TU WIEN Informatics

Automatic Detection and Emoji-Based Communication of Non-Verbal Cues during Online Presentations

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Peter Oberhauser, BSc
Matrikelnummer 1363772

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.-Prof. Dr.-Ing. Dipl.-Ing. Sebastian Schlund
Mitwirkung: Dipl.-Ing. David Kostolani
Univ.Prof. Dipl.-Inf. Dr.sc.techn. Florian Michahelles

SIGNATURE INFORMATION	Signatory	Sebastian Schlund
	Date/Time-UTC	2023-12-06T20:28:08+01:00
	Verification	Information about the verification of the electronic signature can be found at: https://www.signaturpruefung.gv.at
Note	This document is signed with a qualified electronic signature. According to Art. 25 para. 2 of the Regulation (EU) No 910/2014 of 23. July 2014 ("eIDAS-Regulation") it shall have the equivalent legal effect of a handwritten signature.	

Wien, 7. Dezember 2023

Peter Oberhauser | Sebastian Schlund

TU WIEN Informatics

Automatic Detection and Emoji-Based Communication of Non-Verbal Cues during Online Presentations

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Peter Oberhauser, BSc
Registration Number 1363772

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.-Prof. Dr.-Ing. Dipl.-Ing. Sebastian Schlund
Assistance: Dipl.-Ing. David Kostolani
Univ.Prof. Dipl.-Inf. Dr.sc.techn. Florian Michahelles

Vienna, 7th December, 2023

Peter Oberhauser | Sebastian Schlund

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Peter Oberhauser, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 7. Dezember 2023

Peter Oberhauser

Danksagung

Mein Dank gebührt in erster Linie meiner Familie, ohne deren Unterstützung das Absolvieren dieses Studiums und letztlich das Verfassen dieser Arbeit nicht möglich gewesen wäre. Darüberhinaus bedanke ich mich bei all meinen Freunden und meinem Partner, die mir auch während stressigen Zeiten den nötigen Halt gegeben haben, um diese Arbeit erfolgreich zu absolvieren.

Mein besonderer Dank gebührt allen Betreuern dieser Arbeit, allen voran David Kostolani, der mit seiner Expertise, Hilfsbereitschaft und Kreativität maßgeblich zum Erfolg dieser Arbeit beigetragen hat. Ebenfalls bedanken möchte ich mich bei allen Teilnehmerinnen und Teilnehmern der Fokus Gruppen und User Tests. Ihre Einblicke und Erfahrungen sind ein wesentlicher Beitrag zu dieser Arbeit.

Acknowledgements

First and foremost, I would like to thank my family, without whose support completing this degree and ultimately writing this thesis would not have been possible. I would also like to thank all my friends and my partner, who have given me the support I needed to successfully complete this thesis, even during stressful times.

Special thanks are due to all the supervisors of this thesis, above all David Kostolani, whose expertise, helpfulness and creativity contributed significantly to the success of this thesis. I would also like to thank all the participants of the focus groups and user tests. Their insights and experiences are an essential contribution to this work.

Kurzfassung

Die COVID-19-Pandemie hat den Einsatz von Videokonferenz-Tools wie Zoom oder Microsoft Teams für Online-Präsentationen im akademischen und beruflichen Umfeld beschleunigt. Allerdings treten oftmals Bedenken hinsichtlich der Balance zwischen Interaktivität und Komfort auf, insbesondere wenn es um das Aktivieren der Webcam während Video-Konferenzen und Online-Präsentationen geht. In dieser Arbeit werden Möglichkeiten von Künstlicher Intelligenz und moderner Bildanalysetechniken untersucht, um non-verbale Kommunikation von Teilnehmenden an Online-Präsentationen aus deren Webcam-Stream zu erkennen, um dadurch Rückschlüsse u.A. auf die Aufmerksamkeit der anwesenden Personen ziehen zu können. Darüber hinaus wird untersucht, wie diese Informationen effektiv an Präsentierende kommuniziert werden können, um die Einschränkungen, die mit Video-Konferenzen und Online-Präsentationen mit deaktivierter Webcam verbunden sind, zu überwinden.

Diese Arbeit wurde in einer mehrstufigen Methodik umgesetzt: In der ersten Phase wurden Erkenntnisse aus vorhandener Literatur und akademischen Projekten zusammengefasst, um Anforderungen für einen funktionalen Prototyp zu entwickeln. Fokusgruppen mit Personen, die Erfahrung mit Online-Präsentationen haben, wurden durchgeführt und deren Ergebnisse flossen ebenfalls in das Design des Prototypen ein. Dieser Prototyp wurde anschließend mit Hilfe von Fragebögen und Interviews untersucht, um die Benutzerfreundlichkeit und die Zufriedenheit der Benutzer zu bewerten.

Diese Arbeit soll dazu beitragen, die Lücke zwischen Interaktivität und Komfort bei Online-Präsentationen zu verkleinern und Möglichkeiten moderner Bildanalysetechniken in diesem Kontext zu untersuchen. Der entwickelte Prototyp ist ein Lösungsvorschlag, der Feedback für Vortragenden verbessert, wertvolle Erkenntnisse für das Design von Videokonferenz-Tools liefert und einen Ausblick auf Möglichkeiten der Optimierung des Online-Präsentationserlebnisses gibt. Die Evaluation des Prototypen mit 13 Usern in einer simulierten Online-Präsentation ergab, dass ein Großteil der Teilnehmenden den zusätzlichen Feedback-Kanal positiv sah. Außerdem gab ein Großteil der Zuhörenden an, dass sie die automatische Zustands-Erkennung komfortabler als Online-Präsentationen mit aktivierter Webcam empfanden. Diese Arbeit eröffnet zahlreiche Wege für zukünftige Forschung, vor allem im Bereich von Mensch-KI-Interaktion.

Abstract

The COVID-19 pandemic has accelerated the adoption of video conferencing tools, like Zoom and Microsoft Teams, for online presentations in academic and professional settings. However, concerns regarding the balance between interactivity and comfort have arisen, especially when it comes to sharing the video stream during online presentations. This thesis explores an approach of utilizing image analysis techniques to automatically detect non-verbal cues of participants of online presentations from their webcam stream, to infer knowledge about their engagement and confusion while listening to a presentation, reducing the need of having to share their webcam video. Furthermore, the thesis studies how this information can be communicated effectively to the presenter to address the limitations associated with camera-disabled presentations.

The thesis presents a multiphased methodology: the initial phase involves synthesizing insights from existing literature and related projects to develop a functional prototype. Focus groups with participants experienced in online presentations were conducted to further inform the design of this prototype. The prototype was subsequently subjected to user studies comprising questionnaires and interviews to assess its usability and subjective user experience.

This work aims to contribute by bridging the gap between interactivity and comfort in online presentations. It offers a solution proposal that enhances feedback for presenters, provides valuable insights for the design of future video conferencing tools, and offers an outlook to optimize the online presentation experience. Evaluation of the prototype with 13 participants in a simulated online presentation environment showed that experienced presenters value the additional feedback stream. Furthermore, listeners being subjected to automatic detection of engagement and confusion rated the technology to be more comfortable than having to share their webcam video during the presentation. This thesis paves the way for further work, particularly into the field of Human-AI interaction.

Contents

CHAPTER 1

Introduction

1.1 Problem Statement

The COVID-19 pandemic and accompanying measures to contain the virus have generated unprecedented interest in video conferencing tools such as Zoom or Microsoft Teams [Tud22]. Although many measures have since been lifted, video conferencing tools remain widely accepted to communicate in workplaces, schools, and beyond. However, many users feel unease when having a camera turned on, i.e. throughout online classes, for reasons such as fears of being exposed or simply a desire for privacy at home [GSP21, LRRX+22]. Furthermore, due to the sudden increased usage of video conferencing tools during the COVID-19 pandemic, the term *Zoom Fatigue* gained traction. It describes the notion that certain qualities of video conferencing applications make them very exhausting for participants after a while. Jeremy N. Baileson associates Zoom fatigue mostly with different aspects in regards to observing the video stream, i.e. excessive amounts of close-up eye gaze, the effort related to intentionally sending and receiving cues via video, increased self-evaluation by seeing yourself in video or constraints on physical mobility caused by the need to stay in view of the camera [Bai21]. To tackle Zoom fatigue, Baileson even suggested making audio-only Zoom meetings the default.

However, having cameras turned off can result in meetings feeling less interactive and ultimately less practical due to missing information usually transmitted via non-verbal cues. This is especially true for online presentation scenarios, where the person presenting is unable to receive any kind of non-verbal feedback from participants. Furthermore, using cameras in online classes has been shown to increase levels of trust and social presence, which consequently encouraged dialog [SPVA22]. To summarize, there seems to be a trade-off between video-enabled online presentations with feedback for presenters and video-disabled presentations with less feedback for presenters but more comfort for audiences.

Image classification, face recognition, and emotion recognition techniques have evolved significantly in recent times. These techniques hold the potential to be used to extract information from a webcam stream, which in turn could be used to provide nuanced feedback to presenters during online presentation conferences without the need to transmit the video feed itself.

Automatically detecting the affective states of listeners to online presentations is a challenging task in itself. However, effectively communicating this information to the presenter and facilitating comprehension while not being distracting is also not trivial. Several mediums can be used to convey non-verbal information, such as emojis, virtual avatars, typography, or even by coding information with different colors. Emojis, or *picture characters* literally translated from Japanese, are pictograms that have been developed to add emotional cues to text messages [Fre18]. That is, they are an effective way of conveying non-verbal cues, and their usage in text-based, asynchronous communication has already been studied fairly extensively [Man21, BEE20, Eld18, CPR+22]. However, the meaningful visualization of non-verbal information from online presentation listeners, which can provide a helpful feedback channel to presenters, is still an open challenge.

1.2 Expected Results

This work aimed to explore the feasibility of developing a system capable of automatically detecting and transferring non-verbal cues during online presentations. To inform the design of the prototype, focus groups with people experienced in conducting online presentations were conducted to give insights into the kind of information that is usually transmitted non-verbally during online presentations. Subsequently, state-of-the-art techniques to automatically detect non-verbal cues were explored. Moreover, it was studied whether Emojis are a suitable medium to convey non-verbal information during online presentations. Another major question of this work was how to incorporate the resulting technology in a web-based video conferencing tool and whether this provides benefits to the online presentation experience, particularly to presenters.

The aim of this work was finding answers to the following research questions:

RQ1 What information do speakers miss in an online presentation setting when their audience webcams are disabled?

RQ2 Can non-verbal cues of participants of online presentations be detected automatically and communicated to the presenter via emojis?

RQ3 How do users perceive the usefulness of non-verbal cues sent during online presentations?

RQ4 Do users feel comfortable with automatic non-verbal feedback detection when participating in online presentations?

1.3 Methods

1.3.1 Literature & State of the Art Review

First, a literature review was conducted to find information on non-verbal communication in video conferencing tools. Keywords to search academic databases included: *computer mediated communication, emoji, non-verbal cues, video conferencing, emotion detection, online presentation*. Databases that were used include: the Vienna University of Technology research portal *CatalogPlus*[1] and the search engine for scholarly literature *Google Scholar* [2]. For the technological State-of-the-Art review a general web search on Google and platforms such as GitHub was conducted.

[1] https://catalogplus.tuwien.at
[2] https://scholar.google.com/

1.3.2 Focus Groups

To inform the design of the prototype, two focus groups with participants experienced in holding online presentations were conducted. The participants were recruited from two departments of the Vienna University of Technology. The key aim of holding the focus groups was to get first-hand experiences and feedback from potential users of the proposed system to help inform the design of the prototype. The process was inspired by Participatory Design methods. Participatory Design is a method of incorporating different stakeholders, i.e. prospective users, in the design phase of a system to facilitate design decisions. Involvement of stakeholders in this setting should happen recurrently and on several occasions, however, the scope of a Master's thesis limits is a somewhat limiting factor. Nevertheless, participatory methods enjoy increasing popularity when it comes to designing AI-based technologies [ZJWG+22].

1.3.3 Prototype Development

Exploratory development of a prototype online conferencing tool with automatic detection of listeners' engagement. The prototype will be inspired by the existing prototype of Low-Bandwidth Video Chat developed in the course *Building Interaction Interfaces*. However, the non-verbal cue detection will be developed from scratch, and the integration into a prototypical online presentation tool will be completely revamped.

1.3.4 Prototype Evaluation

The resulting prototype will be subjected to testing within a simulated online presentation environment. During this evaluation, a combination of questionnaires and semi-structured interviews will be used to collect feedback from users. The primary objective is to gain insights into the subjective experience regarding the usefulness of the proposed technology. This evaluation aims to provide valuable insights into the potential utility of the new technology as perceived by its target audience, leading to additional feedback and guidelines for future work.

1.4 Structure of the Work

This thesis is organized into eight sections that collectively address the research objectives and aim to contribute to a holistic understanding of the subject. The following summary outlines the key areas covered in each section:

1. Introduction: The opening section provides a contextual introduction to the subject, articulating the problem statement, anticipated outcomes, methodological approach of the thesis, and an overview of the structure of the work.

In chapter **2. Theoretical Foundations**, the theoretical underpinnings of the research area are explored. The discussion encompasses the significance of emotion recognition,

the role of emojis in expressing emotions, and an overview of other related research of interest.

In chapter **3. State of the Art** reviews the contemporary landscape of emotion recognition. It examines advanced techniques in emotion recognition, their applicability to edge computing devices, and the current state of popular video conferencing tools.

Chapter **4. Prototype Design** provides insights into the design phase of the developed prototype. Drawing from literature, the state of the art, and insights garnered from focus groups, the approach of how the concept of the prototype was designed is illustrated.

Details about the technical implementation of the prototype can be found in chapter **5. Prototype Implementation**. The discussion commences with an architectural overview and subsequently delves into the image classification module, a major component of the prototype.

The chapter **6. Evaluation** centers on how the usability of the prototype was evaluated in a user study and presents the results of these tests. Chapter **7. Discussion** encapsulates the study's findings. Additionally, it provides a platform for discussing implications arising from the research as well as potential limitations. The final chapter **8. Conclusion** gives a concluding summary of the contributions of this thesis.

CHAPTER 2

Theoretical Foundations

2.1 Current Challenges in Video Conferencing

In recent years, the widespread adoption of video conferencing platforms like Zoom or Microsoft Teams has revolutionized the way we communicate and collaborate in workplaces, schools or universities. However, with the convenience of these technologies have come several associated challenges. This chapter explains the term Zoom fatigue, a phenomenon characterized by feelings of mental and physical exhaustion following prolonged virtual meetings. While several factors contribute to this phenomenon, one critical aspect is the ubiquitous use of webcams. The chapter explores how constant webcam usage during video conferences can contribute to Zoom fatigue and what other factors of webcam usage bring along negative experiences for participants. Additionally, it contemplates the concept of video conferencing without webcams and the potential advantages of such a shift in virtual communication and online presentations.

2.1.1 Zoom Fatigue

With the increasing use of video conferencing tools such as Zoom, exacerbated by the COVID-19 pandemic, the term *Zoom Fatigue* gained traction. It describes the phenomenon that some people perceive online video conferencing to be more exhausting and tiring than offline face-to-face meetings. In a theoretical argument, professor of communication Jeremy Bailenson attributes this to four characteristics of online meetings [Bai21]:

- **Eye Gaze at a Close Distance**
  The interface of video conferencing tools, which usually shows videos of participants looking straight into the camera from a close distance, can lead to the impression that participants are looking at you at all times. Whereas in a face-to-face scenario, you only receive direct eye gaze sporadically, i.e. when you are speaking.

- **Cognitive Load**
  Certain aspects of online video conferencing tools lead to additional cognitive load, compared to face-to-face conversations. An example would be the extra effort of sending and receiving non-verbal cues. Due to the limited ways of communicating non-verbal cues (i.e. head and body pose, or other contextual information is missing), successful communication can get more tedious.

- **An All Day Mirror**
  Video conferencing tools let you observe your webcam stream similar to watching in a mirror by default. There is no pre-video-conferencing scenario where people are confronted with having to observe images of themselves throughout extended periods.

- **Reduced Mobility**
  Due to the limited view and stationary nature of most webcams, participants of video conferences are more or less locked to their desks. Additionally, temporarily disabling the webcam i.e. to briefly leave your desk can in some cases be interpreted negatively

2.1.2 Privacy Concerns related to Webcam Usage

While video conferencing tools offer an efficient way to communicate irrelevant of the participants' location, they involve several privacy concerns. Many of the concerns are related to the transmission of users' webcam videos since they can hold sensitive information such as the participant's age, gender, race, or a private glimpse into the participant's personal space such as their apartments or unintended appearances of family members. Kagan et al. [KAF22] demonstrated that it is possible to automatically extract the personal information of video conferencing participants using image processing, text recognition, and social network analysis. They showed several ways in which malicious actors can extract valuable and sensitive information from publicly available screenshots of video conferences. In a showcase, they collected screenshots of video conferences from publicly available sources like social media or search engines. From the resulting data set, sensitive information like name, gender, age, geographic location, or information derived from the visible background was extracted using state-of-the-art image processing techniques. Using face recognition, network graphs of connections between different users throughout different sessions were generated. Using the information extracted in the first step, an attempt to link participants to their social media profiles was conducted. This showcase indicates that a concerning amount of information can be automatically derived, by analyzing the webcam images only. A survey among 484 professionals, however, indicates that a privacy paradox among users of video conferencing tools exists. While many users do have privacy concerns, they continue to use the tools when the perceived benefits of the tools outweigh the potential privacy concerns [SVGTO23]. However, the authors of the study suggest that there is significant room for improvements regarding privacy in video conferencing tools. Furthermore, they suggest

that organizations using video conferencing tools should put in place guidelines that mitigate these risks [SVGTO23].

2.2 Emotion Recognition: Theoretical Foundations

Emotion recognition techniques can provide ample opportunities in the realm of online presentations, offering a novel way to understand and respond to participants' non-verbal cues and emotional states without the need to directly share their webcam video. By analyzing and categorizing the webcam input, this technology can decipher emotions, track engagement levels, and provide valuable insights into audience reactions. The following chapter will introduce the topic of emotion recognition and lay a path of what opportunities arise in the context of online presentations.

The ability to recognize the emotional state of fellow human beings is a pivotal skill to thrive in society. That is, humans generally need to be very good at recognizing subtle cues that can be used to make inferences about the emotional state of their counterparts. Even though evidence is inconclusive, some studies suggest, that deficits in facial emotion recognition could be associated with a propensity to violence [BBG+20].

Emotion Recognition and Detection is a popular research field since it can provide many interesting applications for Human-Computer Interaction research. Saxena et al. [SKG20] compiled a review of emotion recognition research and found that most emotion detection methods base their predictions on one or more of these four input factors: *1) Physiological Signals*:, i.e. electrocardiographic signals (ECG) or electrodermal activity (EDA), *2) Text*: textual emotion recognition with Natural Language Processing (NLP), *3) Speech* or *4) Facial Expressions*. The context of conventional video conferencing systems reduces these potential input sources. Capturing ECG or EDA signals would require additional hardware and would be very intrusive. Emotion recognition from participants' text or speech would be feasible during an online conference with balanced collaboration. However, during online presentations, where information flows mostly uni-directional, from the presenter to the listeners, inferring information from text or speech of listeners is not possible due to a lack of input. That is, the only remaining input modality for inferring information about the listeners of online presentations is their facial expressions. Subsequently, research on facial expressions, some historical context as well as important models of categorizing emotion will be introduced.

2.2.1 Research on Facial Expressions of Emotions

The human face is a particularly expressive, but also hard-to-decipher outlet of human emotion. This section will give a brief overview of the development of facial expression research. Furthermore, an introduction to important emotion and affect theories, classification systems, and dimensions will be given.

Early Modern Emotion Research

Human emotions have been of interest to philosophers since ancient times, with both Plato and Aristotle significantly influencing early modern theories of emotions. Descartes shaped early modern understanding of emotions with his book *Passions of the Soul* (1649) [Sch21]. He identified six primitive emotion categories: wonder, love, hatred, desire, joy, and sadness. Additionally, an infinite number of other emotions could be formed by combining any of the primitive emotions, or passions as he called them. According to Descartes, each primitive emotion, with the exception of wonder, has an embedded direction of motion, being either appetitive or aversive towards something. *Love* (appetitive) is opposed to *Hatred*(aversive), *Joy*(appetitive) is the opposite of *Sadness* (aversive), *Desire* (appetitive) has no direct opposite and *Wonder* has no embedded direction of motion at all (figure 2.1) [Sch21].

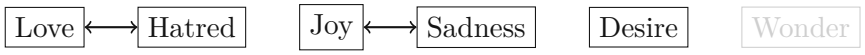


Figure 2.1: Classification of the Passions based on Descartes [Sch21]

Important early modern work on facial expressions was conducted by the French painter and chancellor of the Royal Academy of Painting and Sculpture in 17th-century France, Charles LeBrun. He drew heavily on Descartes' theories of the passions but put more emphasis on the bodily expression of emotions, even though Descartes had reservations when it came to inferring conclusions about the passions from bodily expressions alone [Ros84]. LeBrun, being a painter first and foremost, provided an interesting and immensely popular guide for expressing the passions in paintings. An overview of some of Lebrun's example illustrations can be found in figure 2.2. To give an example, for the passion of *Anger*, LeBrun gave the following explanation:

> When anger possesses the soul, whoever feels this passion has red and inflamed eyes, the pupil distracted and sparkling, the eyebrows sometimes lowered, sometimes raised, one like the other. The forehead will appear strongly creased with folds between the eyes, the nostrils will appear open and enlarged, the lips press against one another and the lower lip surmounts the upper leaving the corners of the mouth a little open, forming a cruel and disdainful laugh.
>
> He will seem to grind his teeth, saliva will appear in his mouth, his face will be pale in some places and inflamed in others and all swollen. The veins of the forehead, temples, and neck will be swollen and taut, the hair bristling. He who feels this passion puffs instead of breathing because the heart is oppressed by the abundance of blood which comes to its rescue [Ros84].

Although LeBrun's work was very influential, his theoretical framework for explaining the mechanics of facial expressions of emotions lacks complexity and has been criticized both in the past and present [Ros84].

Figure 2.2: Charles LeBrun, The Expressions, Public Domain

Modern Emotion Research

Paul Ekman holds significant importance in the field of Emotion Research in modern times specifically when it comes to aspects of facial expressions. Ekman conducted cross-cultural, empirical studies and developed methods to measure facial expression. In his cross-cultural studies, he showed, among others, that members of independent groups of people from around the world were mutually proficient in identifying affective states solely from their respective facial images [Ekm93]. He formulated his theory of basic emotions (BET). The theory states that there are at least seven basic emotion families that are genetically encoded and universal to all humans, irrespective of culture. These basic emotions manifest in so-called *affect programs*, which include bodily or facial expressions, which are executed once the affect program is triggered [Col14, p. 26-28]. Tracy and Randles reviewed four major contemporary basic emotion models, including Ekman's. They found that the authors of all reviewed models have a similar understanding of what qualifies as basic emotion: a) *a basic emotion should be discrete*, b) *entail a fixed set of neural and bodily expressed components*, and c) *have a fixed feeling or motivational component* [TR11]. Consequently, their lists of basic emotions, identified through theoretical reasoning and/or empirical studies, have significant overlaps as well (see table 2.1).

However, it is worth noting that inferring a person's emotional state from facial expressions alone is not free from criticism. Empirical evidence suggests that there is variation in which emotions are expressed in facial movements, which may be influenced by cultural

Izard	Panksepp &Watt	Levenson	Ekman & Cordaro
Happiness	Play	Enjoyment	Happiness
Sadness	Panic/Grief	Sadness	Sadness
Fear	Fear	Fear	Fear
Anger	Rage	Anger	Anger
Disgust		Disgust	Disgust
Interest	Seeking	Interest*	
Contempt*			Contempt
	Lust	Love	
	Care	Relief*	Surprise

*) definitive evidence still outstanding according to resp. author

Table 2.1: Basic emotion models: discrepancies and similarities [TR11]

norms, situational context, or even individual differences [BAM+19].

Valence and Arousal

An even more fundamental framework of affect was introduced by Wilhelm Wundt. He argued that affective states can be categorized in as little as two dimensions: *Valence* and *Arousal*. Valence indicates how pleasant or unpleasant an emotional experience feels. Arousal refers to the level of intensity or activation of an emotional experience.

Throughout the decades, many hypotheses about the relationship between valence and arousal have been suggested. Kuppens et al analyzed 8 datasets to find evidence of a universal relation between valence and arousal in subjective experience. However, while they did identify a non-linear pattern, they concluded that, due to large individual differences, it was unlikely that there is a universal relationship [KTRB12].

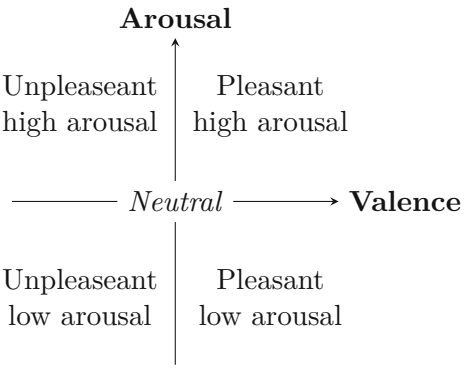


Figure 2.3: Valence and Arousal in Emotion Theory[KTRB12]

2.2.2 Facial Emotion Recognition Critique

Many facial emotion detection approaches, particularly facial recognition techniques, are based on assumptions of Ekman's Theory of Basic Emotion (BET) with the aim of identifying one of Ekman's emotion families.

It should be emphasized that some scholars argue that automatic facial emotion recognition should be questioned in general. They argue, that from a constructivist point of view, facial expressions do not carry intrinsic meaning and that emotions instead are socially constructed. Consequently, the human perceiver infers emotional meaning based on facial expressions and various contextual information [TD21]. This view, however, is in contrast to e.g. Ekman's findings.

Sharon Richardson, a scientist with a focus on the influence of data and technology on cognitive behavior, voiced her concern when it comes to affective computing applications in general. She pointed out that the validity of emotion recognition technology is questionable, and that contemporary research was inconclusive on that topic. Furthermore, affective computing created unprecedented challenges to individual privacy, especially regarding the data necessary to train the systems, according to Richardson [Ric20].

2.2.3 Additional Online Participation Metrics

In the context of online presentations, meaningful metrics of participation extend beyond mere attendance and interaction counts. Interesting factors of the online learning experience include e.g. assessing student engagement and identifying moments of student confusion during presentations. Several systems to estimate engagement have been proposed. Sharma et al. proposed a system to estimate engagement by combining information about emotional states, eye tracking, and head movement of students. Such metrics can be valuable feedback that allows instructors to reflect on their lectures. Cavalcanti et al. have conducted a systematic literature review about automatic feedback collection in online learning environments. They found that proposed tools and approaches are often tailored to support learners by providing insights into their progress and understanding of the material, rather than focusing on generating feedback for instructors [CBC+21]. The reviewed works' approaches to generate feedback were mostly comparing students' works to desired outcomes.

