest levels of sadness and stress, respectively. Kruskal-Wallis tests revealed significant differences after Bonferroni correction ($\alpha = 0.005$) for valence, arousal, dominance, basic emotions, and stress in terms of the different application categories.

Affective states can also fluctuate during the day and during the week. In Figure 6.10

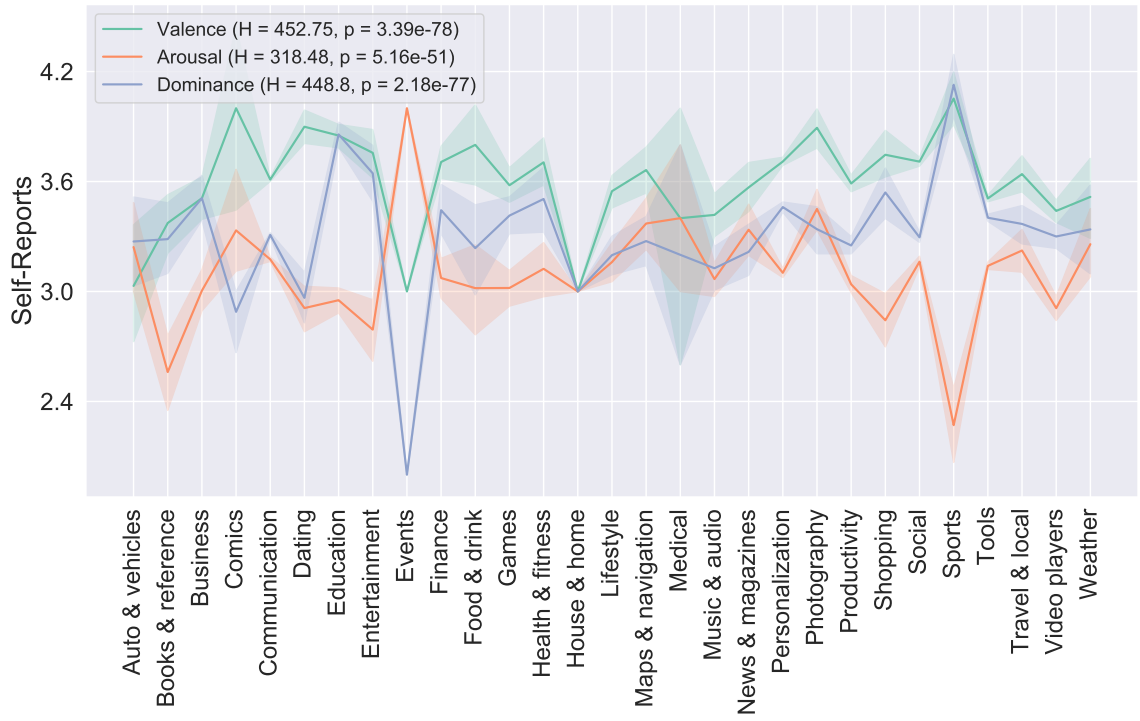


Figure 6.9: *Mean and 95% confidence interval (shaded area) of the reported valence, arousal, and dominance for different application categories. The affective dimensions were encoded in the interval* $[1, 5]$*. The legend discloses the results of Kruskal-Wallis tests to investigate whether there were significant differences in terms of valence, arousal, and dominance for the application categories.*

we see that valence, arousal, and dominance decreased around midnight and started increasing again around 7 a.m. Similar patterns are visible for the basic emotions and stress with peaks of anger, sadness, surprise, disgust, and stress in the night (see Figure B.1 in the appendix). Using Kruskal-Wallis tests we found significant differences after Bonferroni correction ($\alpha = 0.005$) for valence, arousal, dominance, all basic emotions (except anger and sadness), and stress in terms of the hour of the day. However, these results must be taken with a grain of salt because the number of self-reports was low during the night (see Figure 6.7B).

When considering the day of the week, valence and dominance culminated on Saturdays (see Figure B.2 in the appendix). On the other hand, the stress level was lowest at the weekends (i.e., Saturdays and Sundays). For dominance, happiness, sadness, and stress we found significant differences after Bonferroni correction ($\alpha = 0.005$) in terms of the weekdays. In our final model, we did not leverage context data as it did not provide significant performance improvements.

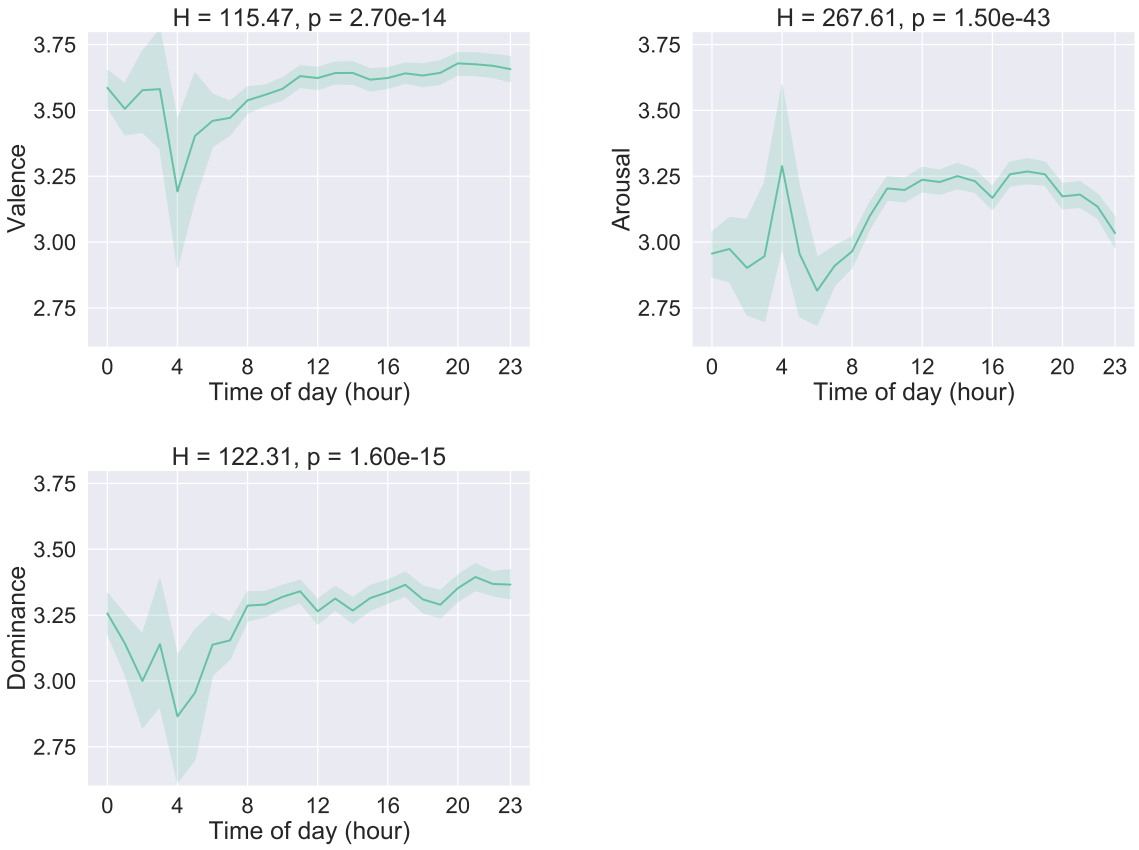


Figure 6.10: *Mean and 95% confidence interval (shaded area) of the reported valence, arousal, and dominance for the time of the day. The affective dimensions were encoded in the interval* $[1, 5]$*. The titles contain the results of Kruskal-Wallis tests to investigate whether there were significant differences in terms of valence, arousal, and dominance during the time of the day.*

6.3.3 Affective State Prediction

To remove noise and foster balanced classes, we simplified the valence, arousal, and dominance measures to three classes (low $\in [1, 2]$, medium $\in [3, 3]$, and high $\in [4, 5]$) of valence (3436, 8071, and 18576 self-reports), arousal (6903, 12572, 10608), and dominance (4870, 12066, 11221). Table 6.3 reveals the performance of our model. See Table B.1 in the appendix for additional metrics.

Classification performance. Using the combination of keystroke and sensor heat maps, for valence, arousal, and dominance, the values for micro-averaged AUC (0.83, 0.85, 0.84) are slightly higher than for macro-averaged AUC (0.78, 0.84, 0.82). When considering the percentage of the most frequent class as baseline (valence $= 62\%$, arousal $= 42\%$, and dominance $= 40\%$), the accuracy is well above the baseline

Table 6.3: *Performance for the prediction of three classes (low, medium, high) of valence, arousal, and dominance. AUC_{micro} and AUC_{macro} represent micro-averaged AUC and macro-averaged AUC, respectively. The chance level of accuracy and AUC is 0.33 and 0.5, respectively.*

Dimension	Heat Map	AUC_{micro}	AUC_{macro}	Accuracy
Valence	Keystrokes	0.82	0.76	66%
	Sensors	0.79	0.73	63%
	Combination	0.83	0.78	70%
Arousal	Keystrokes	0.81	0.80	63%
	Sensors	0.83	0.82	64%
	Combination	0.85	0.84	65%
Dominance	Keystrokes	0.82	0.79	67%
	Sensors	0.81	0.79	63%
	Combination	0.84	0.82	68%

for valence (70%), arousal (65%), and dominance (68%). Figure 6.11 shows the confusion matrices for valence, arousal, and dominance evaluated on the combination of the keystroke and sensor heat maps. Often neighboring classes are confused with each other. For all three dimensions, the high class was most often confused with the medium class and vice versa. For arousal (see Figure 6.11B) and dominance (see Figure 6.11C) the low class was often mispredicted as the medium class. In contrast, for valence (see Figure 6.11A) the low class was more often confused as the high class, which may be attributed to the class imbalance.

Heat map comparison. The keystroke heat maps perform slightly better than the sensor heat maps for valence (+0.03 AUC) and dominance (+0.01 AUC). In contrast, for arousal, the sensor heat maps outperform the keystroke heat maps (+0.02 AUC). The combination of the two types of heat maps provides only a marginal improvement in performance (up to 0.03 AUC).

6.3.4 Basic Emotion and Stress Prediction

Our model achieved a performance of 90% (0.77 AUC) for anger, 75% (0.81 AUC) for happiness, 93% (0.82 AUC) for sadness, 95% (0.86 AUC) for surprise, 93% (0.85 AUC) for fear, 97% (0.86 AUC) for disgust, and 82% (0.83 AUC) for stress. The differences between AUC and accuracy are due to class imbalance. Basic emotions can also be blended to form complex emotions (e.g., the combination of happiness and sadness result in melancholy) [Shoumy et al., 2020]. Table 6.4 presents the F_1-scores and the number of self-reports (in brackets) for the first-order complex emotions. We

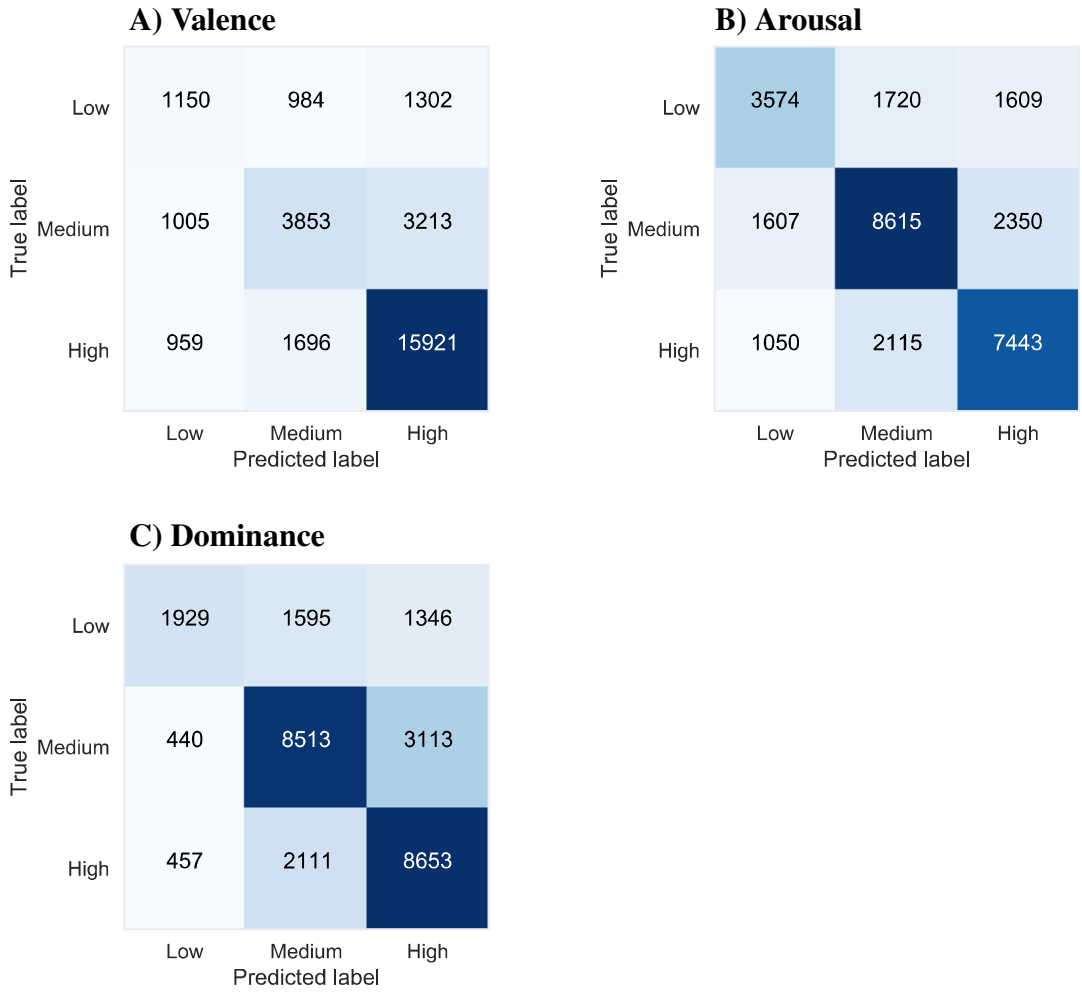


Figure 6.11: *Confusion matrices for the classification of three levels (low, medium, high) of A) valence, B) arousal, and C) dominance. The confusion matrices are calculated by predicting self-reports using the combination of keystroke and sensor heat maps.*

evaluated the performance by combining the predictions of the individual models pertaining to the respective basic emotions. Complex emotions formed by happiness were recognized well. Similarly, the combinations of stress and the basic emotions were identified accurately. Interestingly, the combination of stress and happiness (i.e., positive stress) occured most often (1563 times) and was recognized well (F_1-score of 0.80). Altogether, we can conclude that our model is also predictive for basic emotions and stress and may even be predictive for complex emotions.

Table 6.4: *F_1-scores for complex emotions formed from two basic emotions and stress. We treat the presence of the complex emotion as the positive class. The number of self-reports for each complex emotion is given in brackets.*

	Anger	Happiness	Sadness	Surprise	Fear	Disgust	Stress
Anger	–	0.76 (153)	0.30 (384)	0.24 (101)	0.28 (222)	0.19 (126)	0.46 (429)
Happiness	0.76 (153)	–	0.78 (402)	0.76 (394)	0.77 (518)	0.76 (104)	0.80 (1563)
Sadness	0.30 (384)	0.78 (402)	–	0.31 (89)	0.37 (418)	0.31 (119)	0.49 (561)
Surprise	0.24 (101)	0.76 (394)	0.31 (89)	–	0.31 (98)	0.23 (53)	0.48 (203)
Fear	0.28 (222)	0.77 (518)	0.37 (418)	0.31 (98)	–	0.28 (105)	0.47 (966)
Disgust	0.19 (126)	0.76 (104)	0.31 (119)	0.23 (53)	0.28 (105)	–	0.46 (214)
Stress	0.46 (429)	0.80 (1563)	0.49 (561)	0.48 (203)	0.47 (966)	0.46 (214)	–

6.3.5 Window Size Analysis

The results presented in Table 6.3 are based on keystroke heat maps extracted from 80 characters and sensor heat maps extracted from 30 seconds (i.e., 3000 sensor values). On average participants typed 74 characters (SD = 43 characters) in the 30 seconds window before filling in the self-report. Thus, the two types of windows for extracting the keystroke and sensor heat maps are a close match Nevertheless, considering longer periods can be beneficial for the classification performance, because the model has more data available. As such, we evaluated our model on heat maps extracted on larger windows ranging from 2 minutes to 30 minutes (the minimum time between self-reports was 30 minutes). To analyze different window sizes, we relaxed the constraint of a fixed number of characters (i.e., 80 characters) and sensor measurements (i.e., 3000 samples). Thus, the heat maps contain a different number of characters and sensor measurements depending on the number and the duration of sessions in the corresponding window. Figure 6.12A shows the macro-averaged AUC for valence, arousal, and dominance for the different window sizes. Peak performance is reached with a window size of 5 minutes for valence (0.80 AUC), arousal (0.86 AUC), and dominance (0.83 AUC). Further increasing window sizes leads to a substantial drop in the performance for all three dimensions. Overall, performance improvements are only marginal for all three dimensions (up to 0.02 AUC).

6.3.6 Personalization

Affective states can be individual and can reflect idiosyncrasies in users. While there may be similar typing and sensor patterns between users characterizing similar affective states, leveraging user-specific data can improve the performance of the model. To investigate the extent of performance gain for a participant with N filled in self-reports, we used the first n self-reports to fine-tune the whole model using 5

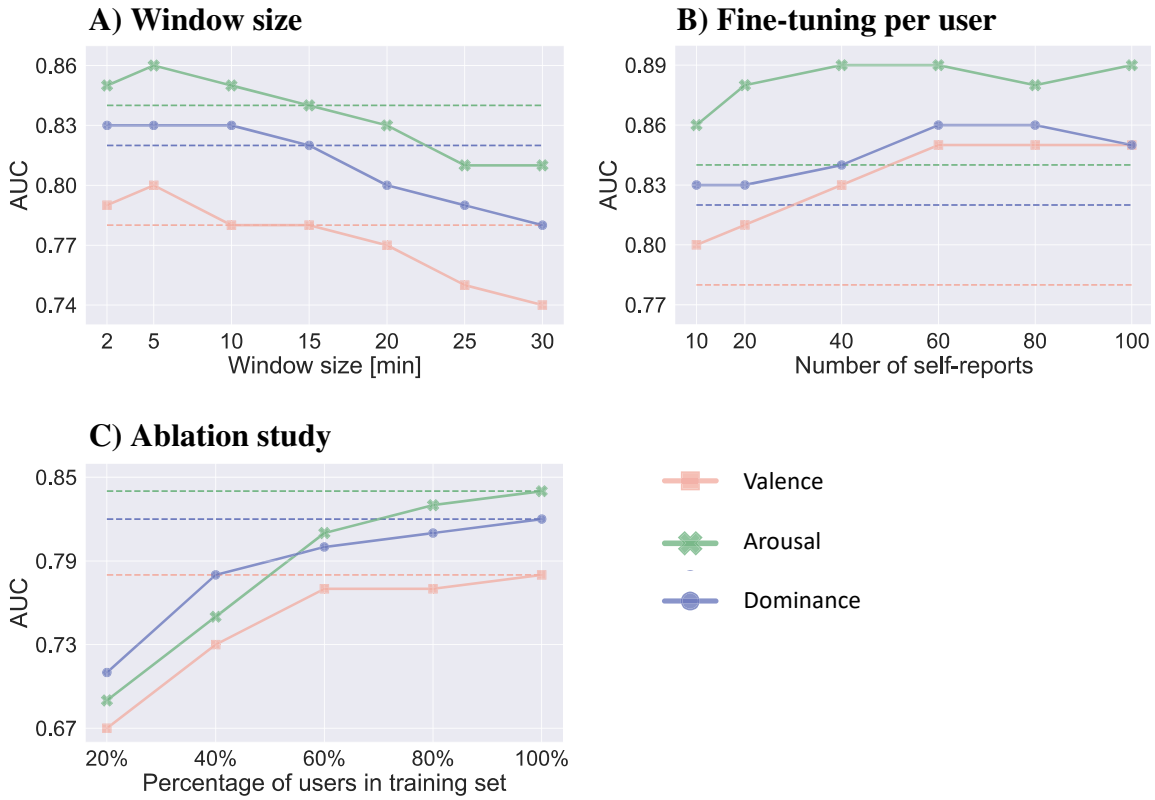


Figure 6.12: *Macro-averaged AUC for the classification of three levels (low, medium, high) of valence, arousal, and dominance using A) different window sizes for the heat map extraction, B) fine-tuning the network per participant on varying number of self-reports, and C) different number of participants in the training set. The dashed lines represent the baseline performance (see Table 6.3).*

epochs and predicted then the $N - n$ remaining self-reports. Figure 6.12B reveals the macro-averaged AUC in terms of n (i.e., the number of self-reports used to fine-tune the model). Fine-tuning on only 10 self-reports provides already a slight performance improvement (up to 0.02 AUC). The performance improvement plateaus at around 40 to 60 self-reports used for fine-tuning. The performance improvements are substantial for valence (+0.07 AUC) and arousal (+0.05 AUC), reaching a performance of 0.85 AUC and 0.89 AUC, respectively. For dominance, the improvements are smaller (+0.03 AUC).

