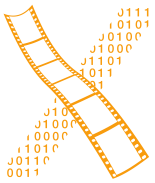

Chair for Computer Science 10 (Media Computing and Human-Computer Interaction)

RWTH AACHEN UNIVERSITY

Are You Still Listening? Measuring Attention in Online Meetings

Master's Thesis
submitted to the
Media Computing Group
Prof. Dr. Jan Borchers
Computer Science Department
RWTH Aachen University

by
Rene Niewianda

Thesis advisor:
Prof. Dr. Jan Borchers

Second examiner:
Prof. Dr. Ulrik Schroeder

Registration date: 07.12.2021
Submission date: 21.01.2022

Zentrales Prüfungsamt/Central Examination Office

Eidesstattliche Versicherung
Statutory Declaration in Lieu of an Oath

Rene Niewianda — Name, Vorname/Last Name, First Name

346905 — Matrikelnummer (freiwillige Angabe) / Matriculation No. (optional)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit/Bachelorarbeit/ Masterarbeit* mit dem Titel
I hereby declare in lieu of an oath that I have completed the present paper/Bachelor thesis/Master thesis* entitled

Are You Still Listening? Measuring Attention in Online Meetings

selbstständig und ohne unzulässige fremde Hilfe (insbes. akademisches Ghostwriting) erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.
independently and without illegitimate assistance from third parties (such as academic ghostwriters). I have used no other than the specified sources and aids. In case that the thesis is additionally submitted in an electronic format, I declare that the written and electronic versions are fully identical. The thesis has not been submitted to any examination body in this, or similar, form.

Aachen, den 21.01.2022 — Ort, Datum/City, Date

Unterschrift/Signature

*Nichtzutreffendes bitte streichen

*Please delete as appropriate

Belehrung:
Official Notification:

§ 156 StGB: Falsche Versicherung an Eides Statt
Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.
Para. 156 StGB (German Criminal Code): False Statutory Declarations
Whoever before a public authority competent to administer statutory declarations falsely makes such a declaration or falsely testifies while referring to such a declaration shall be liable to imprisonment not exceeding three years or a fine.
§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt
(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.
(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.
Para. 161 StGB (German Criminal Code): False Statutory Declarations Due to Negligence
(1) If a person commits one of the offences listed in sections 154 through 156 negligently the penalty shall be imprisonment not exceeding one year or a fine.
(2) The offender shall be exempt from liability if he or she corrects their false testimony in time. The provisions of section 158 (2) and (3) shall apply accordingly.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:
I have read and understood the above official notification:

Aachen, den 21.01.2022 — Ort, Datum/City, Date

Unterschrift/Signature

Contents

List of Figures

List of Tables

Abstract

Seeing and evaluating the audience's reactions is vital for giving good presentations. During in-person lectures or meetings, this is relatively easy. However, this task can be difficult in online sessions since most video conference applications only show very few if any audience members while in presentation mode. This limits who and what the lecturer can see. Therefore, most presenters miss relevant audience feedback like non-verbal social cues. This lack of feedback makes it harder for presenters to "read" the room or to respond to the audience, and can therefore increase the presenters' anxiety level which can lead to a worse experience for the lecturer and the audience.

We developed a feedback application to accommodate these problems by analysing the emotional state of the individual audience members with a Convolutional Neural Network and displaying the attention level of the whole audience to the lecturer. To show the attention value to the presenter, we designed three different interfaces, each focusing on different aspects: A simple interface that uses a single emoji with three states to show the general attention level, an extension of this interface, which also indicates the prevalent emotions of the audience, and a line graph that shows the attention level of the last 2.5 minutes.

We tested the interfaces in a laboratory user study with predefined slides and attention curves.This simulated audience keeps all participants' study conditions similar. Most participants stated that the interface which showed the prevalent emotions was hard to interpret and therefore not as helpful as the other two. The simple interface and the graph were considered easy to read and interpret. Still, most presenters did not know how to react to low attention levels during their presentations because they were too busy holding the talk. They also stated that interfaces like the graph would be more helpful to evaluate lectures in retrospect. We also conducted a field test in an actual lecture whose results agreed with those of the user study.

Überblick

Um gute Präsentationen zu halten, ist es wichtig, die Reaktionen des Publikums zu erkennen und sie zu evaluieren. Bei normalen Vorträgen oder Meetings ist dies vergleichsweise einfach. Bei Online-Sitzungen ist die Aufgabe dadurch erschwert, dass die meisten Videokonferenzanwendungen im Präsentationsmodus, wenn überhaupt, nur sehr wenige Teilnehmer zeigen. Dies schränkt ein, wen und was der Vortragende sehen kann. Daher entgehen den meisten Vortragenden relevante Rückmeldungen aus dem Publikum, wie z.B. nonverbale Signale. Dieser Mangel an Feedback erschwert es den Vortragenden, den Raum zu "lesen" und auf das Publikum zu reagieren. Dies kann zu einer unangenehmeren Vortragsatmosphäre für den Vortragenden und die Zuhörer führen.