Fredericks et al. have characterized student engagement as a multifaceted construct [FBP04]. They proposed three types of engagement: *Behavioral Engagement*, *Emotional Engagement*, and *Cognitive Engagement* (see figure 2.4). Behavioral Engagement is described as observable actions like following rules and classroom norms or even the absence of disruptive behavior. Emotional Engagement is categorized by affective reactions from students towards the school, teacher, or class. Different researchers have identified various reactions that can be shown, i.e. interest, boredom, happiness, anxiety, etc. Cognitive Engagement is categorized by the investments and effort students are willing to take to learn and the motivation and resilience they show even when faced with failure. Fredericks et al. gave examples of measurement techniques for all three engagement

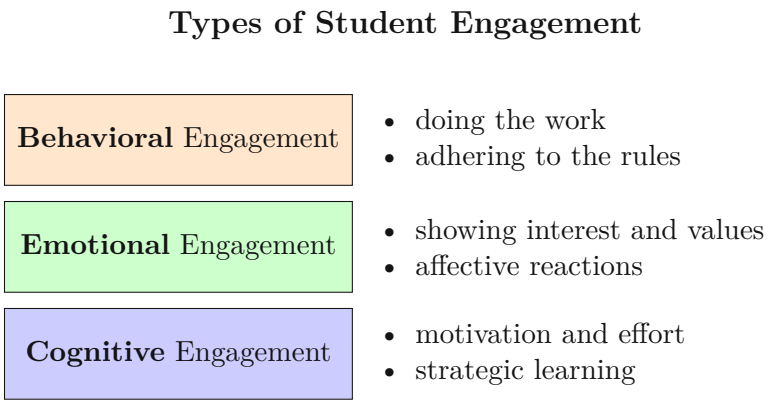


Figure 2.4: Student Engagement Model [FBP04]

types. Behavioral engagement can be measured to some extent by teacher ratings or self-report surveys. Emotional engagement can be assessed by i.e. questionnaires about emotions related to school and people. Measurement of cognitive engagement is the most difficult quality to assess. However, Fredericks et al. emphasize that all techniques measuring student engagement have associated problems and that individuals may have very different needs towards their learning environment to achieve success [FBP04].

Advancements in machine learning (ML) and artificial intelligence (AI), particularly in image processing and computer vision, have opened up new avenues for assessing engagement levels through webcam images. These technologies empower AI systems to analyze facial expressions, eye movements, and other non-verbal cues to gauge student engagement. The training of AI-based systems often relies on datasets like DAiSEE, which offers webcam videos of students participating in online presentations, annotated with precise engagement information. This dataset can be used for training and evaluating AI models, enabling them to distinguish subtle variations in engagement or other important metrics in the context of online presentations. These models could provide valuable feedback for presenters or other stakeholders involved in online presentations.

Figure 2.5: Example of Affective States from DAiSEE Dataset [GDAB16]

2.3 Visualizing and Communicating Emotions

Assuming that a system can detect the desired affective states of users, visualizing and communicating these states to stakeholders is a challenge in itself. The subtlety and ambiguity of emotions are hard to code in discrete representational systems like i.e. written text. To communicate non-verbal information, several representation systems like emojis or avatars have been introduced. Other, even simpler ways of coding such information are different forms of typography or using color schemes. The following section introduces several of those concepts.

2.3.1 Emojis

Emojis, or *picture characters* literally translated from Japanese, are pictograms that have been developed to add emotional cues to text messages [Fre18]. That is, they are an effective way of conveying non-verbal cues and their usage in text-based, asynchronous communication has already been studied fairly extensively. Manganari conducted a literature review on the topic of emoji use in computer-mediated communication (CMC). She found that typical scenarios for using emojis in CMC include: *expressing emotions*, *making communication more informal*, *reducing ambiguity*, *communicating the sender's mood*, *expressing boredom or sarcasm*, etc. [Man21]. Elder argues that because of the distinct role that facial expression recognition has in our brain, adding emojis in text messages can affect relationships built through CMC [Eld18]. Beattie et al. found that chatbots incorporating emojis into their messages are rated more favorable by users compared to bots using text only. The same can be said for human messengers as well [BEE20].

There is a de facto standard set of emojis curated by the Unicode consortium, which currently (v15) contains 1874 distinct emojis[1]. Unicode categorized the emojis into high-level categories. Relevant high-level categories for expressing human emotions or affective states are i.e. the categories *Smileys & Emotion* or *People & Body*.

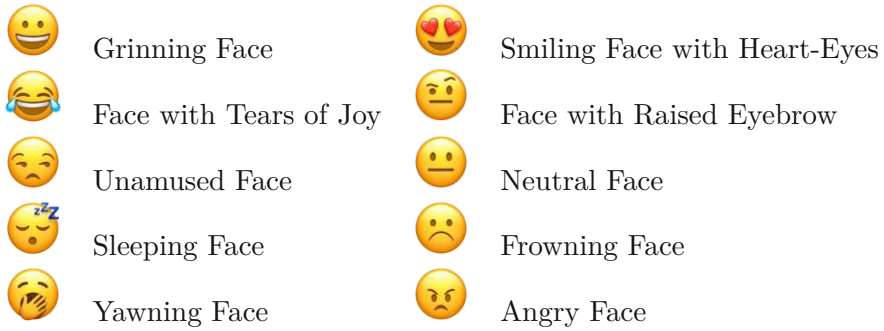


Figure 2.6: Examples from the Unicode emoji list

[1] https://unicode.org/emoji/charts/full-emoji-list.html

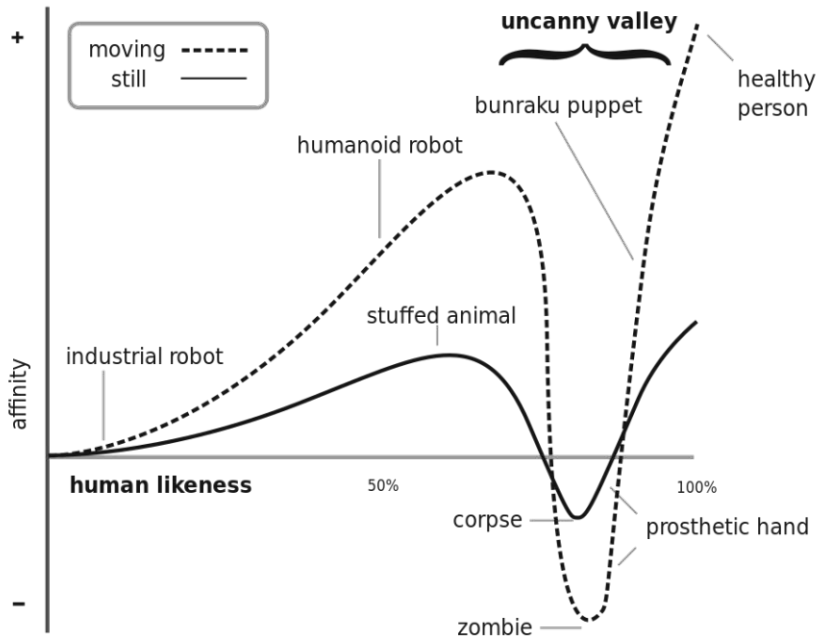


Figure 2.7: The Uncanny Valley, `https://commons.wikimedia.org/wiki/File:Mori_Uncanny_Valley.svg`, CC-BY-SA

2.3.2 Virtual Avatars

Virtual avatars are similar to emojis in being able to transfer non-verbal emotional cues visually. However, they can provide more fine-grained, less discrete ways to communicate non-verbal cues since they can be controlled extremely finely. One way to categorize virtual avatars is by their grade of resemblance to humans or by their realism. For some time there was resistance to creating realistic virtual avatars caused by the phenomenon called *uncanny valley*. The term describes that users showed aversion for avatars that are almost but not quite realistic (see figure 2.7), caused i.e. by prevailing flaws of animation techniques. Recent studies, however, suggest that due to improvements in animation techniques the so-called uncanny valley may be bridged [SYD+21].

2.3.3 Typography and Color

Beyond the use of visual representations like emojis or avatars, simple text formatting such as bold, italics, or capitalization can be a way to emphasize certain words or phrases to convey emotions. Serafini et al. argue that typography in itself is a semiotic resource in itself with several meaningful potentials [SC12]. They identified several typography dimensions that can carry meaning: font weight, color, size, slant, framing, the formality of the font, and font flourishes or additions.

Color is a powerful medium for conveying emotions as well. A study showed that in general brighter colors are associated with positive emotions while darker colors tend to be associated more negatively. Furthermore, it revealed that children tend to attribute color more positively compared to adults [Hem96]. Another more recent, study [KE04] asked college students to associate a given color with a specific emotional response. The results (see table 2.2) showed that green was the primary color with the most positive associations. The color red was associated with both positive and negative emotions. The interpretation of color therefore is very much dependent on the context. Furthermore, color-emotion associations can certainly be subject to cross-cultural differences. To summarize, color as such can be a powerful tool to communicate or underscore emotional messages. Due to potential ambiguity or cross-cultural differences it, however, should not be the sole medium of emotion communication.

	Red	Yellow	Green	Blue	Purple
Angry (a)	28.6	0	0	0	0
Bored (a)	0	0	0	0	5.1
Calm (b)	4.1	0	29.6	61.2	28.6
Comfortable (b)	0	0	15.3	4.1	3.1
Depressed(a)	0	0	0	6.1	0
Energetic (b)	5.1	10.2	0	0	0
Excited (b)	18.4	8.2	2.0	0	4.1
Fearful (a)	0	0	0	0	5.1
Happy (b)	21.4	75.5	28.6	10.2	21.4
Hopeful (b)	0	0	8.2	0	0
Lonely (a)	0	0	0	3	0
Loved (b)	15.3	0	0	0	0
Peaceful (b)	0	0	12.2	4.1	0
Powerful (b)	0	0	0	0	7.1
Sad (a)	4.1	0	0	8.2	13.3
Tired (a)	0	6.1	0	0	9.2
No emotion	3	0	4.1	3.1	3.1

(a) negative emotion, (b) positive emotion

Table 2.2: Color Emotion Association (in % of total, excerpt) [KE04]

2.4 Trustworthiness and Acceptance of Artificial Intelligence

In today's world technology and algorithms affect the way people work and interact with each other. In the case of social media, feed algorithms even influence our perception of the world. The rise of artificial intelligence has led to software that is very effective at solving tasks and problems that previously were hard or impossible to compute. However, understanding how these AI systems work and recognizing their limitations is essential for fostering acceptance and ensuring the appropriate use of such technology. Kaur et al. have proposed five requirements for trustworthy AI based on a review of literature and legislation: Fairness, Explainability, Accountability, Privacy, and Acceptance [KURD22]. Below, these requirements and potential harms to be remedied will be explained in more detail:

Fairness – Fair AI systems must avoid bias and discrimination. They should treat all individuals and groups fairly, ensuring that the outcomes they generate do not disproportionately harm or benefit any particular group of people. What tremendous effect decisions made by AI systems can have, shows a report by ProPublica[2]. The report alleged, that a software used in several US states to predict the probability of defendants to recidivate or commit future crimes and which outputs were influencing bail and court decisions, was biased against black people. Though several aspects of the analysis were methodically criticized [FBL16], it shed light on which critical aspects of our lives are subjected to decisions made by algorithms, sometimes completely unaware to the general public.

Explainability – AI systems decisions should be transparent, understandable, and explainable to establish trust towards systems. Explanations of AI systems can roughly be divided into two categories: *Ex-Ante* and *Ex-Post* explanations. *Ex-Ante* explanations should reflect information about the general working of an AI system and should give users an idea about how well-designed, tested, and validated the system is. *Ex-Post* explanations on the other hand should elaborate how an AI system reached one specific decision. According to Kaur et al. there is a significant amount of research conducted concerning generating explanations for AI systems decisions. However, there is little research when it comes to ways to communicate these explanations to users [KURD22].

Accountability – It is important to ensure accountability of the decisions made by AI systems. On the one hand is important to monitor whether decisions made can have harmful consequences. On the other hand, it is vital to be able to identify accountable people and entities, should harmful behavior occur. The question of accountability of AI systems is especially relevant when it comes to systems with high potential risk, i.e. the control software of self-driving cars. Any problems or miscalculations can lead to serious or even fatal accidents. In case this happens, there need to be procedures in place to hold people accountable. Legal liability is part of this conversation. The European Union

[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

proposed new directives for the liability of AI systems in 2022. However, according to medical law experts, the directives show significant gaps that make predicting the risks of both creating and using AI-based systems in medicine impossible to date [KURD22].

Privacy – Privacy is a crucial concern in all stages of designing, deploying, and using AI-based systems. Since these systems require vast amounts of data for training, the privacy of the underlying data has to be ensured at all times. All processes must adhere to stringent data protection regulations, ensuring that individuals' personal information is handled securely and responsibly. Furthermore, it has to be ensured that user consent is obtained whenever necessary. Privacy measures have to be considered both for the users of the developed AI systems and for individuals whose data was used to train the system.

Acceptance – Trust and acceptance from users and society at large are crucial for the successful integration of AI systems. These systems should be designed with user needs and values in mind, and their deployment should align with norms and ethical standards, ensuring they are utilized and employed appropriately.

Some of these characteristics of trustworthy AI can be supported by putting effort into consciously and carefully introducing users to a new technology or tool. User onboarding is a process of guiding new users through the capabilities of an application and helping them understand how to use an application efficiently. Onboarding processes are of vital importance when confronting users with novel applications or systems. A study conducted at the University of Jönköping showed, that onboarding processes in mobile applications have an impact on users' attitude on continued use of the application [EP19].

CHAPTER 3

State of the Art

3.1 Emotion Recognition: State-of-the-Art

Emotion recognition applications, especially in the field of facial emotion recognition (FER), have improved over the last decades together with the rise of deep learning and artificial intelligence research. Within the last decade, many novel facial emotion recognition techniques have shifted from classical image processing and computer vision approaches to neural network-based techniques. The design of neural networks is inspired by the principles of the human brain. Capabilities that a human brain acquires by being immersed in environments and experiences, can be taught to a neural network by systematically training it with vast amounts of data i.e. using Machine Learning (ML) methods. On a high level, machine learning approaches can be categorized as supervised and unsupervised learning, based on whether the training data is labeled or not. An additional way of categorization is by the type of the problem to be solved [ANK18]:

- *Classification Problems* are categorized by having outputs of a fixed, defined number of classes. For instance, classifying emails as spam or not spam or recognizing handwritten digits as numbers from 0 to 9.
- *Anomaly Detection Problems* are aimed to find anomalies within patterns of large amounts of data. The objective is to identify patterns that deviate significantly from the norm, highlighting potential issues or areas of interest within the dataset.
- *Regression Problems* deal with tasks that have continuous or numerical outputs. Unlike classification, where the output is discrete, regression deals with estimating values along a continuous spectrum.
- *Clustering Problems* aim to find patterns and structures within data and provide ways to add new, unseen data within previously identified clusters. Clustering is a valuable tool for uncovering hidden structures and patterns within data.

- *Reinforcement Problems* facilitate i.e. decision-making based on previous experiences. Agents are being trained by rewarding correct and penalizing incorrect behavior. Reinforcement learning is applied in various domains, such as game playing, autonomous robotics, and recommendation systems, where decisions are influenced by previous experiences and feedback.

Among the proposed neural network-based techniques, most approaches were implemented with Convolutional Neural Networks (CNNs) [CMM+22]. CNNs are very useful when it comes to solving image-driven pattern recognition tasks [ON15]. Recurrent Neural Networks (RNNs) are useful for approaches concerning sequential or time-series data since it is a network model with a memory function [Li22].

Facial Emotion Recognition Challenges

The *Emotion Recognition in the Wild* (EmotiW) challenge is a platform for researchers from the field of emotion recognition. The challenge has been conducted almost every year since 2013 and has led to many interesting novel techniques and publications in the field. The main focus of the challenge is studying techniques to recognize emotional states of humans from data captured in real-world contexts as opposed to data captured in a laboratory setting [DGGS16].

The *Affective Behavior Analysis in-the-wild Competition* (ABAW) has been held since 2017 and aims at studying innovative ways of automatically analyzing affect based on a subject's facial expressions. Much like the EmotiW challenge, ABAW distinguishes itself by focusing on real-world data. Instead of relying on controlled environments like laboratory settings, ABAW embraces studying facial expressions observed in natural settings [Kol22]. The findings and techniques are important for a growing demand for emotion analysis in practical applications, such as human-computer interaction, healthcare, and social robotics, where emotions are often expressed in unscripted and diverse scenarios.

Based on submissions to the aforementioned challenges, Savchenko published several pre-trained models[1] for efficient face identification tasks [SSM22].

[1] https://github.com/HSE-asavchenko/face-emotion-recognition

Facial Emotion Recognition Resources

Most image classification models are trained using the supervised learning paradigm. That is, they require pre-labeled training data. The following is a list of important datasets that are used in the domain of facial emotion recognition:

- **FER-2013**
  The *FER-2013* dataset [GEC+13], consists of 35.000+ face images, annotated with Ekman's seven basic emotions, gathered via Google's search API.
- **EMOTIC**
  The *EMOTIC* dataset is a collection of images showing people in a natural environment, as opposed to face images only. Furthermore, the data set is annotated, among others, with 26 emotion categories [KARL20].
- **DAiSEE**
  The *DAiSEE* dataset consists of 9.000+ video snippets from 112 users in an e-learning environment annotated with four affective states: boredom, confusion, engagement, and frustration [GDAB16].
- **AFEW**
  The *Acted Facial Expressions in the Wild* (AFEW) database was created by extracting video snippets from movies. The snippets were selected and categorized by analyzing keywords from movie subtitles and reviewed by human labelers [DGGS16].
- **SFEW**
  The *Static Facial Expressions in the Wild* (SFEW) database was collected by extracting images from the AFEW database using fiducial points-based clustering technique [DGGS16].
- **AffectNet**
  AffectNet is a database containing more than 1 million images of facial expressions collected via search engine queries conducted with emotion-related keywords. Approximately half of the database was annotated manually with seven discrete emotion categories as well as values for valence and arousal [MHM19].

3.2 Emotion Recognition on Edge Computing Devices

Traditionally, artificial intelligence applications had to be run in dedicated cloud infrastructure for most parts, which has several downsides, most notably challenges related to scalability and cost. The capabilities of modern end devices have created opportunities to use their processing power without the need to perform all the processing over the network. Instead, important processing such as inferencing a machine learning model can be done directly on the users' device. Edge computing applications have several advantages. They can decrease latency and bandwidth costs. Furthermore, they can elevate the security and privacy of applications, since data does not have to be sent over the network [SD16].

3.2.1 ONNX Runtime Web

ONNX Runtime Web was developed and is maintained by Microsoft and enables running machine learning models in the ONNX (Open Neural Network Exchange) format directly in the browser, without the need for a server backend [2]. The framework supports both processing on the computers' CPU (WebAssembly) or if available the devices' GPU (webgl or webgpu). There are, however, limitations i.e. regarding the size of the model that have to be considered.

OS/Browser	Chrome	Edge	Safari	Electron	Node.js
Windows 10	wasm, webgl	wasm, webgl	-	wasm, webgl	wasm
macOS	wasm, webgl	wasm, webgl	wasm, webgl	wasm, webgl	wasm
Ubuntu LTS 18.04	wasm, webgl	wasm, webgl	-	wasm, webgl	wasm
iOS	wasm, webgl	wasm, webgl	wasm, webgl	-	-
Android	wasm, webgl	wasm, webgl	-	-	-

Table 3.1: Backend Compatibility ONNX Runtime Web [3]

The ONNX model zoo is a repository maintained by the ONNX project that holds a collection of several pre-trained, state-of-the-art machine learning models that can be deployed easily. The repository encompasses a number of machine learning models, organized into three distinct categories: Vision, Language and Other.

This Vision category includes models for a variety of computer vision tasks, including image classification, object detection, image segmentation, body analysis, face recognition, and gesture analysis. The pre-trained models can be reused to create image classifiers or to develop a real-time face recognition system.

Many Language-related tasks can be solved using artificial intelligence. The ONNX model zoo entails several state-of-the-art models in that regard. From machine comprehension

[2] official documentation: `https://onnxruntime.ai/docs/tutorials/web/`
[3] `https://github.com/microsoft/onnxruntime/tree/main/js/web`

and machine translation to language modeling, the repository offers models that are trained to process and generate human language. These models enable applications such as chatbots, language translation services, and text summarization.

Other models included in the repository are for visual question-answering, dialog systems, speech and audio processing. The pre-trained models provided in the repository can be readily integrated into any machine-learning project with relatively little effort.

3.2.2 Tensorflow.js

Tensorflow.js is an open-source library developed and maintained by Google. It enables developing machine learning models in JavaScript and running them directly in the browser [4]. Tensorflow.js enables the development of machine learning models using the familiar language of JavaScript, bridging the gap between AI and web development. Similarly to the ONNX model zoo, Tensorflow hosts a repository of pre-trained, state-of-the-art models for easy re-use similarly to the ONNX model zoo. The pre-trained models hosted by Tensorflow are categorized into the following types: Images, Audio, Text, Depth Estimation, and General Utilities.

[4] official documentation: `https://www.tensorflow.org/js`

3.3 Emotion in Video Conferencing Tools

According to a survey conducted by the technology advisory T3 tech hub, published by Statista, the video conferencing market was dominated by the tools Zoom (55.44 %) and Microsoft Teams (20.92 %) in 2022 [BB22].

Both Zoom and Microsoft Teams have incorporated functionalities to communicate with emoji-based non-verbal cues, albeit to varying degrees. In both applications, users have the ability to react to text chat messages or images using a predefined set of emojis, a feature that enables subtle expressions of emotion and agreement or disagreement within conversations (as depicted in Figure 3.1).

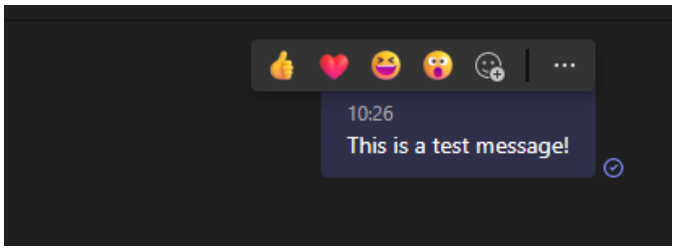


Figure 3.1: Adding Reactions to Text Messages in Microsoft Teams

Zoom even incorporated an emoji-based, animated reaction feature (see figure 3.2). These reactions, whenever manually triggered by participants, are instantly conveyed in real-time to all other members of the video conference, creating an additional non-verbal communication channel. However, none of the popular video conferencing tools have incorporated methods to automatically detect non-verbal cues or emotional cues. Instead, they rely on manual user input for participants to convey emotions, reactions, or non-verbal signals during virtual meetings.

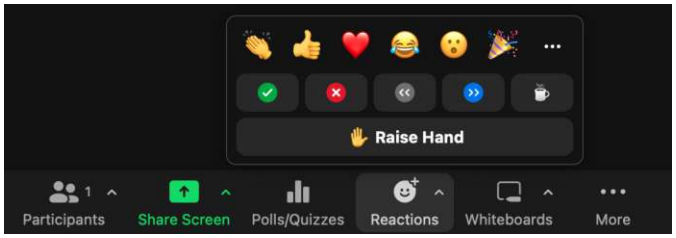
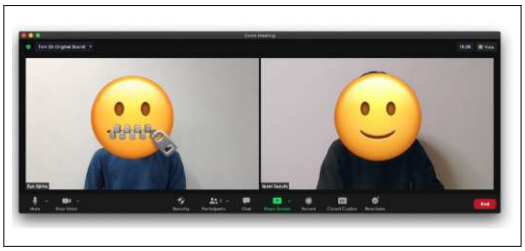


Figure 3.2: Sending Emotion Cues via Reactions in Zoom

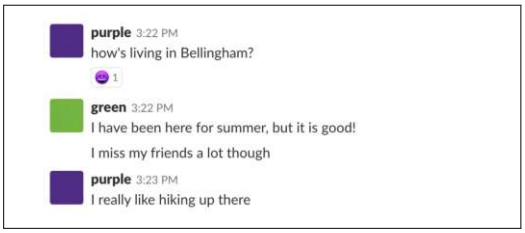
3.4 Emotion Recognition in Video Conferencing

In academia, frameworks and prototypes have been developed that feature automatic detection and communication of non-verbal and affective cues. However, many of them are limited to the detection of Ekman's basic emotions or have other distinct fields of application. Following is a list of projects with similarities to the proposed system:

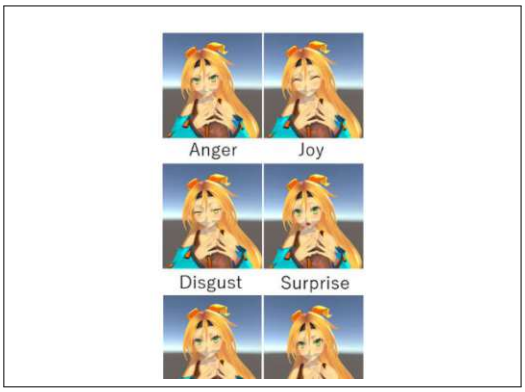
Nawikawa et al. implemented EmojiCam [NSI+21], a facial recognition system that analyzes the video feed of a webcam and overlays an emoji based on seven possible reactions directly onto the video feed. The manipulated video feed was then used to join a Zoom meeting without studying ways to improve usability and optimize interaction in the application itself. Furthermore, the detection of non-verbal cues was limited to Ekman's seven basic emotions. Suzuki et al developed VFep [STT21] which categorizes six basic emotions based on audio input only and generates a 3D model of a face displaying the emotion.



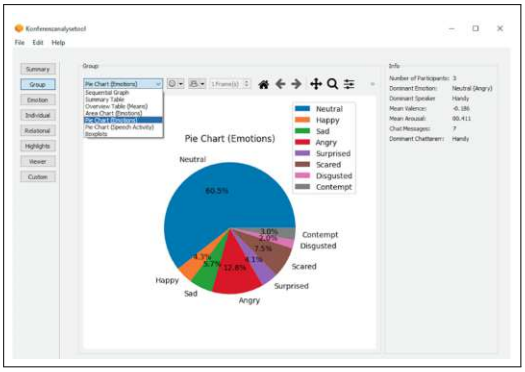
(a) EmojiCam [NSI+21]



(b) ReactionBot [LWP+18]



(c) VFEP [STT21]



(d) Conference Analysis [BMF22]

Figure 3.3: Emotion in Academic Communication Tool Prototypes

Liu et al developed ReactionBot [LWP+18] which similarly uses facial recognition to identify one out of seven emotions via the users' webcam. The detected emotion is then automatically added as a reaction to the latest message in the Slack messaging program. However, so far, implicit transmission of subtle cues in video conferencing using emojis has not been studied in detail and thus remains an open topic.

Lutfallah et al. developed a system that aims to communicate visual, non-verbal cues to the visually impaired [LKHK22]. Their system is able to detect non-verbal emotional cues and categorize portrait videos in the categories *agree*, *neutral*, and *disagree*. They implemented a prototype interface which unfortunately still is visual only, making it not accessible to the visually impaired.