6.3.7 Ablation Study

A model can only be as good as the data that supports it. If the data (i.e., the heat maps) show clear patterns, we can achieve a well-performing model with only a little

amount of data. On the other hand, if the data is noisy, a much larger dataset is needed to achieve the same performance. In our experiment, we collected a homogeneous dataset consisting of mostly bachelor and master students around the age of 23. Thus, we hypothesize that typing and smartphone usage behavior were similar among participants and less training data is needed to achieve a good performance for the classification of affective states. To test our hypothesis, we conducted an ablation study by training the model on data from a subset of the participants. To accomplish this, we selected a percentage of participants at random in each of the 82 training sets of the leave-one-user-out cross-validation. Figure 6.12C shows the macro-averaged AUC for different percentages of users in the training data. Performance plateaus at around 60% (49) of participants for valence (0.77 AUC) and around 80% (66) of participants for arousal (0.84 AUC) and dominance (0.82 AUC). Thus, a subset of the users (i.e., between 60% and 80%) is enough to achieve a performance close to the performance reached when using data from all the users. By linear extrapolation, we can roughly predict that with the double amount of participants (i.e., 164 participants), we could come close to a performance of around 0.83 AUC for valence, 0.89 AUC for arousal, and 0.87 AUC for dominance.

6.3.8 Runtime Analysis

We conducted a runtime analysis of the different parts of our model. Our computing environment consisted of an Intel® Xeon® CPU E5-2630 v4 @ 2.20GHz and an NVIDIA GeForce® GTX 1080 Ti. The prediction of a new data point consisted of extracting keystroke heat maps (mean = 0.84 s, SD = 0.03 s) and sensor heat maps (mean = 0.23 s, SD = 0.18 s), followed by the convolutional neural network and the fully connected layer for the classification of the affective states (mean = 0.0036 s, SD = 0.004 s). Summing up these values leads to a prediction time of 1.07 seconds if considering both types of heat maps. If only the keystroke heat maps and sensor heat maps are used, the prediction time amounts to 0.84 seconds and 0.23 seconds, respectively. The higher runtime for creating the keystroke heat maps compared to the sensor heat maps is due to the preprocessing (i.e., mapping touch positions to keys, calculating the metrics between the key pairs, and sanity checks).

6.4 Discussion

In this chapter, we presented a model that can be used on mobile devices for predicting valence, arousal, and dominance, the basic emotions, and stress. We believe that the ability to run our model on smartphones can improve user experience, provide ubiquitous access to affective state predictions, and can be beneficial for security and privacy.

The predictions of our model were based on heat maps generated from keystroke data and sensor data collected during smartphone usage in real-world environments. We found that both types of heat maps are capable of accurately predicting valence, arousal, and dominance. In particular, the sensor heat maps showed the best performance for predicting arousal (0.83 AUC), while the keystroke heat maps were most predictive for valence (0.82 AUC) and dominance (0.82 AUC). These results are in line with the findings by Olsen and Torresen [2016], reporting that accelerometer data is more predictive for arousal than valence. In Figure 6.8 we saw only a slight effect of the magnitude of acceleration and the rate of rotation on dominance. Potentially, the model benefited from having x-axis, y-axis, and z-axis data separately available and from the combination of both types of sensors in the heat map.

The keystroke heat maps provide an intuitive and compact visualization of typing patterns compared to the heat maps presented in Chapter 5. On the other hand, the recording of sensor data is less privacy-invasive. Sensor data is also less prone to bias than typing data (i.e., users might be more aware of their typing behavior than of their smartphone holding behavior). In addition, the runtime to generate sensor heat maps is substantially lower than that of keystroke heat maps. In conclusion, we suggest the sensor heat maps as most appropriate for the prediction of affective states in real-world applications.

We used 80 characters and 30 seconds of accelerometer and gyroscope data to generate the keystroke and sensor heat maps, respectively. In practice, 30 seconds of sensor data can be stored continuously in the background of the smartphone until the user has typed 80 characters. If only sensor data is used for the prediction, the restriction does not apply anymore and predictions are possible more often (i.e., also when users did not type). For larger window sizes it takes longer until a prediction is possible. We showed that peak performance is reached with a window size of 5 minutes (+0.02 AUC). A potential explanation for the performance improvement is that with larger window sizes the model can implicitly gauge the total time spent on the smartphone from the sparseness of the heat maps (i.e., a sparser heat map implies a less active user). On the other hand, if the window size becomes too large, the heat maps become too dense and noisy which degrades the performance.

We also showed that fine-tuning our model per participant can substantially improve the performance for valence (+0.07 AUC), arousal (+0.05 AUC), and dominance (+0.03 AUC). Peak performance for personalizing the model was reached using the first 40 to 60 self-reports of the participants. After the start of the experiment, it took some time until the participants got used to filling in the self-reports (i.e., the variance tended to be larger for the first self-reports). Thus, 40 to 60 self-reports were necessary for fine-tuning the model to learn the stable self-report pattern of the participants and reaching peak performance. A reason for this performance improvement is that the network can learn keystroke and sensor patterns typical for

a specific participant. Moreover, for participants that reported one class (e.g., high valence) more often than other classes, the model can shift towards predicting this class with a higher probability. Other researchers reported performance improvements for personalized models of up to 6.3% for predicting valence [Dai et al., 2016] and 17.6% for predicting arousal [Ghandeharioun et al., 2019].

For the prediction of the affective states, we employed an architecture that can be used on mobile devices. The ability to run our model on a smartphone can improve user experience, provides anytime and anywhere access to affective state predictions, and can be beneficial for security, privacy, and energy consumption. In addition, our model is easier to tune compared to the autoencoder architecture used in Chapter 5. We also experimented with an LSTM architecture to model the time series of self-reports and the time series of the heat maps in the windows. We did not find significant performance improvements but substantial higher memory consumption and higher runtime. Similarly, joint optimization of the labels (i.e., valence, arousal, and dominance as well as the basic emotions) substantially degraded performance.

We also experimented with enriching our model by context data (i.e., mean ambient light, application type, daytime, and weekday) extracted over the same windows as the heat maps were extracted. We encoded the daytime and weekday as integers ranging from 0 to 23 and 0 (Monday) to 6 (Sunday), respectively. We further encoded the duration in seconds of the used applications as a 29-dimensional vector capturing 29 different application categories provided by the Google Play Store (e.g., communication, entertainment, and events). From the light sensor, we considered only the mean value because ambient light is typically constant over short periods (e.g., 30 seconds). In the context of affective state prediction, the light sensor can be used to gauge the location of the phone. For example, the values are low in the night under dimmed light, higher under normal light during the day, and highest outdoors. Although ambient light correlates to daytime, it still carries additional information which daytime cannot provide (e.g., using the phone in the evening in a dark versus bright environment). Overall, we encoded the context data using a 32-dimensional vector, which we passed through a fully connected network and subsequently concatenated with the output of the global average pooling layer of the heat maps. Although ambient light is connected to emotions and moods [Canazei and Weiss, 2013] and our analysis of the context data presented in Section 6.3.2 was promising, performance did not improve significantly. We believe that the heat maps provide similar or richer information than the context data. In addition, typing-based applications constitute the largest part which decreased the variance of the application types (i.e., prevalence of communication applications).

A direct comparison of the performance of our model is difficult due to differences in the measurement of affective states and experimental setups. Olsen and Torresen [2016] reported slightly higher accuracy for the prediction of arousal ($+10\%$) but

lower accuracy for valence (−19%). In contrast to our work, they captured accelerometer data during sequences of walking from only 10 participants. In comparison to Ruensuk et al. [2019], our model performed similar for valence (+1%) and arousal (−7%), although these authors predicted only two levels of valence and arousal. Contrasted to the personalized model of Ghosh et al. [2019b], our generalized model performed similar for happiness (−0.03 AUC), sadness (−0.02 AUC), and stress (−0.01 AU). In addition to most other works, we also considered dominance, which we believe is an important dimension of affective states.

In comparison to our model evaluated on laboratory data presented in Chapter 5, the performance in terms of macro-averaged AUC was superior for arousal (+0.04 AUC) and dominance (+0.02 AUC) but inferior for valence (−0.05 AUC). The inferior performance for valence may be attributed to class imbalance (11% low, 27% medium, 62% high) or the data collection in the wild. With regard to the basic emotions, our model presented in this chapter performed better for surprise (+0.1 AUC) and stress (+0.03 AUC) but was inferior for anger (−0.07 AUC), happiness (−0.07 AUC), and sadness (−0.05 AUC). Laboratory experiments provide more control while in the wild experiments provide more ecological validity (i.e., less control) and offer the possibility of collecting larger datasets.

We acknowledge potential limitations of the approach presented in this chapter. We analyzed the runtime of our model on a computer. On a mobile device the runtime of generating the heat maps and the inference time of the network might be slightly higher. To keep runtime low, the model could be deployed on a server. Another limitation is the number of data required until a prediction can be made. In our experiment, we ensured that we have enough sensor and keystroke data available by unlocking a self-report when the user typed at least 80 characters and used the smartphone for at least 30 seconds. In practice, using only the sensor heat maps for the prediction of affective states relaxes the constraint of typing 80 characters while at the same time providing a similar performance. In addition, by allowing the participants to fill in self-reports only every 30 minutes, we could have missed finer-grained changes in affective states. Finally, due to the requirement of having typed at least 80 characters for unlocking a self-report, the windows used for creating the sensor heat maps always contained keystrokes. As such, predicting affective states using sensor heat maps during periods with no keyboard input (e.g., watching videos on YouTube) requires further research.

6.5 Conclusion

In this chapter, we presented a pipeline for predicting affective states, basic emotions, and stress based on two-dimensional heat maps generated from 80 keystrokes and 30 seconds of sensor data. We evaluated our pipeline with data collected in an

experiment in the wild with 82 participants over 10 weeks. Using leave-one-user-out cross-validation, we demonstrated that our model can accurately predict three levels (low, medium, high) of valence (up to 0.83 AUC), arousal (up to 0.85 AUC), and dominance (up to 0.84 AUC). By fine-tuning the network per participant we achieved substantial performance improvements for valence (+0.07 AUC), arousal (+0.05 AUC), and dominance (+0.03 AUC). We also showed that 60% to 80% of the training data is enough to achieve a similar performance. When doubling the number of users for training the network, we expect a performance improvement for valence (0.83 AUC), arousal (0.89 AUC), and dominance (0.87 AUC). We further presented results for the prediction of two levels (present vs. absent) of stress (0.83 AUC) and the basic emotions anger (0.77 AUC), happiness (0.81 AUC), sadness (0.82 AUC), surprise (0.86 AUC), fear (0.85 AUC), and disgust (0.86 AUC). Finally, we showed that we could achieve a similar performance using sensor heat maps alone which is beneficial in terms of privacy and runtime (0.23 seconds vs. 0.84 seconds).

In comparison to the typing heat maps presented in Chapter 5, the keystroke heat maps provide an intuitive and compact visualization of the typing characteristics. In addition, the keystroke heat maps allowed us to investigate the distribution of keystroke pairs in relation to the measured affective states (e.g., more frequent occurrence of keystroke pairs with a backspace when experiencing negative emotions). By using sensor heat maps instead of raw sensor data, we are taking into account the relationship between acceleration and rotation and provide a less privacy-invasive way for affective state prediction compared to typing heat maps. In comparison to Chapter 5, we evaluated our model on data collected in real-world environments, demonstrating the applicability of affective state prediction beyond laboratory settings.

CHAPTER 7

Visualization of Affective States

Recognizing the affective state of users is one significant part of an affect-aware system. Another important component is the visualization of the recognized affective states. Such visualizations can be used to communicate affective information to others and for making users aware of their affective states. Despite the variety of benefits, graphical user interfaces (GUIs) often focus on the purely objective interaction with the user and do not visualize affective states [Cernea et al., 2013]. In this chapter, we therefore visualize the affective states by extending existing approaches and optimizing them for intuitiveness and precision. To achieve this objective, we developed two application-specific GUI widgets, which visualize the user's affective state with glyph-based methods [Borgo et al., 2013] in two different ways. The first widget, called the intuitive widget, focuses on users that need an assessment of the current affective state at first sight, e.g., teachers monitoring the widgets of their students. The second widget, called the precise widget, focuses on users that want to track the individual dimensions of the affective states over time to enable a more detailed analysis of affective states. The design space of the glyphs can be seen in Figure 7.1. The widgets are compared to a baseline visualization in a user study with 644 participants for testing them on understandability and intuitiveness. The study shows that, particularly in terms of understandability, our widgets are able to outperform the baseline significantly. Furthermore, the intuitive widget is intuitive and understandable without further explanation of the meaning of the visualization.

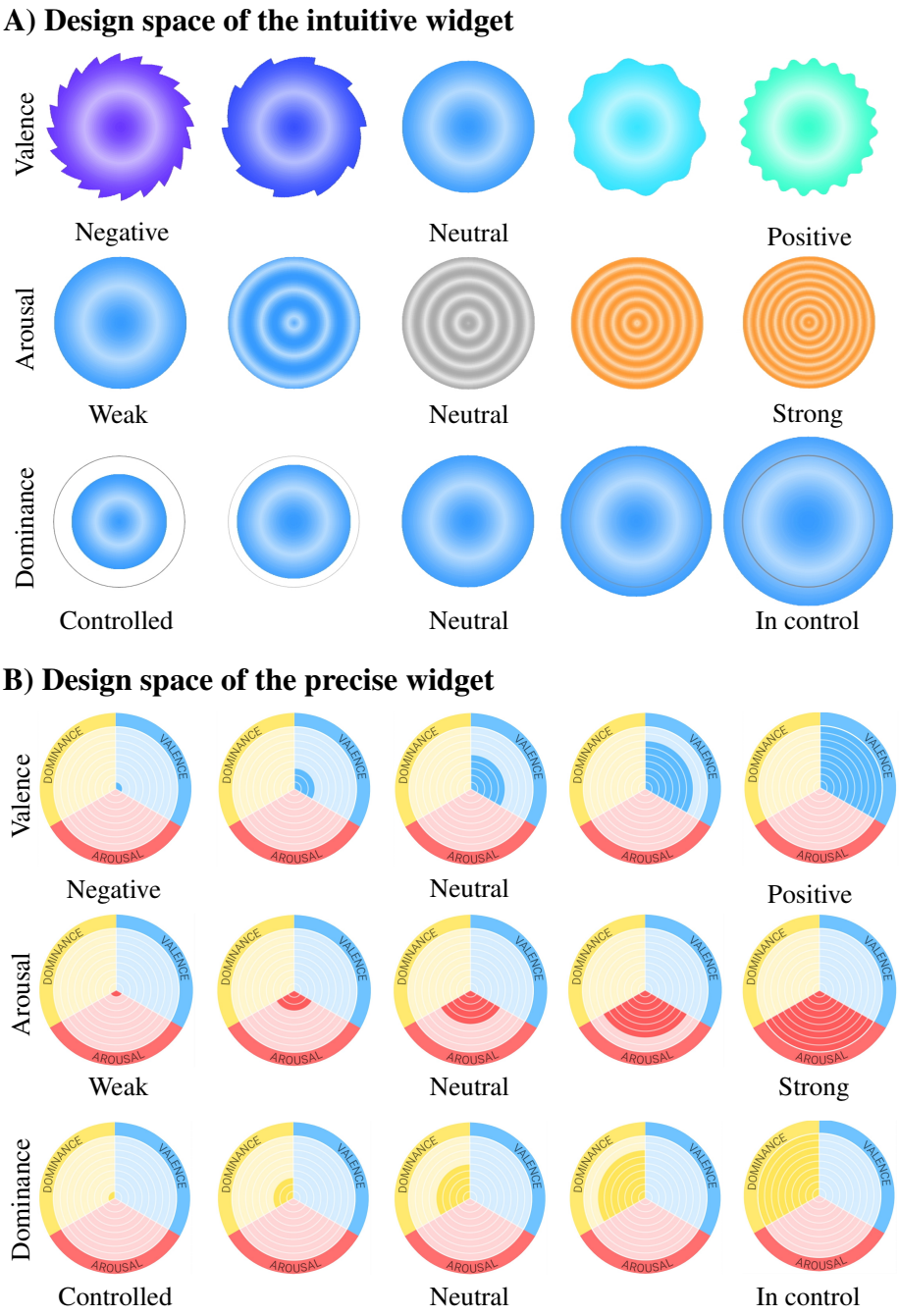


Figure 7.1: *We developed two graphical user interface widgets that visualize affective states in terms of valence, arousal and dominance to users. Our first widget in A) focuses on the intuitiveness of understanding different combinations of valence and arousal. Therefore, they are mapped to color. Our second widget in B) focuses on the precise representation of the levels of the three dimensions using radial bar charts.*

7.1 Method

First, we describe three requirements for the design space of our widgets. We then outline the intuitive widget (fast assessment of affective states) and the precise widget (accurate tracking of affective states over time) in relation to these requirements.

7.1.1 Requirements

The requirements for the visualization are application-dependent. In this chapter, we consider two possible use cases. First, we focus on users that are mainly interested in a fast assessment of the current affective state, e.g., teachers that are monitoring the affective states of their students. Such users may want to have an overview of the affective states at first sight. Second, we focus on users that are interested in precise measurements over time, e.g., a user that is interested in adjusting daily routines based on potential stress causes. Such a user may want a more detailed analysis of the affective dimensions over time in order to draw useful conclusions. From those two use cases, we derive the following requirements.

R1. We aim for a fast assessment of the current affective state. The state should be identifiable at first sight (i.e., intuitive and preattentive [Ware, 2020]). Past affective states are less important.

R2. To track current trends over time, we aim for exact values and an intuitive reading of time. Furthermore, the visualization should keep the dimensions easily separable for clarity.

R3. In both use cases, a core requirement is a small space consumption. Further, the glyphs should be orientation-independent, compact, and transparent, such that users can place the widget wherever it interferes least with other activities on the screen. A circular shape meets these requirements and is therefore used for both widgets.