Wir haben eine Feedback-Anwendung entwickelt, die diesen Problemen entgegenwirkt, indem sie den emotionalen Zustand der einzelnen Zuhörer mit einem Convolutional Neural Network analysiert und dem Vortragenden die Aufmerksamkeit des gesamten Publikums anzeigt. Um diesen Aufmerksamkeitswert darzustellen, haben wir drei verschiedene Benutzeroberflächen entwickelt, die sich jeweils auf unterschiedliche Aspekte konzentrieren: Eine einfache Anzeige, die ein einzelnes Emoji mit drei Zuständen verwendet, um den allgemeinen Aufmerksamkeitsgrad anzuzeigen. Eine Erweiterung der ersten Schnittstelle, die auch die vorherrschenden Emotionen des Publikums anzeigt, und ein Liniendiagramm, das den Aufmerksamkeitsgrad der letzten 2,5 Minuten anzeigt.

Wir haben diese Benutzeroberflächen in einer Laborstudie mit vordefinierten Folien und Aufmerksamkeitskurven getestet. Dieses simulierte Publikum stellt sicher, dass die Studienbedingungen für alle Teilnehmer möglichst gleich sind. Die meisten Teilnehmer gaben an, dass die Anzeige, welche die vorherrschenden Emotionen anzeigte, schwer zu interpretieren sei und daher nicht so hilfreich war wie die beiden anderen. Die einfache Anzeige und das Diagramm waren laut den Teilnehmern leicht zu lesen und zu interpretieren. Dennoch wussten die meisten Vortragenden nicht, wie sie während ihres Vortrags auf niedrige Aufmerksamkeitswerte reagieren sollen, da sie zu sehr mit dem Halten des Vortrags beschäftigt waren. Einige Teilnehmer merkten an, dass Benutzeroberflächen wie das Diagramm hilfreicher für die Nachbereitung von Vorträgen wären. Die Ergebnisse unseres Praxistests in einer realen Vorlesung stimmten mit den Ergebnissen der Nutzerstudie überein.

Acknowledgements

First of all, I would like to thank Prof. Dr. Borchers and Prof. Dr. Schroeder for examining this thesis.

Secondly, I want to thank my family and friends who supported me during this thesis. Especially, I would like to thank Franca auf der Heiden, Tobias Möhring, René Schäfer and Nicole Schops for their support and advice.

Lastly, I want to thank my supervisor Sebastian Hueber for his help and feedback.

Conventions

Throughout this thesis we use the following conventions.

Text conventions

Definitions of technical terms or short excursus are set off in coloured boxes.

EXCURSUS:
Excursus are detailed discussions of a particular point in a book, usually in an appendix, or digressions in a written text.

Definition:
Excursus

The whole thesis is written in British English.

For reasons of politeness, unidentified third persons are described in male form.

Chapter 1

Introduction

Giving lectures and presentations is an essential part of the work of teachers, researchers, and many other professions [Ataei et al., 2020, Murali et al., 2021, Rivera-Pelayo et al., 2013]. For the presenter, one major part of those activities is the evaluation or ”reading” of the room’s atmosphere. This means that the lecturer can interpret gestures and facial expressions to gauge the attention and engagement of his audience [Chen, 2003, Sun et al., 2019].

”Reading” the room is important during talks

Due to the recent global health pandemic (COVID-19), more and more these lectures and talks are given online instead of in-person [Bennett et al., 2021, Chen et al., 2021]. This transition to remote teaching makes the task of ”reading” the room way more complicated. Only a few students share their video feed, and for most lectures, the students’ audio feeds are disabled except when they want to ask or remark something. Since they neither hear nor see their audience, some lecturers state that online lectures feel like they talk into a void [Yarmand et al., 2021]. The missing (positive) feedback and audience responsiveness can also increase the stress and anxiety levels of the speaker [Bassett et al., 1973, MacIntyre et al., 1997].

Almost no audience feedback during online lectures

Presenters miss non-verbal social cues

But even if large parts of the audience share their video, most video conference applications like *Zoom*[1] and *Microsoft Teams*[2] only show very few audience members while in presentation mode. In addition, these video streams are small, so that it is hard to recognise facial expressions or gestures. Murali et al. [2021] found in their survey that 83.11% of the interviewed lecturers miss this audience feedback in online meetings.