Bissinger et al. explored the possibilities of emotion recognition technology in communication software. They created a prototype conference analysis tool that can visualize detected basic emotions throughout an online communication session. The developed prototype was introduced. However, experiments and evaluation of the prototype were announced to be conducted in future work [BMF22].

Hassib et al. developed EngageMeter [HSE+17], a prototypical system for implicit audience engagement sensing. The system used different brain-computer interfaces (BCIs) to capture information about the brain activity of listeners to a live presentation. The captured data was then used to infer the participants' engagement levels and provide both real-time and post-hoc feedback to the presenter.

CHAPTER 4

Prototype Design

To study the feasibility of automatic non-verbal cue detection and whether a proposed system would be accepted by users, a prototype was developed and subsequently evaluated. The prototype development was conducted in an explorative manner, meaning that the outcome was uncertain, and experimentation with different technologies and approaches was encouraged. The design and development process involved three stages (see figure 4.1). Each ultimately with the goal of formulating requirements and inform the design of a prototype or a future iteration of the prototype, respectively.

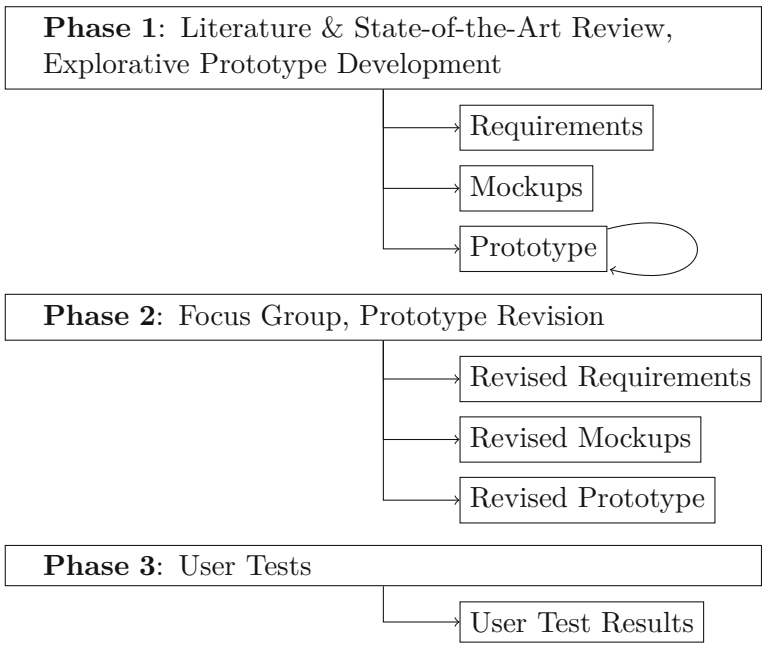


Figure 4.1: Prototype Design & Development Process with Outputs

4.1 High Level Concept

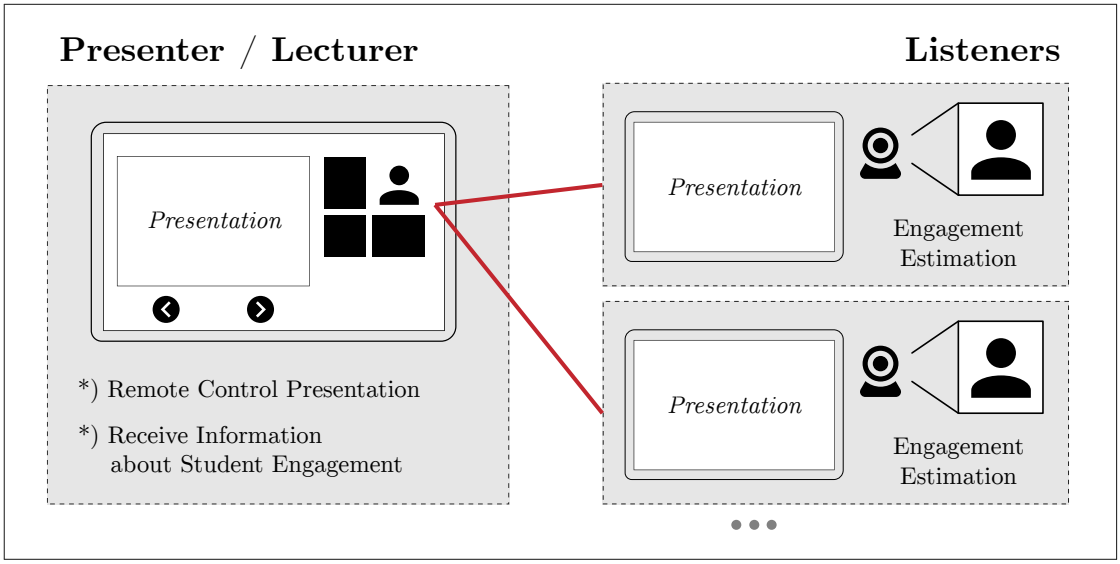


Figure 4.2: Prototype High-Level Concept

The prototype's overall objective is the exploration of innovative techniques for the automatic detection of listener engagement during online presentations. Another major question to be answered by the prototype development is how the captured information may be communicated to the presenters. The underlying goal is to amplify the feedback loop between presenters and their audience without introducing additional distractions or disruptions.

The prototype has two overall user groups: the person holding an online presentation and the listeners invited to the presentation, i.e. in the case of a university lecture, the students. Thus, prototype requires the development of two distinct modes, one tailored to presenters and another one focussing on the needs of listeners. The presenter mode should enable the person presenting to control presentation slides, and invite listeners to their presentation session. Meanwhile, the listener mode places emphasis on providing an unobtrusive view of the presentation, controlled by the presenter. Additionally, it should facilitate the listener to set up their webcam to enable automatic engagement detection. The detection process adheres to the principles of edge computing, ensuring that computations are performed directly on the listeners' end devices. This approach minimizes latency and preserves privacy by transmitting only minimal information to the presenter.

The resulting interface provides presenters with an overview of the current engagement status of their listeners, empowering them to adapt their presentation delivery in real-time. The exact affordances of this overview will be compiled and refined in subsequent design steps.

4.2 Insights from Literature & State-of-the-Art Review

A literature review revealed some interesting impediments of current video conferencing tools, especially several factors related to so-called *Zoom Fatigue* (see section 2.1.1). To mitigate these effects, several requirements in regard to reducing the exposure of users to their webcam video were formulated. Other requirements were aimed at increasing the acceptance of users by providing explanation of the technology and how it works as well as minimizing the amount of information that is transferred among users.

4.2.1 Prototype Requirements

1) Provide functionalities for communicating non-verbal cues without a camera stream as additional communication layer.

2) Hide the webcam view once the user confirms positioning in front of the webcam, to mitigate Zoom fatigue.

3) Provide onboarding to guide users through the prototype and explain functioning to facilitate acceptance.

4) Limit the amount of information transferred to the minimum to facilitate acceptance and privacy by locally processing video data.

4.2.2 Low Fidelity Mockups

Mockups, in particular low-fidelity mockups, are a cheap but effective tool to facilitate the early stages of a development process. Hence, mockups that entail the most important functionalities of the prototype, based on the requirements specified from insights from the literature and state-of-the-art review, were produced.

Since an onboarding process with a brief explanation of the functionality of the prototype was deemed important, the mockups entailed examples of key information relevant to the layout (see figure 4.3). Additionally, the screen represents a webcam view that is shown to the user after the onboarding process is completed, to facilitate proper positioning in front of the webcam. This view will be hidden, however, after the user confirms appropriate positioning, to avoid negative consequences associated with Zoom-Fatigue.

The main view of the application needs to have two different modes. One for the person presenting and one for the participants listening to the presentation. The presenter's view should include the presentation slides and affordances to control the slides, as well as the information gathered from the participants' webcams. When the low-fidelity mockups were drafted, it was still unclear how the information gathered from the participant's webcam could be communicated to the presenter, or even which kind of information would be available. However, one major design decision we were aware of but still undecided about was whether this information should be communicated to the presenter on a granular level, with individual information for each participant, or somehow aggregated

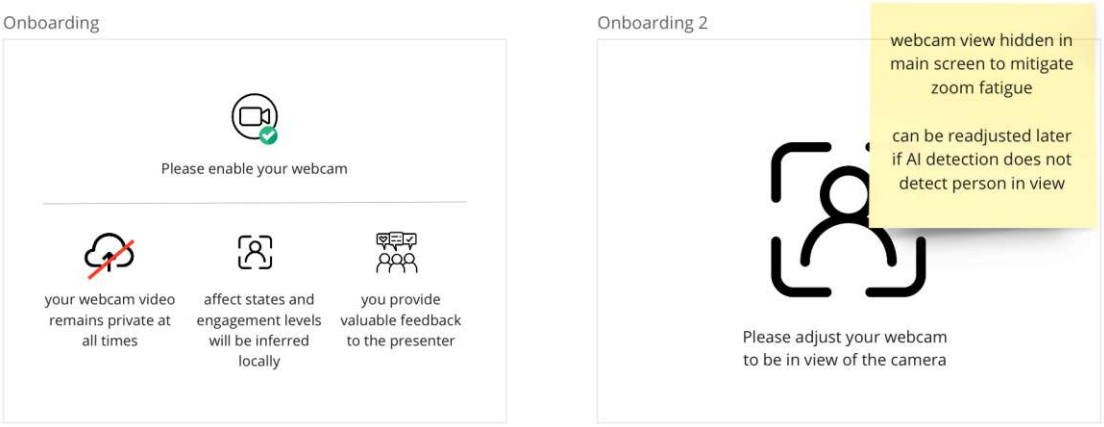


Figure 4.3: Low Fidelity Mockup: Onboarding View

to provide an overview across the entire digital presentation room (see figure 4.6 visualizing the two different approaches). The initial design idea was visualizing each participant with an emoji and communicating the current state of the participant with a discrete emoji. However, during the design phase and particularly after insights from the focus groups (see section 1.3.2), avenues with less discrete coding elements like colors and progress bars were explored.

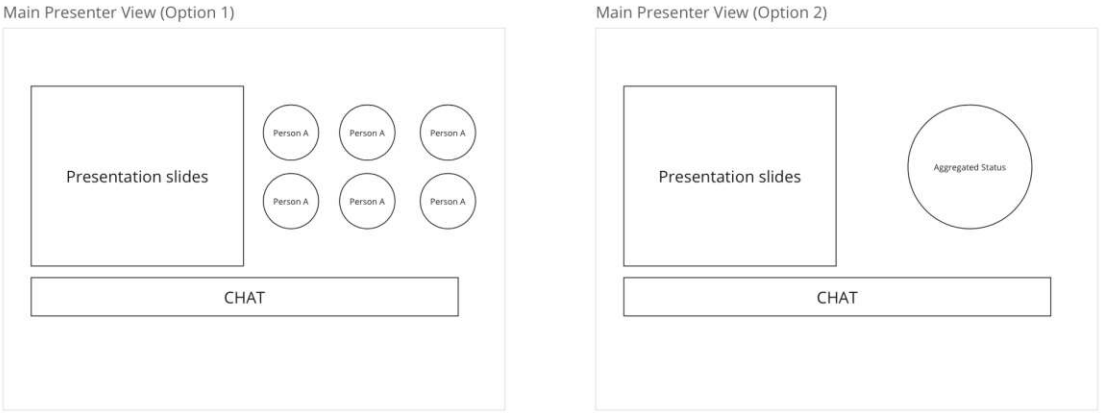


Figure 4.4: Low Fidelity Mockup: Presenter View

The listener view was more straightforward, only containing a view of the presentation slides or the current slide respectively. What was yet unclear regarding the listener view was what should happen, when the automatic detection yielded a status that is associated with low attention or engagement with the presentation. Possible scenarios included notifying the user about the detected status and possibly providing tips for concentration or providing ways for validating the detected state by manual user input.

The mockups of both the presenter and the listener view included a chat functionality

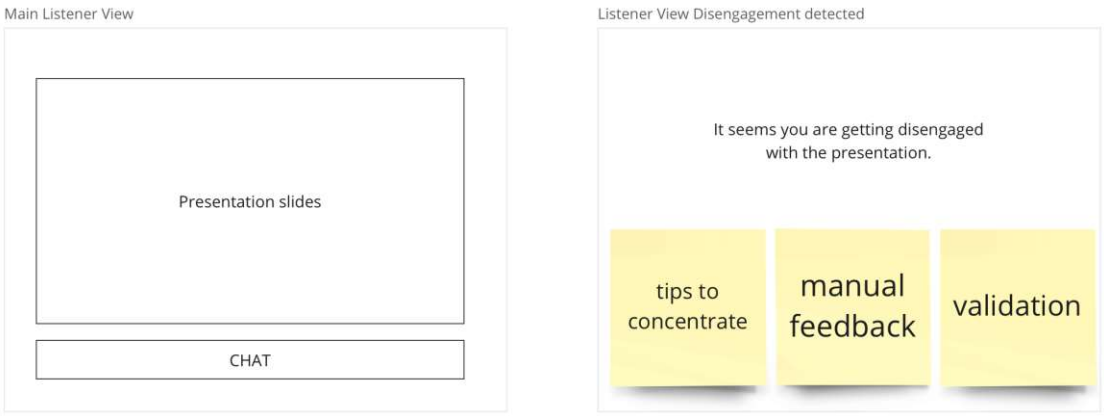


Figure 4.5: Low Fidelity Mockup: Listener View

in the low-fidelity mockups. However, it was concluded that the functionality was not necessary for the prototype and decided that it would not be implemented.

Another feature that was deemed interesting for the prototype was a report functionality (see figure 4.6) which can be used by the presenter to retrospectively analyze the detected information of the presentation session. In the mockup, a line chart was proposed to visualize the participants' properties on a timeline throughout the duration of the presentation. However, the exact type of visualization was yet to be defined, since the information that will be captured was still unclear.

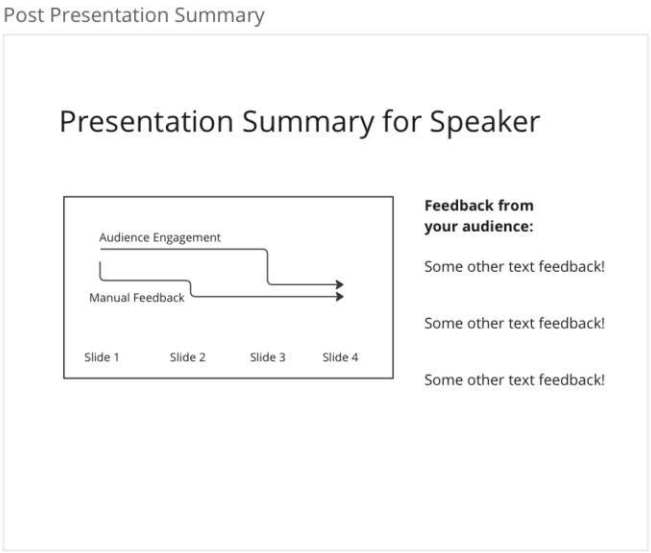


Figure 4.6: Low Fidelity Mockup: Presentation Summary

4.3 Insights from Focus Groups

In phase two of the prototype design and development phase, two focus groups were conducted with the overall aim of informing design revisions for the prototype. The focus groups were held with groups from two institutes of Vienna University of Technology. The focus group design was inspired by suggestions by Adams and Cox[AC08]. The focus group participant count was between three and eight. Furthermore, we invited rather homogeneous groups of people, recruited from one research group or department respectively, since homogeneous groups of people usually find it easier to talk to one another.

Focus Group	No. Participants	Affiliation (Research Group)	Duration
Focus Group 1	4	Human-Machine Interaction	60 mins
Focus Group 2	5	Artifact-based Computing and User Research	80 mins

Table 4.1: Focus Group Details

The focus groups were moderated by two people. The roles of the moderators were ensuring that all participants had equal opportunity to share their views, making sure that data and insights were properly recorded, and keeping the focus of the discussion on the topic. To further guide the discussion in a direction, the participants were faced with three slides consecutively with the following questions on them:

- **Introduction**
  - Can you briefly introduce yourself?
  - What are your experiences with online teaching?
  - What challenges did you face?
- **Situation with traditional online presentation tool**
  A screenshot of a presentation session with Zoom was shown. All participants had their cameras disabled, some did not even provide their proper name.
  - What kind of audience feedback is relevant to you during online presentations?
  - How do you receive feedback from the audience?
  - What kind of feedback do you miss?
- **Prototype Design**
  A screenshot of a preliminary prototype design was shown.
  - What kind of information would you find relevant?
  - How would you like to receive this information
    * Granular vs Aggregated?
    * Synchronous vs Asynchronous?

The focus group sessions were audio-recorded to facilitate subsequent analysis. The thematic analysis was guided by Braun and Clarke's step-by-step guide on *Using thematic analysis in psychology* [BC06], summarized in figure 4.7.

- **Phase 1: Familiarizing yourself with your Data**
  It is vital to be familiar with the collected data before the start of any analysis. Immersing yourself in the data, i.e. listen to the recordings or read the transcript of data repeatedly.
- **Phase 2: Generating Initial Codes**
  This phase involves the production of initial codes for segments of the data. A code can be any information regarding or feature of a segment of data that seems relevant to the analyst.
- **Phase 3: Searching for Themes**
  After successful coding of the data, the resulting codes can be organized more broadly into underlying themes. This process can be facilitated with techniques such as mind-maps or visualization tools.
- **Phase 4: Reviewing Themes**
  Upon identification of the initial themes, a review of the results should be performed to assess whether the identified classes indeed qualify to be themes.
- **Phase 5: Defining and Naming Themes**
  When the selection of themes is completed, the naming and definition of themes should be finalized. It is important to create a clear and concise definition and narrative of what the underlying data of a theme entails.
- **Phase 6: Producing the Report**
  The report should provide a convincing and comprehensible account of the essence of the underlying data and its themes. The report can include excerpts of important and distinctive passages of the data. Furthermore, it must entail valid arguments to justify the selection of themes.

Figure 4.7: Thematic Analysis Steps [BC06]

After the transcription of the conversations was completed, the individual conversation parts were coded. Subsequently, the codes were collected in a Miro board which enabled clustering the codes into groups in a collaborative effort among the two moderators of the focus groups. A visual overview of how the codes were grouped can be seen in screenshots from the Miro board in figures 4.8, 4.9 and 4.10. The identified groups formed the basis for the selection of underlying themes of the conversations. In the following sections, the identified themes will be discussed. Furthermore, interesting excerpts from the discussion will be given as examples to underscore the statements of the participants.

4.3.1 Theme I: Limited Value of Webcams in Video Conferencing

An interesting insight from the focus group was that webcams in traditional online presentation scenarios are not deemed useful in general by several participants. One participant highlighted this by mentioning:

> Usually, you barely see a person, often in poor quality. That's not really useful at all... except for knowing that there is someone you can talk to. (Appendix 1, #25, translated from German)

Mentioned by several participants was the fact that, irrespective of whether participants have their webcams enabled or not, they are missing both explicit and implicit feedback from the audience during a presentation.

> In an in-person lecture, you can really gauge whether people are listening to you. In an online presentation that is missing completely even if you have all cameras enabled. Also asking questions is really hard. … you have to rely on very proactive people. (Appendix 2,#11)

Similarly, another participant states that they are usually not capable of reading facial expressions from the webcam video. Instead, the only valuable information they can read from a webcam video is that a student is still physically in front of their computer and in principle approachable for collaboration:

> I do it simply to see that they're in the meantime not engaging with another course or went to the kitchen. So actually the minimum threshold is just to see that they didn't go away. It doesn't help to see their facial expression, also it's quite annoying to the students but I at least, I think, they feel more pressure to be involved in the course. (Appendix 2,#23)

Figure 4.8: Focus Group Analysis: Organized Codes, Insights Video Conferencing Tools

4.3.2 Theme II: Lack of and Desire for More Feedback

A common theme among all participants of the focus groups was a perceived lack of feedback when holding online presentations. This feeling was emphasized by multiple anecdotes of occasions where technical difficulties during online presentations occurred, but due to the expected lack of feedback, the presenter failed to realize the defect.

> I once even had the situation that my internet connection broke and I didn't

> realize for probably about 15 min. because nothing changed really and then one moment I received a message on my phone from a colleague that students were reaching out to them that actually the connection broke down. In some way there was just missing feedback. That moment, however, I realized, ok - there's actually someone listening and is interested in the connection to work which was quite positive feedback in a way. (Appendix 2, #8)

Several participants also mentioned similar experiences when they were attending lectures themselves and failed to communicate with the person presenting.

> I had a similar experience but the other way round. The lecturer muted himself but didn't notice. So he continued to go throughout 40 minutes, but we didn't have any sound. People were also trying to let him know, but he didn't hear us. (Appendix 2, #7)

> I taught a big lecture during covid. The interaction part did not change so much because in a big lecture you do not have that much interaction even if offline. What I really felt was the difference of not getting any direct feedback. You were somewhat seeing all these tiny black boxes that never had any reaction. (Appendix 2, #8)

Many participants noted that they would appreciate more feedback when holding online presentations. They pointed out that they believe it would be helpful to revise and improve presentations and that it would be especially useful for people with little experience in presentation.

> ... as means for self-reflection that could be useful for sure - especially for junior lecturers. ... for lecturers it would be very useful. (Appendix 1, #36)

> ... it would be more information than what we have today. If it helps lecturers to revise their presentation it would certainly be a good thing. (Appendix 1, #40)

Relevant feedback parameters by participants of the focus groups were the dimensions of *Engagement* and *Confusion* of listeners. Though there were varying degrees of confidence that automatic detection could capture this information proficiently. That is, the suggestion to use a combination of automatic detection and manual feedback was proposed.

> The parameter that I'm interested in is Engagement. In a good lecture, people don't always understand 100% of the content (Appendix 1, #30)

> I think with confusion... this could be easily self-reported. With engagement on the other hand students probably would not answer truthfully, here AI could be involved. (Appendix 2, #41)

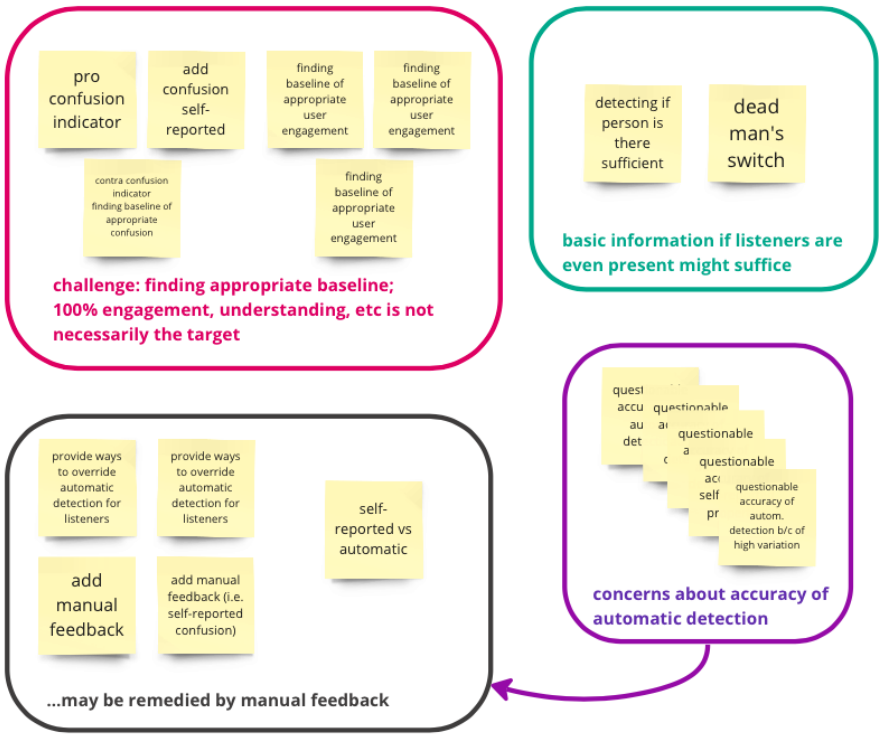


Figure 4.9: Focus Group Analysis: Organized Codes, Automatic Detection of Non-Verbal Cues

4.3.3 Theme III: Expected Challenges of Human-AI Interaction

A reoccurring theme among most participants was a feeling that interaction between users and automatic non-verbal cue detection would be a challenge to implement. On the one hand, concerns were expressed about how the application should communicate detected non-verbal cues to the user or listener. From the point of view of presenters - which was the main purpose of the focus groups - many participants voiced concern that a system that constantly informs them about the state of their listeners could be a distraction when presenting or could be confusing if the results are not valid.

> ... if I get many signals from a system... what can I improve? Maybe it would be better to get a summary retrospectively. Synchronously during a lecture, there might not be much I can improve, because I have to concentrate on my slides. (Appendix 1, #19, translated from German)

> While detailed, synchronous feedback would likely be interesting, I would prefer aggregated key indicators. It would likely be distracting if I'd be shown the information from i.e. 40 listeners all the time. A personal setting to control the presentation would be useful. (Appendix 1, #23, translated from German)

An obstacle identified by many participants would be identifying a threshold when the automatic detection should trigger a status change. Several participants mentioned, that even during a face-to-face lecture, it is normal that not all participants are fully engaged all the time or that all participants understand everything all the time since lectures were meant to stimulate curiosity to engage yourself with a topic apart from the lecture as well.

> I believe it is not important that every student is 100% motivated until the end of the lecture. There are fundamental topics that just have to be taught and learned. However, it should be evaluated if there is something that can be improved. (Appendix 1, #20, translated from German)

Most participants suggested that the system should only provide feedback to the lecture in a summarized manner or once a certain threshold is hit, in order not to be too distracting to the presenter.

> I do not care if a 100% or 90% of the people are engaged. I need to have that minimum threshold, i.e. 60% and now the system tells me it's time to worry… of course, not everyone is going to be fully engaged throughout the lecture.(Appendix 2, #35)

Several participants also suggested that a retrospective summary of the captured information during the lecture could be very interesting for optimizing their presentation and particularly useful for junior lecturers. Though one participant mentioned that if such a system would be used in a real-world scenario, a particular focus would have to be laid on complying with privacy regulations such as GDPR.