7.1.2 The Intuitive Widget

The widget that focuses on the first use case from Section 7.1.1 should meet R1 and R3. Since it is more self-explanatory than the second widget, we call it the intuitive widget. For its design, we extended the idea by Cernea et al. [2015] and improved the visualization for the use on small screens, see Figure 7.2 for an overview.

Valence. Because of the similarity of spikes and waves for small widget sizes, we replaced the spikes by a curve that resembles a saw blade because it maintains a sharp appearance that is better distinguishable from the waves (see Figure 7.1A).

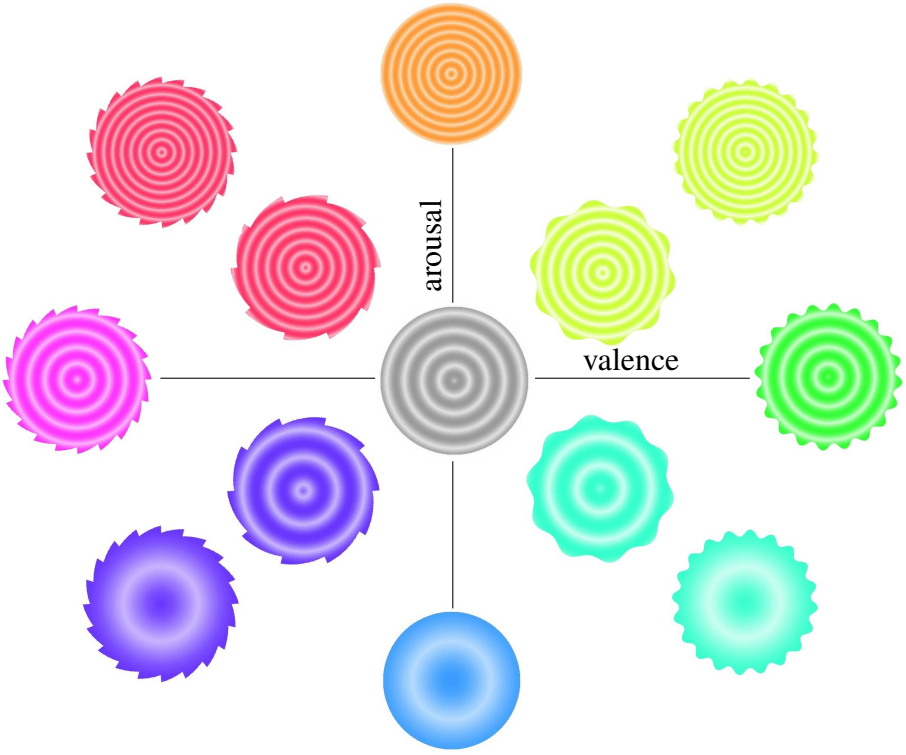


Figure 7.2: *As in the literature [Ståhl et al., 2005], valence and arousal are mapped to color. Valence increases from left to right, arousal increases from bottom to top.*

Dominance. Dominance is incorporated by smoothly adapting the radius of the base circle proportional to the change in dominance. In addition, a thin gray reference circle denoting medium dominance is constantly displayed (see Figure 7.1A).

Arousal. Since pulsation affects the widget size—which would interfere with dominance—we used concentric waves flowing with constant speed from the widget center to the border (see Figure 7.1A). The number of waves increases with increasing arousal.

Color. The widget is color-coded to enable affective state identification even at peripheral vision. The color is determined by the level of valence and arousal based on a unification [Ståhl et al., 2005] of Itten's color wheel [Itten, 1970] and Russell's Circumplex Model of Affect [Russell, 1980]. A measurement of valence and arousal can be seen as a two-dimensional point living in the discrete space $\{1,\ldots,9\}\times\{1,\ldots,9\}$. Given a certain level of valence v and arousal a, we map the corresponding point to its polar angle $\phi = \text{atan2}(a,v)$. The angle is then mapped to hue in HSV color space, similar to Ståhl et al. [2005]. For the neutral point $(v = 5, a = 5)$ the color is mapped to gray. Thereby, saturation and value are set

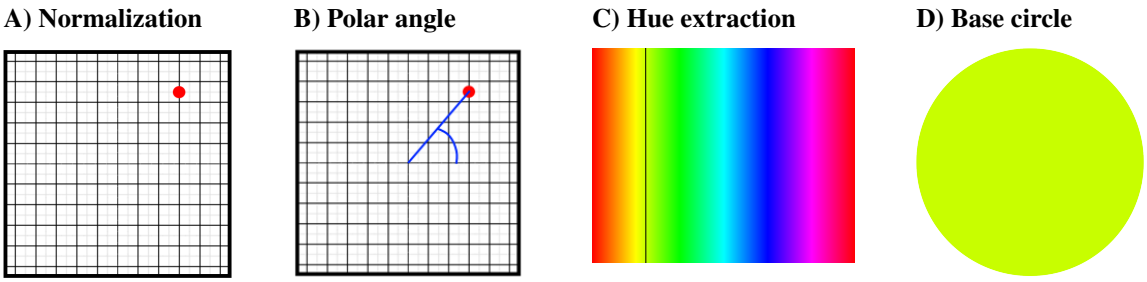


Figure 7.3: *Color mapping example. The affective space is centered around the origin and is normalized (A). Then, the polar angle of the current point of interest (red dot) is calculated (B). From the polar angle, the hue is extracted (C). The base circle is filled using the color having this hue, setting the saturation and the value to one in HSV color format (D).*

to one (see Figure 7.3). For colors like yellow or green, we adapt the saturation to enhance contrast.

7.1.3 The Precise Widget

The precise widget focuses on the second use case from Section 7.1.1. It should meet R2 and R3. A simple bar chart fulfills R2. For R3, we visualize the bar chart as a circle such that three sectors are formed, each of which corresponds to one affective dimension (see Figure 7.1B). Thus, all dimensions are adjacent, which enables direct comparability among dimensions. The number of filled parts per segment denotes the current level of the corresponding affective dimension. This representation is unambiguous and lossless with respect to the underlying space of valence, arousal, and dominance since each affective state is displayed precisely. However, this is less intuitive and demands more time for interpretation because the user has to parse first which affective state is represented by combining the dimensions.

The time dimension is incorporated by dividing each sector into equal pieces (see Figure 7.4). The pieces in a sector show the development of a dimension over the last few affective states. This allows the user to track current trends in each dimension.

7.2 User Study

To evaluate our intuitive and precise widgets we conducted a user study. First, we outline the study setup showing the task design which consisted of images and sentences. Then, we evaluate the intuitiveness of the first (intuitive) widget, followed

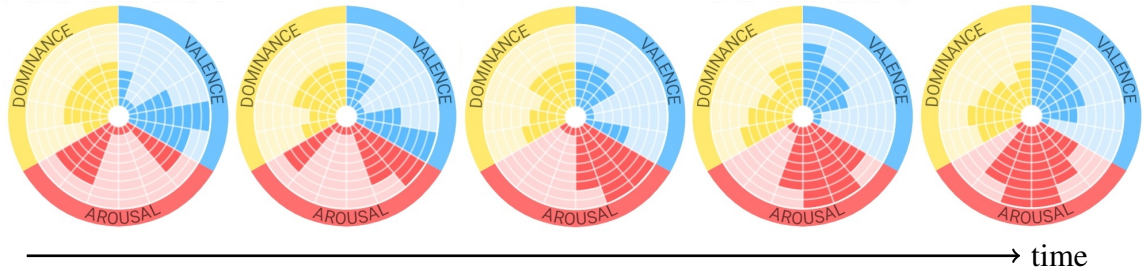


Figure 7.4: *The precise widget shown in five consecutive time steps from left to right. The new affective state is put in the first segment in clockwise direction, and all other affective states are shifted in this direction whereby the oldest one is not shown anymore.*

by comparing both widgets to a baseline widget. Finally, we present results on the understandability of our widgets.

7.2.1 Study Setup

The user study was carried out using an online survey platform. We recruited 644 participants between the ages of 18 and 52 (mean = 23.8 years, standard deviation SD = 4.2 years), mostly undergraduate students. The study was split into three parts. The first part assessed the intuitiveness of the first widget, while the other two parts compared our two widgets to the baseline. As baseline, we used the visualization by Cernea et al. [2013] (see Section 2.3). Since their approach does not include dominance, we extended it by a third bar using a different sequential color map (ochre). The last part focused on demographical questions and subjective ratings of understandability. We enclosed the experimental setup in Appendix C.

Task design. In Part 1 and Part 2, we presented six images and six short sentences. Each image showed a person, expressing a specific affective state. Alternatively, each sentence consisted of a statement about a person and contained one affective keyword. For each image and each sentence, the participants were shown three different visualizations of each widget visualizing different affective states. The participants were asked to choose the correct one. For the sentences, the keywords were mapped to valence, arousal, and dominance using the lexicon by Mohammad [2018]. While there are databases for images such as the International Affective Picture System [Lang et al., 2008] that trigger certain affective states, there is no database for images that expresses affective states analogously to the lexicon used for the sentences. Thus, we mapped affective keywords to the images instead, and then tested and improved the choice of images as well as sentences in a pilot study ($n = 10$). To obtain affective states, the keywords were mapped to valence, arousal, and dominance using the same lexicon as used for the sentences.

7.2.2 Intuitiveness of Widget 1

Hypotheses. Since the color mapping was shown to be intuitive [Ståhl et al., 2005], we expect that the design of our widget is intuitive.

Experimental setup. The participants were shown the images described above, and three different states of the intuitive widget. For this part, no explanation about the widget nor any background information about affective states was provided to the participants.

Results. 54.9% of the responses were correct, which is above random level (33%). Furthermore, the state of the widget, which differed the most from the correct state in terms of valence, was often ruled out by the participants. A more detailed analysis of the results showed that in 82.9% of the cases either the correct version or the most similar version in terms of valence was chosen. Hence, the visualization of valence seems self-explanatory (64.0% for arousal, 65.9% for dominance).

7.2.3 Baseline Comparison

Research question. How do our two widgets compare to the baseline in terms of performance in the tasks presented above?

Experimental setup. In this part, background information about affective states and a detailed explanation about the widgets were provided. Afterwards, both images and sentences were presented to the participants. In order to ensure comparability, the three states shown per widget always encoded the same affective state across the widgets. The order of the appearance of the widgets and the order of the states was randomized.

Results. In the sentence-based part the average number of correct answers was 86.6% (SD = 18%) for the intuitive widget, 83.5% (SD = 19%) for the precise widget, and 82.8% (SD = 20%) for the baseline. For the image-based part, the average number of correct answers was 71% (SD = 18%) for the intuitive widget, 76.2% (SD = 18%) for the precise widget, and 70.6% (SD = 17%) for the baseline. Our widgets provide a small but statistically not significant improvement over the baseline.

7.2.4 Questionnaire

In the last part of the study, the participants were asked to rate the understandability of the widgets on a 5-point scale. The average ratings were 3.6 (SD = 1.0) for the intuitive widget, 4.05 (SD = 1.0) for the precise widget, and 2.45 (SD = 1.25) for the baseline. An independent Welch's t-test showed that the differences of the means are pairwise significant ($t = -7.8, p = 1.62 \cdot 10^{-14}$ for our two widgets,

$t = 17.86, p = 8.09 \cdot 10^{-64}$ for the intuitive widget and the baseline). Hence, the precise widget is the most understandable, followed by the intuitive widget and the baseline. In addition, the participants were asked to indicate on a 5-point scale how important each dimension was when solving the tasks. The results showed a mean score of 4.5 (SD = 0.7) for valence, 3.95 (SD = 0.9) for arousal, and 2.9 (SD = 1.15) for dominance. An independent Welch's t-test showed a statistically significant difference of the means ($t = 12.74, p = 5.1 \cdot 10^{-35}$ for valence and arousal, $t = 18.13, p = 3.0 \cdot 10^{-65}$ for arousal and dominance).

7.2.5 Discussion

At the end of the study, participants had the opportunity to leave comments. It was repeatedly pointed out that two images were ambiguous, and that the participants struggled in guessing the affective state of the person in the image. For one of those two images the number of correct answers was in fact random.

7.3 Conclusions

Measuring the affective state of users and visualizing it to the users themselves can have various benefits such as increasing the learning experience [Grawemeyer et al., 2015] or improving the user's well-being [Lane et al., 2012]. However, GUIs seldom make use of those benefits. We developed two application-dependent GUI widgets that provide affective feedback. The intuitive widget focuses on an intuitive and fast assessment of the current affective state. The precise widget concentrates on an exact, clear, and time-dependent visualization. We tested the widgets on intuitiveness and understandability, and compared them with a baseline in a user study with 644 participants. The results showed that the intuitive widget is indeed self-explanatory, especially the valence dimension. In terms of subjective understandability ratings, both widgets outperformed the baseline widget. The widgets were designed to be compact and transparent such that they interfere as little as possible with other activities on the user's screen.

CHAPTER 8

Conclusion

In this thesis, we investigated different data modalities to predict affective states. We evaluated the different data modalities in terms of the protection of privacy and the applicability in real-world environments. These two factors are important to increase the users' acceptance of novel affective technologies and enable the embedding of affective components in real-world applications. The technical foundation is based on data-driven concepts. We employed novel data representations and features extracted from the data and used machine learning to predict affective states enabling innovative applications in different fields (e.g., psychology and education).

In the following, we review the principal contributions of the thesis, exemplify the limitations of our work, and outline some potential future work.

8.1 Principal Contributions

In Chapter 3, we leveraged video data to predict affective states from recordings of external cameras and built-in cameras of mobile devices. From video recordings, we extracted novel features from user movement, eyes, and face. Using data from a laboratory experiment with 88 participants, we demonstrated that our Random Forest classifier can accurately predict two levels (low and high) of valence (up to 0.80 AUC) and arousal (up to 0.73 AUC) while participants were solving math tasks (active part) and were looking at pictures (passive part). In general, video recordings pose a strain on privacy due to the disclosure of the identity of the user and other people in the vicinity of the user. On the other hand, usability in real-world environments depends on the used cameras. Built-in cameras of mobile devices are usually preferable in terms of unobtrusiveness and costs. Typically, recordings from built-in cameras suffer

from only partially visible faces due to the angle of view of the camera. Thus, we also developed a novel cheap and unobtrusive mirror construction and neural inpainting pipeline to improve the visibility of the user's face on cameras of mobile devices. Our setup improved confidence in facial landmark detection by up to 88% and provided comparable performance to external cameras in terms of affective state prediction.

In Chapter 4, we employed handwriting data recorded from a stylus and biosensors measuring skin conductance, heart rate, and skin temperature to predict affective states. In comparison to video data, both biosensor and handwriting data are less privacy-invasive which fosters user acceptance. In our model, we fused different existing approaches with novel features extracted from biosensor data and handwriting data. We showed that our Random Forest classifier is capable of accurately predicting (0.88 AUC) five regions in the valence and arousal space while participants solved math tasks on a tablet. Interestingly, using only data acquired by the stylus provided a comparable performance (0.84 AUC). Thus, we conclude that stylus data is a viable alternative to biosensors in light of the higher usability in real-world environments. In addition, we also investigated if our model can generalize from solving math tasks (active) to looking at pictures (passive). We reported a satisfying performance of 0.68 AUC considering the very different domains.

In Chapter 5, we presented a semi-supervised classification pipeline based on novel pressure and speed of typing heat maps generated from touch input from the smartphone's on-screen keyboard. In comparison to raw data, these two-dimensional heat maps also consider the spatial distribution of the data. To take advantage of the vast amount of unlabeled touch data, we employed a variational autoencoder to learn a low-dimensional representation of the heat maps. We then used fully connected layers on this low-dimensional representation for the classification of affective states. To evaluate our model, we conducted a laboratory experiment with 70 participants engaged in a chat conversation. We demonstrated that our model can accurately predict three levels (low, medium, high) of valence (up to 0.84 AUC), arousal (up to 0.82 AUC), and dominance (up to 0.82 AUC). In addition, our model can also predict two levels (present vs. absent) of anger (0.84 AUC), happiness (0.88 AUC), sadness (0.87 AUC), surprise (0.76 AUC), and stress (0.80 AUC). Our lightweight and less invasive model can be used on different types of mobile devices and has the potential to enable emotion-aware chat conversations.

In Chapter 6, we introduced two-dimensional heat maps which encode different typing metrics of keystroke pairs. The heat maps provide an intuitive visualization of typing characteristics. Using a MobileNetV2 network architecture, we extracted meaningful features from the heat maps and used a classification network for the prediction of affective states. We evaluated our model based on data collected in the wild in a large-scale user study with 82 participants over 10 weeks. We showed that our model can accurately predict three levels (low, medium, high) of valence

(up to 0.83 AUC), arousal (up to 0.85 AUC), and dominance (up to 0.84 AUC). Performance could be further increased by personalizing the model (up to +0.07 AUC). In addition, our model showed also decent performance for predicting two levels (present vs. absent) of stress (0.83 AUC) and the basic emotions anger (0.77 AUC), happiness (0.81 AUC), sadness (0.82 AUC), surprise (0.86 AUC), fear (0.85 AUC), and disgust (0.86 AUC). We also showed that by using two-dimensional heat maps built from gyroscope and accelerometer measurements, a comparable performance can be achieved. Thus, sensor heat maps are a viable alternative to keystroke heat maps considering the higher privacy protection and faster generation (0.23 seconds vs. 0.84 seconds). Another noteworthy contribution is our novel emoji-based questionnaire for measuring valence, arousal, and dominance on five levels. Our questionnaire is particularly useful on mobile devices and provides a modern and easy-to-understand visualization for a fast and accurate assessment of affective states.

Finally, in Chapter 7, we presented two compact and transparent widgets for visualizing valence, arousal, and dominance on mobile devices. The first widget provides a fast and intuitive assessment of affective states, which can be useful for conceiving affective states of several people at the same time (e.g., students in a classroom). The second widget provides an exact and time-dependent visualization, which is valuable when observing the affective state of a single person over time. We proved our design choices and the requirements imposed on the widgets by conducting a user study with 644 participants.

In conclusion, we showed that deep learning and traditional machine learning techniques can predict affective states accurately using a variety of data modalities. In general, our models performed well on laboratory data but also the model for predicting affective states in a real-world environment based on smartphone data showed a decent performance. We conclude, that it is possible to build a non-invasive and privacy-protecting system that can reliably predict affective states in the wild by carefully selecting the source and representation of the data. In this thesis, we presented such a model based on heat maps extracted from smartphone sensor data. Other data modalities (i.e., video data, biosensors, and handwriting data) can provide a comparable performance but with typically lower usability in real-world environments and higher invasion of privacy. On the other hand, these modalities are not tied to a specific context (i.e., smartphones). To conclude, we believe that machine learning for affective computing has a huge amount of potential with unexplored areas in different fields. We hope that this thesis has contributed a step forward in predicting affective states in the wild and has raised awareness of the importance of privacy protection for affect-aware systems.

8.2 Limitations

In the following, we discuss the core limitations of the presented models. A discussion of the specific limitations of the different models can be found in the corresponding chapters. We collected the ground truth (i.e., the affective states) for training our models by using self-reports of people. Directly measuring affect or emotions is difficult [Rosenthal and Rosnow, 1991]. Thus, our results are restricted to the specific conceptualization of affect that we chose (i.e., basic emotions and valence, arousal, and dominance). We identify three potential sources of bias associated with the self-reports:

- **Truthfulness.** Potentially, participants filled in the self-reports not always truthfully due to getting used to filling in self-reports or embarrassment to report certain emotions (e.g., reporting happiness when looking at a cruel picture). In post-experiment questionnaires, participants reported having filled in the self-reports truthfully. For the self-reports collected in the wild, we did not find any patterns of irregularities in the self-reports. Similarly, the self-reports collected in the laboratory experiments were in accordance with the emotions we wanted to trigger with the different tasks.
- **Subjectivity.** Self-reports always have a subjective component because perceptions of emotions are not universal but they are dependent on different factors such as culture and social background [Gendron et al., 2014; Jack et al., 2012; Mesquita and Frijda, 1992]. In our experiments, we recruited a homogeneous population of participants in terms of culture (i.e., most participants were European). The generalization of our models to other cultures needs further investigation. To reduce the subjective component, we provided the participants a detailed explanation of the self-reports including examples at the beginning of the experiments. By merging the self-reports into classes or regions, we also tried to alleviate the subjective bias.
- **Burn-in phase.**[1] It takes some time until participants are getting used to filling in self-reports. During such a burn-in phase self-reports may be biased (e.g., fluctuating too much or too little). Similarly, self-reports can be biased if participants are getting bored of filling in self-reports.