Presenters cannot infer the audience's attention level

The lack of feedback creates a communication gap between the lecturer and the audience. The presenter cannot see if his listeners are confused, bored or interested. Thus, it is challenging to notice if the audience pays attention to the lecture and to react accordingly. This circumstance makes it harder to identify which sections were understood and which were not.

We want to design an application to help presenters

We decided to accommodate these problems by designing an application that supports a presenter during online talks and lectures by displaying the audience's attention level. However, holding a lecture is already a taxing task by itself. Therefore we want to focus on what information is most beneficial to the lecturer and how we present it effectively without overwhelming or distracting him.

1.1 Research Questions

We divided the App development process into several steps

To develop such an application, we first need to determine what features can be extracted from the audience's video feeds. In the next step, we must derive meaningful information from these features. Then we need to decide what information we want to display to the presenter and how we want to show it to him. Therefore, we need to design multiple interfaces that differ in what information they provide and how they display it. Lastly, we must test these interfaces to see how useful the information is for the presenter and how it impacts him.

[1] `https://zoom.us/` Accessed: 18.12.21

[2] `https://www.microsoft.com/teams` Accessed: 18.12.21

Hence, we want to answer these research questions in this thesis:

RQ1 Which features from a video stream can we extract and use to track attentiveness?

RQ2 How can we effectively relay this information to the lecturer?

RQ3 How is the lecturer impacted by this information?

1.2 Thesis Outline

In the next chapter of this thesis, we will begin by reviewing related work. We start by showing and classifying different existing feedback applications. This classification includes explicit applications, where the audience needs to give feedback actively, and implicit applications, which gather feedback independently. As an example for such software, one application is dissected in detail: The *AffectiveSpotlight* app designed by Murali et al. [2021]. In the third chapter, we will describe how we used this knowledge to develop our own implicit feedback application. We designed three different interfaces during the development of this application, which we then tested with a user study in Chapter 4. Since we conducted this user study in a lab setting without a real audience, we also tested a refined interface in a small field study. In the last chapter of this thesis, we will summarise our conclusions and suggest a course for future work in this field.

Chapter 2

Related Work

Getting feedback from the audience and interpreting it to assess the audience's attentiveness is an integral part of public speaking [Murali et al., 2021]. In the following chapter, we present various methods that allow a speaker or presenter to gather and display different types of audience feedback, for example the *EngageMeter* [Hassib et al., 2017] and the Live Interest Meter App [Rivera-Pelayo et al., 2013]. We start by categorising the feedback methods and showing some example applications for explicit and implicit feedback collection. Then we are going to dissect one implicit feedback software in detail, namely *AffectiveSpotlight* developed by Murali et al. [2021]. We choose this application since the authors precisely describe their engagement measuring process. This computer program uses, among other metrics, the audience's emotional state to find the most active and attentive member of an audience. Therefore, we will conclude this chapter by exploring the connections between emotions, attentiveness, and learning behaviour.

We will discuss explicit and implicit feedback applications as well as the effects of emotions on a person's attentiveness

2.1 Feedback Applications

The audience feedback design space

The interaction of presenters with their audiences depends on the type of meeting. For example, presenters have other options to interact with their audience during an in-person lecture than in an online meeting. This diversity leads to different categories of audience feedback. Hassib et al. [2018] defined this audience feedback design space and divided it into four dimensions: *Sender and Receiver Cardinality* (one-to-one, one-to-many, and many-to-many), *Location of the Audience* (collocated or distributed), *Feedback Synchronicity* (synchronous or asynchronous) and *Feedback Sensing Style* (implicitly or explicitly). Depending on the classification of a lecture or meeting in this design space, the available feedback methods can change drastically.

Online meetings provide little to no audience feedback

Previous researchers developed multiple applications for different situations to help presenters get feedback from and on their audience. Those programs are even more critical for online presentations than they are for in-person meetings because, unlike those regular meetings, online meetings provide little to no audience feedback. [Murali et al., 2021]. The following sections will focus on one-to-many audience feedback systems since these were primarily developed for lectures or presentations. We will show explicit feedback systems, where the audience has to actively give feedback, and implicit systems, which can collect feedback without the need for the audience to take action.

Figure 2.1: Classroom Performance Keypad (right), with which students can answer questions or polls and Clicker Receiver (left) which can wirelessly collect data from all keypads [Barber and Njus, 2007].