> I think it would be a good design choice to have more detailed information after the lecture and less information during the lecture. (Appendix 2, #49)

> I think it would be very interesting to see the summary of i.e. the engagement per slide… that would be very useful for revising your slides (Appendix 2, #50)

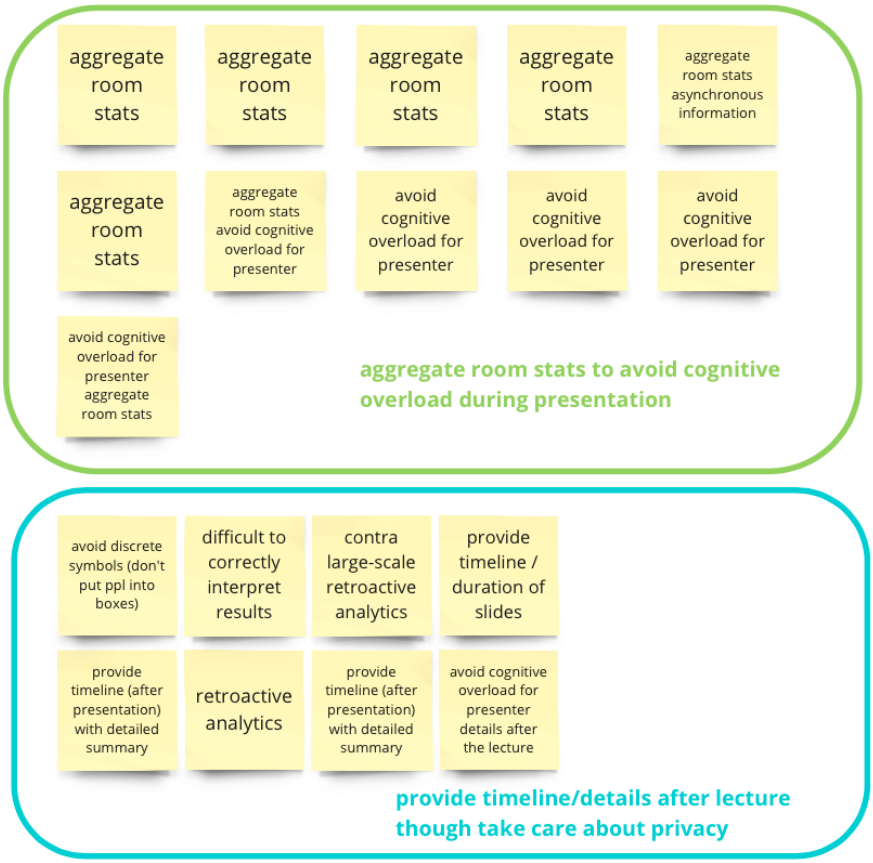


Figure 4.10: Focus Group Analysis: Organized Codes, Communication of Detected Features

4.3.4 Miscellaneous Interesting Excerpts

Since several participants mentioned that the primary use case of webcams during online presentations for them was to check if participants are ”still there”, one participant explained this would be achieved in an industrial machine: the dead man’s switch. This could be an interesting and simple way to check attendance without the need for sophisticated image recognition techniques.

> A nice analogy for checking attention is the dead man’s switch... it would be a nice analogy to implement checking attention i.e. through some kind of interaction like mouse tracking. (Appendix 1, #34, translated from German)

Another interesting input mentioned by several participants was the suggestion to improve anonymity by hiding the names of the participants since it is not deemed relevant for

them. This idea was mainly pitched to potentially increase the acceptance of using the system from a listener's perspective.

> I was wondering… do we even need the names. Maybe we can show like a lecture room with fixed positions, so we know i.e. the guy in the right is always sleeping, but he's always there, so we have an association with specific spots, and then we do not actually need the name. ... I think displaying the names would be too much information. And this would only work for a seminar with 15-20 people but for a big lecture it would be too much information. (Appendix 2, #35)

4.3.5 Revised Prototype Requirements

Based on insights from the focus group, the requirements were revised or extended respectively accordingly:

1) Provide functionalities for communicating non-verbal cues without a camera stream as additional communication layer.
2) Hide the webcam view once the user confirms positioning in front of the webcam, to mitigate Zoom fatigue.
3) Provide onboarding to guide users through the prototype and explain functioning to facilitate acceptance.
4) Limit the amount of information transferred to the minimum to facilitate acceptance and privacy by locally processing video data, and hiding names of participants.
5) Implement non-verbal cue detection with ways for a manual override to avoid invalid detection.
6) Avoid distraction of the presenter by keeping notifications balanced and communicating information summarized.

4.4 Revised Mockups

There are two major changes in the revised mockups. First, the live information visible to the presenter during the presentation will be aggregated and show average values from all participants. This shift is aimed at providing a more consolidated and comprehensive overview of the participants' feedback without being distracting to the presenter. Furthermore, there are notable modification in the visual representation of this information. The information will be communicated with less discrete, color coded progress bars, instead of discrete emojis. This transition was made due to the complexity associated with translating averaged values into discrete emoji representations, as this was found to be challenging and undesirable by some participants during the focus group discussions. Furthermore, the two parameters to be captured by the automatic detection, engagement

and confusion cannot easily be translated into discrete emojis. However, it's important to acknowledge that the current color-coding system, which ranges from red to green, may pose accessibility challenges for individuals with red-green color blindness. For future adaptions this color-coding scheme should be optimized to increase accessibility.

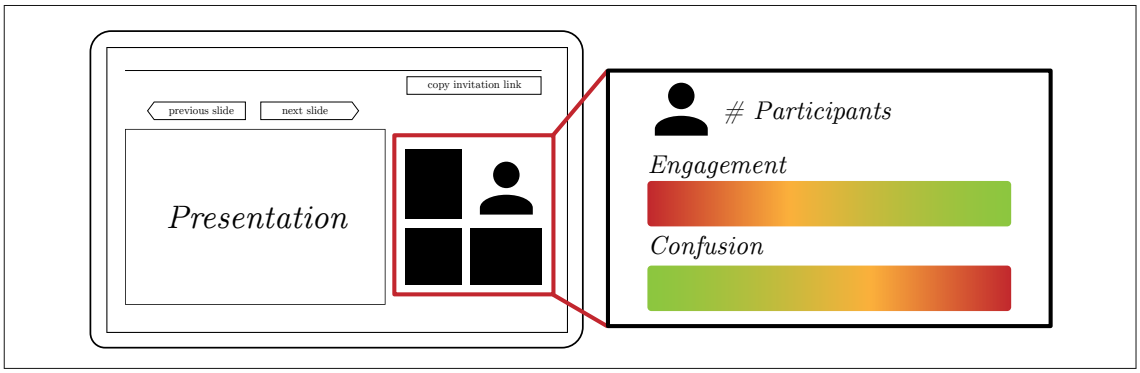


Figure 4.11: Revised Mockup, Presenter View

Additionally, a major change from the first mockups is the added functionality of giving manual or overriding automatic feedback detected by the image processing component to account for a potential proneness to error regarding the detection process. It was decided that in listener mode, a feedback picker will be visible below the presentation slides. The picker enables providing feedback in the two automatic detection dimensions engagement and confusion. The provided feedback will override any automatically detected value for the entire duration of the slide.

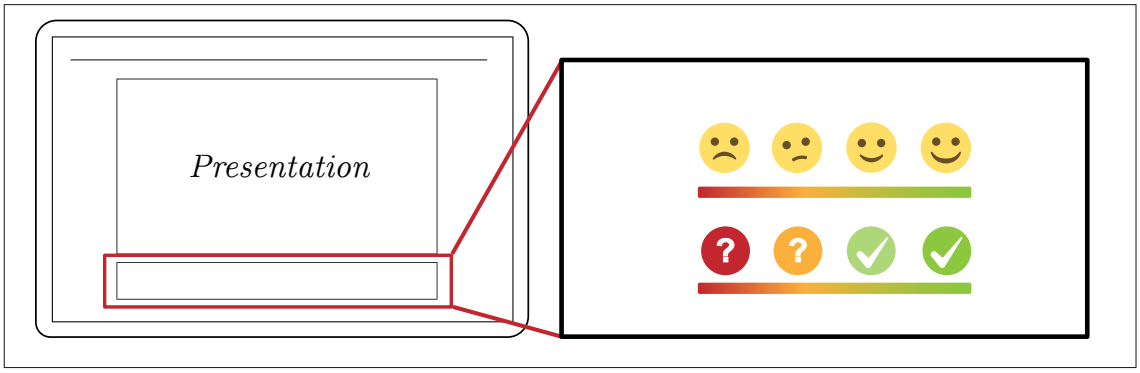


Figure 4.12: Revised Mockup, Listener View

Since many of the focus group participants stated that a retrospective summary of the detected presentation statistics would be useful for analyzing and optimizing their presentation, a summary component was added to the mockups. The component includes the most important statistics such as the total duration of the presentation, the number of participants as well as line graph indicating the measured detection results throughout the presentation.

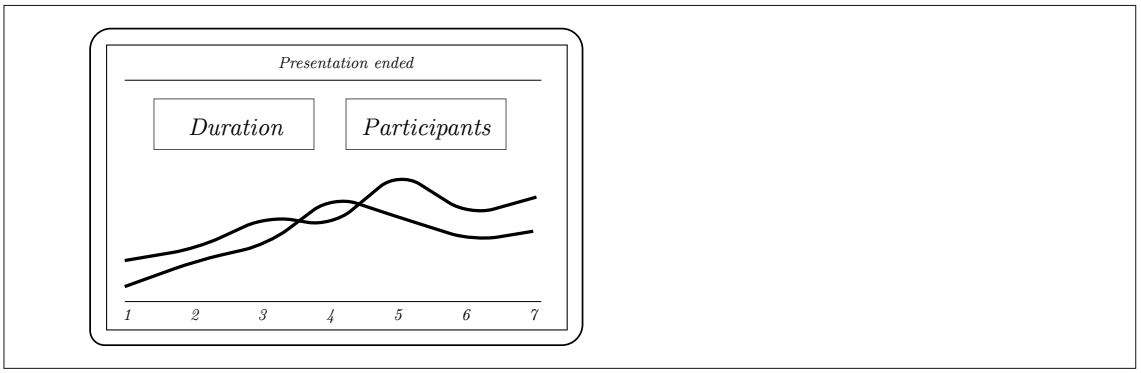


Figure 4.13: Revised Mockup, RetrospectivePresentation Summary

CHAPTER 5

Prototype Implementation

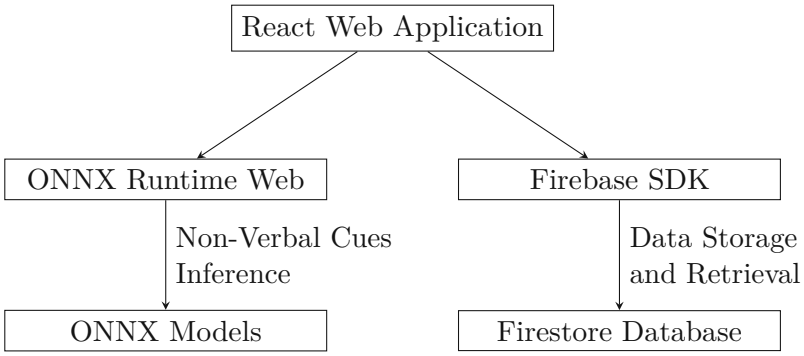


Figure 5.1: Prototype Web Application Overview

To study the feasibility and acceptance of a system with automatic detection of non-verbal cues, a prototype web application was developed. The purpose of the prototype was on the one hand demonstrating technical feasibility and on the other hand having a system that can be evaluated in a realistic online presentation scenario. The prototype was developed as a web application for reasons such as:

- **Rapid Prototyping**: Modern web development frameworks offer modules and reusable features that may be used to quickly implement and test research ideas.
- **Accessibility**: Web applications can be accessed by any end device with a browser, irrespective of its operating system. That is, it is a very resourceful and cost-effective way to test prototypes even with large crowds of participants.
- **Remote Sessions**: Since the web prototype can be accessed from any private end device, it is an ideal facilitator for remote experiments. The possibility to conduct test sessions remotely reduces coordination effort, because one does not

have to coordinate in-person meetings with participants. However, it may also lead to a more authentic test experience compared to studies conducted in a lab environment.

- **Data Collection**: Various tracking and data collection methods can be built into the prototype which, depending on the evaluation method, can lead to valuable insights.

5.1 Overview and Architecture

The prototype was developed utilizing modern web development technologies and tools including React, TypeScript, Ant Design, a Firestore real-time database, and Redux Toolkit (RTK) for real-time updates. The machine learning integration was realized with ONNX Runtime Web. This chapter delves into the key architectural decisions and provides more information on the technologies that were used to create a working prototype.

5.1.1 Project Setup and Architecture

The web application was developed using React, a popular JavaScript library for building web applications. The project was set up utilizing Create React App (CRA)[1] which creates a bootstrapped single-page React web application with several pre-configured state-of-the-art development tools included. Since the prototype application was designed with manageable complexity, we decided against the use of a React framework, like Next.js. We did, however, choose to use a Typescript template for setting up the project due to its static typing capabilities, which enhance code quality and maintainability. Additionally, to get started with the ONNX runtime web with the React Javascript library, an example implementation by Wahyu Setianto[2] showcasing using a Yolov8 model to detect objects on an uploaded image served as inspiration especially some parts of image pre- and post-processing were re-used.

5.1.2 Database and Real-Time Updates

Firebase[3] is a set of cloud computing services for app development developed by Google. Firestore is a NoSQL cloud database and part of Firebase. It was selected as the cloud storage and communication medium for the prototype. Firebase offers real-time synchronization and seamless integration with frontend technologies, making it an ideal choice for managing application data and facilitating communication between users. The Firebase JavaScript SDK[4] provides a client to easily interact with the Firestore database.

[1] https://create-react-app.dev
[2] https://hyuto.github.io/yolov8-onnxruntime-web/
[3] https://firebase.google.com
[4] https://firebase.google.com/docs/web/setup

Firestore's NoSQL data model provided a flexible framework to store and query the data relevant to the prototype.

State management in the web application was implemented using the JavaScript library Redux[5]. Real-time updates within the prototype were implemented using Redux Toolkit's RTK Queries. RTK Queries[6] simplify the process of making API calls and managing data within the Redux store. By utilizing RTK Queries, the prototype was able to seamlessly integrate with the Firebase Firestore Realtime Database, enabling automatic data synchronization and reducing the complexity of manual state management.

5.1.3 Machine Learning Integration

The web prototype integrated an image classification module using ONNX Runtime Web. This technology enables real-time image classification using the user's webcam feed as input. ONNX Runtime Web is a JavaScript library that supports executing ONNX (Open Neural Network Exchange) models directly within the browser. Information with regard to the training of the Machine Learning model can be found in section 5.2. The trained model was exported in the ONNX format and integrated into the web application, enabling the prototype to analyze webcam images and locally infer predictions without the need to send the webcam image to a server.

5.1.4 UI Framework

The Ant Design[7] framework was employed to create a cohesive and visually pleasing user interface. Ant Design offers a wide array of pre-designed components that simplify the UI development process while ensuring a consistent design language throughout the application.

[5]https://redux.js.org
[6]https://redux-toolkit.js.org/rtk-query/overview
[7]https://ant.design

5.1.5 Prototype Communication Flow

The communication flow within the prototype involves multiple layers of interaction. The automatic detection module combined with user inputs and interface interactions on the frontend trigger requests to the Firebase Firestore real-time database through RTK Queries. These requests initiate data updates or retrievals, which are then synchronized with other connected clients in real-time.

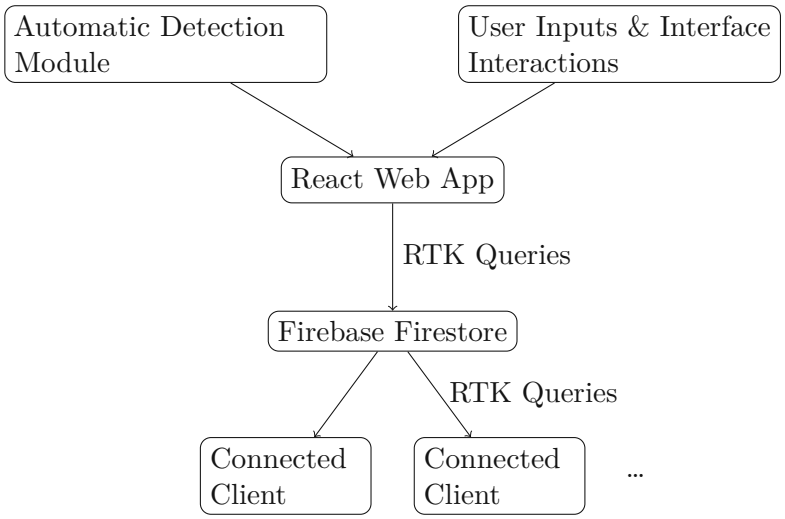


Figure 5.2: Prototype Communication Flow

5.2 Emotion Classification Module

Based on the results from the focus groups, listeners' engagement and content understanding were deemed as the two most useful feedback mechanisms. Hence, we opted for engagement and confusion states as the recognition objectives. During the literature analysis we uncovered the dataset DAiSEE, which contains labels of both of these properties. Considering the architecture of the algorithm, 3D CNN networks [GDAB16] and transformers [ASLC22] have proven to be effective in prior works on engagement recognition. However, none of these works performed both, simultaneous engagement and confusion recognition, hence we propose a new approach in this work. The final image classification module was implemented using a Residual Network (ResNet) and a Temporal Convolutional Network (TCN) architecture. The architecture was developed in collaboration with the Dept. of Human-Machine Interaction and is heavily inspired by a proposal from Abedi and Khan [AK21].

The advantage of the proposed architecture is that it allows for capturing spatio-temporal features from a video input, while enabling us to distribute the processing of the video stream sequentially, processing image by image. Given the limited computing capabilities available on end devices, computationally-heavy methods that require multiple frames as

input such as transformers or 3D CNN networks may not provide recognition that is fast enough for immediate feedback. Using these approaches, outsourcing the processing to the cloud might be necessary, which would reduce privacy and would require additional infrastructure. Our method works by first extracting spatial features from consecutive video frames using a 2D ResNet, and then analyzing the temporal changes in these frames with a light-weight TCN to detect the participants' engagement and confusion level. The Residual Network was trained with the AffectNet [MHM19] dataset containing images of facial expressions annotated with emotion categories. After pretraining the 2D CNN, we modified the architecture from Abedi and Khan [AK21]. In the first step, we replaced linear layers connecting the backbone and the network head with 1x1 1D CNNs to improve computational efficiency. Moreover, we reimplemented the classification head using a similar strategy, replacing summing operators proposed by Abedi and Khan with 1x1 1D CNNs for an improved backward pass. We also extended a second head, which allows us to perform simultaneous training and classification for both engagement and confusion. Lastly, we have reworked the training strategy of the network. As the DAiSEE dataset is highly imbalanced [GDAB16], we opted for stratified sampling to improve the predicting capabilities. While the original paper from Abedi and Khan elaborated on the strategy, we found the implementation was not consistent with the paper. We implemented stratified sampling in the form of weighted random sampling based on the class occurrence within the dataset.

Finally, the pre-trained 2D ResNet and the Temporal Convolutional Network were trained using the DAiSEE [GDAB16] dataset which entails video sequences of students listening to online lectures. Overall, we reached accuracies of 52 and 55 percent respectively, which were also consistent for the classes which were sparsely represented in the dataset.

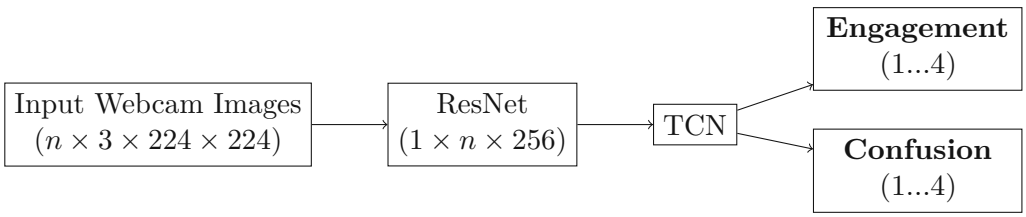


Figure 5.3: Engagement & Confusion Detection Architecture

For the deployment, since the engagement detection model expects a cropped face image as input, the webcam image is being cropped with a state-of-the-art Yolov8 face detection model[8] before using the feature detection with ResNet. In the deployment, we opted for a network split, exporting both the 2D ResNet and the TCN separately, which enables us to process the data from the video stream sequentially. Hence, each frame of the video stream is processed directly after acquisition, saving the computed features in a stack. After accumulating 30 frames we pass the stack into the TCN classifier. In our tests, we found that the processing speed was sufficient to enable almost immediate emotion classification.

[8] https://github.com/akanametov/yolov8-face

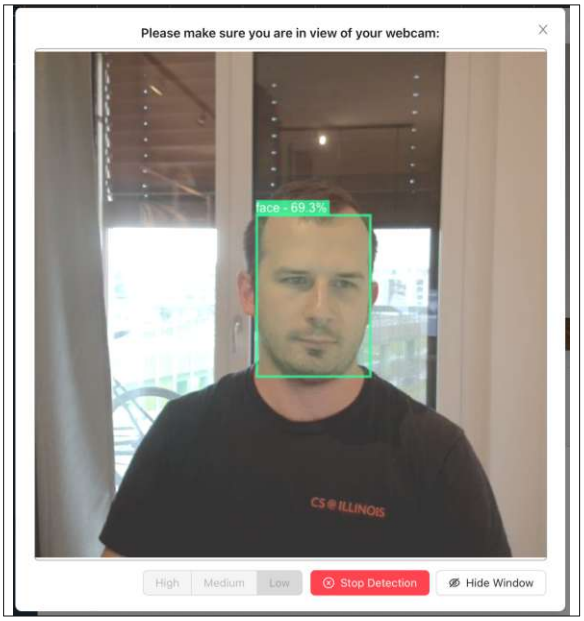


Figure 5.4: Prototype, Face Detection with Yolov8

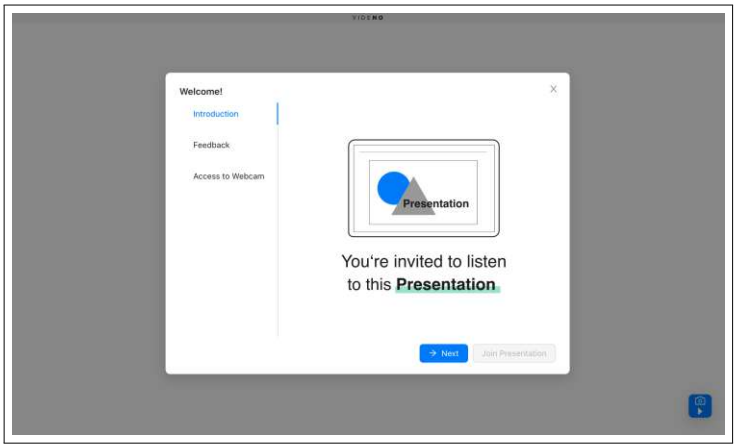
5.3 Prototype Walk-Through

This chapter provides a Walk-Through through the final prototype web application. The application provides two entry points: one for the person holding the presentation, and another one for participants listening to the presentation. The presenter view's main objective is to provide an easily digestible visual overview of the current state of the audience in terms of engagement or confusion. The listener view's main objective is to provide an unobstructed view of the presentation slides, to conduct the automatic inference of the engagement state and to provide ways to override whatever was inferred automatically.

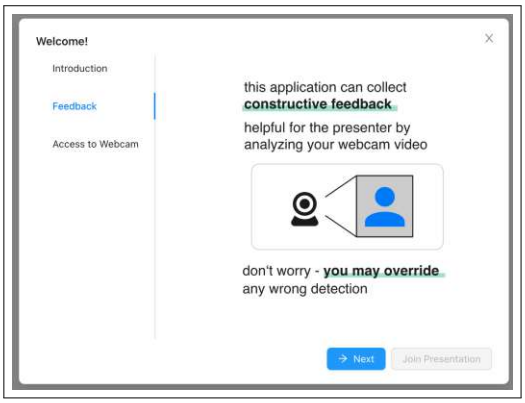
5.3.1 Listener Mode

The listener mode displays an onboarding module upon starting to introduce the user to the concept of the prototype (see figure 5.5). The main objective of the onboarding module was to explain that the automatic engagement detection is based on the user's webcam input and that the inference process is being carried out exclusively locally, on the user's device. Furthermore, pointing out that users have the possibility to override the automatic detection results, was important.

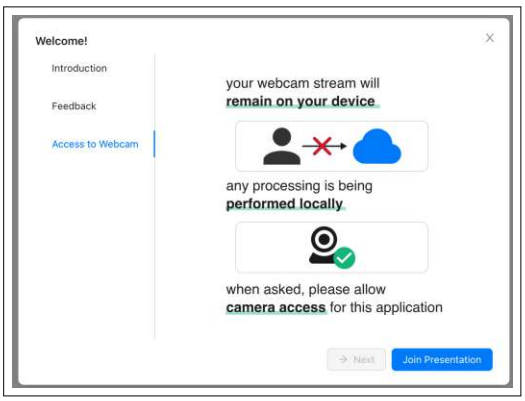
Upon exiting the onboarding module, a window with the user's webcam image is automatically opened (see figure 5.6), asking the user to position themselves in view of the camera. For this to work, the user may also have to approve webcam access for the web application in advance. The window and webcam view close automatically, once a face is



(a) Onboarding 1



(b) Onboarding 2



(c) Onboarding 3

Figure 5.5: Prototype, Participant Onboarding

detected in view, to avoid the effects of Zoom fatigue of constantly having to watch an image of yourself during online presentations [KAF22].

Once a face in the webcam stream is detected, or the user manually closes the webcam window, they have an unobstructed view of the presentation slides provided by the presenter (see figure 5.7). On the left-hand side, a button that can be expanded shows the name of the user, which due to insights from the focus groups was universally changed to "Anonymous User". Upon clicking the button, the current results of the engagement inference can be inspected. To facilitate the presentation, an emoji of the presenter with the provided name is shown. Additionally, a presentation stepper, visualizing the number of slides and indicating the current position within the presentation was implemented.

Below the presentation slides, there is the manual feedback picker where users can choose to override the manual detection results of the two parameters engagement and confusion. The selected input will override the results throughout the duration of the current slide. The feedback picker indicates the current, automatically detected state of engagement

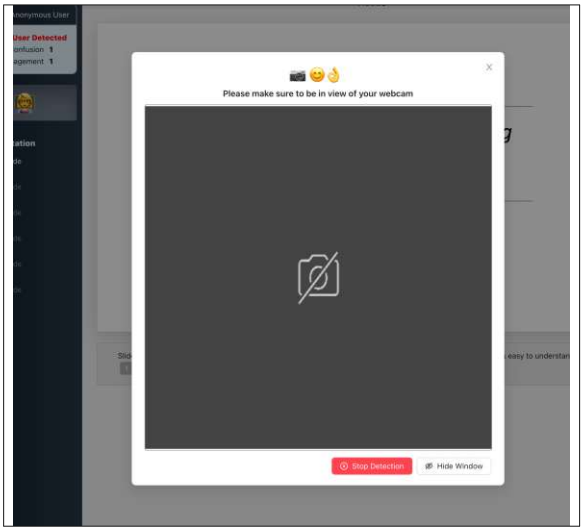


Figure 5.6: Prototype, Webcam Modal in Listener View

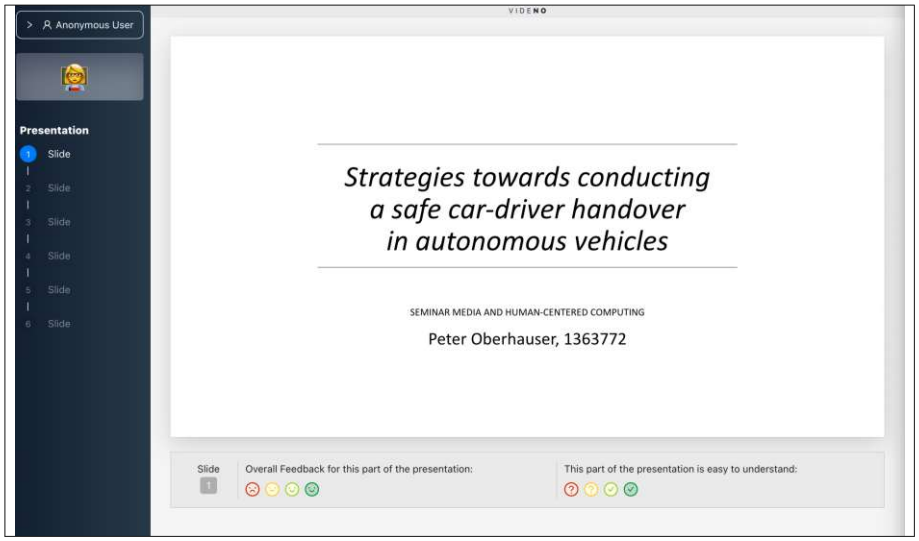


Figure 5.7: Prototype, Listener View

and confusion by slightly increasing the size of the respective emoji (see figure 5.8).