Another limitation is the missing interpersonal relations in laboratory experiments (i.e., the experimenter does not know the participants beforehand). This is different from interactions in the real world, where communication often happens between people knowing each other. It is known that interpersonal relationships can influence emotional reactions [Calvo and D'Mello, 2010]. In particular, it is a limitation for chat conversations on smartphones. Solving math tasks, looking at pictures or other

[1] Suggested by R.M. Weber (personal communication, March 25, 2021).

egocentric activities are less affected. In addition, the presence of the experimenter in laboratory experiments and the awareness of being recorded can also influence the participants [Calvo et al., 2015].

In our experiments, we recruited a homogeneous population of bachelor and master students which may limit the generalization to the non-academic world and people of other ages. We are optimistic that our models also work for a broader population given proper baseline normalization of the signals.

Finally, we trained and evaluated our models offline using resources (i.e., memory, CPU, and GPU) exceeding the resources of most mobile devices. In particular, smartphones and tablets rarely have built-in GPUs. A workaround for incorporating the models in real-world applications would be to deploy the models on a server instead of directly on mobile devices. This workaround may cause additional privacy issues and can lower the acceptance of users. On the other hand, deploying the model on mobile devices requires some adaptations of the model. Our model presented in Chapter 6 for predicting affective states based on smartphone keystroke and sensor data leverages the MobileNetV2 architecture. We chose this architecture keeping in mind the potential deployment of the model on mobile devices. Using depthwise separable convolutions, MobileNetV2 models can be run efficiently on mobile devices [Sandler et al., 2018].

8.3 Future Work

Future research comprises refining and extending our hardware setup and inpainting pipeline presented in Chapter 3. The CelebA-HQ dataset, which we used to train our inpainting model, contained only images with a frontal view of faces. In our recordings, individuals are captured at different angles. Thus, rotation of the recordings or using a dataset providing faces at different angles can improve the neural inpainting model. In addition, the feature set could be extended by gesture-based features. Such features have shown to be promising for predicting affective states [Bustos et al., 2011].

Potential refinements for our model based on biosensors and handwriting shown in Chapter 4 include using non-linear IBI features and frequency features for skin temperature. Additionally, an in-depth analysis of handwriting that takes into account the slant of the handwriting could further improve the classification performance. Another interesting direction would be to make use of existing large biosensor databases for semi-supervised learning by using autoencoders to infer an efficient feature embedding similar to the semi-supervised pipeline presented in Chapter 5.

Further, personality traits and the depression level could be modeled. We measured personality traits and the depression level of the participants in the experiments pre-

sented in Chapter 5 (laboratory experiment) and Chapter 6 (in the wild experiment). This would complement ongoing research that has shown the feasibility of predicting personality based on keyboard input [Khan et al., 2008], touchscreen-based interactions [Küster et al., 2018], and smartphone accelerometer data [Gao et al., 2019]. Knowing a person's personality traits can help context-aware recommender systems to provide recommendations tailored to the personality of the person [Braunhofer et al., 2015; Recio-Garcia et al., 2009]. Automatic inference of the personality traits and the depression level can also be useful in therapeutic settings to change the personality or coping with depression [Martin et al., 2014]. Further, predicting feelings, moods, attitudes, and temperament in addition to emotions could be beneficial to model affective states holistically [Calvo and D'Mello, 2010].

The deep learning models presented in Chapter 5 and Chapter 6 could be refined by transfer learning from other existing large-scale datasets such as ImageNet [Deng et al., 2009], which contains millions of images. Transfer learning was already successfully applied for emotion recognition from video data [Ng et al., 2015] and stress prediction from smartphone sensor data [Maxhuni et al., 2016].

In this thesis, we evaluated our models on a homogeneous population of bachelor and master students. Future research should consider a more heterogeneous population consisting of people of different ages, professions, and cultural backgrounds. Recently, methods capturing similarity in human behavior such as community similarity networks [Lane et al., 2011] have attracted attention. The underlying idea of such methods is to generalize affective state prediction by identifying individuals who can be treated as uniform in terms of affective state inference. Another possibility to foster generalization is to fuse different data modalities for building competent multimodal affective state prediction systems. Several techniques to fuse data have been proposed such as feature-level fusion, decision-level fusion, and data-level fusion [Wagner et al., 2009]. On that ground, several works presented models fusing different data sources for affective state prediction [Shoumy et al., 2020]. Although combining different data modalities can improve performance, issues related to privacy and usability in the real world might be amplified.

Another extension of our work could broaden the affective space by extending the set of basic emotions to include more distinct categorical states (e.g., bored, contented, excited, nervous, relaxed, and upset). Affective ground truth labels can also be targeted to the specific context. For example, in an educational setting, confusion, frustration, flow, curiosity, and anxiety can be of interest [Calvo and D'Mello, 2010]. In addition, instead of performing classification on the dimensional states (i.e., valence, arousal, and dominance), a regression model is an alternative providing a more fine-grained distinction between affective states.

Potential future work for the visualization presented in Chapter 7 is to investigate the cognitive processing time for both widgets, as well as the cognitive processing

stages when translating from sentences and images to visualizations. Furthermore, the precise widget can be improved in two ways. The different levels of the radial bar chart show varying prominence, which could be solved by using varying thickness of the rings. Also, the medium level could be highlighted such that the sign of the dimension is identifiable. One could investigate other choices than a radial bar chart for the precise widget. A final step would consist of testing the widgets on different devices such as smartwatches and using a perceptually normalized hue for the color mapping of the intuitive widget to eradicate contrast issues.

Finally, another future direction is to connect our visualization of affective states with our models to provide visual feedback in real-time. Such feedback can be useful for teachers in classrooms to observe student's emotions and for status messages on mobile phones to communicate the emotional state to others. Aside from real-time feedback, retrospective feedback can provide useful information about the dynamics of affective states in the past days or weeks (e.g., in a calendar application), which can foster self-regulation. A connection between our models and our visualization widgets can be achieved by deploying the models on a server and sending the predicted affective state back to the mobile device or by conducting the inference on the mobile device itself.

APPENDIX A

Affective State Prediction Using Smartphones in the Lab

Table A.1: *Performance for the prediction of three classes (low, medium, high) of valence, arousal, and dominance. AUC_{micro} and $F1_{micro}$ represent micro-averaged AUC and F_1-score, respectively. AUC_{macro} and $F1_{macro}$ represent macro-averaged AUC and F_1-score, respectively. The chance level is 0.33 for accuracy and F_1-score, 0.5 for AUC, and 0 for Cohen's kappa.*

Dimension	Heat Map	AUC_{micro}	AUC_{macro}	Accuracy	$F1_{micro}$	$F1_{macro}$	Cohen's kappa
Valence	Pressure	0.75	0.74	56%	0.56	0.55	0.33
	Down-down	0.81	0.81	64%	0.64	0.64	0.45
	Up-down	0.79	0.79	61%	0.61	0.61	0.41
	Combination	0.84	0.83	67%	0.67	0.66	0.49
Arousal	Pressure	0.80	0.78	62%	0.62	0.59	0.38
	Down-down	0.75	0.73	55%	0.55	0.51	0.28
	Up-down	0.73	0.70	53%	0.53	0.50	0.26
	Combination	0.82	0.80	63%	0.63	0.61	0.41
Dominance	Pressure	0.79	0.77	63%	0.63	0.57	0.37
	Down-down	0.80	0.78	63%	0.63	0.56	0.37
	Up-down	0.78	0.76	61%	0.61	0.55	0.34
	Combination	0.82	0.80	65%	0.65	0.59	0.41

A.1 Additional Statistics Supporting Experimental Validation

The tables provide the p-values of the Pearson correlations between the chat conversations (exciting, shocking, rude, and confusing) based on the self-reports (SAM,

four basic emotions, and stress). For the SAM, we calculated the mean of the values per participant and chat conversation. For the four basic emotions and stress, we added the times that participants reported a specific basic emotion or stress during each of the conversations. Values in bold represent statistical significant correlations after Bonferroni correction.

Table A.2: *The p-values of the Pearson correlations between the chat conversations based on the reported valence. Bonferroni correction with $\alpha = 0.003$ (18 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		**< 0.001**	**< 0.001**	**< 0.001**
Shocking	**< 0.001**		**< 0.001**	**< 0.001**
Rude	**< 0.001**	**< 0.001**		0.108
Confusing	**< 0.001**	**< 0.001**	0.108	

Table A.3: *The p-values of the Pearson correlations between the chat conversations based on the reported arousal. Bonferroni correction with $\alpha = 0.003$ (18 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		0.369	0.964	0.066
Shocking	0.369		0.718	**< 0.001**
Rude	0.964	0.718		0.082
Confusing	0.066	**< 0.001**	0.082	

Table A.4: *The p-values of the Pearson correlations between the chat conversations based on the reported dominance. Bonferroni correction with $\alpha = 0.003$ (18 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		**< 0.001**	0.003	**< 0.001**
Shocking	**< 0.001**		0.231	0.834
Rude	0.003	0.231		0.495
Confusing	**< 0.001**	0.834	0.495	

Table A.5: *The p-values of the Pearson correlations between the chat conversations based on the reported anger. Bonferroni correction with $\alpha = 0.002$ (30 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		**< 0.001**	**< 0.001**	0.22
Shocking	**< 0.001**		0.004	**< 0.001**
Rude	**< 0.001**	0.004		0.007
Confusing	0.22	**< 0.001**	0.007	

Table A.6: *The p-values of the Pearson correlations between the chat conversations based on the reported happiness. Bonferroni correction with $\alpha = 0.002$ (30 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		**< 0.001**	**< 0.001**	**< 0.001**
Shocking	**< 0.001**		0.015	0.006
Rude	**< 0.001**	0.015		0.92
Confusing	**< 0.001**	0.006	0.92	

Table A.7: *The p-values of the Pearson correlations between the chat conversations based on the reported sadness. Bonferroni correction with $\alpha = 0.002$ (30 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		< 0.002	0.339	0.693
Shocking	< 0.002		**< 0.001**	**< 0.001**
Rude	0.339	**< 0.001**		0.192
Confusing	0.693	**< 0.001**	0.192	

Table A.8: *The p-values of the Pearson correlations between the chat conversations based on the reported surprise. Bonferroni correction with $\alpha = 0.002$ (30 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		0.068	0.111	**< 0.001**
Shocking	0.068		0.944	0.014
Rude	0.111	0.944		0.021
Confusing	**< 0.001**	0.014	0.021	

Table A.9: *The p-values of the Pearson correlations between the chat conversations based on the reported stress. Bonferroni correction with* $\alpha = 0.002$ *(30 comparisons).*

	Exciting	**Shocking**	**Rude**	**Confusing**
Exciting		0.022	**< 0.001**	0.022
Shocking	0.022		0.057	0.972
Rude	**< 0.001**	0.057		0.065
Confusing	0.022	0.972	0.065	

APPENDIX B

Affective State Prediction Using Smartphones in the Wild

Table B.1: *Performance for the prediction of three classes (low, medium, high) of valence, arousal, and dominance. AUC_{micro} and $F1_{micro}$ represent micro-averaged AUC and F_1-score, respectively. AUC_{macro} and $F1_{macro}$ represent macro-averaged AUC and F_1-score, respectively. The chance level is 0.33 for accuracy and F_1-score, 0.5 for AUC, and 0 for Cohen's kappa.*

Dimension	Heat Map	AUC_{micro}	AUC_{macro}	Accuracy	$F1_{micro}$	$F1_{macro}$	Cohen's kappa
Valence	Keystrokes	0.82	0.76	66%	0.66	0.53	0.33
	Sensors	0.79	0.73	64%	0.64	0.48	0.27
	Combination	0.83	0.78	70%	0.70	0.57	0.40
Arousal	Keystrokes	0.81	0.80	63%	0.63	0.60	0.42
	Sensors	0.83	0.82	64%	0.64	0.61	0.43
	Combination	0.85	0.84	65%	0.65	0.64	0.46
Dominance	Keystrokes	0.82	0.79	67%	0.67	0.62	0.46
	Sensors	0.81	0.79	63%	0.63	0.59	0.39
	Combination	0.84	0.82	68%	0.68	0.64	0.47

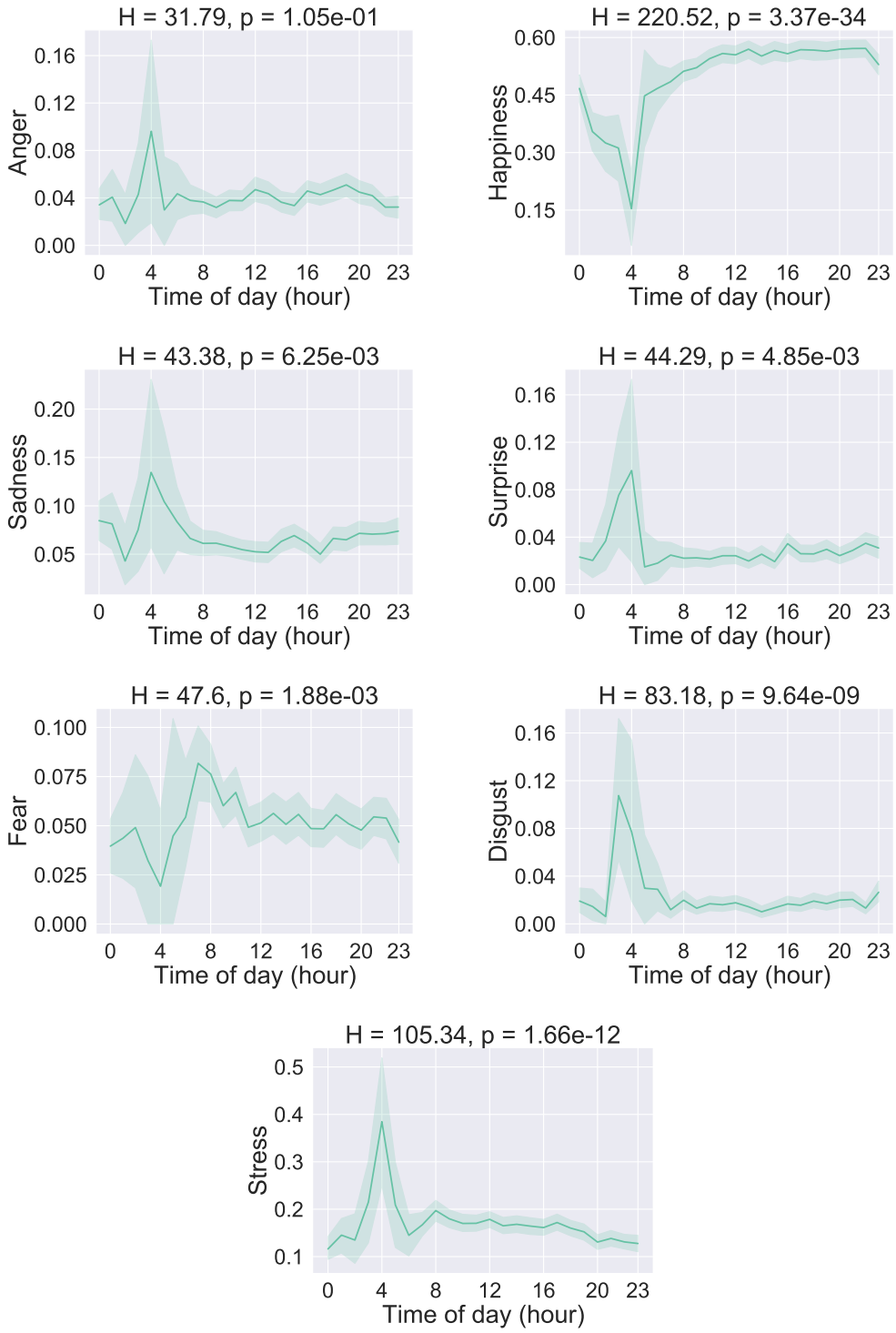


Figure B.1: *Mean and 95% confidence interval (shaded area) of reported basic emotions and stress level (interval $[0, 1]$) over the time of the day. The titles contain the results of Kruskal-Wallis tests to investigate whether there were significant differences during the time of the day.*

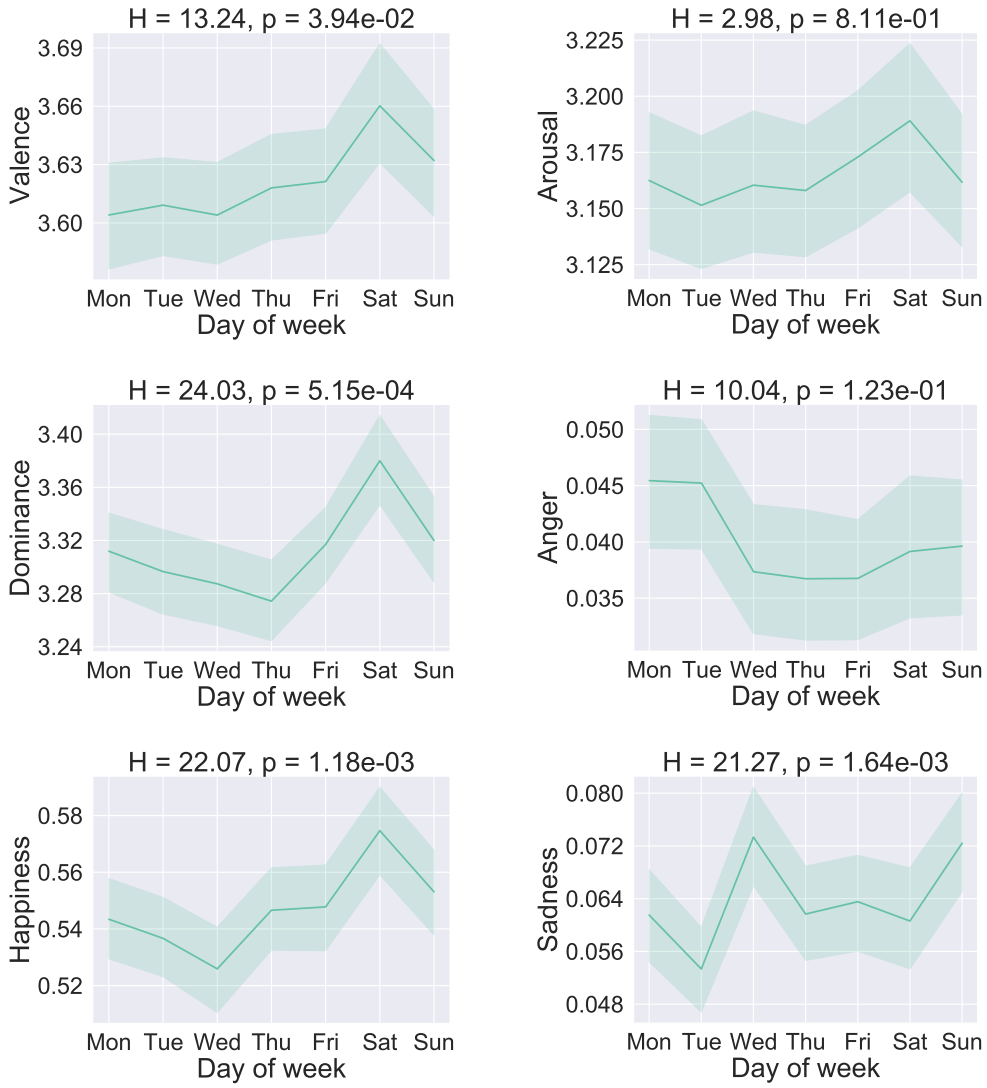


Figure B.2: *Mean and* 95% *confidence interval (shaded area) of reported valence, arousal, and dominance (interval* $[1, 5]$*) as well as the basic emotions and stress level (interval* $[0, 1]$*) over the days of the week. The titles contain the results of Kruskal-Wallis tests to investigate whether there were significant differences during the day of the week.*