2.1.1 Explicit Feedback Applications

Clickers are the most common explicit ARS

One of the most common explicit Audience Response Systems (ARS) are Clickers [Caldwell, 2007]. These devices allow presenters to collect responses to a posted question during a lecture without asking specific students. The first Clickers were small handheld devices that had one [Poulis et al., 1998] or multiple buttons like the Classroom Performance Clicker [Barber and Njus, 2007], shown in Figure 2.1. Those clickers allowed the audience to send messages to the lecturer. Modern Clicker systems can provide two-way communication, which helps audience members to pay more attention by directly involving them in the lecture [Teevan et al., 2012]. This engagement strongly correlates with the learning success of students [Chamillard, 2011].

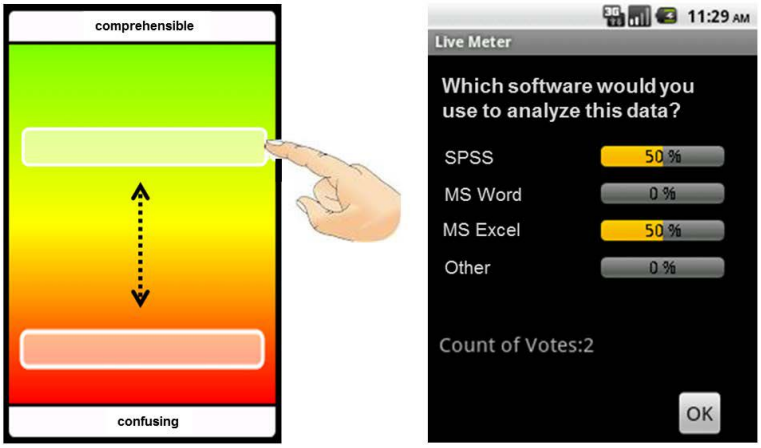

Figure 2.2: Gauge (left) and poll (right) screens of the *Live Interest Meter* App. The gauge gives users the opportunity to select a value from a continuous range for their vote while the poll only allows the selection of predefined answers [Rivera-Pelayo et al., 2013].

The LIM App provides more detailed feedback options than normal Clickers

More sophisticated ARS than Clickers can provide additional interaction methods for presenters and listeners. An example of such a system is the *Live Interest Meter* App (*LIM* App), developed by Rivera-Pelayo et al. [2013], which supports speakers presenting for large audiences (250+). The simple gauge of the *LIM* App (see Figure 2.2) allows the audience to provide and quantify their feedback. The presenter has to relay a question and possible answers to the audience to get a response. The app then aggregates, visualises, and saves the feedback so the presenter can use the feedback during his talk or analyse the talk in retrospect. During the user study, the participants considered the functions of the *LIM* App useful for gathering audience feedback during presentations.

MiRA needs two lecturers to be operated properly

Akbari et al. [2010] introduced the microblog system *MiRA* (**Mi**croblog **R**WTH **A**achen) to further increase the communication between students and lecturers. This application offers *Twitter*[1] inspired communication methods for users to communicate during and after a lecture. The optimal scenario for *MiRA* is that the lecturer has an assistant who

[1] `https://twitter.com` Accessed: 12.12.21

filters and relays the information given through the app. This division of labour is advised since using the blog and chat functions would divert the presenter's attention from the lecture itself.

Although these systems encourage student engagement and communication inside of the classroom, they may distract students from the actual lecture content [Rivera-Pelayo et al., 2013]. Furthermore, most of these systems like *MiRA* and the *LIM* app require the lecturer to present questions and polls to their audience. This additional task of motivating students to participate in the lecture actively introduces a higher cognitive load to the presenter [Rivera-Pelayo et al., 2013, Teevan et al., 2012]. Implicit feedback gathering techniques can reduce this load by removing the need for active participation by the presenter and the audience.

The use of explicit ARS increase the cognitive loads of lecturers and students

2.1.2 Implicit Feedback Applications

These Audience Response Systems (ARS) can automatically gather feedback on the audience [Hassib et al., 2018]. They can, for example, monitor the emotional state and the behaviour of the audience. Physiological signals like skin conductivity can also be monitored to infer the arousal or engagement of a person. Picard and Scheirer [2001] developed the *Galvactivator*, a glove that measures this conductivity. The glove (see 2.3) was equipped with a conductivity sensor and an LED. The brightness of this LED increases when the skin conductivity increases. Therefore a bright LED signals a high state of arousal and a dim LED a low state. To explore the *Galvactivators* communication potential and to test if it works correctly, Picard and Scheirer [2001] distributed the glove to 1200 attendees of a symposium. They found that the brightness LEDs increased during live demonstrations or when a new speaker entered the stage. Therefore speakers could see if the audience was attentive during their talks. One speaker stated that he felt disheartened when he saw that the LEDs got dim during his talk.