When the presentation is ended by the presenter, the listener is being notified. From this point on, listeners can not interact with the web application anymore. Currently, listeners have no way to restart the presentation or download the slides. This could potentially be improved in the future.

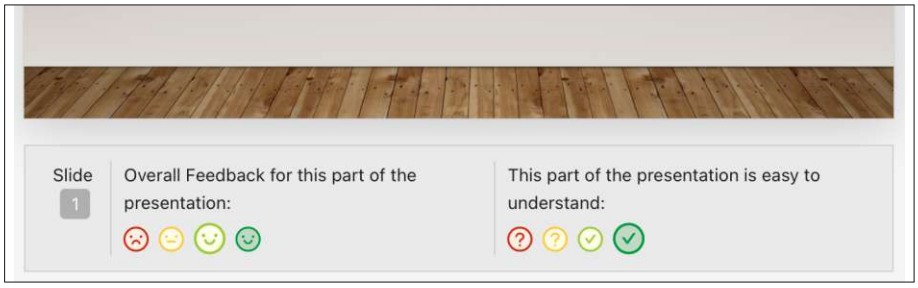


Figure 5.8: Prototype, Listener View

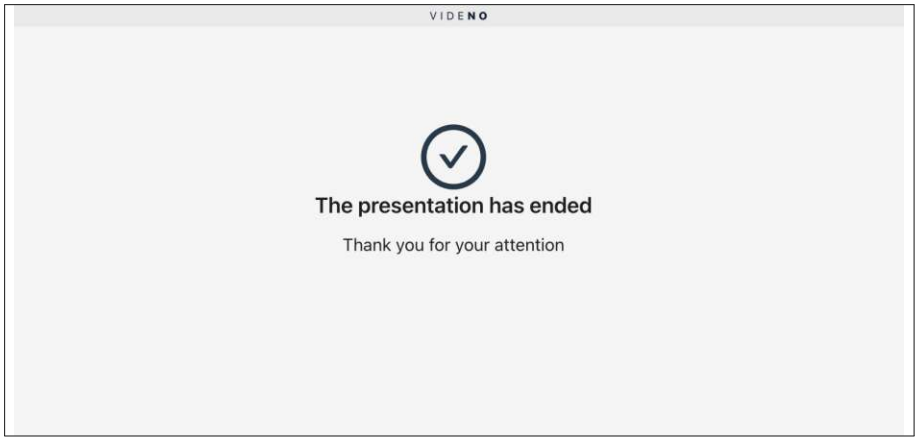


Figure 5.9: Prototype, Presentation Ended

5.3.2 Presenter Mode

In presenter mode, the left-hand side of the application shows the same affordances as in the listener mode. The main view similarly includes a view of the presentation slides, though without the manual feedback pickers. At the top of the screen, presenters additionally have a button to start recording the aggregated presentation room information as well as to copy an initiation link to the current presentation session, which can be shared with the audience. Crucially, above the presentation slides, there are buttons that enable the presenter to remote control which presentation slide is currently being shown to the audience.

The main difference to the listener mode, however, is in the right-hand side of the screen. Here, the presenter has an aggregated overview of the automatic detection and manual override results. The presenter can see, how many users logged in to the presentation session and how many of the logged-in users were successfully detected in their webcam stream. Below, the two colored bars indicate how the detected or manually overridden engagement and confusion scores averaged throughout all participants in the room. The bars are color-coded with a green-to-red gradient, with green indicating positive values and red indicating negative connotations. In the case of engagement, high engagement is

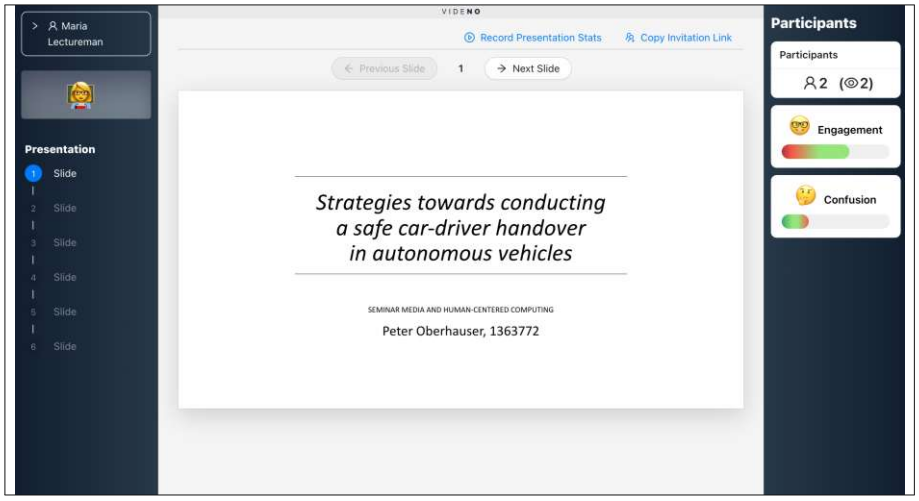


Figure 5.10: Prototype, Presenter View

coded with green color and low engagement with red. In the case of confusion, the color coding was reversed. As already mentioned, the use of emojis as a primary communication medium was replaced with color-coded progress bars during the design phase to have a less discrete visualization of averaged values (see chapter 4) since coding these values in Emojis proved to be very difficult. The purpose of the two remaining emojis is mainly aesthetic and to provide an additional visual cue of the two dimensions that are detected.

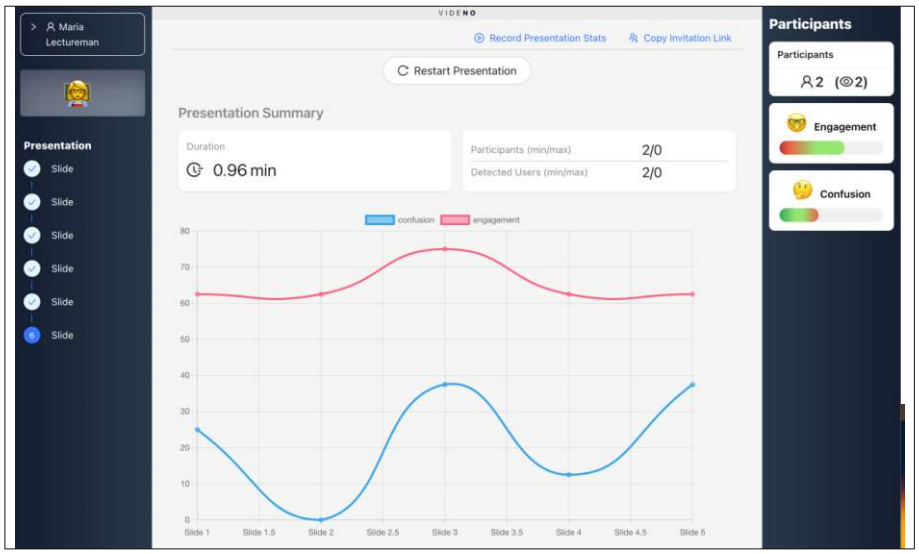


Figure 5.11: Prototype, Presentation Summary

Upon concluding the presentation by clicking the ”End Presentation” button, which becomes visible when the final slide is active, the presenter is directed to the presentation summary component. Here, a concise yet informative summary screen awaits the presenter.

The component features a simple line graph, that visually illustrates the recorded average values for each of the presentation slides. Moreover, a selection of critical presentation statistics is presented, including the overall duration of the presentation, as well as the minimum and maximum recorded number of participants throughout the presentation. Additionally, the minimum and maximum number of participants that were detected in their webcam stream is provided. This summary component can provide insights to presenters about how the different parts of their presentation were perceived by their audience and may help optimize their presentation. Something that was desired by several participants of the focus groups with university lecturers that were conducted in this work.

5.4 Prototype GitHub Repository

The source code of the prototype is available under the following GitHub repository:

</>	**Prototype GitHub Repository** `https://github.com/oberpete/vide-no.git`

5.5 Prototype Deployment

The prototype is deployed with the GitHub pages[9] service since the prototype is a Single Page Application (SPA) which does not require any backend.

[9] https://pages.github.com

CHAPTER 6

Evaluation

The research employed in this study is conducted in a mixed-methods approach. During the prototype's design phase, a focus group session to get real-world insights from people experienced in holding online presentations was conducted. The outcome of the focus group was analyzed using thematic analysis and was used to inform the design of the prototype to be developed. More information about the focus groups in the prototype design phase can be found in section 4.3.

To assess the prototype that was developed and the usability of the proposed system, a user study was carried out, focusing on capturing participants' subjective impressions of the prototype. This assessment included the use of both questionnaires and interviews. When defining the study design, we considered adding objectively measurable variables like comparing the quantitative metrics of the prototype with those of traditional tools as a baseline. However, we concluded that any results gathered in such a study were prone to various order effects [HZ20], such as participants being more familiar with traditional tools or fatigue that kicks in for presentations held at later stages of the tests. That is, we decided to only consider the subjective experiences of participants and furthermore ask them to compare their experiences made with the prototype to experiences made in the past with traditional tools for online presentations.

The open-ended results from the questionnaires and interviews were qualitatively analyzed through thematic analysis, similar to the data from the focus group in the design phase of the prototype. Additionally, a subset of the questionnaire containing Likert scale questions was analyzed quantitatively. It is however important to interpret these quantitative findings cautiously, given the lack of a baseline for the comparison. Hence, we opted for a qualitative analysis for the evaluation of the perceived usefulness of the prototype.

6.1 Participants

The initial goal was to recruit approximately 12 participants. In the end, the user tests had 13 participants. The number of participants was defined on the one hand based on what seemed feasible to recruit within the timeframe available for the user study. Furthermore, the participant count was inspired by local standards of HCI studies. For instance, a participant count of 12 is the most common number within studies published in the Conference for Human Factors (CHI) community according to an analysis by Caine published in 2016 [Cai16], a major conference in the field. Participants were recruited to match these requirements:

- Participants should have some experience in attending and holding online presentations with traditional video conferencing tools such as Zoom or Microsoft Teams. Figure 6.1 shows the self-reported experience levels of the participants. 10 out of 13 participants reported to be very experienced in listening to online presentations. Similarly, 10 out of 13 participants reported to be very experienced or experienced in conducting online presentations.
- The participant pool should be as diverse as possible when it comes to gender, age, etc. to ensure validity and generalizability. Considering the size of the test user pool, however, generalizability beyond the participant pool is limited.
- At least 6 people should be willing to present, ideally, a presentation that they personally created and that is relevant to all participants. The duration of the presentation should be between 5 and 10 minutes. In the end, 8 of 13 participants held a presentation during the user tests.

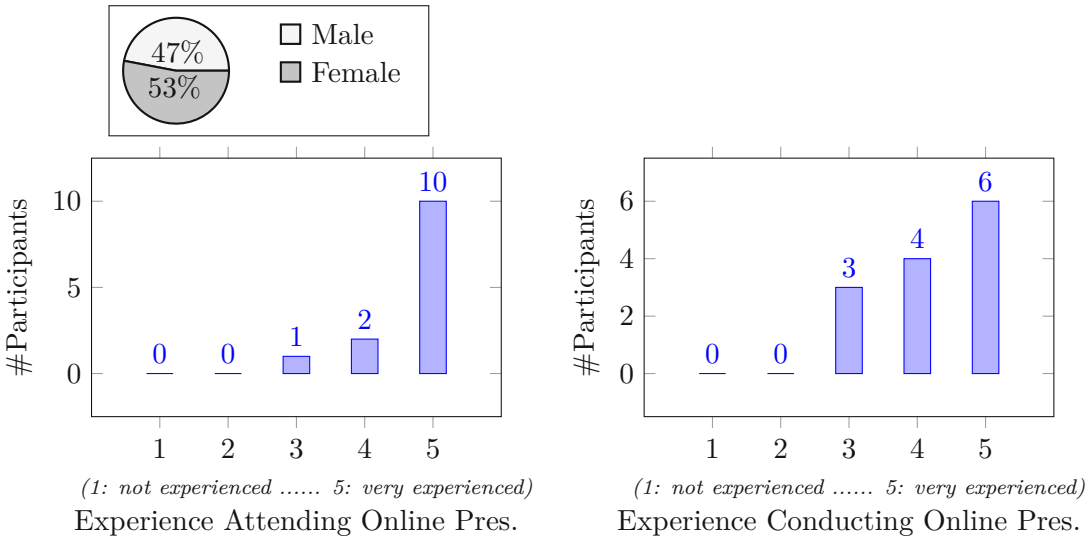


Figure 6.1: Participants Demographic Overview

6.2 Procedure Summary

The resulting prototype underwent a user study in a context similar to a real-world online presentation. During the study, participants took turns conducting short online presentations using the prototype, accessible from their personal computers. While presenting, they received both automatic and manual feedback from other participants. Following the test presentations, participants were asked to complete a questionnaire. The user tests started by participants accessing a survey link with their personal computers. The pre-study questionnaire aimed to gather basic demographic information about the participants as well as inform the participants about the study's modalities and capture their consent to participate in the study. Subsequently, the participants were asked to access the web-based prototype application. When all participants managed to access the application, a short onboarding session was conducted, outlining the main functionality of the prototype. Since the prototype does not support audio communication at the moment, the audio communication was conducted via a parallel Zoom session that remained active throughout the test. Participants were instructed to keep the Zoom windows minimized and hidden at all times during a presentation cycle. After completion of the prior steps, the actual presentation rotation commenced. The schedule was determined randomly. Technically, for each presentation rotation, a session with a specific entry URL was prepared. The session was set up with the slides of the respective presenter. This ensured that the generated presentation summary remained separated for each presentation. After all presentation cycles were finished, the participants were asked to fill out a post-study questionnaire. Additionally, participants who presented were asked to take part in a short interview, which aimed to capture in-depth information about their experiences using the prototype.

6.3 Questionnaire

To capture the participants' subjective experiences with the prototype during the user tests, the participants were asked to fill out a questionnaire. The questionnaire was split into pre- and post-study segments, to ensure that consent is acquired and participants are informed before the study. However, another reason for this division was to keep the questionnaires concise and avoid overwhelming the participants. The questionnaires were conducted with Microsoft Forms.

6.3.1 Pre-Study Questionnaire

The primary purpose of the pre-study questionnaire was to gather basic demographic information from the participants. Given that the study was conducted remotely, with participants accessing the prototype via their laptops, the pre-study questionnaire also served the purpose of informing participants about the study's modalities, the type of information that will be collected, how it will be stored and processed, as well as obtaining their consent to participate in the study.

No.	Question	Answer Type
0.	*Information about Study and Consent Form*	
1.	What is your age?	Number
2.	What is your gender?	[female, male, other]
3.	I am experienced in holding online presentations (i.e. in school, work, etc).	Likert Scale
4.	I am experienced in listening to online presentations.	Likert Scale

Table 6.1: Pre-Study Questionnaire

6.3.2 Post-Study Questionnaire

In the pursuit of evaluating the resulting prototype, standardized questionnaires like the System Usability Score (SUS), Usability Metric for User Experience LITE (UMUX-Lite) [LUM15] and others were taken into consideration for usability assessments. However, it became apparent that none of the existing questionnaires aligned with the unique attributes of this project and the resulting scores were not deemed helpful for evaluation due to the lack of comparability. Consequently, a pragmatic custom questionnaire tailored specifically to the intricacies of the prototype was created. The questionnaire included many open questions and was subsequently evaluated qualitatively.

6.4 Interviews

After the experiments, interviews with six participants who presented were conducted in a semi-structured manner. The goal of these interviews was to get insights into the experiences of the participants with the prototype and any issues they encountered without biasing them with leading questions. That is, the interviews were planned with minimal structure. The cornerstones of the interviews encompassed talking about problems that occurred during the test, general thoughts about the principles of the technology used and suggestions for future improvements. The interviews were recorded and evaluated similarly to the focus groups, guided by Braun and Clarke's step-by-step guide on *Using thematic analysis in psychology*[BC06].

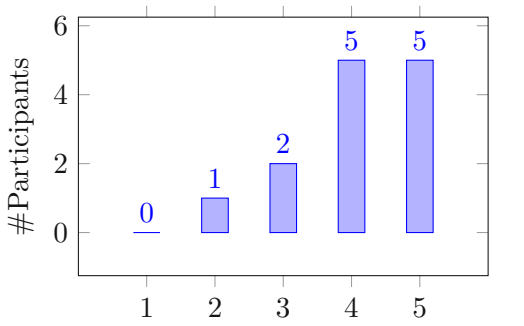
No.	Question	Answer Type
A	*Presenter-specific Questions*	
1.	How does your online presentation experience with the prototype compare to your experiences with i.e. Zoom?	Open Text
2.	Do you feel that the metrics analyzed by the prototype (engagement, confusion) were helpful? How did the provided feedback affect you during the presentation?	Open Text
3.	After ending the presentation, you received a summary outlining engagement and confusion levels throughout the presentation. How would you use this information, if at all?	Open Text
4.	Would you prefer conducting your online presentations with a system similar to the prototype, or would you prefer a video-based conference tool? Please provide reasoning for your choice.	Open Text
5.	Please share any other feedback regarding your experience as a presenter with the prototype.	Open Text
B	*Listener-specific Questions*	
1.	Do you feel that automatic detection of engagement/confusion as demonstrated in the prototype is favorable compared to online presentations with camera enabled?	Likert Scale
2.	Please explain your decision from the previous question.	Open Text
3.	Did you observe the output of automatic engagement detection? Did you override by providing manual feedback? If so, why?	Open Text
4.	How comfortable do you feel being subjected to automatic detection of engagement and confusion as demonstrated in the prototype?	Likert Scale
5.	Please elaborate based on your previous answer.	Open Text
6.	How eager are you to use the tool to listen to online presentations compared to traditional conferencing tools such as Zoom?	Likert Scale
7.	Do you usually feel comfortable having your camera enabled during online presentations?	Likert Scale
8.	Please share any other feedback regarding your experience as a listener with the prototype.	Open Text

Table 6.2: Evaluation Questionnaire

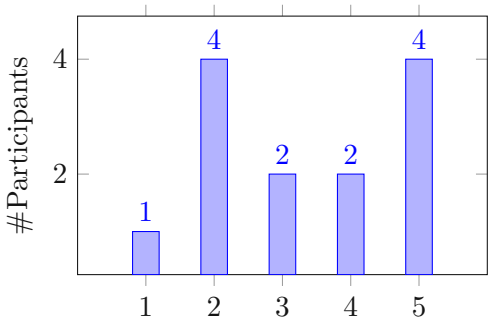
6.5 Results

In this section, results from the questionnaires and interviews will be elaborated. Since the questionnaire was split into questions related to the presenter experience and listener experience, results will also be presented similarly. The questionnaires revealed interesting insights regarding the experiences participants had using the prototype. The retrospective interviews were intentionally kept short because it became apparent that responses were very similar to what was submitted in the questionnaire. Nevertheless, the interviews did yield some further valuable insights. This chapter will showcase noteworthy themes and excerpts from both the questionnaires and interviews.

6.5.1 Listener Perspective



How comfortable do you feel being subjected to automatic detection of engagement and confusion as demonstrated in the prototype?



Do you usually feel comfortable having your camera enabled during online presentations?

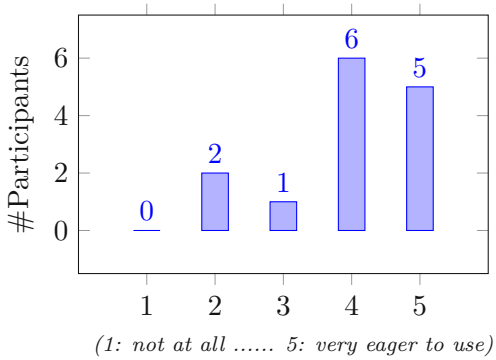
Figure 6.2: Presentation Listening Experience Poll, 1

The primary research question from a listeners' perspective was getting insights about how comfortable listeners felt being subjected to automatic emotion detection compared to an online presentation situation with webcam enabled. The subjective, self-reported results from the participants were quite favorable in that regard. 10 out of 13 participants responded that they felt comfortable being subjected to automatic detection (see figure 6.2). Participants stated that they would value giving the presenter more feedback, if accurate. Furthermore, several participants acknowledged elevated levels of anonymity, on the one hand by not having to share their webcam video with a potentially large audience and by only being visible to the presenter in an aggregated manner.

> I don't mind because I´m anonymous, so the presenter doesn't know who is confused for example. (Post-Study Questionnaire, LQ5)

> As a lecturer, you can see if people are really present… but participants can remain kind of anonymous and they don't have to reveal their appearance. As a listener, I believe, this would be more comfortable for me. (Interview Excerpt, Interview 6)

Another key finding from the listener questionnaire is that most participants observed the outputs of the automatic detection closely and most tried to either adjust their appearance or provide manual overrides if they felt the result was inaccurate. Due to the novelty of the approach for participants, this was to be expected, but in general, it should be studied in the future how much effort it takes to mediate the AI results compared to mediating your appearance on a webcam stream that is shared with participants.



How eager are you to use the tool to listen to online presentations compared to traditional conferencing tools such as Zoom?

Figure 6.3: Presentation Listening Experience Poll, 2

6.5.2 Presenter Perspective

Many participants described positive experiences when conducting a presentation with the prototype and particularly noted that it seemed easier to focus on the presentation due to minimal distractions and a summarized view of the audience.

> ... It was motivating to receive real-time feedback, in contrast to conventional programs where the audience is invisible. (Post-Study Questionnaire, PQ1)

All participants except one stated that the results from the automatic detection were useful to them and that they tried to incorporate the received feedback to adapt their presentation right away while holding the presentation.

> You can even during the presentation react to the audience or ask questions if there appear to be ambiguities. That's something that other [online presentation] tools don't offer because you cannot focus on all i.e. faces during a presentation. That I found great... (Interview Excerpt, Interview 2)

Similarly, the presentation summary, which shows a summary of average detection results and manual feedback after finishing the presentation (see 5.11) was mentioned positively by all but one participant. They mentioned that the information seemed suitable to adapt the presentation retrospectively or to identify critical parts of the presentation where engagement or confusion is particularly high or low, respectively. However, some participants mentioned, that the exact visualization of the summary plot could be improved. For example, reading the plot proved difficult to some participants, especially since the two detected metrics need to be read inversely because high engagement is positive while high confusion would be a negative finding.

> I also found the summary after the presentation very interesting because in my presentation, I first had a theoretical part introducing the topic and in the second part, I presented my results with Excel tables. And on the slide where I had the first Excel table, there was a noticeable increase in the confusion levels... that was a great insight. (Interview Excerpt, Interview 5)

Several participants noted that while they found the additional feedback from the prototype helpful for their presentation, they still missed seeing the participants' faces at times. Though many of them acknowledged that in practice the same is true for video conferencing tools, where many participants choose not to enable their camera. Nevertheless, several participants noted that they would prefer a solution that offers both the option of having a webcam enabled and automatic feedback detection. There might be situations where seeing each other is important and others where it is not, but some kind of feedback is still appreciated.

> As I often teach online courses in the field of teacher training, I know that the majority of participants do not switch on the camera during presentations. It takes some time for lecturers to get used to speaking with "black screens". I can well imagine that a combination of video-based conference tools and the prototype would bring many advantages: At the beginning of the lectures, you could meet and greet each other using video, and during the lecture, cameras can be switched off, as the parameters (possibly others) provide information about the mood. During presentations, you don't usually see the participants' cameras anyway ... (Post-Study Questionnaire, PQ4)

> I would integrate the tool as an additional option for online presentations. It can't completely replace Zoom or Teams, as I find the cameras, chat function and room allocation necessary. However, the prototype with its analyses helps the examiner and presenter to get quick feedback. Post-Study Questionnaire, PQ4)

CHAPTER 7

Discussion

The primary objective of this work was to study the current shortcomings of video conferencing tools in the context of holding online presentations. Subsequently, the potential of advancements in artificial intelligence and image processing techniques to address some of these deficiencies was explored. The main focus of this work was problems related to the phenomenon called Zoom Fatigue, which describes the phenomenon that online video conferences feel rather exhausting, compared to attending in-person presentations. These feelings are associated with using webcams and being subjected to a constant mirror of yourself, among others. Some researchers therefore suggest conducting video conferences without webcams enabled [Bai21]. To mitigate the loss of information that might otherwise be transmitted non-verbally via video, opportunities presented by state-of-the-art image processing techniques and affective computing were explored. The primary objective of this work was to address four key research questions. Subsequently, these research questions will be addressed, and the insights uncovered in the study will be discussed.

[**RQ1**] *What information do speakers miss in an online presentation setting when their audience webcams are disabled?*

Focus groups conducted with nine people experienced in holding online presentations revealed that information conveyed via webcams during online presentations was in fact minimal. While some participants stated that non-verbal cues can be transferred via webcams, most of the time this was limited to conferences with a small amount of people. When holding online presentations, however, participants used webcam videos merely to see that people are still in fact present, appear to be listening to the lecture or are available for interaction. This minimal information, however, was still deemed relevant to many participants. Furthermore, the attending lecturers pointed out that additional feedback while holding online presentations would be considered useful because there is a perceived lack of feedback compared to in-person presentations.

[**RQ2**] *Can non-verbal cues of participants of online presentations be detected automatically and communicated to the presenter via emojis?*

A literature and state-of-the-art review revealed opportunities and advancements in image processing and affective computing that promise capabilities to infer emotional and other information from webcam images. Modern machine learning algorithms and runtimes are efficient enough to process this inference on users' end devices which would ease some concerns when it comes to the security and privacy of such applications. However, the validity of affective computing applications is questioned by some researchers [TD21, Ric20]. Technology with affective computing components must be developed responsibly, with that proneness to error in mind.