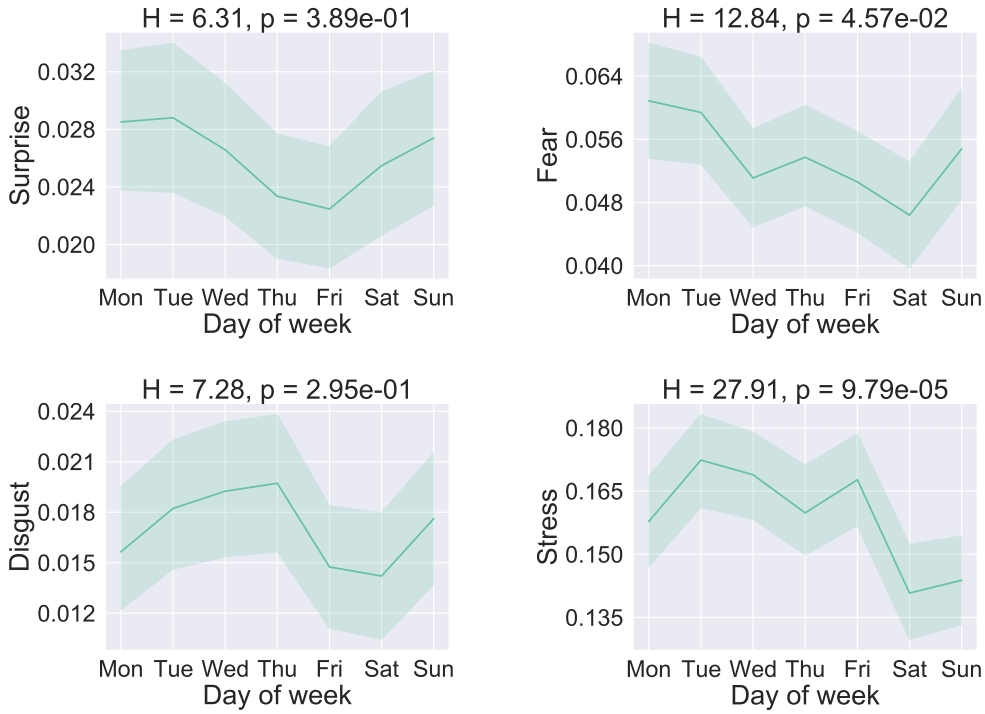


Figure B.2 ***(cont.)*****:** *Mean and* 95% *confidence interval (shaded area) of reported valence, arousal, and dominance (interval* $[1, 5]$*) as well as the basic emotions and stress level (interval* $[0, 1]$*) over the days of the week. The titles contain the results of Kruskal-Wallis tests to investigate whether there were significant differences during the day of the week.*

Figure B.3: *Mean and 95% confidence interval (shaded area) of reported basic emotions and stress level (interval $[0, 1]$) over different application categories. The titles contain the results of Kruskal-Wallis tests to investigate whether there were significant differences for the application categories.*

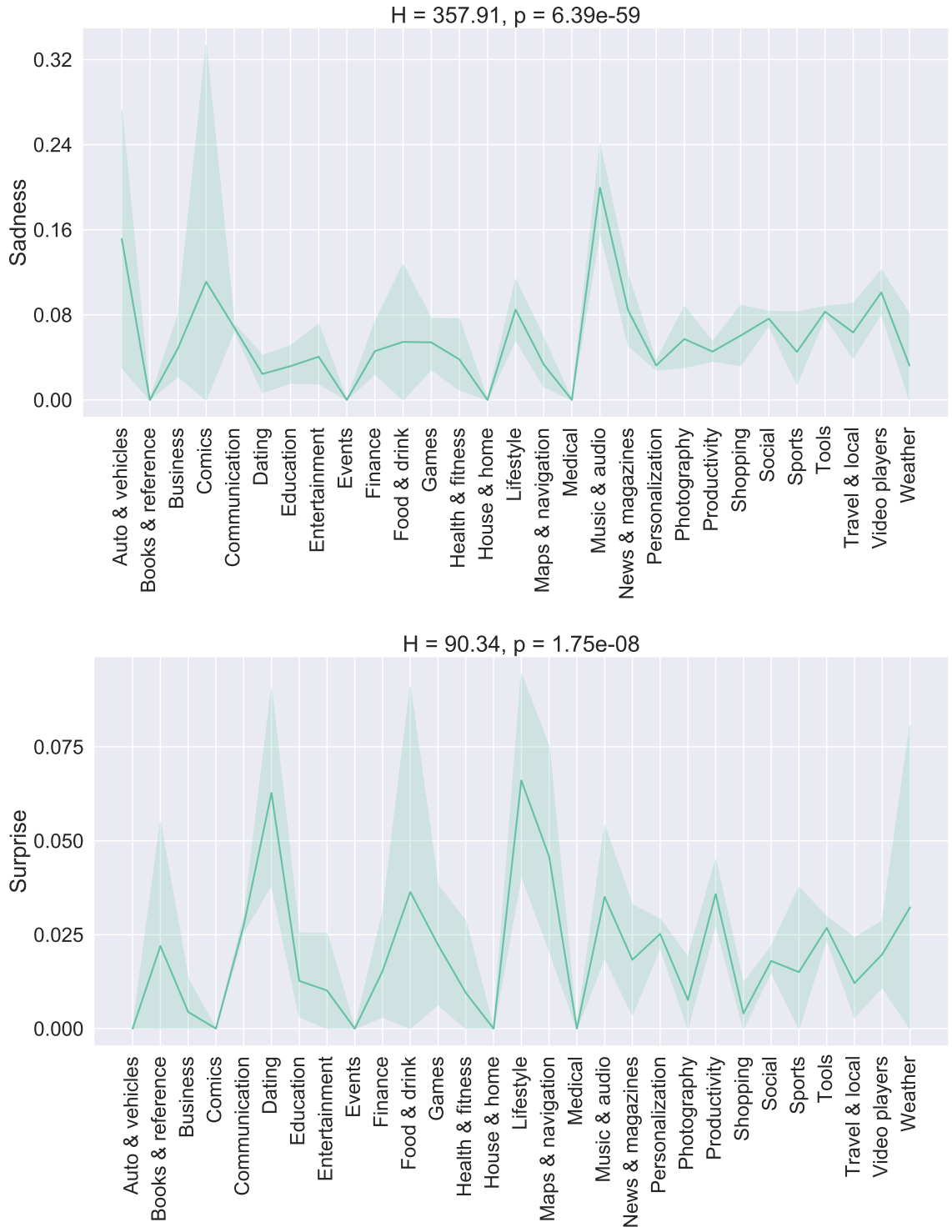


Figure B.3 ***(cont.)*****:** *Mean and 95% confidence interval (shaded area) of reported basic emotions and stress level (interval* $[0, 1]$*) over different application categories. The titles contain the results of Kruskal-Wallis tests to investigate whether there were significant differences for the application categories.*

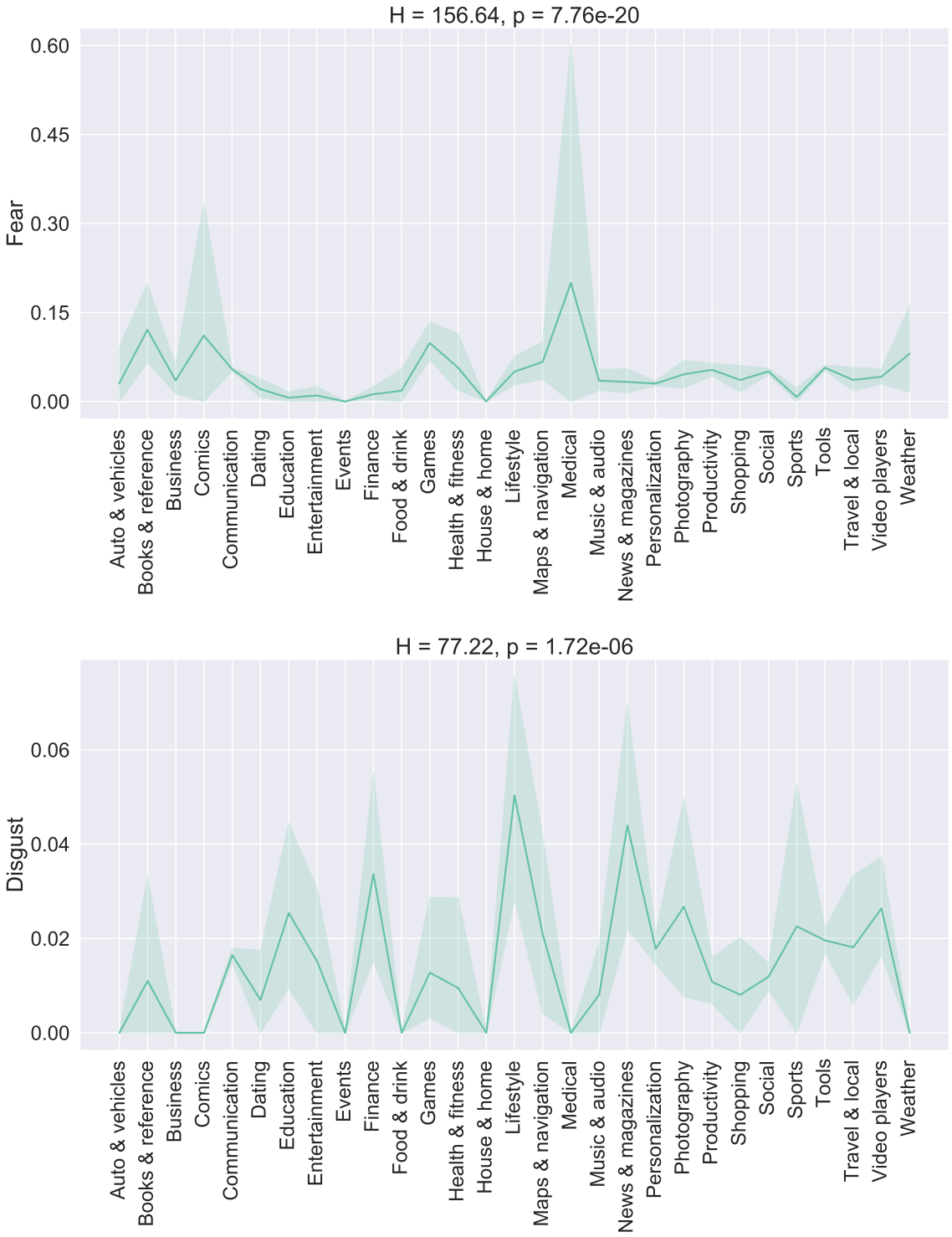


Figure B.3 ***(cont.)*****:** *Mean and 95% confidence interval (shaded area) of reported basic emotions and stress level (interval* $[0, 1]$*) over different application categories. The titles contain the results of Kruskal-Wallis tests to investigate whether there were significant differences for the application categories.*

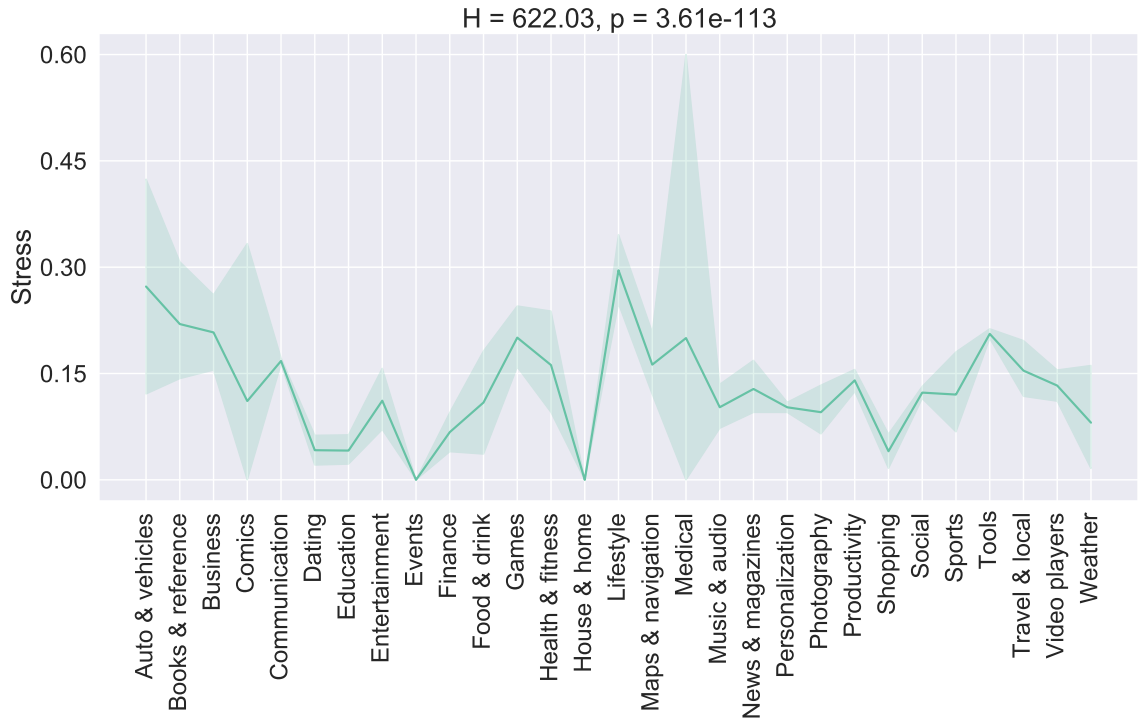


Figure B.3 ***(cont.)*****:** *Mean and 95% confidence interval (shaded area) of reported basic emotions and stress level (interval* $[0, 1]$*) over different application categories. The title contains the result of a Kruskal-Wallis test to investigate whether there were significant differences for the application categories.*

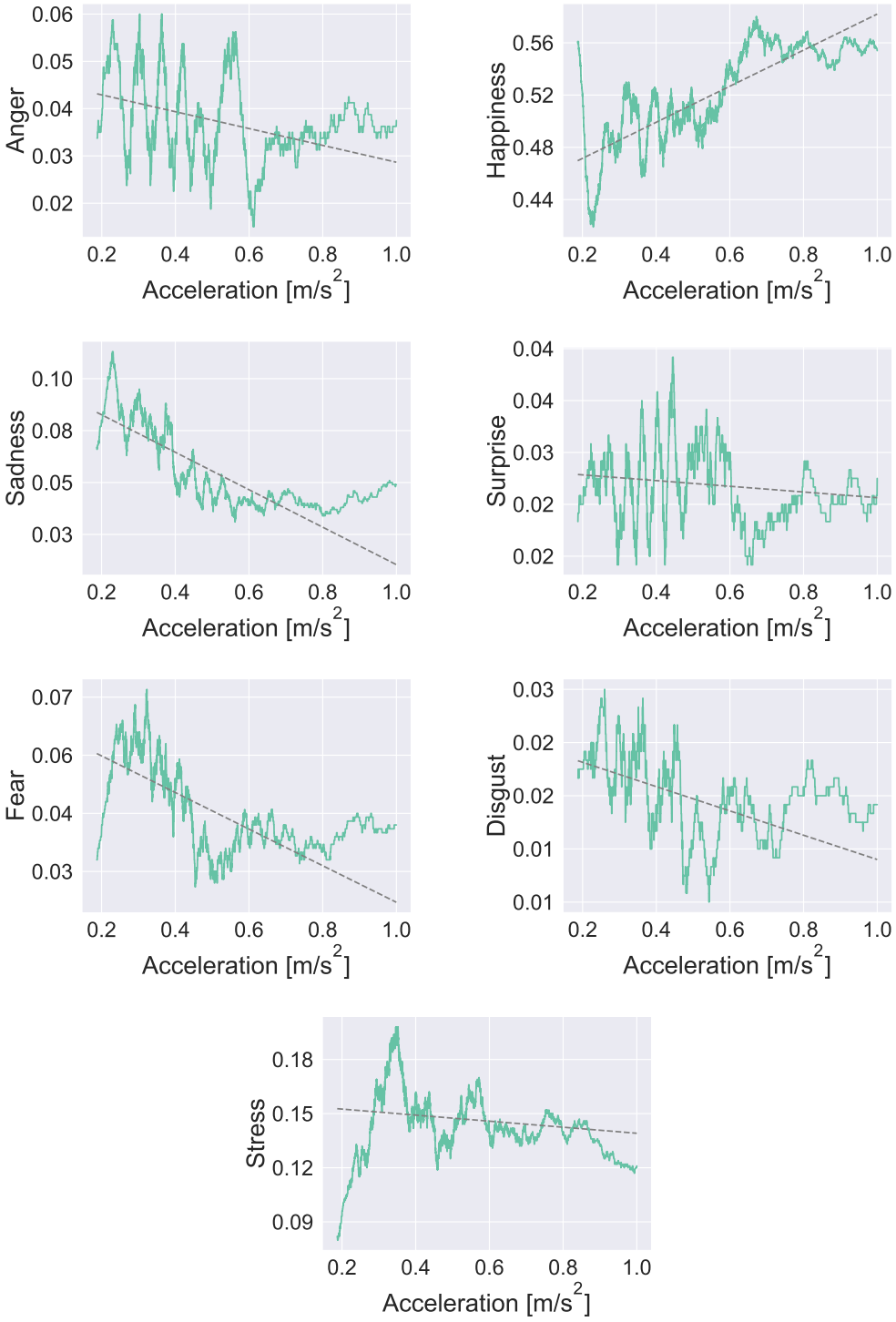


Figure B.4: *The basic emotions and stress level (interval $[0, 1]$) in relation to the magnitude of linear acceleration. The dashed regression line shows the linear trend in the data.*