Implicit ARS automatically gather feedback on the audience

Skin conductivity is tied to a person's arousal

Figure 2.3: The glove of the *Galvactivator* [Picard and Scheirer, 2001].

Another gadget was developed by Hassib et al. [2017] in the form of the *EngageMeter*. This device features an electroencephalography (EEG) headset that the students have to wear (see Figure 2.4). This headset can measure the brainwaves of the wearer. Research has shown that cognitive engagement can be directly calculated from these signals [Pope et al., 1995]. The audience's attention was visualised for the presenter with three different interfaces (see Figure 2.5). These interfaces are divided into real-time and post-hoc views. To monitor the engagement during a presentation, *EngageMeter* provides a gauge as well as a graph that displays the audience's engagement as percentages. Later, the presenter can use this graph and an additional interface with the average attention for each slide for a post-hoc lecture evaluation.

Cognitive engagement can be inferred from EEG signals

Figure 2.4: Three different students wearing the *EngageMeter* headset [Hassib et al., 2017].

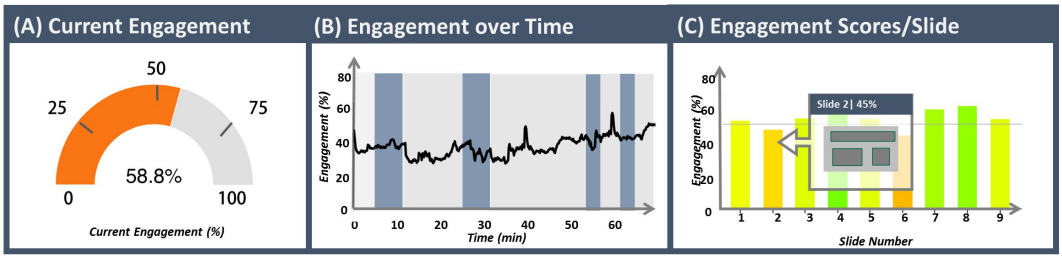

Figure 2.5: *EngageMeter* Interfaces: (A) Real-time gauge which shows the current engagement of the audience in percent. (B) Real-time engagement graph with vertical sections to indicate slide changes. (This interface can also be viewed in post-hoc) (C) Slide score showing the average engagement for each slide in post-hoc [Hassib et al., 2017].

After using *EngageMeter* in a lecture, presenters stated that the real-time and the post-hoc interfaces were both helpful. However, they described that the gauge was more challenging to interpret than the graph because they had to remember previous engagement levels to put the displayed value into perspective. This interpreting could deter lecturers from using this tool because it would increase their already high cognitive load while presenting. The claim was also backed up by the statement of one presenter who said that he was so “in the zone” that he did not want to check the interface at all [Hassib et al., 2017].

The gauge was more challenging to interpret than the graph

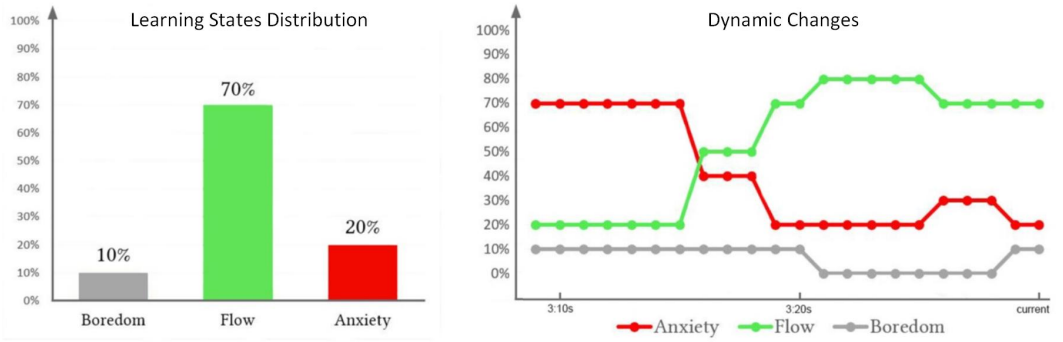

Figure 2.6: Presenter's dashboard of the audience flow prediction system. The interface shows the learning state distribution as a bar graph (left) and the changes in the distributions as a line graph (right) [Sun et al., 2019].

Using only video feeds instead of special devices is more feasible

Although these two methods provide a relatively accurate measurement for the engagement and the audience's attention, they both have a significant drawback: They need special measuring devices. To circumvent the need for additional devices Sun et al. [2019] only used webcam footage to predict a student's psychological state. This system uses real-time facial expression analysis to detect anxiety, flow, or boredom. Contrary to anxiety and boredom, the state of flow is a state of high concentration and improves one's learning capabilities [Buil et al., 2019].