Furthermore, effective engagement detection of listeners to online presentations solely by means of processing their webcam images entails several serious –potentially irreconcilable –impediments. The fact that webcam setups vary greatly can lead to various side effects, that may affect accurate detection. Some of those side effects may be remedied by instructing users to use consistent setups. Other side effects may be tackled automatically, e.g. for this prototype, we used face detection techniques to detect and crop the faces of listeners before any subsequent processing. However, the most severe constraint of this approach is, that it cannot be ensured that listeners lay their focus solely on the presentation the entire time. That is, even if a detection technique could correctly estimate listeners' properties at all times, occasions when listeners divert their attention to other targets, would render the estimation irrelevant to the presentation. Detecting the target of the listener's attention at all times would require more intrusive techniques which would likely be inappropriate.

During the design phase of the prototype, the initially envisioned communication medium of emojis was replaced by a less discrete system of color-coded progress bars. This decision was made based on the desire of participants to see a summary of averaged room statistics rather than individual values for all participants represented by discrete emojis. Furthermore, it proved hard to code the automatically captured parameters of engagement and confusion into discrete emojis. Instead, we decided to use progress bars to visually communicate the average value of engagement and confusion of the audience. To further facilitate the intake of the information, the progress bars were color-coded to highlight positive or negative detections.

[**RQ3**] *How do users perceive the usefulness of non-verbal cues sent during online presentations?*

Two focus groups conducted with lecturers and people experienced in holding online presentations revealed that there is a desire for more feedback and information regarding the state of the audience during online presentations. The evaluation of the prototype revealed that the information based on automatic detection of engagement and confusion and manual feedback from the audience was appreciated by most participants. Furthermore, most participants mentioned, that they instantly reacted to the feedback during the presentation e.g. by trying to adapt their presentation. However, it is crucial to

acknowledge that some participants noted a learning curve in seamlessly integrating the tool's insights, emphasizing the need for experience in using the tool and learning how to interpret the provided information.

The feedback also revealed a diversity of opinions on the visualization and communication of the gathered information. This suggests that a highly customizable approach to presenting this information to the presenter would be optimal, to be able to cater to varying preferences of users. In essence, the feedback emphasizes the potential of using visual analysis techniques as a useful additional feedback stream for enhancing the presentation experience while highlighting the importance of adaptability and user customization.

[**RQ4**] *Do users feel comfortable with automatic non-verbal feedback detection when participating in online presentations?*

In the scope of our user study, 13 participants were intentionally exposed to our approach of incorporating automatic detection of engagement and confusion with manual override capabilities within a reasonably realistic setting. Participants were asked to provide insights about their subjective experience of being exposed to automatic detection. The post-study questionnaire illuminates a generally favorable sentiment among participants regarding this approach, with 10 out of 13 participants feeling comfortable in the simulated situation. Notably, respondents expressed a positive outlook, particularly when comparing this approach against the prospect of having the webcam active throughout the entire duration of a presentation. Furthermore, participants acknowledged elevated levels of anonymity using the prototype, knowing that their information was only visible to the presenter in an average value of the presentation room.

7.1 Limitations

The main contribution of this work, next to demonstrating technical feasibility, was studying the practicability in a real-world scenario and capturing the experiences and opinions of people experienced in holding and attending online presentations. The developed prototype proved to be a helpful tool to study the usability of this technology in a realistic scenario and capture the subjective experiences of the participants. However, to be able to generalize findings beyond the participant pool, tests on a larger scale with a bigger participant pool need to be considered, to assess how well these findings translate beyond the participant pool and the specific situation simulated. For future tests, we would also suggest incorporating objective measures in addition to the subjective results that have been captured in this work.

Furthermore, we would also like to address a couple of limitations, especially emerging from the dataset used to train the automatic detection module. First, as noted in chapter 5.2, the distribution of the classes in the DAiSEE dataset is heavily skewed. For example, while 4477 samples for high engagement are present in the dataset, it contains only 61 for very low engagement. Similarly, 6024 samples for very low confusion are present, but

only 101 for very high confusion. This is even worse for class combinations. For example, there were only 5 samples of high confusion and low engagement in the dataset. Moreover, we want to note that previous methods have not achieved simultaneous engagement and confusion recognition, hence our experimental work was limited when it comes to transferring architectures from prior works to this setting. Overall, this presents a big challenge for robust recognition.

Moreover, cultural properties of emotions may present an additional limitation. The dataset that we used was based on video snippets of Indian students listening to a lecture. Cultural differences between how engagement and confusion manifest in facial expression may exist among various groups of people around the world. Overall, the availability of high-quality data sets on emotion recognition is limited.

7.2 Future Work

In this thesis, possibilities of how state-of-the-art image processing techniques can be used to enhance feedback in the online presentation experience were studied. Given the exploratory nature of this study, numerous avenues for future work have surfaced.

7.2.1 Research on Human-AI Interaction

An interesting insight from the user tests was the fact that some users attempted to influence the results from the automatic detection by e.g. changing their facial expressions. This behavior runs counter to the intended purpose of the prototype, which aimed to reduce the cognitive load of participants compared to webcam-enabled presentations. That being said, users don't always interact with systems the way developers and designers intended. Having AI-powered components as actors within systems complicates this even more. The growing prevalence of AI-driven systems underscores the need for further research into designing seamless Human-AI interaction. Furthermore, to facilitate acceptance and transparency of AI-enabled systems, it is vital, that users can comprehend the decision-making process of the AI. While the research field of explainable AI has gained traction in the last years, there are still many open topics to be explored and there is a lack of applicable guidelines for practitioners building novel systems.

7.2.2 Implications for the Design of Online Collaboration Tools

Throughout both the design and evaluation phases, diverse requirements surfaced among participants regarding, for example, the communication of automatic detection results in the prototype. When it comes to conveying detection results, some participants favored highly aggregated and summarized information, while others leaned towards more granular details. Some participants mentioned they prefer seeing numerical results of the detected audience's state while others preferred the more visual approach that was taken in the implementation of this prototype. A suggestion for future work would be the implementation of a customizable system, allowing users to fine-tune their experience

based on their specific preferences. Furthermore, participants voiced diverse opinions when it comes to whether or not is necessary to see webcam videos of listeners to online presentations. A pragmatic approach to satisfy the various preferences could entail implementing a system that initially shares webcam footage at the presentation's outset, then transitions to automatic feedback while deactivating video transmission. When it comes to using emojis as a communication medium for this feedback, we faced difficulties. Despite emojis' simplicity and effectiveness in non-verbal communication, visualizing the aggregated engagement and confusion data from listeners was better achieved using standard UI components like color-coded progress bars. This approach offered a more direct and comprehensive representation in the context of our use case.

CHAPTER 8

Conclusion

This thesis primarily sought to explore the technical feasibility of automating the detection of engagement and confusion within the context of online presentations while also delving into exploring its practicality in real-world scenarios. Based on state-of-the-art examples and techniques, a functional prototype incorporating automatic detection capabilities was developed. The design of the prototype was informed by insights from experienced presenters and lecturers. The resulting prototype was evaluated with user tests simulating a realistic online presentation scenario. The majority of participants rated the prototype favorably and noted its potential to enhance the otherwise limited feedback available for online presentations. Moreover, participants subjected to automatic detection of engagement and confusion expressed interest in this technology, particularly over the alternative of having to share their webcam video throughout online presentations. The thesis offers a solution proposal that enhances feedback for presenters and provides valuable insights for the design and enhancement of future online presentation and video conferencing tools. Key takeaways from this work are that advancements in artificial intelligence and image processing techniques offer avenues to improve the online presentation experience by gathering additional feedback, while not having to rely on transmitting webcam video. However, the use of emojis as a communication medium as initially conceived was found inadequate for conveying averaged, summarized information about online presentation sessions. In contrast, employing straightforward and color-coded UI components emerged as a more effective method for visualizing such averaged data. This thesis paves the way for further exploration, particularly into the field of Human-AI interaction.

List of Figures

List of Tables

Bibliography

[AC08] Anne Adams and Anna L. Cox. Questionnaires, in-depth interviews and focus groups. In Paul Cairns and Anna L. Cox, editors, *Research Methods for Human Computer Interaction*, pages 17–34. Cambridge University Press, Cambridge, UK, 2008.

[AK21] Ali Abedi and Shehroz S Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In *2021 18th Conference on Robots and Vision (CRV)*, pages 151–157. IEEE, 2021.

[ANK18] Jafar Alzubi, Anand Nayyar, and Akshi Kumar. Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1142(1):012012, November 2018. Publisher: IOP Publishing.

[ASLC22] Xusheng Ai, Victor S. Sheng, Chunhua Li, and Zhiming Cui. Class-attention Video Transformer for Engagement Intensity Prediction, 2022. _eprint: 2208.07216.

[Bai21] Jeremy N. Bailenson. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior*, 2(1):No Pagination Specified–No Pagination Specified, 2021. Place: US Publisher: American Psychological Association.

[BAM+19] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest*, 20(1):1–68, 2019. Publisher: Sage Publications Sage CA: Los Angeles, CA.

[BB22] Joel Bruckenstein and Veres Bob. Global market share of videoconferencing software 2022, by program, 2022.

[BBG+20] Viola Bulgari, Mattia Bava, Giulia Gamba, Francesco Bartoli, Alessandra Ornaghi, Valentina Candini, Maria Teresa Ferla, Marta Cricelli, Giorgio Bianconi, Cesare Cavalera, and others. Facial emotion recognition in people

with schizophrenia and a history of violence: a mediation analysis. *European archives of psychiatry and clinical neuroscience*, 270(6):761–769, 2020. Publisher: Springer.

[BC06] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[BEE20] Austin Beattie, Autumn P. Edwards, and Chad Edwards. A Bot and a Smile: Interpersonal Impressions of Chatbots and Humans Using Emoji in Computer-mediated Communication. *Communication Studies*, 71(3):409–427, 2020.

[BMF22] Bärbel Bissinger, Christian Märtin, and Michael Fellmann. Support of Virtual Human Interactions Based on Facial Emotion Recognition Software. In *International Conference on Human-Computer Interaction*, pages 329–339. Springer, 2022.

[Cai16] Kelly Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 981–992, 2016.

[CBC+21] Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027, 2021. Publisher: Elsevier.

[CMM+22] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

[Col14] Giovanna Colombetti. *The feeling body: Affective science meets the enactive mind*. MIT press, 2014.

[CPR+22] Bernardo P. Cavalheiro, Marília Prada, David L. Rodrigues, Diniz Lopes, and Margarida V. Garrido. Evaluating the Adequacy of Emoji Use in Positive and Negative Messages from Close and Distant Senders. *Cyberpsychology, Behavior, and Social Networking*, 25(3):194–199, 2022. _eprint: https://doi.org/10.1089/cyber.2021.0157.

[DGGS16] Abhinav Dhall, Roland Goecke, Tom Gedeon, and Nicu Sebe. Emotion recognition in the wild. *Journal on Multimodal User Interfaces*, 10, March 2016.

[Ekm93] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. Publisher: American Psychological Association.

[Eld18] Alexis M Elder. What words can't say: Emoji and other non-verbal elements of technologically-mediated communication. *Journal of Information, Communication and Ethics in Society*, 2018. Publisher: Emerald Publishing Limited.

[EP19] Hanna Eriksson and Emelie Parflo. Mobile application onboarding processes effect on user attitude towards continued use of applications, 2019. Backup Publisher: Jönköping University, JTH, Computer Science and Informatics.

[FBL16] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016. Publisher: HeinOnline.

[FBP04] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1):59–109, 2004. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

[Fre18] Alisa Freedman. Cultural literacy in the empire of emoji signs: Who is crying with joy? *First Monday*, 2018.

[GDAB16] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. DAiSEE: Towards User Engagement Recognition in the Wild, 2016.

[GEC+13] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, and others. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013.

[GSP21] Vasile Gherhes, Simona Simon, and Iulia Para. Analysing students' reasons for keeping their webcams on or off during online classes. *Sustainability*, 13(6):3203, 2021. Publisher: MDPI.

[Hem96] Michael Hemphill. A note on adults' color–emotion associations. *The Journal of genetic psychology*, 157(3):275–280, 1996. Publisher: Taylor & Francis.

[HSE+17] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. EngageMeter: A System for Implicit Audience Engagement Sensing Using Electroencephalography. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 5114–5119, New York, NY, USA, 2017. Association for Computing Machinery. event-place: Denver, Colorado, USA.

[HZ20] Guy Hoffman and Xuan Zhao. A Primer for Conducting Experiments in Human–Robot Interaction. *J. Hum.-Robot Interact.*, 10(1), October 2020. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[KAF22] Dima Kagan, Galit Fuhrmann Alpert, and Michael Fire. Zooming Into Video Conferencing Privacy. *IEEE Transactions on Computational Social Systems*, pages 1–12, 2022.

[KARL20] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Context Based Emotion Recognition Using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2020.

[KE04] Naz Kaya and Helen H Epps. Relationship between color and emotion: A study of college students. *College student journal*, 38(3):396–405, 2004. Publisher: PROJECT INNOVATION INC.

[Kol22] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336, 2022.

[KTRB12] Peter Kuppens, Francis Tuerlinckx, James Russell, and Lisa Barrett. The Relation Between Valence and Arousal in Subjective Experience. *Psychological bulletin*, 139, December 2012.

[KURD22] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.*, 55(2), January 2022. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[Li22] Yinglong Li. Research and Application of Deep Learning in Image Recognition. In *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, pages 994–999, 2022.

[LKHK22] Mathieu Lutfallah, Benno Käch, Christian Hirt, and Andreas Kunz. Emotion recognition-a tool to improve meeting experience for visually impaired. In *International Conference on Computers Helping People with Special Needs*, pages 305–312. Springer, 2022.

[LRRX+22] Na Li, Guillermo Romera Rodriguez, Yuqiao Xu, Parth Bhatt, Huy A. Nguyen, Alex Serpi, Chunhua Tsai, and John M. Carroll. Picturing One's Self: Camera Use in Zoom Classes during the COVID-19 Pandemic. In *Proceedings of the Ninth ACM Conference on Learning @ Scale*, L@S '22, pages 151–162, New York, NY, USA, 2022. Association for Computing Machinery. event-place: New York City, NY, USA.

[LUM15] James R Lewis, Brian S Utesch, and Deborah E Maher. Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction*, 31(8):496–505, 2015. Publisher: Taylor & Francis.

[LWP+18] Miki Liu, Austin Wong, Ruhi Pudipeddi, Betty Hou, David Wang, and Gary Hsieh. ReactionBot: Exploring the Effects of Expression-Triggered Emoji in Text Messages. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[Man21] Emmanouela E. Manganari. Emoji Use in Computer-Mediated Communication. *The International Technology Management Review*, 10(1):1–11, 2021.

[MHM19] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.

[NSI+21] Kosaku Namikawa, Ippei Suzuki, Ryo Iijima, Sayan Sarcar, and Yoichi Ochiai. EmojiCam: Emoji-Assisted Video Communication System Leveraging Facial Expressions. In Masaaki Kurosu, editor, *Human-Computer Interaction. Design and User Experience Case Studies*, pages 611–625, Cham, 2021. Springer International Publishing.

[ON15] Keiron O'Shea and Ryan Nash. An Introduction to Convolutional Neural Networks. *CoRR*, abs/1511.08458, 2015. arXiv: 1511.08458.

[Ric20] Sharon Richardson. Affective computing in the modern workplace. *Business Information Review*, 37(2):78–85, 2020. Publisher: SAGE Publications Sage UK: London, England.

[Ros84] Stephanie Ross. Painting the Passions: Charles LeBrun's Conference Sur L'Expression. *Journal of the History of Ideas*, 45(1):25–47, 1984. Publisher: University of Pennsylvania Press.

[SC12] Frank Serafini and Jennifer Clausen. Typography as Semiotic Resource. *Journal of Visual Literacy*, 31(2):1–16, 2012. Publisher: Routledge _eprint: https://doi.org/10.1080/23796529.2012.11674697.

[Sch21] Amy M. Schmitter. 17th and 18th Century Theories of Emotions. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2021 edition, 2021.

[SD16] Weisong Shi and Schahram Dustdar. The Promise of Edge Computing. *Computer*, 49(5):78–81, 2016.

[SKG20] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020. Publisher: Institute of Electronics and Computer.

[SPVA22] Živilė Sederevičiūtė-Pačiauskienė, Ilona Valantinaitė, and Vaida Asakavičiūtė. 'Should I Turn on My Video Camera?'The Students' Perceptions of the use of Video Cameras in Synchronous Distant Learning. *Electronics*, 11(5):813, 2022. Publisher: MDPI.

[SSM22] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022. Publisher: IEEE.

[STT21] Tomoya Suzuki, Akihito Taya, and Yoshito Tobe. VFep: 3D Graphic Face Representation Based on Voice-based Emotion Recognition. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 74–79, 2021.

[SVGTO23] Ramandeep Kaur Sandhu, João Vasconcelos-Gomes, Manoj A. Thomas, and Tiago Oliveira. Unfolding the popularity of video conferencing apps – A privacy calculus perspective. *International Journal of Information Management*, 68:102569, 2023.

[SYD+21] Mike Seymour, Lingyao Ivy Yuan, Alan Dennis, Kai Riemer, and others. Have we crossed the uncanny valley? Understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *Journal of the Association for Information Systems*, 22(3):9, 2021.

[TD21] Anna Tcherkassof and Damien Dupré. The emotion–facial expression link: evidence from human and automatic expression recognition. *Psychological Research*, 85(8):2954–2969, 2021. Publisher: Springer.

[TR11] Jessica L. Tracy and Daniel Randles. Four Models of Basic Emotions: A Review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, 3:397 – 405, 2011.

[Tud22] Cristiana Tudor. The Impact of the COVID-19 Pandemic on the Global Web and Video Conferencing SaaS Market. *Electronics*, 11(16):2633, 2022. Publisher: MDPI.

[ZJWG+22] Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in*

Computing Systems, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery. event-place: New Orleans, LA, USA.

Appendix

Focus Group 1 - Coded Transcript

#	Part.	Statement	Codes
1	B	Die großen Herausforderungen [bei Online Teaching] sind mangelndes Commitment und Interaktion. Ich kann nicht genau sagen woran es liegt - im Zweifel sage ich an mir - aber ich glaube das Setting ist viel größer. Wir hatten in den Freiwilligenkursen viele Leute dabei, die wenig oder nicht vorbereitet gekommen sind.	lack of commitment lack of interaction in online pres
2	B	Die Interaktion ist anders als im physischen Setting aus mehreren Gründen. 1) teilweise sieht man die Leute nicht, weil die Kamera ausgeschalten ist. 2) selbst wenn man die Leute sieht, sind sie nicht immer präsent. Da ist es schwer einzuschätzen bzw. in einem Interaktionsmodus überhaupt darauf zu referenzieren ob jemand gerade dabei ist. 3) Es ist gefühlt eine andere Art von Setting... man interagiert gefühlt zu 80% mit seinem Foliensatz und zu 20% mit der Auswahl an Bildern oder Buchstabenkürzeln die durch den Algorithmus nach vorne gespült werden.	lack of interaction in online pres
3	B	Die größte Herausforderung ist meine Motivation, weil die mittlerweile durch diese Erfahrungen [Online Presenting/Teaching] auf ein Minimalmaß gesunken ist. … Ich habe mittlerweile gar keine Lust mehr Vorlesungen online zu halten.	loss in motivation b/c of online pres
4	C	Als Unterstützerin wurde die Erfahrung gemacht dass sehr wenig Interaktion und Teilnahme an Online Vorlesung im Laufe des Semesters stattfindet. Die Studierenden sind nicht bereit die Kamera einzuschalten. Wir wissen nicht was sie machen, wir reden ins Leere. Das ist eine große Herausforderung für den Vortragenden. ...	lack of interaction in online pres lack of transparency for presenter
5	C	Umgekehrt wurden in einer LV eher positive Eindrücke im Bereich Online Vortrag gesammelt. Da war es relativ spannend weil die Interaktion relativ gut funktioniert hatte - das war mit einer Universität in Luxemburg gemeinsam. Dadurch, dass das eine LV-Übung war, war das Commitment vl ein bisschen höher, dass man da interaktiv teilnimmt. Am Anfang hatten wir das Problem mit dem Kamera-einschalten. Sobald sie eingeschalten waren konnte man allerdings sehen, dass die Leute aufpassen und es wurden auch Fragen gestellt. Zumindest hatte man das Gefühl, dass ein Großteil dabei war und aufgepasst hat während der LV.	rare good experiences with online collaboration
6	C	Wenn man davon ausgeht, dass alle Kameras ausgeschalten sind, weiß man nicht ob die Person überhaupt vorm Laptop sitzt oder nur teilgenommen hat und im Nachbarraum ist… Dann falls die Person anwesend ist weiß man nicht, ob die Person gerade etwas anderes macht oder aktiv der Präsentation folgt. Man weiß nicht ob die Person das interessant findet was vorgetragen wird, oder ob die Person gelangweilt ist. Das ist die Information die fehlt...	lack of transparency are people even present? lack of feedback reg. content of presentation
7	A	Für mich ist die Frage wie viele Teilnehmer im Raum sind. Sind es wenige, wäre es einfach jede Person zu addressieren und Feedback zu erfragen. … Wenn im Termin z.B. 40 Leute sind, … dann hat man nicht so viel Flexibilität. Dann kann man in die Runde fragen ob jemand Feedback hat und meistens melden sich 1,2 Personen.	situation dependent of room size little response when asking questions
8	A	Bei anderen Beteiligungsformen, z.B. Liken, Frage im Chat, Umfrage, … ist die Beteiligung [meiner Erfahrung nach] ca 50 %. … Im Vergleich zu offline Vorlesung fehlt der Augenkontakt …. in der Vorlesung bekommt man aus der Atmosphäre einen Eindruck ob das heute funktioniert hat oder nicht. Bei Online Vorlesungen glaube ich, wenn es keine Beschwerden gibt, dass alles gut gelaufen ist. Und wenn viele Rückmeldungen kommen ... dann bekommt man einen Eindruck [was man verbessern kann das nächste mal]	moderate response i.e. for poll/reaction only extreme feedback (no feedback or lots of)

9	B	Was ich schwierig finde ist, dass Mimik und Gestik als Interkationskanal nicht vorhanden sind. Das ist nicht nur online ein Thema, das hat mach genauso massiv in der [Corona] Übergangsphase gestört, als die Menschen mit Masken in der Vorlesung saßen. Es ist ganz schwer erkennbar z.b. ob jemand lacht wenn er eine Maske auf hat. Das nimmt einen Teil der Interaktion raus. Ich glaube das betrifft mich besonder stark weil ich versuche viel spontan mit dem Publikum zu interagieren und dafür braucht man diesen Kanal irgendwie, ob man aufmerksam und dabei ist.	lack of facial expressions lack of instant feedback
10	D	Es gibt diese Handheben Funktionen oder den Chat aber ich habe das Gefühl es werden online weniger Fragen gestellt oder die Hand gehoben als vor Ort. … Online heben die Leute weniger die Hand und stellen Fragen sondern warten oft auf eine explizite Sektion/Aufforderung für Fragen. Und bis diese kommt, haben sie die Frage oft schon vergessen. ... Außerdem merkt man als Vortragender oft nicht wenn eine Hand gehoben wurde	lack of questions lack of "connectedness"
11	D	Es kommt zu wenig Signal von den Systemen dass es Fragen gibt im Chat.	lack of signifiers i.e. hand raised or question asked
12	A	Ich hatte 3 Bildischirme für Online Präsentation weil mit 2 habe ich es nicht geschafft, dass ich alles sehe was ich sehen musste.	difficulty to find right technical setup
13		[Ansprechen dass Kameras eingeschalten werden sollen], machen 3 von 4 Teilnehmern. Die Frage wie man es forciert.	request to turn on cameras
14	B	Bei Vorlesungen kann man es ansprechen aber nicht forcieren. Bei den Übungen - wo ein Teil der Leistungsüberprüfung Mitarbeit ist - schon eher. Da ist auch die Bereitschaft von den Teilnehmern höher.	lack of motivation to turn on cameras
15	A	Während Covid Zeit hat es funktioniert, dass am Beginn der Einheit alle die Kamera eingeschaltet haben. Wir wollten uns sehen und Beziehung aufbauen… aber wenn wir heute nachfragen, gibt es für viele kein Motiv dafür. [Erfolgsquote ist gesunken]	lack of motivation to turn on cameras
16	B	Ich würde unterstützen - das die Erfolgsquote die Kamera einzuschalten - gesunken ist. … ich glaube am Anfang [der Covid-Pandemie] war jeder froh dass es überhaupt weitergegangen ist. Da war die Bereitschaft selbst was zu investieren höher… jetzt ist der Erwartungshaltung an den Lehrenden höher	lack of motivation to turn on cameras
17	B	[Die Ansicht der User] sollte nicht so starr rechts sein, sondern vielleicht in Format eines Vorlesungsraumes.	visualize users like in a lecture hall
18	B	Wenn es eine größere Anzahl an Zuhörenden ist wäre eine aggregierte Version der Darstellung gut. Ich glaub das ist auch das was man in einem großen Hörsaal wahrnimmt, bzw. was die Erwartung der Interaktion ist.	aggregate room stats
19	A	Normalerweise in der Vorlesung sehe ich zwei Leute aufstehen und denke mir mmhmm… Dann bekomme ich indirekt den Eindruck etwas geht schief (auch wenn die vielleicht andere Gründe haben) Oder Leute schauen in andere Richtungen. Was kann ich verbessern/ändern? ... ich kann evtl eine Pause anbieten oder eine Zwischenfrage stellen. Aber online wenn ich so viele Signale bekomme vom System ... was kann ich verbessern? Vielleicht wäre es besser eine Zusammenfasung nachträglich zu bekommen. Aber simultan in der Vorlesung kann ich wahrscheinlich nicht viel ändern weil ich muss mich auf die Folien konzentrieren.	avoid cognitive overload for presenter
20	A	Ich glaube nicht jede Studierende muss bis am Ende 100%ig motiviert sein… manche Themen sind Grundlage und müssen einfach gelernt werden. Es muss evaluaiert werden in wie weit ich überhaupt etwas verbessern kann.	finding baseline of appropriate user engagement
21	D	Synchrones Feedback könnte einen stressen. Die Frage ist ob ich nicht sogar mehr informationen bekomme als ich live bekomme…	avoid cognitive overload for presenter