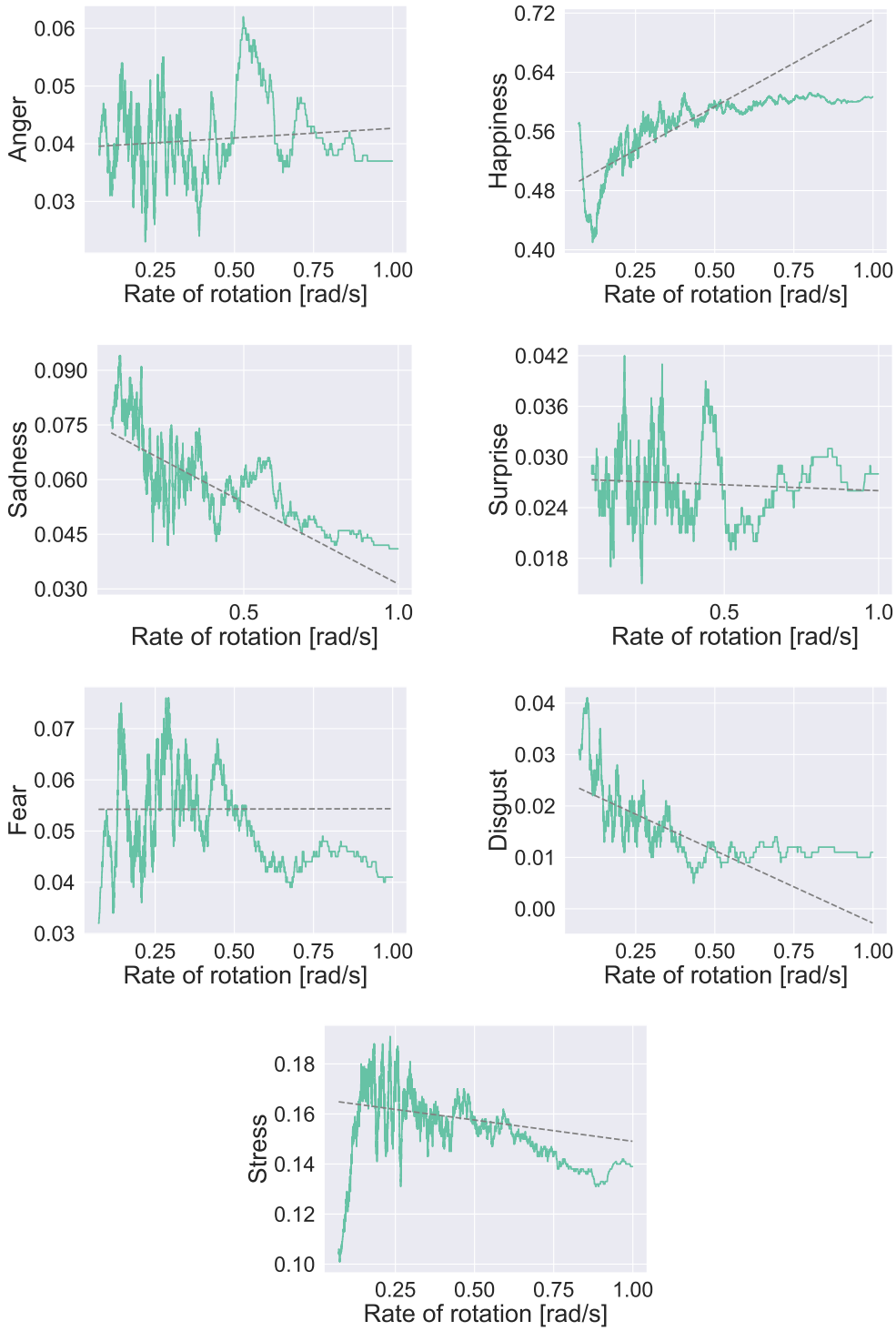


Figure B.5: *The basic emotions and stress level (interval $[0, 1]$) in relation to the magnitude of the rate of rotation. The dashed regression line shows the linear trend in the data.*

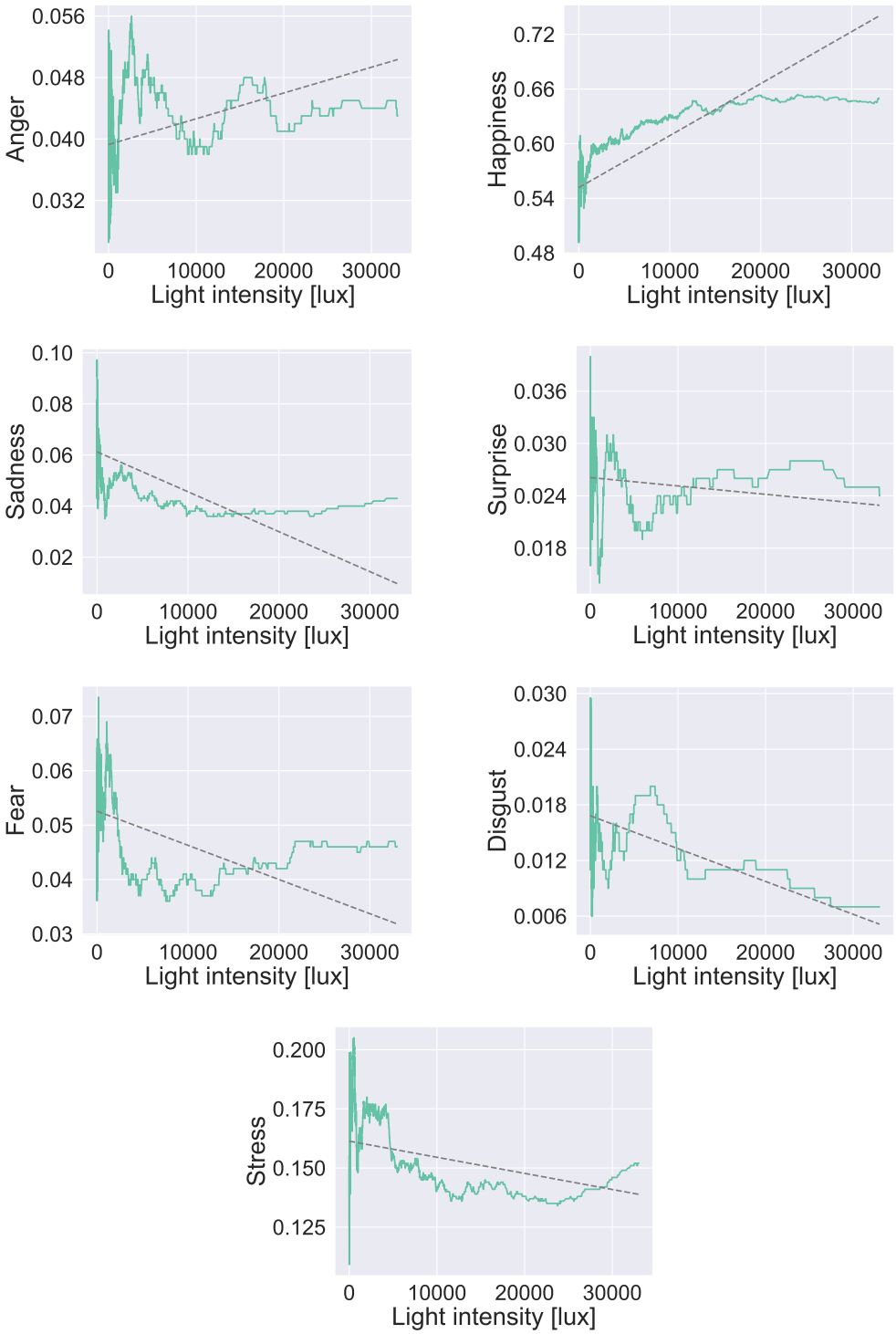


Figure B.6: *The basic emotions and stress level (interval* $[0, 1]$*) in relation to light intensity. The dashed regression line shows the linear trend in the data.*

APPENDIX C

Affective State Visualization

C.1 User Study Examples

Two task examples from the study are shown. The first example shown in Figure C.1 is taken from part 1 where the intuitiveness of our first widget has been investigated. The participants were not given any additional information about affective states or the widgets at this point. The second example presented in Figure C.2 is taken from part 2 where our widgets were compared to the baseline. Additional explanations regarding the widgets and the concept of affective states were provided previously to this task.

C.2 Sentences and Images Used in the Study

In Table C.1 we present an overview of the sentences that have been used in the study. The sentences are provided together with the keyword that defines the corresponding levels of valence, arousal and dominance. VAD stands for valence, arousal and dominance and indicates the level of each dimension on a 9-point scale. Furthermore, the alternatives listed on the right indicate which keyword has been used in the other two state visualizations shown in a particular task. The same information is provided for the images in Table C.2. Figure C.3 shows the six pictures used in the study.

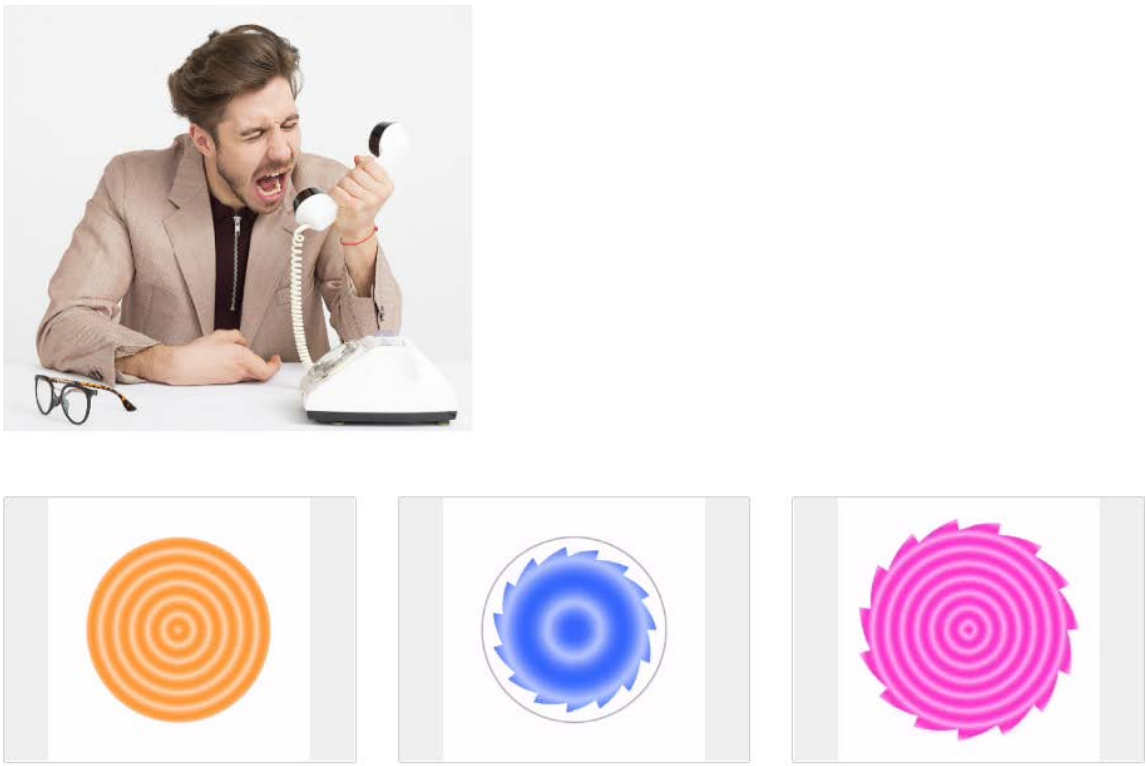


Figure C.1: *Example from the study for investigating the intuitiveness of our first widget.*

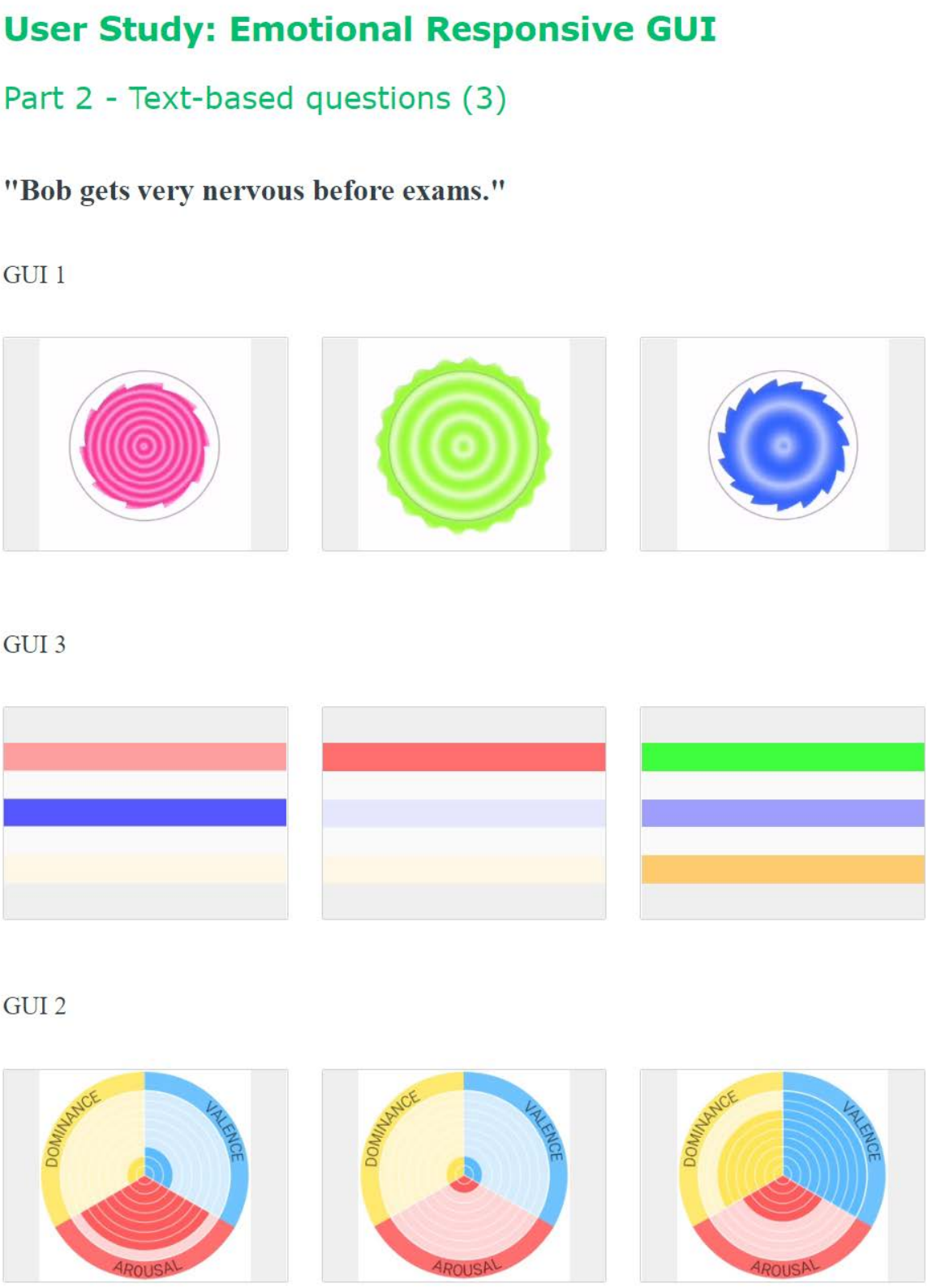


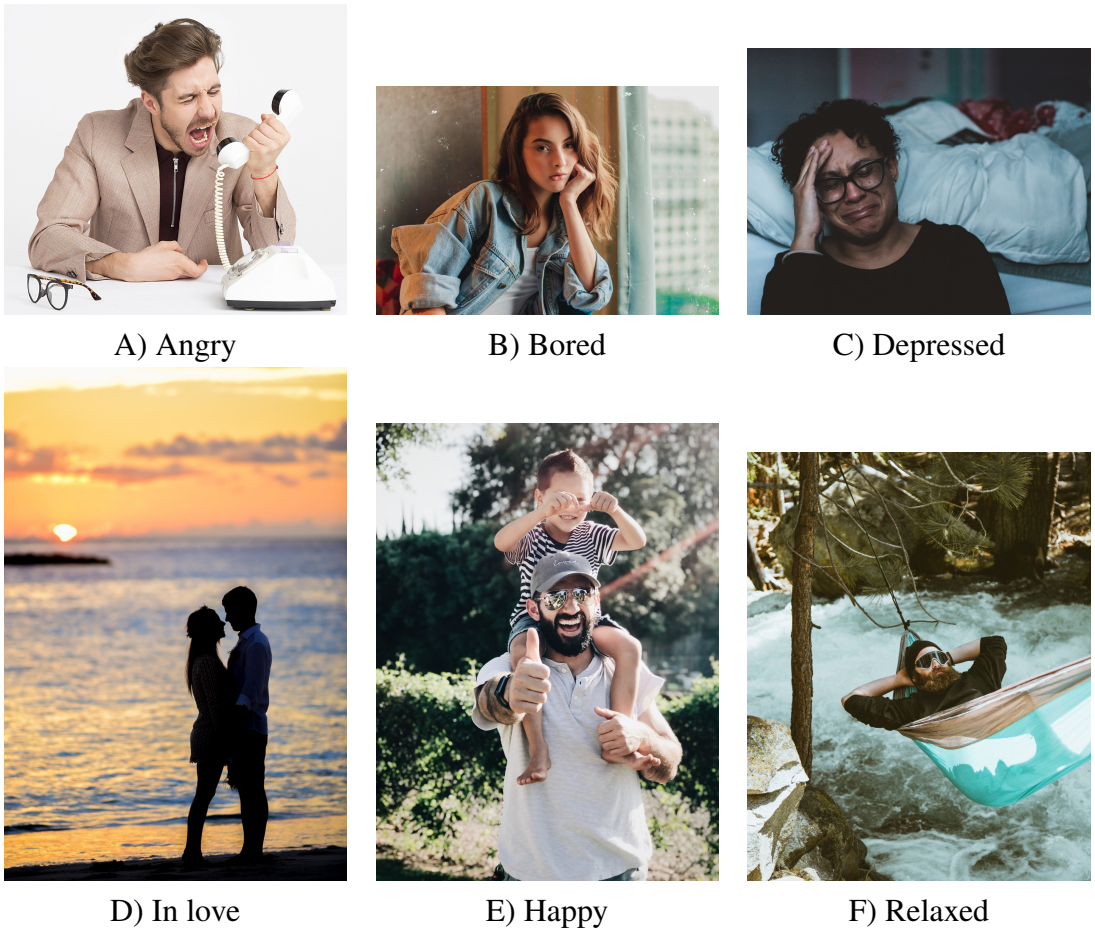
Figure C.2: *Example from the study for comparing our widgets to the baseline (GUI 3).*

Table C.1: *The sentences used in the study. The keyword defines the corresponding valence, arousal and dominance (VAD) level on a 9-point scale. The alternatives indicate which keyword has been used in the other two state visualizations shown in a particular task.*

Sentence	Keyword	VAD Score	Alternative 1	Alternative 2
Alice feels **sick** today.	sick	1, 4, 2	amused	sleepy
Charlie's team **won** the hacking competition.	win	9, 7, 8	scared	surprised
Bob gets very **nervous** before tests.	nervous	3, 8, 2	bored	in love
Charlie **wasn't** really **interested** in the seminar.	disinterested	4, 3, 4	depressed	angry
Alice and Bob finally reached an **agreement**.	agreement	7, 4, 7	excited	sad
Charlie was **shocked** after hearing the news.	shocked	3, 8, 4	bored	relaxed

Table C.2: *The images used in the study and the corresponding valence, arousal and dominance (VAD) level on a 9-point scale. The alternatives indicate which keyword has been used in the other two state visualizations shown in a particular task.*

Image	VAD Score	Alternative 1	Alternative 2
angry	2, 8, 6	surprised	bored
happy	9, 7, 8	sleepy	disinterested
relaxed	8, 1, 4	excited	in love
bored	2, 2, 2	sick	sleepy
in love	9, 5, 7	relaxed	agreement
sad	2, 7, 2	disintested	amused



A) Angry B) Bored C) Depressed

D) In love E) Happy F) Relaxed

Figure C.3: *The six pictures used in the study. The pictures express the emotional states angry (A), bored (B), depressed (C), in love, (D), happy (E) and relaxed (F). The images are taken from* `http://www.pexels.com` *and* `http://www.unsplash.com`.

References

[ACT, 2017] ACT. The ACT technical manual, 2017.

[Aksan et al., 2018] Emre Aksan, Fabrizio Pece, and Otmar Hilliges. Deepwriting: Making digital ink editable via deep generative modeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 2018.