The presenter is alarmed when the audience is bored or anxious

The system features two different session types: The first session type only displays the bar and line graphs shown in Figure 2.6. The second type actively intervenes if more than 50% of the listeners are bored or anxious. This intervention contains a sound alarm and a text prompt to inform the presenter. These cues remove the need to continuously check and analyse the audience's state.

The alarm reduces the cognitive load of the presenter

Sun et al. [2019] found during their user study that the system, in fact, increases the cognitive load of the presenter. Still, the intervention cues can reduce this load since they remove the need to monitor the interface constantly. Their results also showed that presenters considered interpreting the bar graph to be more intuitive than the line graph.
This statement stands in direct contrast to the findings of Hassib et al. [2017].

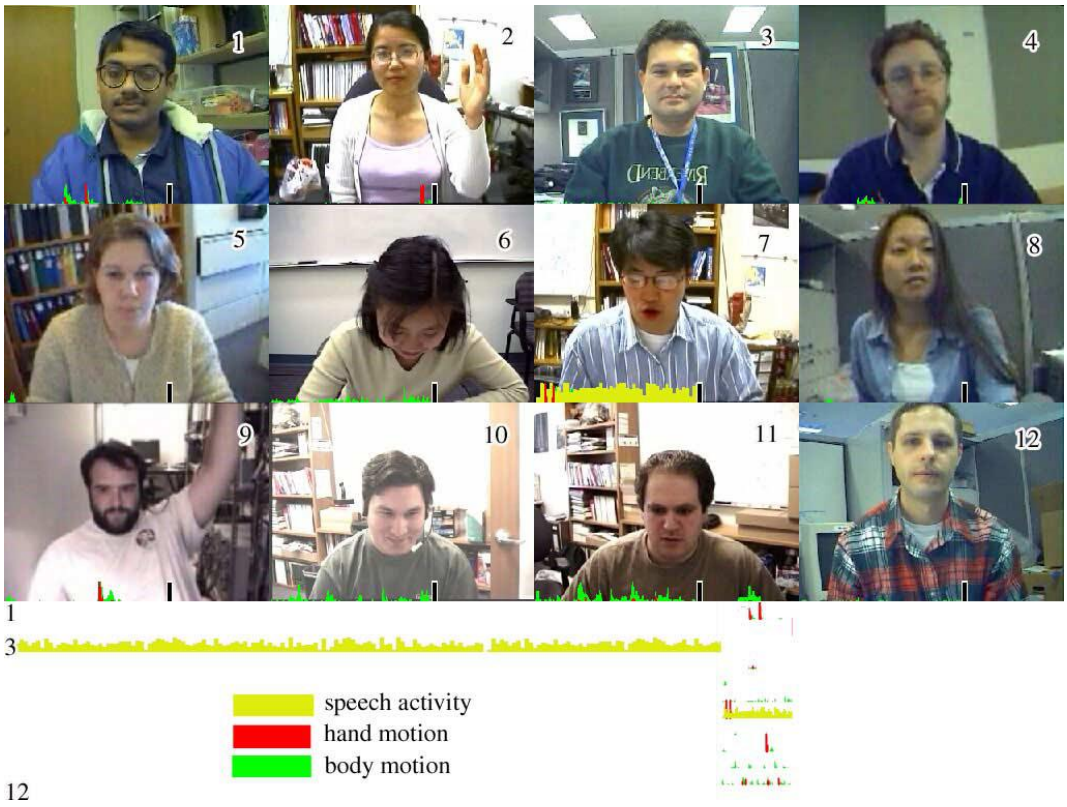

Figure 2.7: Screenshot of the activity history interface. The short-term speech and motion history are displayed bottom left of each stream. Yellow bars indicate speech activity; hand motion is shown in red and body motion in yellow. A black vertical line denotes the current time. The grouped timelines below the videos streams show the long-term activity history (The authors used the numbers on each video stream in the original paper to refer to an individual student) [Chen, 2003].

A student's movement, gestures and speaking activity is monitored

Chen [2003] proposed a different approach for gadgetless attention and engagement measurement in online meetings: His system determines if a student speaks, makes gestures, or moves in his seat. The recent activities of each student are then plotted to the bottom left of his video steam using coloured bars (see Figure 2.7). All individual activity feeds are grouped at the bottom of the interface to allow the presenter to "see" the overall classroom dynamics. The teachers interviewed during his user study stated that the system is somewhat helpful for in-class teaching since it is convenient to see events they may have missed.