22	A	Ich glaube aggregiert ist sehr sinnvoll - evtl sogar über ganzes Semester… dann könnte man z.g. sagen ganz am anfang waren alle motiviert, am ende nicht. ist das normal oder hat sich dieses Semester etwas geändert. … solche Informationen wären sehr hilfreich. Aber ganz simultan hätte es Effekte aber ist vielleicht nicht so sinnvoll.	aggregate room stats asynchronous information
23	C	Detailliertes Simultan Feedback wäre ws interessant aber besser wären aggregierte Kennzahlen, weil sonst würde es ws ablenken, wenn ich von z.b. 40 zuhörern die einzelnen Information permanent eingebelendet bekommen. Aber vielleicht wäre eine persönliche Einstellung sinnvoll.	avoid cognitive overload for presenter aggregate room stats
24	C	Ob das seitlich eingeblendet wird oder aggregierte Kennzahlen ist wahrscheinlich persönliche Präferenz und auch abhängig von der Teilnehmeranzahl.	provide different visualization modes
25	B	Für mich als Vortragender wären die Kennzahlen motivierender als die Namen. Und ich glaube es wäre auch geschickter als die Videos der Personen. Weil im Normalfall sieht man eine Person, oft ganz schlechte Qualität und das hilft mir eigentlich gar nicht... außer das ich weiß da sitzt jemand in dem ich hineinreden kann. Ich kann aber nicht sagen ob das mittelfristig [sinnvoll ist] bzw. ob ich das zur Interaktion miteinbeziehen würde.	anonymity, hide names aggregate room stats webcam video not useful
26	B	Die Basislinie muss man hinkriegen.. es ist bei einer normalen Vorlesung auch nicht üblich dass 100% aufmerksam sind	finding baseline of appropriate user engagement
27	B	Das Kriterium für eine gute LV ist nicht dass alle Studierenden 100% glücklich und zufrieden sind.... sondern dass etwas gelernt wurde	
28	A	Es gibt Bereich Learning Analytics… wo man Daten nutzen kann… die Informationen wäre dazu sinnvoll	retroactive analytics
29	D	Wenn jemand eine Frage hat oder etwas nicht versteht, wäre es gut wenn man es gleich sieht. Vl traut sich die Person auch nicht eine Frage zu stellen. Es wäre auch gut wenn das vl. anonym wäre und wenn eine bestimmte Grenze überschritten wurde könnte man als Vortragender vl. noch einmal erklären.	pro confusion indicator anonymity
30	B	Bei einer Vorlesung würde ich das nicht haben wollen. Die Ebene die mich interessiert ist Aufmerksamkeit. … bei einer guten Vorlesung verstehen die Leute auch nicht 100%. .. max. 80% weil den Rest möchte ich so haben dass die Leute aktiviert sind… und sich selber auf-schlauen. ... ich empfinde es als übergriffig	contra confusion indicator finding baseline of appropriate confusion
31	B	Ich halte das Thema Analytic für gefährlich… bei solchen Sachen sollte ein Mehrwert für die beteiligten Personen unmittelbar erkenntlich sein. Das sehe ich beim Thema Interaktion gegeben aber nicht bei Analytic.	contra large-scale retroactive analytics
32	A	Man müsste von den Studierenden auch zusätzlichen Consten einfordern. … wen man die Daten nicht speichert sonder nur unmittelbar in der Vorlesung zeigt, könnte es mit der DSGVO konform sein.	privacy concerns
33	C	Was auch spannend ist, nur in die Kamera schauen bedeuted nicht unbedingt aufmerksam zu sein… ich schreibe z.b. oft gerne mit (und schaue dann weg von der Kamera)	questionable accuracy of autom. detection
34	B	Eine schöne Analogie zum Prüfen der Aufmerksamkeit ist der Totmann-Schalter… das wäre eine schöne Analogie zum Umsetzen, z.B. über Mausbewegung oder irgendeine Art von Interaktion.	dead man's switch
35	D	[Zusammenfassung Chart am Ende] Wäre es nicht besser wenn man das anonymisiert?	anonymity

36	A	(Zusammenfassung Chart am Ende] Das ist ein Feedback-Kanal am Ende des Tages… als Selbst-Feedback an Dozenten könnte das auf jeden Fall gut sein - vor allem für Junior Vortragende. Aber das würde ich niemals mit den Studierenden teilen. … für die Vortagenden finde ich das sehr sinnvoll.	valuable feedback for optimizing lecture
37	B	(Zusammenfassung Chart am Ende] Das würde ich sicher nutzen um nochmal die Folien quer-zu-checken… aber dann muss man überlegen was sind wirklich die Erfolgs-Kriterien einer guten LV. z.B. wenn nebenbei irgendeine Breaking-News ist und plötzlich sackt die Aufmerksamkeit aller Zuhörer ab.	valuable feedback for optimizing slides
38	D	Es wäre gut wenn die Zeit dabei stehen würde (Uhrzeit, Dauer)	provide timeline / duration of slides
39	C	Es lässt viel Interpretationsspielraum..	difficult to correctly interpret results
40	B	Auf der anderen Seite ist es mehr Information als das was wir jetzt haben. Wenn es den Vortragenden hilft da nochmal drüberzuschauen ist das sicher gut. Ich würde das auf jeden Fall nutzen. Ob man die richtigen Rückschlüsse zieht hängt dann von anderen Faktoren ab.	more information than status quo difficult to correctly interpret results

Focus Group 2 - Coded Transcript

#	Part.	Statement	Codes
1	E	I thaught two semesters online during covid... I was trying to involve students in discussion but it was a nightmare to do it. ... It is difficult to keep track of whether people who are there but really don't have anything to add or whether they really actually left already. So I had to ask them to turn on the camera mandatory.	lack of feedback request to turn on cameras
2	F	It was the same issue that we faced that student would shut their cameras off and I think there is a policy that you cannot force students to switch on their cameras when you're recording the lectures. ... we did not know whether or not people were engaged during our lectures. only people who responded in the poll, or some (3 or 4) would be proactive. so what we did was just randomly calling out people: you answer this question. Which was kind of uncomfortable for students but I think it was a challenge to understand the engagement in general.	lack of transparency for presenter forcing people to participate by calling them out handful of proactive people who were "present"
3	G	The challenges were mostly about deciding whether to record something offline and giving the students the option to watch it whenever they want vs. having the teaching being done in real time i.e. with Zoom with the downside of having this missing information about engagement b/c it would be missing this information regardless. My previous university tried to keep record of who at least has watched the video, who has opened the file or how long they watched it so that they could make some kind of statistics about the engagement. But I would agree that engagement was the most challenging factor.	lack of engagement
4	H	I had one interesting online class taught via Twitch. It was a very fancy and very cool online class, more students joined the course compared to the previous years. I am trying to say, maybe online teaching does not have to be boring. It really depends how much effort the instructor or the people who join the meeting want to put into the meeting.	interesting teaching platforms: twitch
5	H	I enjoy this type of online course because sometimes I am a little bit afraid of asking questions in the physical class... but in an online course you can ask what you want because nobody knows who you are.	online lecture: low barrier for shy people
6	H	The challenging part about online classes was the missing eye contact. Sometimes the instructor cannot motivate people without that.	missing eye contact
7	I	I thaught a big lecture during covid. The interaction part did not change so much because in a big lecture you do not have that much interaction even if offline. What I really felt was the difference of not getting any direct feedback. You were somewhat seeing all these tiny black boxes that never had any reaction.	also in big in-person lecture little immediate interaction lack of feedback in online lecture
8	I	I once even had the situation that my internet connection broke and I didn't realize for probably about 15 min. because nothing changed really and then one moment I received a message on my phone from a colleague that students were reaching out to them that actually the connection broke down. In some way there was just missing feedback. That moment, however, I realized, ok - there's actually someone listening and is interested in the connection to work which was quite positive feedback in a way. But in a lecture context this not really having any feedback just by watching students or yawning, or students focussing you - normally there is at least a little bit of feedback.	funny anectode
9	F	I have a similar story. I was attending a course last year - it was taught hybrid - and we were 6 or seven students. He muted us and he thought that he was sharing his screen, but he did not. The whole lecture we just saw half of him and no screen. We kept writing in the chat to notify him but he only realized towards the end of the lecture.	funny anectode
10	E	I had a similar experience but the other way round. The lecturer muted himself but didn't noticed. So sliced continued to go throughout 40 minutes but we didn't have any sound. People were also trying to let him know but he didn't hear.	funny anectode

11	F	In an in-person lecture you can really gauge whether people are listening to you. In an online presentation that is missing completely even if you have all cameras enabled. Also asking questions is really hard. … you have to rely on very proactive people.	lack of feedback webcam video not useful
12	E	People do not like to talk in their screens b/c it feels like everybody else is staring at you. In an in-person lecture only some people stare at you while you are talking while online it feels like everyone is watching you when cameras are enabled. …	webcam problems (ppl are staring at you)
13	I	I think it also has to do with missing eye contact because even though we see the videos of other people it seems they always look away slightly.	webcam problems (not directo eye contat)
14		[question: are there less follow up questions online compared to offline?] not conclusive	
15	I	When you have bad connection you do not know whether you can talk or if you are talking over somebody since there is a slight delay sometimes.	difficulty to collaborate online
16	E	A really cool feature of online video conferencing tools are polls. In in-person setting it is really hard to accomplish but online it's very easy to do, also anonymously.	manual feedback / polls very useful
17	G	We incorporated feedback pools in some previous courses but mostly in the end of the semester and about technical aspects. We never asked people: did you understand this or that.	currently no immediate feedback / only end of semester
18	G	If you are holding a lecture in-person you can use your own intonation, your eye-contact with someone, etc to try to in a way encourage them to stay with you. … Those are all things that are missing on the screen. That's not feedback but feed-forward though.	
19	E	Sometimes I use the reaction-functionalities to make lectures more interactive by asking people a questions and letting them react with specific symbols. … it's a good opportunity to see if they are engaged …	manual feedback (instant) by reactions useful
20	G	With a lecture there would be - as a student - no real benefit of turning on your camera and this is why I think a lot of students are like "I'm not really gaining anything from this, all this is putting on me is more pressure because I need to always be mindful of where I'm looking and what I'm looking at even if I'm paying attention I i.e. cannot look at my phone because this will be a sign that I'm not paying attention even if I have."	lack of motivation to turn on camera
21	G	So there are some people who have different kind of information processing. There are people who mostly rely on audio. They would choose, even if they were sitting in the lecture hall, to not look at the presenter or slides but just look down and listen to them, and this is how they pay attention. Measuring engagement i.e. by detecting whether or not they look on the screen takes away from the whole experience at least for them.	questionable accuracy of autom. detection b/c of high variation
22	I	I think where some additional pressure comes from [of turning on webcam] is the fact that in virtual meetings you constantly see yourself and immediatley get feedback about how you look/behave/act… while when you're sitting in a lecture hall you can sit wherever you want and you don't have a mirror of how you behave.	webcam: constant mirror
23	E	[question: why do you encourage students to turn on camera? what information do you gain?] I do it simply to see that they're in the meantime are not enaging with another course or went to the kitchen. So actually the minimum threshold is just to see that they didn't go away. It doesn't help to see their facial expression and also it's quite annoying to the students but I least, I think, they feel more pressure to be involved in the course. ... From my point of of view, seeing the webcam stream for non verbal cues does not help at all.	webcam primary focus: seeing if someone is here

24	F	Yeah, I think it's just building pressure for them [the students] to be available and not just log in and go away.	webcam primary focus: seeing if someone is here pressure for students to stay here
25	I	It's a bit similar to mandatory lectures, [via the camera] as a lecturere you can see that students are at least physically there but it also leads to a bad feedback-loop since it increases the level of discomfort for the students.	
26	E	So it's some AI-thing that decides whether I'm engaged or not? [yes] wow.. that's even more intrusive	privacy concerns
27	I	I think it would be very important for the students [to see their own emoji/detected state] to give them possibilities to change their attitude.	make process transparent for students
28	G	With having a webcam enabled traditionally, students have pressure to constantly present themselves. This, however, adds another layer of confusion. Because this model is not perfect, it can due to any kind of error misdetect the current state I'm in and now I also have the feeling that I'm also kind of auditioning for this AI. And then you will eventually run in this situation where you have: "oh.. the best way to trick the AI is by doing so and so... i.e. if I play a game on my phone but I look at the camera every 3 seconds I'm fine"	enjoyability concerns for listeners
29	E	Or with a fake smile…	questionable accuracy of autom. detection
30	G	I think a granular view is not the best for either the presenter or the student. I think it's not about individuals but whether you still "have the room" whether they are still with you or not. And I think in this case an aggregated view is more helpful and it removes the pressure on individuals because I'm now part of a whole, the whole room is mostly engaged or the whole room is mostly distracted. And I don't think having this aggregated view actually removes something from the instructor because thinking about it from the instructors perspective: you will not run through this list of say 100 people to gauge whether everyone is engaged so I guess just one figure or one piece of information is actually more meaningful because you have to spend less effort to get the meaning or figure it out.	aggregate room stats avoid cognitive overload for presenter
31	E	I agree on the notion of that it is intrusive, because the system puts you into a box with a limited amount of symbols. I love the idea that an AI detects that no person is there because the only reason why I watch the video is to see if a person is there. ... I would like it to be granular but maybe students can decide what emoji is put there.	detecting if person is there sufficient
32	G	It would be more useful to have this kind of voting systems to have on i.e. different kind of slides. So before changing the slide you can vote if everything is understandable on this slide. … For the instructor to kind of estimate how much of the content is understood. ... this removes some of the cognitive load from the instructor to scan the crowd and try to read that from the faces.	add manual feedback
33	I	I was wondering… do we even need the names. Maybe we can show like a lecture room with fixed position so we know well the guy in the right is always sleeping, but he's always there so we have an association with specific spots and we do not actually need the name. ... I think here putting the names would be too much information. And this would only work for a seminar with 15-20 people but for a big lecture it would be too much information.	anonymity (remove names)
34	I	I think here a bigger granularity like an average would be better, I'm not interested if a single person switches between two states, that's confusing when you're having a lecture.	aggregate room stats

35	F	I do not care if a 100% or 90% of people are engaged. I need to have that minimum threshold, i.e. 60% and now the system tells me it's time to worry about it. … of course not everyone is going to be fully engaged all the time of the lecture.	finding baseline of appropriate user engagement
36	F	Granular would probably break the flow of me teaching … it would be very disturbing for me as an instructor	aggregate room stats
37	E	Would would be interesting to have as aggregated would be some kind of question mark to show the percentage - but this is not an AI thing ... - how many people pressed this question mark and then I can see oh… most of the people just didn't get what I said. ... and this question mark can grow and maybe get more red depending on how many people pressed it	add manual feedback (i.e. self-reported confusion)
38	G	You could also show a timeline after the presentation finished and show i.e. slide one you had everyones attention, slide two people started asking questions, slide three was difficult to understand, slide four was OK. … and then you as a lecturerer can revise your slides ... this could be very useful.	provide timeline (after presentation) with detailed summary
39	E	I really do not like discrete symbols that put boxes in specific boxes… no matter how many boxes you have. I prefer to have continuous scales, whether it's color or something else. … then it's less intrusive for the people	avoid discrete symbols (don't put ppl into boxes)
40	G	The question is how you capture that information. if it is self-reported then it puts more effort on the students but if it is automatically detected then it's up to whatever mechanism you use and if is accurate or not. …	self-reported vs automatic
41	E	I think with confusion this could be easily self-reported. With engagement on the other hand student probably would not answer truthfully, here AI could be involved. But for confusion I think it would be better to ask students.	add confusion self-reported
42	G	I wouldn't think that something like that could be detected universally, so I think there should be some mechanism to override that …	questionable accuracy of autom. detection
43	I	I would very much prefer aggregated information, i.e. listening, interest, energy-level, to actually be able to react to that immediately.	aggregate room stats avoid cognitive overload for presenter
44	I	Maybe you could also give the students options to directly give feedback and in some way ovverride what you are inferring from the webcam, so that the users can directly say, "Im very tired now" or "this is tough, I really don't understand" so that you don't only rely on the automatic inference which might be prone to some error. ... this would probably improve the quality of the feedback	provide ways to override automatic detection for listeners
45	G	You also wanna avoid having just to rely on the inference because it could be more or less accurate depending on cultural differences. without the power to override you basically put yourself under the mercy of how accurate your system is.	provide ways to override automatic detection for listeners
46	E	The problem with engagement is, that students will probably always override saying: "no no, I'm totally engaged"	questionable accurary of self-reported properties
47	I	I would actually disagree, the moment a student is engaged they will use manual feedback tools. if they are sleeping an not engaged at all they just won't use it at all	
48	I	Another problem of virtual meetings is also that you have less feedback for the lecturer but also you do not have any interaction with your peers	lack of interaction with peers
49	G	I think it would be a good design choice to have more detailed information after the lecture and less information during the lecture	avoid cognitive overload for presenter details after the lecture
50	I	I think it would be very interesting to see the summary of i.e. the engagement per slide… that would be very useful for revising your slides	provide timeline (after presentation) with detailed summary

Questionnaire - Coded Results

Post-Study Questionnaire, Presenter Q1

P1) How does your online presentation experience with the prototype compare to your experiences i.e. with Zoom?

Responses ↓; Codes →	positive remarks	better focus	neutral response	lack of some sort
I prefer it when I can see participants. I find it confusing when I can't see the participants.				x
The user interface was very practical, including being able to see where you are in the presentation. It was motivating to receive real-time feedback, in contrast to conventional programs where the audience is invisible.	x	x		
This tool is very manageable and helpful for getting the mood of the audience and receiving feedback. Very user-friendly and well prepared.	x	x		
I quite liked the interface and I would say that the presentation experience was quite similar. The only thing I was missing was the information on the elapsed time of my presentation since this information is collected anyway.			x	x
The experience was positive, everything worked and it was nice to see that more people listening seemed to be engaged than confused. The tool itself is quite handy, there were no problems using it.	x			
I found presenting very pleasant. In particular, I found the fact that you don't have to share the screen when giving Power Point presentations very positive and stress-free. I also kept looking at the parameters for engagement and confusion that were displayed. If there was too much confusion, I tried to present my content in a more understandable way. Of course, there is still a lot of functionality that needs to be built in so that it can compete with Zoom etc.	x		x	
Zoom contains much more information than the prototype used. I am missing the names of the participants. I didn't have the opportunity to Chat to the participants. I thought the feedback in the prototype was very good as you could get a sense of how interested the participants were.	x			x
It was a different approach compared to Zoom. The video windows that are usually on the screen didn't distract the me during my presentation. I was able to focus more on the slides and what I was talking about.		x		
	5	3	2	3

Post-Study Questionnaire, Presenter Q2

P2) Do you feel that the metrics analyzed by the prototype (engagement, confusion) were helpful? How did the provided feedback affect you during the presentation?

Responses ↓; Codes →	feedback helpful	incorporation of feedback	distraction	neutral	visualization critique
I would have had to ask the participants at the end.				x	
It was interesting to see how the respective information was perceived by the audience. Above all, seeing the mean values of the room helps with the presentation, as you get real-time feedback.	x				
The analysis of the key figures was very helpful because it meant that incomprehensible topics could be discussed again and you could receive immediate feedback as to whether a new explanation would help.	x	x			
The live feedback was a nice touch and made me more conscious about engaging my audience. It was nice to have that info without being distracted by the attendees' videos. I would maybe change either the engagement or confusion bar since one is going from "red" to "green" and the other one from "green" to "red" which was slightly confusing to me. Maybe it would be easier if both of them had the same order :).	x	x			x
I think the metrics were helpful. As stated before, it is a good feeling to see that the audience seems to be engaged. I´m not sure how it would have felt for me if it said that the people were confused. I think I would start talking faster or explain in more detail, so that the level of confusion would decrease.	x	x			
As already described above, I always paid attention to the parameters displayed and orientated myself to them. I paid more attention to the confusion parameter than the engagement parameter. If there was too much confusion, I tried to present my content in a more understandable way.	x	x			
The feedback is very interesting. It helped me understand which slides were better or worse received. So I can improve the potential of my presentation. Unfortunately, I paid a lot of attention to the feedback, so I lost my rhythm a few times because I was surprised by the feedback. I also tried to change my pitch and presentation style based on the feedback in order to get better feedback, but unfortunately this had the opposite effect.	x	x	x		
The progress bar with confusion and engagement points was helpful for me during the presentation. Since the points are next to the slide you can easily see that e.g., one slide had more audience members confused, so I immediately tried to explain the context of the slide better and in detail and it didn't affect me emotionally. Usually when you see that the audience is not motivated, yawning, rolling their eyes, it has a negative effect on the presenter, I for one would take it personally and couldn't control my emotions and the whole presentation would be disappointment for me. Increasing the engagement points during the presentation has made me more relaxed and confident in my presentation performance.	x	x			
	7	6	1	1	1

Post-Study Questionnaire, Presenter Q3

P3) After ending the presentation, you received a summary outlining engagement and confusion levels throughout the presentation. How would you use this information, if at all?

Responses ↓; Codes →	feedback	helpful	adjustment of presentation	visualization critique
I would focus more specifically on the interaction between the participants.				
The information clearly shows which slides/sections of the presentation were perceived as interesting, as well as those that triggered less enthusiasm. This makes it possible to adapt the presentation to the audience in the future or to revise or omit less interesting or confusing content.	x	x		
In a new presentation, I would address the answers and reprocess them or provide deeper insights.	x	x		
I think the information is really useful to see when the participants were paying the least amount of attention or were really confused. If I were to present the same slides again I can see how useful these metrics are to know what I can improve in particular. The only problem for me was the lack of labels in the charts. I think it was a percentage scale from 0-100 but maybe this could have been better explained to me :). Or maybe the chart can always show the full range/"height" from 0-100.	x	x	x	
In my case I saw a noticable drop of engagement at one slide. After seeing this I immediately had to check which slide that effect caused.		x		
After reviewing the summary, I focussed on the peak where the confusion was at its highest. I then looked for the slide that was presented at that point. I thought about why there was so much confusion about this particular slide. Perhaps it was because the participants were already bored or it was actually due to the complexity of the content.	x	x		
This is a subjective behaviour that evaluates little of the content of the presentation. A technically very well prepared presentation can be better or worse received by the participants, depending on the time, place and topicality of the subject. I can therefore use the data to interpret the reaction of sketches and models and possibly make them more vivid. I would use the information to redesign and rethink individual slides with very negative feedback.	x	x		
This information was helpful for me as it relates to the individual slides of the presentation. For example, if the level of the slide was marked with hight confusion level, I would look at my slides and see how I could rewrite them for better understanding for further presentations. Maybe I would list simple examples for better and easier understanding. For the level of engagement, I would use this summary to improve my voice and pronunciation to help the audience listen better. I was also surprised by the length of the entire presentation. I expected to be done in less than 6 minutes, but it took me more which integrated timer has showed. I really like this timer because I usually have my cell phone with a timer next to me. By using this prototype I have two device in one, which is great.	x	x		
	6	7	1	

Post-Study Questionnaire, Presenter Q4

P4) Would you prefer conducting your online-presentations with a system similar to the prototype, or would you prefer a video-based conference tool? Please provide reasoning for your choice.

Responses ↓; Codes →	prototype: improved feedback	video missing	prototype: less distraction	prototype good for large audience
I would always prefer a video-based format, as interaction is also important in online formats and I tend to rely more on the facial expressions of the participants in the picture.		x		
I would use the new tools from this prototype in combination with a video conferencing system. I think it's important to be able to see the video of participants, but since it is difficult to have an overview of the overall mood, this new tool could be a very helpful addition.	x	x	x	
I don't prefer a video-based conference tool because I can better immerse myself in the role of the speaker and therefore not be observed.			x	
I think it might depend on the context. For a smaller, more "intimate" group, e.g. during a workshop, seeing the faces can be nice and would let me engage more with the audience. But for a lot of use cases, e.g. a presentation at university or at work with many participants that I don't know, this could be really useful and I would be happy to use it!		x		x
I think I would prefer a system similar to that prototype because here I can see, how many people really sit in front of their laptops. I am not sure if I would trust the emojis 100% but in my opinion it´s helpful. It gives you a good feeling if it says that the participants are engaged and I think it would also have an impact if it says that the audience is confused. I would try to explain better or maybe leave some not necessary details out. With video-based tools I made the experience that many people just switch off their cameras so you don´t know if there is somebody listening at all. Personally, I also did a lot of different things like cooking, cleaning etc. while listening to presentations in the past. So with this tool it could also be more comfortable for listeners not having to show their faces but be there listening and also give reactions for the person presenting. Showing your face all the time can make you feel uncomfortable.	x			
As I often teach online courses in the field of teacher training, I know that the majority of participants do not switch on the camera during presentations. It takes some time for lecturers to get used to speaking with "black screens". I can well imagine that a combination of video-based conference tools and the prototype would bring many advantages: At the beginning of the lectures, you could meet and greet each other using video, and during the lecture, cameras can be switched off, as the parameters (possibly others) provide information about the mood. During presentations, you don't usually see the participants' cameras anyway. Above all, the display of how many people are actually sitting in front of the camera can be a great treasure - one that could possibly be expanded. However, I would like to see video-based tools for interaction with each other.	x	x		
I would integrate the tool as an additional option for online presentations. It can't completely replace Zoom or Teams, as I find the cameras, chat function and room allocation necessary. However, the prototype with its analyses helps the examiner and presenter to get quick feedback. I miss some features in the analyses, such as a textual summary of the results. What value does the Y axis represent? What does 10-60 mean? I would like to know which users were interested and which were not in order to get feedback on what the participant did not understand. In addition, you could build in questions that the participants ask anonymously. This is another way to get feedback and rethink your slides.	x			
Although I have use it for the first time I would use this prototype-tool, because: 1. It was easy to use and I didn't had any troubles with the set-up 2. This tool is self-explained so you don't need the instructions 3. You are not distracted with the videos of the audience 4. You are not affected emotionally 5. You can use the engagement and confusion levels to immediately improve your presentation	x		x	
	5	4	3	1

Post-Study Questionnaire, Presenter Q5

P5) Please share any other feedback regarding your experience as presenter with the prototype.

Responses ↓; Codes →	zoom	preferable	add feature	positive experience	improved feedback
I find the tool well structured but I definitely prefer Zoom!	x		x		
All in all, a very user-friendly program that can be used to support the evaluation of feedback.			x	x	
This tool is already prepared in such detail that I would immediately use it in everyday life.			x		
I loved the clean and simple interface :). Maybe some more explanations/tooltips at the end could have been nice.		x	x		
To me the most interesting thing was to see the summary in the end and to check at what slides the confusion raised up. It also gave me a good feeling to see the engagement of the audience. All in all the tool worked well for me, although I have to say, that I´m not sure if I would trust the emojis fully, to be honest. Nonetheless, this feedback would definitely have an impact on the way I present - at least subconsciously.			x	x	
I am already familiar with AI-based systems from a university course and find it very exciting that there are further developments in this area in connection with presentation tools. This harbours great potential, especially for teacher training, which will increasingly take place online. I would like to see more parameters that can be detected, e.g. boredom, tiredness, excitement (possibly to create opportunities for interposed questions). The parameter of how many people are recognised could become even more prominent.		x	x		
The prototype is a very cool thing. Unfortunately, the cameras have to stay switched on, which is not so easy due to the frequent internet fluctuations. In my current work environment, colleagues often switch off the camera because otherwise they can't follow the meeting properly. This could be a limitation of the tool. The analysis of the presentation should also be expanded to include some KPIs, such as the names of the individual participants and their feedback, a chat function to anonymously ask questions about certain slides or dragging emoji's onto certain parts of a slide to signal incomprehensible or easily understandable elements of a slide. The prototype works amazingly well and is able to recognise the facial expressions of the participants very well. I think the tool is very useful and user-friendly in everyday school life.		x	x		
I have one point that can be taken as an improvement to the prototype for the future. I was missing some kind of green light and GO when I can start my presentation. As you can't see on the videos if all participants are there. So if every registered user is recognized, a short information that you can start the presentation would be helpful.		x			
	1	4	7	2	

Post-Study Questionnaire, Listener Q2

L1) Do you feel that automatic detection of engagement/confusion as demonstrated in the prototype is favorable compared to online presentations with camera enabled? Please explain your decision from the previous question.