[Andreassi, 2010] John L. Andreassi. *Psychophysiology: Human Behavior and Physiological Response*. Psychology Press, 2010.

[Androidrank, 2021] Website, 2021. Retrieved May 13, 2021 from `https://www.androidrank.org`.

[Araújo et al., 2005] Lívia C. F. Araújo, Luiz H. R. Sucupira, Miguel Gustavo Lizarraga, Lee Luan Ling, and Joao Baptista T. Yabu-Uti. User authentication through typing biometrics features. *IEEE Transactions on Signal Processing*, 53(2):851–855, 2005.

[Arroyo et al., 2009] Ivon Arroyo, David G. Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. Emotion sensors go to school. In *Proceedings of the Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, pages 17–24, NLD, 2009. IOS Press.

[Ayi and El-Sharkawy, 2020] Maneesh Ayi and Mohamed El-Sharkawy. RMNv2: Reduced Mobilenet V2 for CIFAR10. In *10th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 287–292. IEEE, 2020.

[Bachmann et al., 2015] Anja Bachmann, Christoph Klebsattel, Matthias Budde, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer.

How to use smartphones for less obtrusive ambulatory mood assessment and mood recognition. In *Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers*, pages 693–702. ACM, 2015.

[Baker et al., 2012] Ryan S. J. D. Baker, Sujith M. Gowda, Michael Wixon, Jessica Kalka, Angela Z. Wagner, Aatish Salvi, Vincent Aleven, Gail W. Kusbit, Jaclyn Ocumpaugh, and Lisa Rossi. Towards sensor-free affect detection in cognitive tutor algebra. In *International Conference on Educational Data Mining (EDM)*, June 2012.

[Baltrusaitis et al., 2018] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 59–66. IEEE, 2018.

[Bauer and Lukowicz, 2012] Gerald Bauer and Paul Lukowicz. Can smartphones detect stress-related changes in the behaviour of individuals? In *IEEE International Conference on Pervasive Computing and Communications Workshops*, pages 423–426. IEEE, 2012.

[Benedek and Kaernbach, 2010] Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80–91, 2010.

[Bengio, 2012] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer, 2012.

[Bertacchini et al., 2013] Francesca Bertacchini, Eleonora Bilotta, Lorella Gabriele, Diana Elizabeth Olmedo Vizueta, Pietro Pantano, Francesco Rosa, Assunta Tavernise, Stefano Vena, and Antonella Valenti. An emotional learning environment for subjects with Autism Spectrum Disorder. In *International Conference on Interactive Collaborative Learning (ICL)*, pages 653–659. IEEE, 2013.

[Best, 2005] Karl-Heinz Best. Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics*, 11:9–31, 2005.

[Betella et al., 2014] Alberto Betella, Riccardo Zucca, Ryszard Cetnarski, Alberto Greco, Antonio Lanatà, Daniele Mazzei, Alessandro Tognetti, Xerxes D. Arsiwalla, Pedro Omedas, Danilo De Rossi, et al. Inference of human affective states from psychophysiological measurements extracted under ecologically valid conditions. *Frontiers in Neuroscience*, 8:286, 2014.

[Beutelspacher, 1996] Albrecht Beutelspacher. *Kryptologie*, volume 7. Springer, 1996.

[Blanchard et al., 2014] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. Automated physiological-based detection of mind wandering during learning.

In *12th International Conference on Intelligent Tutoring Systems*, pages 55–60, Berlin, Heidelberg, 2014. Springer-Verlag.

[Bogomolov et al., 2013] Andrey Bogomolov, Bruno Lepri, and Fabio Pianesi. Happiness recognition from mobile phone data. In *International Conference on Social Computing*, pages 790–795. IEEE, 2013.

[Bogomolov et al., 2014] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, pages 477–486, New York, NY, USA, 2014. ACM.

[Borgo et al., 2013] Rita Borgo, Johannes Kehrer, David H. S. Chung, Eamonn Maguire, Robert S. Laramee, Helwig Hauser, Matthew Ward, and Min Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics - State of the Art Reports*. The Eurographics Association, 2013.

[Bosch et al., 2015] Nigel Bosch, Sidney D'Mello, Ryan S. J. D. Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 379–388, 2015.

[Boucsein et al., 2012] Wolfram Boucsein, Don C. Fowles, Sverre Grimnes, Gershon Ben-Shakhar, Walton T. roth, Michael E. Dawson, and Diane L. Filion. Publication recommendations for electrodermal measurements. *Psychophysiology*, pages 1017–34, 2012.

[Bradley and Lang, 1994] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994.

[Bradski, 2000] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000.

[Braunhofer et al., 2015] Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. User personality and the new user problem in a context-aware point of interest recommender system. In *Information and Communication Technologies in Tourism*, pages 537–549. Springer, 2015.

[Breazeal, 2011] Cynthia Breazeal. Social robots for health applications. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5368–5371. IEEE, 2011.

[Buschek et al., 2018] Daniel Buschek, Benjamin Bisinger, and Florian Alt. Researchime: A mobile keyboard application for studying free typing behaviour in the wild. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[Bustos et al., 2011] Dana May Bustos, Geoffrey Loren Chua, Richard Thomas Cruz, Jose Miguel Santos, and Merlin Teodosia Suarez. Gesture-based affect modeling for intelligent tutoring systems. In *International Conference on Artificial Intelligence in Education*, pages 426–428. Springer, 2011.

[Cabestrero et al., 2018] Raul Cabestrero, Pilar Quirós, Olga C. Santos, Sergio Salmeron-Majadas, Raul Uria-Rivas, Jesus G. Boticario, David Arnau, Miguel Arevalillo-Herráez, and Francesc J. Ferri. Some insights into the impact of affective information when delivering feedback to students. *Behaviour & Information Technology*, 37(12):1252–1263, 2018.

[Calvo and D'Mello, 2010] Rafael A. Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[Calvo et al., 2015] Rafael A. Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas. *The Oxford Handbook of Affective Computing*. Oxford Library of Psychology, 2015.

[Canazei and Weiss, 2013] M. Canazei and E. Weiss. The influence of light on mood and emotion. *Handbook of Psychology of Emotions: Recent Theoretical Perspectives and Novel Empirical Findings; Nova Science Publishers: Hauppauge, NY, USA*, 1:297–306, 2013.

[Carneiro et al., 2012] Davide Carneiro, José Carlos Castillo, Paulo Novais, Antonio Fernández-Caballero, and José Neves. Multimodal behavioral analysis for non-invasive stress detection. *Expert Systems with Applications*, 39(18):13376–13389, 2012.

[Carpenter, 2011] Rollo Carpenter. Cleverbot, 2011.

[Cernea et al., 2013] Daniel Cernea, Christopher Weber, Achim Ebert, and Andreas Kerren. Emotion scents: A method of representing user emotions on gui widgets. In *Visualization and Data Analysis*, volume 8654, pages 168–181. International Society for Optics and Photonics, SPIE, 2013.

[Cernea et al., 2015] Daniel Cernea, Christopher Weber, Achim Ebert, and Andreas Kerren. Emotion-prints: Interaction-driven emotion visualization on multi-touch interfaces. In *Visualization and Data Analysis*, volume 9397, pages 82–96. International Society for Optics and Photonics, SPIE, 2015.

[Chen et al., 2015] Yuxuan Chen, Nigel Bosch, and Sidney D'Mello. Video-based affect detection in noninteractive learning environments. In *Proceedings of the 8th International Conference on Educational Data Mining*, pages 440–443. International Educational Data Mining Society (IEDMS), 2015.

[Christin et al., 2011] Delphine Christin, Andreas Reinhardt, Salil S. Kanhere, and Matthias Hollick. A survey on privacy in mobile participatory sensing applications. *Journal of Systems and Software*, 84(11):1928–1946, November 2011.

[Colombetti, 2009] Giovanna Colombetti. From affect programs to dynamical discrete emotions. *Philosophical Psychology*, 22(4):407–425, 2009.

[Conati and Maclaren, 2009] Cristina Conati and Heather Maclaren. Modeling user affect from causes and effects. In *User Modeling, Adaptation, and Personalization*, pages 4–15. Springer, 2009.

[Critchley, 2002] Hugo D. Critchley. Electrodermal responses: What happens in the brain. *The Neuroscientist*, 8(2):132–142, 2002.

[Csikszentmihalyi, 2008] Mihaly Csikszentmihalyi. *Flow: The Psychology of Optimal Experience*. Harper Perennial, New York, NY, 2008.

[Dai et al., 2016] Daxiang Dai, Qun Liu, and Hongying Meng. Can your smartphone detect your emotion? In *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1704–1709. IEEE, 2016.

[Danner et al., 2016] Daniel Danner, Beatrice Rammstedt, Matthias Bluemke, Lisa Treiber, Sabrina Berres, Christopher J. Soto, and Oliver P. John. *Die deutsche Version des Big Five Inventory 2 (BFI-2)*. GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim, 2016.

[De Luca et al., 2012] Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. Touch me once and I know it's you!: Implicit authentication based on touch screen patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 987–996. ACM, 2012.

[Deng et al., 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[Dhakal et al., 2018] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. Observations on typing from 136 million keystrokes. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[Ditzler et al., 2016] Christine Ditzler, Eunsook Hong, and Neal Strudler. How tablets are utilized in the classroom. *Journal of Research on Technology in Education*, 48(3):181–193, 2016.

[D'Mello et al., 2018] Sidney K. D'Mello, Nigel Bosch, and Huili Chen. *Multimodal-Multisensor Affect Detection*, pages 167–202. Association for Computing Machinery and Morgan & Claypool, 2018.

[Dzedzickis et al., 2020] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

[Ekman and Friesen, 1978] Paul Ekman and Wallace Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement: Investigator's Guide 2 Part*. Consulting Psychologists Press, 1978.

[Ekman and Rosenberg, 1997] Paul Ekman and Erika L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[Ekman et al., 1987] 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, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712, 1987.

[Ekman, 1999] Paul Ekman. Basic emotions. In *Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons, 1999.

[Elfenbein and Ambady, 2003] Hillary Anger Elfenbein and Nalini Ambady. Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science*, 12(5):159–164, 2003.

[Epp et al., 2011] Clayton Epp, Michael Lippold, and Regan L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 715–724. ACM, 2011.

[Exposito et al., 2018] Marc Exposito, Javier Hernandez, and Rosalind W. Picard. Affective keys: Towards unobtrusive stress sensing of smartphone users. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, pages 139–145. ACM, 2018.

[Fairhurst et al., 2015] Michael Fairhurst, Meryem Erbilek, and Cheng Li. Study of automatic prediction of emotion from handwriting samples. *IET Biometrics*, pages 90–97, 2015.

[Falloon, 2013] Garry Falloon. Young students using iPads: App design and content influences on their learning pathways. *Computers & Education*, 68:505–521, 2013.

[Ferdous et al., 2015] Raihana Ferdous, Venet Osmani, and Oscar Mayora. Smartphone app usage as a predictor of perceived stress levels at workplace. In *9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 225–228. IEEE, 2015.

[Fitzpatrick et al., 2017] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2):e19, 2017.

[Folkman and Moskowitz, 2000] Susan Folkman and Judith Tedlie Moskowitz. Stress, positive emotion, and coping. *Current Directions in Psychological Science*, 9(4):115–118, 2000.

[Folkman, 2008] Susan Folkman. The case for positive emotions in the stress process. *Anxiety, Stress, and Coping*, 21(1):3–14, 2008.

[Fridlund and Duchaine, 1996] Alan J. Fridlund and Bradley Duchaine. "Facial expressions of emotion" and the delusion of the hermetic self. *The Emotions: Social, Cultural and Biological Dimensions*, pages 259–284, 1996.

[Fritz et al., 2014] Thomas Fritz, Andrew Begel, Sebastian C. Müller, Serap Yigit-Elliott, and Manuela Züger. Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th International Conference on Software Engineering*, pages 402–413. ACM, 2014.

[Gao et al., 2012] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction*, 19(4), 2012.

[Gao et al., 2019] Nan Gao, Wei Shao, and Flora D. Salim. Predicting personality traits from physical activity intensity. *Computer*, 52(7):47–56, 2019.

[Garcia-Ceja et al., 2015] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. Automatic stress detection in working environments from smartphones' accelerometer data: A first step. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1053–1060, 2015.

[Gendron et al., 2014] Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, 14(2):251, 2014.

[Ghandeharioun et al., 2019] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. Emma: An emotion-aware wellbeing chatbot. In *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

[Ghosh et al., 2017] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. Tapsense: Combining self-report patterns and typing characteristics for smartphone based emotion detection. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 2017.

[Ghosh et al., 2019a] Surjya Ghosh, Niloy Ganguly, Bivas Mitra, and Pradipta De. Designing an experience sampling method for smartphone based emotion detection. *IEEE Transactions on Affective Computing*, 2019.

[Ghosh et al., 2019b] Surjya Ghosh, Shivam Goenka, Niloy Ganguly, Bivas Mitra, and Pradipta De. Representation Learning for Emotion Recognition from Smartphone Keyboard Interactions. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 704–710. IEEE, 2019.

[Grafsgaard et al., 2013] Joseph Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Proceedings of the 6th International Conference on Educational Data*, pages 43–50. International Educational Data Mining Society, 2013.

[Grawemeyer et al., 2015] Beate Grawemeyer, Manolis Mavrikis, Wayne Holmes, Alice Hansen, Katharina Loibl, and Sergio Gutiérrez-Santos. The impact of feedback on students' affective states. In *CEUR Workshop Proceedings*, volume 1432. CEUR Workshop Proceedings, 2015.

[Grawemeyer et al., 2016] Beate Grawemeyer, Manolis Mavrikis, Wayne Holmes, Sergio Gutierrez-Santos, Michael Wiedmann, and Nikol Rummel. Affecting off-task behaviour: How affect-aware feedback can improve student learning. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge*, pages 104–113, New York, NY, USA, 2016. Association for Computing Machinery.

[Greco et al., 2016] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 63(4):797–804, 2016.

[Guillemot and Le Meur, 2013] Christine Guillemot and Olivier Le Meur. Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine*, 31(1):127–144, 2013.

[Gunes and Pantic, 2010] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, 2010.

[Guo et al., 2019] Zongyu Guo, Zhibo Chen, Tao Yu, Jiale Chen, and Sen Liu. Progressive image inpainting with full-resolution residual network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2496–2504, 2019.

[Haak et al., 2009] Martijn Haak, Steven Bos, Sacha Panic, and Léon J. M. Rothkrantz. Detecting stress using eye blinks and brain activity from EEG signals. *Proceeding of the 1st Driver Car Interaction and Interface*, pages 35–60, 2009.

[Hayashi et al., 2016] Elaine C. S. Hayashi, Julián E. Gutiérrez Posada, Vanessa R. M. L. Maike, and M. Cecília C. Baranauskas. Exploring new formats of the Self-Assessment Manikin in the design with children. In *Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems*, pages 1–10, 2016.

[Healey et al., 2010] Jennifer Healey, Lama Nachman, Sushmita Subramanian, Junaith Shahabdeen, and Margaret Morris. Out of the lab and into the fray: Towards modeling emotion in everyday life. In *International Conference on Pervasive Computing*, pages 156–173. Springer, 2010.

[Hernandez et al., 2014] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. Under pressure: Sensing stress of computer users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 51–60. ACM, 2014.

[Higgins et al., 2016] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.

[Higgins et al., 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *In Proceedings of the 8th International Conference on Learning Representations*, 2(5), 2017.

[Hill et al., 2015] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49:245–250, 2015.

[Hovsepian et al., 2015] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. cStress: Towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*, pages 493–504. ACM, 2015.

[Huang et al., 2018] He Huang, Bokai Cao, Philip S. Yu, Chang-Dong Wang, and Alex D. Leow. dpMood: Exploiting local and periodic typing dynamics for personalized mood prediction. In *IEEE International Conference on Data Mining (ICDM)*, pages 157–166. IEEE, 2018.

[Hupperich et al., 2016] Thomas Hupperich, Henry Hosseini, and Thorsten Holz. Leveraging sensor fingerprinting for mobile device authentication. In *Proceedings of the 13th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 377–396, Berlin, Heidelberg, 2016. Springer-Verlag.

[Iizuka et al., 2017] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):1–14, 2017.

[Itten, 1970] Johannes Itten. *The Elements of Color*, page 73. Van Nostrand Reinhold Company, 1970.

[Jack et al., 2012] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[Jaques et al., 2014] Natasha Jaques, Cristina Conati, Jason M. Harley, and Roger Azevedo. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems*, pages 29–38. Springer, 2014.

[John et al., 1991] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. The big five inventory - versions 4a and 54, 1991.

[John et al., 2008] Oliver P. John, Laura P. Naumann, and Christopher J. Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of Personality: Theory and Research*, 3(2):114–158, 2008.

[Jraidi et al., 2014] Imène Jraidi, Maher Chaouachi, and Claude Frasson. A hierarchical probabilistic framework for recognizing learners' interaction experience trends and emotions. *Advances in Human-Computer Interaction*, 2014:6, 2014.

[Kai et al., 2015] Shiming Kai, Luc Paquette, Ryan S. J. D. Baker, Nigel Bosch, Sidney D'Mello, Jaclyn Ocumpaugh, Valerie Shute, and Matthew Ventura. A comparison of video-based and interaction-based affect detectors in physics playground. In *Proceedings of the 8th International Conference on Educational Data Mining*, pages 77–84. International Educational Data Mining Society (IEDMS), 2015.

[Kanjo et al., 2015] Eiman Kanjo, Luluah Al-Husain, and Alan Chamberlain. Emotions in context: Examining pervasive affective sensing systems, applications, and analyses. *Personal and Ubiquitous Computing*, 19(7):1197–1212, 2015.

[Karras et al., 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations ICLR*, 2018.

[Karras et al., 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[Käser et al., 2013] Tanja Käser, Gian-Marco Baschera, Alberto Giovanni Busetto, Severin Klingler, Barbara Solenthaler, Joachim M. Buhmann, and Markus Gross. Towards a framework for modelling engagement dynamics in multiple learning domains. *International Journal of Artificial Intelligence in Education*, 22(1-2):59–83, 2013.

[Khan et al., 2008] Iftikhar Ahmed Khan, Willem-Paul Brinkman, Nick Fine, and Robert M. Hierons. Measuring personality from keyboard and mouse use. In *Proceedings of the 15th European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction*, pages 1–8, New York, NY, USA, 2008. Association for Computing Machinery.

[Kim et al., 2004] Kyung Hwan Kim, Seok Won Bang, and Sang Ryong Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, 2004.

[King, 2009] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference for Learning Representations*, 2015.

[Kingma and Welling, 2013] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kingma et al., 2014] Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

[Klingler et al., 2017] Severin Klingler, Rafael Wampfler, Tanja Käser, Barbara Solenthaler, and Markus H. Gross. Efficient feature embeddings for student classification with variational auto-encoders. In *Proceedings of the 10th International Conference on Educational Data Mining*, pages 72–79, 2017.

[Kołakowska et al., 2020] Agata Kołakowska, Wioleta Szwoch, and Mariusz Szwoch. A Review of Emotion Recognition Methods Based on Data Acquired via Smartphone Sensors. *Sensors*, 20(21), 2020.

[Kołakowska, 2013] Agata Kołakowska. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *6th International Conference on Human System Interactions (HSI)*, pages 548–555. IEEE, 2013.

[Kostyuk et al., 2018] Victor Kostyuk, Ma Victoria Almeda, and Ryan S. J. D. Baker. Correlating affect and behavior in reasoning mind with state test achievement. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 26–30. ACM, 2018.

[Kozhevnikov et al., 2013] Maria Kozhevnikov, James Elliott, Jennifer Shephard, and Klaus Gramann. Neurocognitive and somatic components of temperature increases during g-tummo meditation: Legend and reality. *PLOS ONE*, 8(3):1–12, 2013.