This system may be more useful to review lectures

However, they also reported that teaching is a mentally taxing task. Therefore, they found it challenging to interpret and use all indicators while presenting. Furthermore, some participants mentioned that the long-term activity history could be more beneficial for self-improvement and formal teacher training than during the class.

2.1.3 AffectiveSpotlight

AffectiveSpotlight shows the most affective and active members of the audience to the presenter

In this section we want to dissect the *AffectiveSpotlight* feedback application. We chose to analyse this particular application since the authors described their development process in detail and because the application combines the ideas of using facial expressions (e.g. [Sun et al., 2019]) and behaviour (e.g. [Chen, 2003]). *AffectiveSpotlight* is a *Microsoft Teams*[2] plugin which was designed to aid speakers during talks in online meetings by spotlighting the most affective and active audience member. The authors stated that since in *Microsoft Teams* the space for displaying the audience members' video feeds is limited to three to four feeds, only the most interesting feeds should be shown to the presenter.

Presenters considered confusion, engagement, raised hands, speaking and head-nods as interesting audience feedback

To find out which feeds are the most interesting, Murali et al. [2021] conducted an exploratory survey where they asked 175 presenters and lecturers which type of audience feedback they considered to be most important. The participants stated that cognitive states, like confusion and engagement, are more interesting than the other emotions. Regarding behaviours, raised hands, speaking and head-nods were considered more interesting than any other behaviour. However, only head gesture recognition was incorporated into the final application. The detailed results can be found in Figure 2.8.

[2] `https://www.microsoft.com/teams` Accessed: 12.12.21

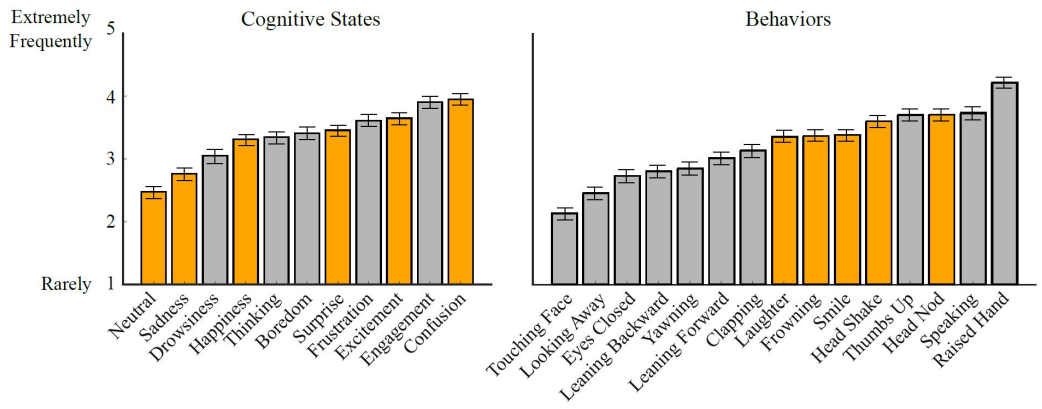

Figure 2.8: Presenters' preferences of audience reactions and cognitive states during online presentations found by Murali et al. in their exploratory survey. Cognitive and behavioural markers that are marked orange were later (partially) used to determine which members of the audience are shown to the presenter [Murali et al., 2021].

The AffectiveSpotlight Application

Murali et al. [2021] used a CNN and a HMM to analyse the video feeds

Murali et al. [2021] then used the results of their exploratory survey to develop the *AffectiveSpotlight* application itself. They utilised the *Microsoft Face API*[3] to detect faces and face landmarks in the participant's video frames. The detected faces and landmarks were then analysed by a Convolutional Neural Network (CNN) and a Hidden Markov Model (HMM). The CNN developed by Barsoum et al. [2016] was used to determine the emotional state of a person. From the available emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) Murali et al. [2021] only included happiness, sadness, surprise, and neutral into their system. The authors made this selection because these emotions were similar to the cognitive states mentioned in the exploratory survey (see Figure 2.8). The HMM trained and developed by Kapoor and Picard [2001] calculated and recognised the head gestures.

[3] `https://azure.microsoft.com//services/cognitive-services/face/` Accessed: 12.12.21

A weighed average determines the most affective and active members of the audience

To find the most affective and active audience members, Murali et al. [2021] computed a score for each video frame using a weighted average composed of the features extracted previously. The authors used the preferences discovered in the exploratory survey and an additional pilot evaluation to fine-tune the weights of each feature. The scores for each participant were gathered over a 15-second interval. Then, the system spotlighted and showed the audience members with the highest activity score to the presenter for the next 15 seconds.