Responses ↓; Codes →	video missed ~~prototype~~	provides targeted ~~feedback~~	comfort not showing face	doubts abt. accuracy
I cannot give a right answer because I didn´t talk to them after the presentation.	x			
On one hand I like to see the person that gives the presentation. On the other hand I sometime like to watch the other listeners (for example if I´m bored during the presentation).	x			
It is more pleasant to transmit feedback in a targeted manner than to be visible in front of an audience.		x		
The detection is very helpful because you can react immediately and get feedback quickly.		x		
I like the fact that I don't have to commit to using a webcam and having to show my face, e.g. for presentations with large groups and many people that I don't know.			x	
Would be very useful to use in homeschooling...				
It is a great Alternative to the Camera-Enabled Presentation, if you don't have good equipment (Camera, etc.). If it is a long Presentation it could be a little monotonous to not see a face to the voice of the presentation.		x		
I find it more pleasant to see the faces of the other participants, especially the presenter. The automatic recognition of engagement and confusion could be very helpful, especially because you are not so focused on the faces of the others during the presentation, but I am not entirely convinced that my engagement or confusion was always recognized correctly.	x			x
I am more engaged in listening to the presenter when I know that he / she receives this kind of feedback. I would like to show respect to the presenter and would try to concentrate more on the presentation.		x		
As already answered in the previous questions, the automatic recognition of motivation and confusion can be very helpful in courses. It is easy to recognise when more detailed explanations are needed and actions can be taken to counteract a continuous drop in motivation. This provides an overall view of motivation and confusion in the entire group (assuming these are detected correctly).		x		
Yes, very advantageous, because you can respond spontaneously to your participants during the presentation. Here you can easily see when you should take a break as a lecturer or whether you should go through the topic again in an extra session for certain reasons, because it's just too much of a challenge.		x		
You have more time to concentrate on the voice of the presenter and not on the facial expression, which for me was good.			x	
It was very difficult to follow the presentation and take a look/correct the autum. detection in unknown fields - it was easier during presentations of my own special knowledge. But however, as a listener it is difficult to do both at the same time (listening & take a look if the automatic feedback is similar to your own perception).				x
	3	6	2	2

Post-Study Questionnaire, Listener Q3

L3) Did you observe the output of automatic engagement detection? Did you override by providing manual feedback? If so, why?

Responses ↓; Codes →	no observation	observation	minor observation	overriding	felt accurate	propse metrics change
No	x					
I sometimes clicked on the button to see the engagement detection, but it closed again automatically. I did not override, because I couldn´t see the detection and the override button at the same time (I had to scroll down to see it).	x					
On the whole, the automatic recognition was correct, but in some situations I corrected it. Especially regarding confusion, I noticed that it was often assumed incorrectly that I would understand it completely.		x	x			
No, because the tool interpreted my answers correctly.		x		x		
At some points. I didn't always agree when I was not paying attention according to the algorithm :D		x	x			
Nein	x					
Yes I have observed the output of automatic engagement detection.		x				
Yes, I often looked at how my engagement was automatically recognized and rated, and sometimes I overwrote it because it wasn't quite right. Sometimes, for example, it was rated as not so good over a longer period of time (yellow emoji) although I was smiling - in such a case I sometimes changed it manually. What I also noticed is that mostly a rather happy face (light green emoji) was rated, but I think rather rarely the even happier face (medium/dark green emoji), even though I was relly laughing sometimes. The assessment of whether I had understood the content was usually with the two green check marks, which was mostly correct, but two times I looked (extra exaggerated) confused, and I don't think the yellow or red emoji was rated.		x	x			
I overrode several times because I think it was not always accurate. I also don´t really get the point of the two different feedbacks - overall and understanding. If the overall feedback is positive the feedback concerning the understanding of the contents has to be positive as well, in my opinion. Because if I don´t understand a slide or what´s being explained/said at the moment, it can´t be positive overall. It would work the other way round, though. I can understand something but also not be engaged because the content is boring to me e.g. Sometimes one of those feedback sections was accurate and the second one I had to override.		x	x			x
Yes, I have overwritten the automatic detection in presentations that were rather boring because I wasn't interested in the topic at all. In this presentation it indicated that I was very satisfied with this part of the presentation.		x	x	x		
Yes, I used the override function because I wanted to try it out. I intentionally changed my facial expressions or didn't look at the monitor because I wanted to test the prototype. I then made a correction.		x	x			
I watched the automatic engagement detection and didn't override it because it was correct and I had no need to override it.		x		x		
I tried to observe the output of automatic engagement detection during the first presentation, but I was not familiar with the topic of the presentation. So I often felt confused and not comfortable with the situation. During a presentation with a topic I am familiar with, the automatic engagement detection did not work, so I could concentrate myself better on the content.		x				
	3	10	6	3	1	

Post-Study Questionnaire, Listener Q5

L5) How comfortable do you feel being subjected to automatic detection of engagement and confusion as demonstrated in the prototype? Please elaborate based on your previous answer.

Responses ↓; Codes →	comfortable	value	anonymity	doubts abt. accuracy	uncomfortable
everything fine	x				
I don´t mind because I´m anonymous, so the presenter doesn´t know who is confused for example.	x	x			
I felt very comfortable not being visible during the presentation.	x	x			
I felt very comfortable with the automatic recognition because I didn't have to concentrate on each individual reaction like I did in online conferences and this gave me a good overview of understanding.	x				
Since it was promised to be locally processed it felt safer to me. I would however be critical of its accuracy and might not always feel that it is representing my accurate reactions, but for that I have to manual override anyway :)	x		x		
I felt a little bit watched - but I guess that's how it should be					x
I felt very comfortable because it gives great feedback to the presenter, on how interesting his presentation is and how many are still engaged or confused on the content or how it is presented.	x				
It doesn't bother me, but it's also a bit of a strange feeling. For example, I sometimes concentrated more on my facial expression and the scoring than on the presentation, but that could also be because it was still unfamiliar. So in summary, I see it as neutral - it doesn't bother me, but it still takes a bit of getting used to.	x				x
I feel comfortable being detected. I think it´s interesting how this tool works. I tried to smile or look sad during the presentation just to see if anything changes. :-) This tool would force me to sit still in front of my laptop while listening to a presentation and not doing anything else. So I think I would automatically be more engaged. I think it would also be more comfortable being detected like this than showing myself to the camera for a full presentation.	x	x			
I don't find it strange when these moods are automatically raised in me. I always switch my camera on during presentations. If the automatically recognised images are stored in compliance with the DSGVO, I even find this more anonymous.	x	x			
As long as the data is only processed locally during the presentation and is not stored on a server in China, I can live with that very well. Otherwise it would be unpleasant to be constantly filmed.					
Although I am a big fan of innovation, it was strange for me at first (it felt strange) to see how accurate the tool was and I checked every slide to see if the tool would fail a detection or not. So I didn't have full confidence as it was my first time dealing with such a tool. But after hearing a few presentation, I was feeling more and more trusting.			x		
In general I feel not uncomfortable being "watched" during a presentation as a listener and on one hand it felt quiet interesting how the system automatic detected engagement and confusion. But I felt unable to cope with checking whether the automatic detection is similar to my own perception and listening to the topics at the same time. I would prefer some seconds between the slides after the presenter explained the topic to look at the automatic detection results and correct them if necessary. From my point of view this would increase the informative value of the feedback significantly.			x		
	9	4	3	2	

Post-Study Questionnaire, Listener Q8

L8) Please share any other feedback regarding your experience as listener with the prototype.

Responses ↓; Codes →	other	more comfort	missed video of presenter	feature proposal
I found it very pleasant, but I think Zoom with video is better!	x			
I think it´s a useful tool for presenters because then they automatically have an interaction with their audience. I also think it can be quite useful for presentations with a bigger audience. From my experience it´s more uncomfortable and therefor unlikelier to ask questions in big lectures if you´re confused. But if the presenter automatically sees that a certain percentage of the audience is confused, he/she can try to clarify.		x		
I think it can be an important tool for the presenter to get real-time feedback. For me as a listener it is more comfortable not to have the camera on, even though I also cannot see who else is in the auditorium, but I miss seeing the presenter a bit. I think it can be more valuable to see the presenter as their facial expressions and emotions add to the content of the presentation.		x	x	
As a listener, I often missed the facial expressions and gestures of the speakers, but I was still able to follow the content of the presentations well.			x	x
I think some more explanation for the manual overriding of the feedback would have been nice since I would not have known/discovered that without you explaining it to us :)				
Great idea, but should the summary of the detection not also be visible to the listener?				x
I think it is a great tool. It could bring a great response to the presenter on how interesting his presentation is viewed by the crowd who is listening.	x			
I think it's a good idea and can help presenters in particular, because I think it gives them more feedback for their presentation. I think the feedback given verbally would also differ to some extent from the facial expressions. So it would be better - if facial expressions are well recognized - because this way very honest, genuine feedback can be obtained per slide. So it's a great idea, with a little room for improvement, but I see a lot of opportunities in it. In my opinion, it would be even better if you could at least see the presenter, i.e. not the other participants, but simply the picture of the presenter, so that it is easier to listen. Without any image, I find it easier to wander off (concentration/attention drops a little for me).			x	x
As described, I tried different facial expressions to see if anything changes and it really did. When I smiled I seemed to be more engaged. I had to override the information several times. All in all I think the idea of this prototype is very interesting but people should also override the information if it´s not accurate. Otherwise the presenter would get wrong information which could impact his presentation style e.g.	x			
At the beginning of a presentation, I left my seat briefly and so the system didn't recognise me. It was very exciting that the presenter immediately realised that one person - me - had been missing for a few minutes.	x			
Unfortunately, as a listener you have few options for feedback other than looking into the camera and perhaps nodding. A few more features like raising hands for questions or marking slides for deeper discussions would be helpful. This means you can give feedback without disturbing the presenter				x
Maybe for a future design, it would be nice to have somehow an interaction with the presenter, for sharing a feedback. e.g., chat, a microphone or something like that if you have a question about the presentation content.				x
During the presentation I was sitting alone in an office - nevertheless there were distractions from outside the office that influenced my visual reaction. I guess it happens quiet a lot during online lessons - so the question is, how does this interactions influence the results of the automated detection. Same thing should be questioned for taking notes, to give one's nose a blow, to drink a glas of water,... I felt not comfortable with: listening to the presentation, looking at the automated generated result of the questions during the presentation an thinking whether they are correct the way I see it - all at the same time. I felt uncomfortable as a listener - I guess I would feel uncomfortable as a presenter as well if I could see the results of the reactions of my listeners during my presentation. On the other hand it's not completely clear if you could trust the results because you don't know the circumstances and distractions of every listener. Furthermore I guess it's nearly impossible to react during the presentation (as presenter) on the results of the feedback. I would prefer giving and receiving feedback at the end of the presentation. I am not convinced of the tool because it takes ones eye off the ball of the main topic - the presentation. I could imagine using the tool for simple presentations in expert groups in special situations.	x			x
	5	2	3	6

Interviews - Coded Results

Post-Study Interview 1

	Interview 1)	
	How did you feel holding your presentation with the prototype?	
1	I found the additional feedback very interesting and I did have the feeling that the feedback is helpful.	feedback helpful
2	However, it could also be that it increases nervousness or feelings of doubt when in a stress situation like haveing a presentation. But in our case - we had a relaxed presentation environment - I did not have the feeling that it makes me more nervous.	might increase nervousness
3	It was a nice support having an overview about the feelings of the audience while presenting.	better overview of audience
4	I think when everything [the engagement and confusion indicator] are visually red, I feel that could increase pressure for the presenter.	
5	Maybe you could think of making the visualization of the audience state more positive to reduce that. Use less red color or make it more encouraging. But I think it very much depends on the type of the presenter what they would prefer.	more positive framing
6	At the same time [with the explicit visualization] it could be great to really stimulate making a concious effort to i.e. speak slowlier or that you reduce speed if it appears that people are currently confused or you could ask the audience if possible. So it could definitely be helpful to improve yourself, for sure.	
7	One could think of making i.e. a notification "maybe you could talk a little slower" or "take a breath and try to speak slower" but it could also affect concentration negatively. One would need to look how to make it more encouraging.	encourage speaker
8	One thing that kind of confused me was ... that one indicator was from red to green and the other from green to red. Maybe one could just reverse it to make it similar.	more consistent labels
9	[How did you interpret the results of the autmatic detection? Was there great variation?] I didn't notice great variation of the detection results.	
10	confusing if there would be many more parameters. Also the description was understandable to me. As said, maybe you could make both positive, i.e. engagement and understanding or something like that. If you want to avoid a negative adjective ... to avoid discomforting the presenter	feedback helpful
11	suggestions "this is what you could improve" ... but the tool currently is not really negative	encourage speaker
12	In general maybe it would be good to include more tooltips or explanation about how the tool works. I.e. explanation texts for some features.	provide more explanation
13	without indicating if it in percent or some other metrics. I.e. in my case the graph was from 0 to 80 and that may be confusing, that needs to be labelled and explained better in my opinion.	improve summary graph
14	Also for non-technical users there are some aspects of the technology that could be hard to understand regarding general understanding of how this works... but that's a common problem. I also sometimes for things that I developed that they are clear to users because I understand them that well, but often it is not.	provide more explanation (technology)
15	the top automatically opend and closed several times if your face is not detected currently.... I think that could be confusing or distracting. ... Maybe I'd suggest to always have it closed. Also, without prior explanantion I would not have guessed that I can manually override this.	indicator that user was not detected confusing

Post-Study Interview 2

	Interview 2)	
	How did you feel holding your presentation with the prototype?	
1	I found the tool very well made, very user-friendly... it was clear to me how to use it, i.e. how to start the presentation. It was well structured.	easy to use
2	I found it nice that you can see after the presentation - or also during the presentation - if the participants were confused or engaged. You can even during the presentation react to the audience or ask questions if there appear to be ambiguities. That's something that other [online presentation] tools don't offer because you cannot focus on all i.e. faces during a presentation. That I found great...	feedback helpful, real time incorporation of feedback
3	The feedback in the end I found great as well...	summary Useful
4	It was a bit difficult because the presentation audience was not versed in the topic I held the presentation, so I couldn't adapt my presentation clearly... but I think if there's a speicific type of audience listening to the presentation the information would be very useful to me.	user-test scenario not realistic
5	**How useful was the information on-the-spot while holding the presentation compared to after the presentation with the summary?**	
6	The summary was a bit more useful to me... but the instant, real-time feedback was useful as well. I just would need to have a bit more experience in order to really instantly use this information while holding the presentation.	summary useful; more experience w tool required
7	You could try experimenting with even more visual cues that highlight specific scenarios, like "the thing you said right now had a reallly great impact on the audience"... in a way that you don't have to specifically "read" the bars regarding engagement and cconfusion. But one would have to see how well	highlight extreme cases positively
8	**How did you feel receiving this additional feedback? Did you feel more nervous or encouraged?**	
9	It didn't make me feel more nervous at all, I even felt more encouraged. I felt it was honest... or at least I hope. It didn't make me more nervous for sure... also it didn't stress me in any way.	comfortable presentation experience
10	I could also see other metcis that would be helpful to capture automatically, i.e. agreement. And in general - maybe it's a question of type - but I'm kind of a numerical type. I'd like to see actual numbers of the detection. Maybe this could be configurable to suit different types of people. If you ask 5 questions about their preference with regards to this you'll probably get 5 different answers.	use positive metrics
11	**How did you feel listening to the presentation compared to a listening in a Zoom session?**	
12	Compared to listening to an online presentation in a zoom session, I felt that I could concentrate more on the presenation. On some occassions, however, I have to say I missed the mimics and gestics of the lecturer at least. I cannot say how exactly... but I kind of missed that. But overall, I could really well	lecturer video missing
13	**How did you feel about automatic detection compared to using a webcam?**	
14	I sometimes feel uncomfortable using a webcam because I feel watched by the other participants. In the case of automatic detection I did not feel that way, even though I also was kind of watched. Probably because I did not see myself.	autom. detection less intrusive than video
15	**How eager are you to use such a system in the future?**	
16	I would definitely use this. And I think it would be suitable to different settings, i.e. if I have a lecture about CPR in a small group I think the feedback would be useful. But also for large-scale lectures I think it would be useful.	suitable for many settings

Post-Study Interview 3

	Interview 3)	
	How did you feel holding your presentation with the prototype?	
1	It was an interesting experience using the prototype. I have to admit that I didn't closely watch the results of the feedback while holding the presentation because I had to concentrate on my presentation.	more experience required to use on-the-spot
2	However, I do think that the summary of the results after the presentation can be helpful for revising your presentation. One thing that could be re-considered is dividing this feedback or summary into slides. I am not sure if this is so useful.... if you're holding a presentation series I'd rather be able to compare the results between the different presentations than comparing individual slides to one another.	summary useful; division metric (slide) questionable
3	However, one thing that was interesting to see... I had some slightly shocking images on some of my slides and I saw that these were visible in the presentation summary. That was interesting.	automatic detection felt accurate
4	To be able to interpret these results I think will require some experience with the tool. I'm not sure how exactly I would use the information now. But I guess this would become easier once you get used to the system, after using it for some time.	more experience required
5	**How did you feel using the prototype as a listener?**	
6	As a listener what kind of confused me was the window in the top that popped open from time to time. I think it was when face was not detected, but I didn't really understand that in the beginning. This was distracting and probably it's not necessary to always show that to the user.	indicator that user was not detected confusing
7	Sometimes I also tried to play around with my mimics to try to influence the result of the automatic detection, but I was not really able to influence it very much.	aims to influence autom. detection
8	One presentation, I really didn't understand much and I didn't really find it interesting to be honest. But I tried really hard to look engaged to be polite to the presenter.	
9	Overall, I think the prototype is an interesting showcase of what's possible but I'm not sure if or how I would like to use it in my professional environment	

Post-Study Interview 4

	Interview 4)	
	How did you feel holding your presentation with the prototype?	
1	I'm personally a bit hesitant to use this system because I generally refuse to do presentations or online conferences without webcam enabled.	
2	I have several courses online each week and I'm adamant that everyone enables there cameras otherwise they cannot join.	misses video
3	**How many participants do you usally have in theses sessions?**	
4	I usually have between 20 and 40 participants and I tell them before that it is a requirement to have the webcam enabled, otherwise they cannot join.	adamant about webcam
5	I think the sessions are just much more interactive when cameras are enabled. For me as a lecturer it is important to actually see the faces of the participants. I have a screen for the presentation and another screen for the webcam videos and when looking at the participants' faces I imemdiatley recognize if there is something wrong.	
6	One thing that I find problematic using the prototype as a listener is not being able to move away from the webcam. Usually when I listen to presentations I also sometimes move away from the webcam to briefly do other stuff. And I wonder how that affects the results of i.e. the presentation summary.	reduced mobility problematic
7	For me I would not do online presentations or conferences without camera. I just wouldn't do it....	adamant about webcam
8	I'm very adamaent and confrontative about this because if some participants disable their camera usuallly that leads to a chain reaction.	
9	Once I noticed that some participants had their camera disabled and then I made a screenshot of the black, empty boxes I see. So I sent the screenshot to the participants to show them how it feels to me and that it's not fun for me seeing nobody. After some time people enabled their webcams again.	
10	For me seeing someones face is also something that leads to trust. I don't know if I would trust the results of this automatic detection... I usually trust people if I see their faces.	lack-of-trust if no face visible
11	I do, however, realize that there are some limitations.. or that there is a threshold of participants where you are able to visually process all the facial information. If it's a presentation of a CEO of a large corporation i.e. with several hundreds of listeners then something like the prototype could be useful for this situation.	suitable for large-scale settings
12	I just don't think that for groups of 20-30 people it would be very helpful.	not suitable for small group
14	In general it would have been intersting for met to see both the automatic detection results combined with the webcam video.	
15	I think if this system is useful for people very much depends on the type of person you are. People have very different requirements and approaches about how they hold courses or presentations...	suitability based on personality type
17	**How did you feel using the prototype as a listener?**	
18	It's a great way to indicate presence and from the perspective of a lecturer it's also a way to check presence and see if participants are really here.	good to check attendance

Post-Study Interview 5

	Interview 5)	
	How did you feel holding your presentation with the prototype?	
1	In general I was very open about holding the presentation... i'm quite experienced with holding online presentations... I found it very intersting to have this overview about the emotions of the listeners and I had a look at the results every now and then but I admit not constantly. Otherwise I probably would not have been able to concentrate on my presentation.	feedback interesting; slightly distracting
2	But it did definitely influence me... i.e. when the results of, what were the metrics again, engagement and confusion were not that favorable I tried to adjust my presentation by i.e. do it faster or leave out details or explain it even simpler. So I tried to use the results but only to some extent	real time incorporatio n of feedback
3	It did not feel that different than presentations in conventional tools to me...	
4	**Did it seeing the results affect you? Did it make you feel nervous or did it encourage you?**	
5	Hmm... I believe it didn't affect me in that way but it did affect the way I present and if I kept it shorter or more detailed. But it did not affect whether I feel nervous or not.	did not affect nervousness
6	**Do you believe that you'd need more experience with the tool in order to comprehend the results while holding a presentation?**	
7	Yes definitely, I think it has to do with how much experience you have in general in holding online presentations or how well you know your presentation. This affects how much capacity you have to comprehend the signals of this tool while presenting. E.g. if you are inexperienced with holding your presentation you can only concentrate on holding your presentation. If I'd present a topic that I barely know I don't think I could incorporate the results of the automatic detection.	experience required to incorporate feedback
8	**Do you think it would be useful if the tool would summarize the detected information even more for the presenter?**	
9	Yes I think that would be useful... it would be great to have this feedback for the presenter when something goes really wrong, i.e. you are caught up in this one explanation, move on.. people aren't listening any more.	
10	This could also be useful the other way around to give the presenter confidence if some parts of the presentation were preceived very well by the audience.	
11	**How did you feel listening to the presentation compared to a listening in a Zoom session?**	
12	I had to correct the automatic detection a couple of times. But I didn't perceive that to be negative because I had the feeling that I was focusing more on the presentation because I wanted to give proper feedback. Sometimes I got a bit distracted because some of the topics weren't interesting to me but I thought I want to give correct feedback to the presenter so I sticked to listening to the presentation and gave manual feedback.	manual feedback helped staying focused
13	I think it helped me concentrating on the presentation because the tool was demanding participation from me in a way. This I liked... because I like giving presenters feedback, aso positive feedback when the presentation is really nice.	manual feedback helped staying focused
14	I also found the summary after the presentation very interesting because in my presentation I first had a theoretical part introducing the topic and in the second part I presented my results with Excel tables. And on the slide where I had the first Excel table there was a noticable increase of the confusion levels... that was a great insight. ... and retrospectively I thought that - especially when presenting the topic to non-experts in the field - there's a need to give more explanation of how these results are to be read.	summary useful

Post-Study Interview 6

	Interview 6)	
	How did you feel holding your presentation with the prototype?	
1	Holding the presentation with the prototype was quite different compared to my experience with convential tools...	
2	On the one hand I felt a bit more relaxed because it was not video based and I didn't have the feeling that people are watching me... instead I jusft felt that a software is watching me.	relaxed feeling
3	And seeing these two metrics - I believe it was confusion and motivation... or engagement - I really tried to react immediately when I saw that the results changed, e.g. the confusion level has risen. I tried to speak more clearly, slower or tried to elaborate more. I think in general the confusion parameter was more relevant to me than the engagement, at least it was more clear to how I can react to that.	feedback helpful, real time incorporatio n of feedback
4	**How did you manage to simultaneously hold your presentation and watch the results of the automatic detection?**	
5	That was very easy... it was no problem for me. It did not distract me or anything. On the contrary, I felt that it helped me.	
6	I found the prototype very interesting and I would quite like to somehow use this in my teaching. To me it would be very interesting to see which other parameters one could measure. I think I made some suggestions in the questionnaire, I can't remember what parameters I suggested.	additional detection parameters
7	**How was your experience using the tool as a listener?**	
8	Using the tool as a listener was really interesting because I briefly left to do something in the kitchen, and it immediatley was apparent to the presenter that somebody was missing.	
9	I think that can be very useful especially if you are presenting to a large crowd and you want to know if people are really present. I would find this much better compared to asking people to have their cameras enabeld at all times.	good to check attendance
10	As a lecturer, you can see if people are really present... but participants can remain kind of anonymous and they don't have to reveal their appearance. As a listener, I believe, this would be more comfortable for me.	

User Study - Consent Form

TECHNISCHE UNIVERSITÄT WIEN

User Study & Interview Procedure Overview

Thank you for your interest in taking part in this user study to evaluate a prototype exploring opportunities of state-of-the-art image processing and affective computing techniques in the context of conducting online presentations. Participating in this user study will involve attending a simulated online presentation and, on a voluntary basis, conducting a short presentation during this simulation. During the online presentation session, the prototype will access participants' webcams and try to infer information about their engagement and confusion levels as an additional feedback stream for the presenter. The captured data will be deleted after the user test.

All participants will be asked to fill in questionnaires before and after the online presentation session. Participants willing to hold a presentation will additionally be asked to take part in a short retrospective interview. The data gathered by these research methods will form the basis for subsequent evaluation. All data will be anonymized for the purpose of data evaluation and analysis, ensuring that no conclusions about your identity can be drawn.

Consent to Take Part in User Study & Interview

- I ..... voluntarily agree to participate in this research study.
- I understand that even if I agree to participate now, I can withdraw at any time or refuse to answer any question without any consequences of any kind.
- I have had the purpose and nature of the study explained to me in writing and I have had the opportunity to ask questions about the study.
- I agree to my interview being audio-recorded.
- I understand that all information I provide for this study will be treated confidentially.
- I understand that in any report on the results of this research my identity will remain anonymous. This will be done by changing my name and disguising any details of my interview which may reveal my identity or the identity of people I speak about.
- I understand that disguised extracts from my interview may be quoted in the thesis: "Automatic Detection and Emoji-Based Communication of Non-Verbal Cues during Online Presentations" or related research papers.

- I understand that signed consent forms and original audio recordings will be stored until the thesis is successfully finished.
- I understand that a transcript of my interview in which all identifying information has been removed will be retained as part of the thesis appendix.
- I understand that I am free to contact any of the people involved in the research to seek further clarification and information.

_____

Signature, Date

CONTACT DETAILS
Main Contact: Peter Oberhauser
[redacted]
+4366[redacted]

Thesis Advisor: David Kostolani
[redacted]@tuwien.ac.at

Consent form partially derived from: https://www.tcd.ie/swsp/assets/pdf/Participant%20consent%20form%20template.pdf