[Kroenke et al., 2001] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[Kroenke et al., 2009] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. The PHQ-8 as a measure of current

depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173, 2009.

[Küster et al., 2018] Ludwig Küster, Carola Trahms, and Jan-Niklas Voigt-Antons. Predicting personality traits from touchscreen based interactions. In *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018.

[Lane et al., 2010] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. A survey of mobile phone sensing. *IEEE Communications Magazine*, 48(9):140–150, 2010.

[Lane et al., 2011] Nicholas D. Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T. Campbell, and Feng Zhao. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proceedings of the 13th International Conference on Ubiquitous Computing*, pages 355–364, New York, NY, USA, 2011. Association for Computing Machinery.

[Lane et al., 2012] Nicholas Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. BeWell: A smartphone application to monitor, model and promote wellbeing. In *5th International ICST Conference on Pervasive Computing Technologies for Healthcare*. IEEE, April 2012.

[Lang et al., 2008] Peter J. Lang, Margaret M. Bradley, and B. N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, 2008.

[Larradet et al., 2020] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G. Caldwell, and Leonardo S. Mattos. Toward emotion recognition from physiological signals in the wild: Approaching the methodological issues in real-life data collection. *Frontiers in Psychology*, 11:1111, 2020.

[Lee et al., 2012] Hosub Lee, Young Sang Choi, Sunjae Lee, and I. P. Park. Towards unobtrusive emotion recognition for affective social communication. In *IEEE Consumer Communications and Networking Conference (CCNC)*, pages 260–264. IEEE, 2012.

[Lee et al., 2015] Poming Lee, Wei-Hsuan Tsui, and Tzu-Chien Hsiao. The influence of emotion on keyboard typing: An experimental study using auditory stimuli. *PLOS ONE*, 10, 2015.

[Leijdekkers et al., 2013] Peter Leijdekkers, Valerie Gay, and Frederick Wong. CaptureMyEmotion: A mobile app to improve emotion learning for autistic children using sensors. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 381–384. IEEE, 2013.

[Leon et al., 2011] Enrique Leon, Manuel Montejo, and Inigo Dorronsoro. Prospect of smart home-based detection of subclinical depressive disorders. In *5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 452–457. IEEE, 2011.

[Leow et al., 2019] Alex Leow, Jonathan Stange, John Zulueta, Olusola Ajilore, Faraz Hussain, Andrea Piscitello, Kelly Ryan, Jennifer Duffecy, Scott Langenecker, Peter Nelson, and Melvin McInnis. BiAffect: Passive monitoring of psychomotor activity in mood disorders using mobile keystroke kinematics. *Biological Psychiatry*, 85(10):S102–S103, 2019.

[Leys et al., 2013] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

[Li et al., 2017] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017.

[Liao et al., 2018] Haofu Liao, Gareth Funka-Lea, Yefeng Zheng, Jiebo Luo, and Kevin S. Zhou. Face completion with semantic knowledge and collaborative adversarial learning. In *Asian Conference on Computer Vision*, pages 382–397. Springer, 2018.

[LiKamWa et al., 2013] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, pages 389–402. ACM, 2013.

[Likforman-Sulem et al., 2017] Laurence Likforman-Sulem, Anna Esposito, Marcos Faundez-Zanuy, Stéphan Clémençon, and Gennaro Cordasco. EMOTHAW: A novel database for emotional state recognition from handwriting and drawing. *IEEE Transactions on Human-Machine Systems*, 47(2):273–284, 2017.

[Litman and Forbes-Riley, 2006] Diane J. Litman and Kate Forbes-Riley. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590, 2006.

[Liu et al., 2018] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[Liu et al., 2019] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4170–4179, 2019.

[Lu et al., 2012] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the ACM Conference on Ubiquitous Computing*, pages 351–360. ACM, 2012.

[Lv et al., 2008] Hai-Rong Lv, Zhong-Lin Lin, Wen-Jun Yin, and Jin Dong. Emotion recognition based on pressure sensor keyboards. In *IEEE International Conference on Multimedia and Expo*, pages 1089–1092. IEEE, 2008.

[Ma et al., 2012] Yuanchao Ma, Bin Xu, Yin Bai, Guodong Sun, and Run Zhu. Daily mood assessment based on mobile phone sensing. In *9th International Conference on Wearable and Implantable Body Sensor Networks*, pages 142–147. IEEE, 2012.

[Macias et al., 2013] Elsa Macias, Alvaro Suarez, and Jaime Lloret. Mobile sensing systems. *Sensors*, 13(12):17292–17321, 2013.

[Mahfouz et al., 2017] Ahmed Mahfouz, Tarek M. Mahmoud, and Ahmed Sharaf Eldin. A survey on behavioral biometric authentication on smartphones. *Journal of Information Security and Applications*, 37:28–37, 2017.

[Malesevic et al., 2019] Dejan Malesevic, Christoph Mayer, Shuhang Gu, and Radu Timofte. Photo-realistic and robust inpainting of faces using refinement GANs. In *Inpainting and Denoising Challenges*, pages 129–144. Springer, 2019.

[Malik et al., 1996] Marek Malik, J. Thomas Bigger, A. John Camm, Robert E. Kleiger, Alberto Malliani, Arthur J. Moss, and Peter J. Schwartz. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17(3):354–381, 1996.

[Martin et al., 2014] Lesley S. Martin, Lindsay G. Oades, and Peter Caputi. Intentional personality change coaching: A randomised controlled trial of participant selected personality facet change using the Five-Factor Model of Personality. *International Coaching Psychology Review*, 9(2):196–209, 2014.

[Maxhuni et al., 2016] Alban Maxhuni, Pablo Hernandez-Leal, L. Enrique Sucar, Venet Osmani, Eduardo F. Morales, and Oscar Mayora. Stress modelling and prediction in presence of scarce data. *Journal of Biomedical Informatics*, 63:344–356, 2016.

[McCallister et al., 2010] Erika McCallister, Timothy Grance, and Karen A. Scarfone. Guide to protecting the confidentiality of personally identifiable information, June 2010.

[McDaniel et al., 2007] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. Facial features for affective state detection in learning environments. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, pages 467–472, 2007.

[McDuff et al., 2016] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana Kaliouby. Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3723–3726, 2016.

[Mehrabian, 1996] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[Mesquita and Frijda, 1992] Batja Mesquita and Nico H. Frijda. Cultural variations in emotions: A review. *Psychological Bulletin*, 112(2):179, 1992.

[Miserandino, 1996] Marianne Miserandino. Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, 88(2):203, 1996.

[Mohammad, 2018] Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 174–184, 2018.

[Monrose et al., 2002] Fabian Monrose, Michael K. Reiter, and Susanne Wetzel. Password hardening based on keystroke dynamics. *International Journal of Information Security*, 1(2):69–83, 2002.

[Navarathna et al., 2014] Rajitha Navarathna, Patrick Lucey, Peter Carr, Elizabeth Carter, Sridha Sridharan, and Iain Matthews. Predicting movie ratings from audience behaviors. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1058–1065. IEEE, 2014.

[Newell, 1995] Patricia Brierley Newell. Perspectives on privacy. *Journal of Environmental Psychology*, 15(2):87–104, 1995.

[Ng et al., 2015] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the ACM on International Conference on Multimodal Interaction*, pages 443–449, New York, NY, USA, 2015. Association for Computing Machinery.

[Nguyen et al., 2017] Binh T. Nguyen, Minh H. Trinh, Tan V. Phan, and Hien D. Nguyen. An efficient real-time emotion detection using camera and facial landmarks. In *7th International Conference on Information Science and Technology (ICIST)*, pages 251–255. IEEE, 2017.

[Olsen and Torresen, 2016] Andreas Fsrøvig Olsen and Jim Torresen. Smartphone accelerometer data used for detecting human emotions. In *3rd International Conference on Systems and Informatics (ICSAI)*, pages 410–415. IEEE, 2016.

[Palin et al., 2019] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. How do people type on mobile devices? observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, New York, NY, USA, 2019. Association for Computing Machinery.

[Pathak et al., 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[Pham and Wang, 2018] Phuong Pham and Jingtao Wang. Predicting learners' emotions in mobile MOOC learning via a multimodal intelligent tutor. In *International Conference on Intelligent Tutoring Systems*, pages 150–159. Springer, 2018.

[Picard et al., 2004] Rosalind W. Picard, Seymour Papert, Walter Bender, Bruce Blumberg, Cynthia Breazeal, David Cavallo, Tod Machover, Mitchel Resnick, Deb Roy, and Carol Strohecker. Affective learning - a manifesto. *BT Technology Journal*, 22(4):253–269, 2004.

[Picard, 2000] Rosalind W. Picard. *Affective Computing*. MIT Press, 2000.

[Pielot et al., 2015] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. When attention is not scarce-detecting boredom from mobile phone usage. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 825–836. ACM, 2015.

[Plews et al., 2017] Daniel J. Plews, Ben Scott, Marco Altini, Matt Wood, Andrew E. Kilding, and Paul B. Laursen. Comparison of heart-rate-variability recording with smartphone photoplethysmography, Polar H7 chest strap, and electrocardiography. *International Journal of Sports Physiology and Performance*, 12(10):1324–1328, 2017.

[Pokrovskii and Polischuk, 2012] Vladimir M. Pokrovskii and Lily V. Polischuk. On the conscious control of the human heart. *Journal of Integrative Neuroscience*, 11(2):213–223, 2012.

[Politou et al., 2017] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. A survey on mobile affective computing. *Computer Science Review*, 25:79–100, 2017.

[Pudāne and Lavendelis, 2017] Māra Pudāne and Egons Lavendelis. General guidelines for design of affective multi-agent systems. *Applied Computer Systems*, 22(1):5–12, 2017.

[Pudane et al., 2019] Mara Pudane, Sintija Petrovica, Egons Lavendelis, and Hazım Kemal Ekenel. Towards truly affective AAL systems. In *Enhanced Living Environments*, pages 152–176. Springer, 2019.

[Rachuri et al., 2010] Kiran K. Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J. Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: A mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 281–290. ACM, 2010.

[Recio-Garcia et al., 2009] Juan A. Recio-Garcia, Guillermo Jimenez-Diaz, Antonio A. Sanchez-Ruiz, and Belen Diaz-Agudo. Personality aware recommendations to groups. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, pages 325–328, New York, NY, USA, 2009. Association for Computing Machinery.

[Reid and Ostashewski, 2011] Doug Reid and Nathaniel Ostashewski. iPads in the classroom - new technologies, old issues: Are they worth the effort? In *EdMedia+ Innovate Learning*, pages 1689–1694. Association for the Advancement of Computing in Education (AACE), 2011.

[Remaida et al., 2020] Ahmed Remaida, Aniss Moumen, Younes El Bouzekri El Idrissi, and Benyoussef Abdellaoui. Handwriting personality recognition with machine learning: A comparative study. In *IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, pages 1–6. IEEE, 2020.

[Riva et al., 2015] Giuseppe Riva, Rafael A. Calvo, and Christine Lisetti. Cyberpsychology and affective computing. *The Oxford Handbook of Affective Computing*, pages 547–558, 2015.

[Rosenthal and Rosnow, 1991] Robert Rosenthal and Ralph L. Rosnow. *Essentials of Behavioral Research: Methods and Data Analysis*, volume 2. McGraw-Hill New York, 1991.

[Ruensuk et al., 2019] Mintra Ruensuk, Hyunmi Oh, Eunyong Cheon, Ian Oakley, and Hwajung Hong. Detecting negative emotions during social media use on smartphones. In *Proceedings of Asian CHI Symposium: Emerging HCI Research Collection*, pages 73–79, New York, NY, USA, 2019. Association for Computing Machinery.

[Russell and Mehrabian, 1977] James A. Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[Russell, 1980] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[Sahlol et al., 2020] Ahmed T. Sahlol, Mohamed Abd Elaziz, Amani Tariq Jamal, Robertas Damaševičius, and Osama Farouk Hassan. A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features. *Symmetry*, 12(7):1146, 2020.

[Salmeron-Majadas et al., 2015] Sergio Salmeron-Majadas, Miguel Arevalillo-Herráez, Olga C. Santos, Mar Saneiro, Raúl Cabestrero, Pilar Quirós, David Arnau, and Jesus G. Boticario. Filtering of spontaneous and low intensity emotions in educational contexts. In *Proceedings of the 17th International Learning Analytics & Knowledge Conference*, pages 429–438. Springer, 2015.

[Salmeron-Majadas et al., 2018] Sergio Salmeron-Majadas, Ryan S. J. D. Baker, Olga C. Santos, and Jesus G. Boticario. A machine learning approach to leverage individual keyboard and mouse interaction behavior from multiple users in real-world learning scenarios. *IEEE Access*, 6:39154–39179, 2018.

[Sandler et al., 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[Saneiro et al., 2014] Mar Saneiro, Olga C. Santos, Sergio Salmeron-Majadas, and Jesus G. Boticario. Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. *The Scientific World Journal*, 2014.

[Sanghvi et al., 2011] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, pages 305–312, 2011.

[Santos et al., 2016] Olga C. Santos, Mar Saneiro, Jesus G. Boticario, and María Cristina Rodriguez-Sanchez. Toward interactive context-aware affective educational recommendations in computer-assisted language learning. *New Review of Hypermedia and Multimedia*, 22(1-2):27–57, 2016.

[Santos, 2016] Olga C. Santos. Emotions and personality in adaptive e-learning systems: An affective computing perspective. In *Emotions and Personality in Personalized Services*, pages 263–285. Springer, 2016.

[Sarpate and Guru, 2014] Geeta K. Sarpate and Shanti K. Guru. Image inpainting on satellite image using texture synthesis & region filling algorithm. In *International Conference on Advances in Communication and Computing Technologies (ICACACT)*, pages 1–5. IEEE, 2014.

[Sarsenbayeva et al., 2019] Zhanna Sarsenbayeva, Niels van Berkel, Danula Hettiachchi, Weiwei Jiang, Tilman Dingler, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. Measuring the effects of stress on mobile interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1), 2019.

[Scherer, 2005] Klaus R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[Schmidt et al., 2018] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. Labelling affective states "in the wild": Practical guidelines and lessons learned. In *Proceedings of the ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 654–659, 2018.

[Shaffer and Ginsberg, 2017] Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258, 2017.

[Shi et al., 2010] Yuan Shi, Minh Hoai Nguyen, Patrick Blitz, Brian French, Scott Fisk, Fernando De la Torre, Asim Smailagic, Daniel P. Siewiorek, Mustafa al'Absi, Emre Ertin, et al. Personalized stress detection from physiological measurements. In *International Symposium on Quality of Life Technology*, pages 28–29, 2010.

[Shoumy et al., 2020] Nusrat J. Shoumy, Li Minn Ang, Kah Phooi Seng, D. M. Rahaman, Motiur, and Tanveer Zia. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, 149:1–26, 2020.

[Sidney et al., 2005] K. Dmello Sidney, Scotty D. Craig, Barry Gholson, Stan Franklin, Rosalind Picard, and Arthur C. Graesser. Integrating affect sensors in an intelligent tutoring system. In *Affective Interactions: The Computer in the Affective Loop Workshop at Intl. Conf. on Intelligent User Interfaces*, pages 7–13. AMC Press, 2005.

[Singh et al., 2018] Ashlesha Singh, Chandrakant Chandewar, and Pranav Pattarkine. Driver drowsiness alert system with effective feature extraction. *International Journal for Research in Emerging Science and Technology*, 5(4):26–31, 2018.

[Smith and Smith, 1991] Karl U. Smith and Thomas J. Smith. A study of handwriting and its implications for cognitive considerations in human-computer interactions. *International Journal of Human-Computer Interaction*, 3(1):1–30, 1991.

[Smith, 2017] Leslie N. Smith. Cyclical learning rates for training neural networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, 2017.

[Solso and King, 1976] Robert L. Solso and Joseph F. King. Frequency and versatility of letters in the English language. *Behavior Research Methods & Instrumentation*, 8(3):283–286, 1976.

[Ståhl et al., 2005] Anna Ståhl, Petra Sundström, and Kristina Höök. A foundation for emotional expressivity. In *Proceedings of the Conference on Designing for User Experiences*, page 6. AIGA: American Institute of Graphic Arts, 2005.

[Stieger et al., 2018] Mirjam Stieger, Marcia Nißen, Dominik Rüegger, Tobias Kowatsch, Christoph Flückiger, and Mathias Allemand. PEACH, a smartphone- and conversational agent-based coaching intervention for intentional personality change: Study protocol of a randomized, wait-list controlled trial. *BMC Psychology*, 6(1):1–15, 2018.

[Sun et al., 2016] Wenjun Sun, Siyu Shao, Rui Zhao, Ruqiang Yan, Xingwu Zhang, and Xuefeng Chen. A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement*, 89:171–178, 2016.

[Tomar, 2006] Suramya Tomar. Converting video formats with FFmpeg. *Linux Journal*, 2006(146):10, 2006.

[Tomkins, 1962] Silvan S. Tomkins. *Affect Imagery Consciousness: Volume I: The Positive Affects*. Springer publishing company, 1962.

[Voigt and Von dem Bussche, 2017] Paul Voigt and Axel Von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 2017.

[Wagner et al., 2009] Johannes Wagner, Elisabeth André, and Frank Jung. Smart sensor integration: A framework for multimodal emotion recognition in real-time. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE, 2009.

[Walk and Walters, 1988] R. D. Walk and K. L. Walters. Perception of the smile and other emotions of the body and face at different distances. *Bulletin of the Psychonomic Society*, 26(6):510, 1988.

[Wang et al., 2006] Chen Wang, Xiaoyan Sun, Feng Wu, and Hongkai Xiong. Image compression with structure-aware inpainting. In *IEEE International Symposium on Circuits and Systems*. IEEE, 2006.

[Wang et al., 2014] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14, 2014.

[Ware, 2020] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2020.

[Woebot, 2021] Website, 2021. Retrieved May 13, 2021 from `https://woebot.io`.

[Xiang et al., 2019] Qian Xiang, Xiaodan Wang, Rui Li, Guoling Zhang, Jie Lai, and Qingshuang Hu. Fruit image classification based on Mobilenetv2 with transfer learning technique. In *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, pages 1–7, 2019.

[Yeh et al., 2017] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017.

[Yoo and Yi, 2004] Cheol-Sung Yoo and Sang-Hoon Yi. Effects of detrending for analysis of heart rate variability and applications to the estimation of depth of anesthesia. *Journal of the Korean Physical Society*, 44:561, 2004.

[Zaletelj and Košir, 2017] Janez Zaletelj and Andrej Košir. Predicting students' attention in the classroom from Kinect facial and body features. *Journal on Image and Video Processing*, 2017(1):80, 2017.

[Zautra, 2006] Alex J. Zautra. *Emotions, Stress, and Health*. Oxford University Press, USA, 2006.

[Zendesk, 2021] Website, 2021. Retrieved May 13, 2021 from `https://www.zendesk.com`.

[Zeng et al., 2008] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2008.

[Zhou et al., 2014] Jianlong Zhou, Kevin Hang, Sharon Oviatt, Kun Yu, and Fang Chen. Combining empirical and machine learning techniques to predict math expertise using pen signal features. In *Proceedings of the ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, pages 29–36. ACM, 2014.

[Zhuang et al., 2009] Y. Zhuang, Y. Wang, Timothy K. Shih, and Nick C. Tang. Patch-guided facial image inpainting by shape propagation. *Journal of Zhejiang University*, 10(2):232–238, 2009.

[Züger and Fritz, 2015] Manuela Züger and Thomas Fritz. Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 2981–2990. ACM, 2015.