User Study and Evaluation

AffectiveSpotlight was tested against *RandomSpotlight* and the *DefaultUI*

To test *AffectiveSpotlight* Murali et al. [2021] conducted a user study with 117 participants which were randomly divided into 13 groups. In each group, a randomly chosen presenter had to hold three different five-minute talks in front of the remaining eight people. According to the exploratory survey [Murali et al., 2021], this size is typical for *Microsoft Teams* meetings. The presenters had eight minutes to prepare each talk and had to use a different interface for each talk. These interfaces were *AffectiveSpotlight*, *RandomSpotlight* and *DelaultUI*. *DelaultUI* was the basic Microsoft Teams Interface (see Figure 2.9 right side) with the presenter's slides and a limited set of audience members on the bottom. *AffectiveSpotlight* and *RandomSpotlight* both displayed the same interface (see Figure 2.9 left side) with a spotlighted audience member on the left side of the screen and the slides on the right side. The difference between these two conditions was the selection process of the spotlighted person. The *AffectiveSpotlight* interface chose the spotlighted person by the algorithm described above, while the *RandomSpotlight* interface randomly selects a person. However, the presenters were not told the spotlight selection criteria before their talks. The performance of each interface was then evaluated by the presenters' self-reports and interviews with presenters and members of the audience.

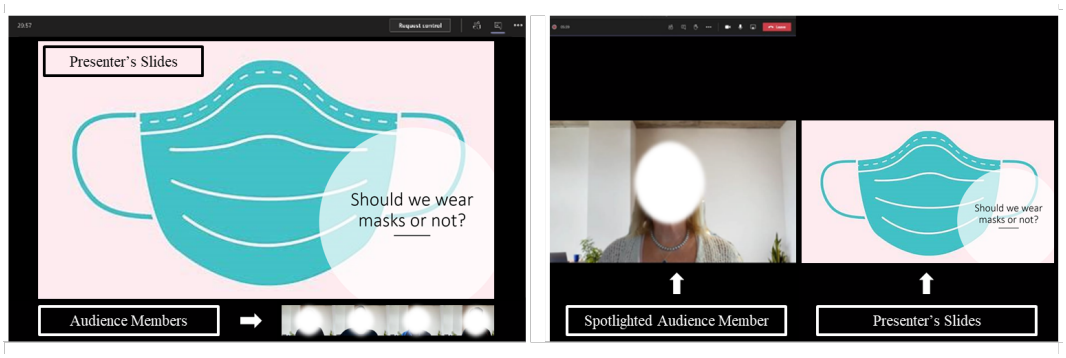

Figure 2.9: Left: Standard Microsoft Teams presentation interface with the presenter's slides and small pictures of the audience members. Right: *AffectiveSpotlight* interface with the highlighted audience member on the left and the presenters slides on the right [Murali et al., 2021].

The evaluation of the system showed that *AffectiveSpotlight* was rated significantly higher in terms of system satisfaction, ease-of-use, and future potential use than the other two interfaces. Presenters additionally stated that the system made them more aware of their audience. This increased awareness also led to a more similar evaluation of presentation quality by the audience and the presenter. In addition to these findings, the results suggested that *AffectiveSpotlight* impacted the presenter's anxiety and confidence in a positive way.

AffectiveSpotlight made the speakers more aware of their audience

Murali et al. [2021] also mention some problems and limitations of their system. For example, the AI systems used to analyse the video streams are far from perfect. They only work reliably when the video stream fulfils specific requirements. To ensure that those requirements were met during their study, the researchers provided camera calibration guidelines which asked for a neutral background, good lighting, no face occlusions (e.g. hats/sunglasses), and a centered frontal face orientation (see Figure 2.10). To further prevent misclassifications, expressions and emotions were not explicitly labelled by the AI systems. Instead, the AI only influenced which signals were shown to the presenter since a human can interpret the expressions more effectively [Murali et al., 2021].

The algorithm has strict requirements regarding camera angle, lighting etc. to work properly

Figure 2.10: Background and camera orientation guidelines provided by Murali et al. [2021] for their user study.

AffectiveSpotlight was only tested in small meetings

Furthermore, *AffectiveSpotlight* was developed and tested on small meetings of around eight people. During the user study, the system spotlighted only 40% of the audience [Murali et al., 2021]. It needs to be confirmed if this percentage holds for larger meetings sizes (e.g., 50+ participants). If the system only shows the same three active people in an online lecture of around 50 students, it is hard for the presenter to "read the room". This problem may make *AffectiveSpotlight* unsuitable for larger online lectures.

2.2 Effects of Emotions on Attention

We want to use a person's emotional state to infer his attentiveness in our own application. Therefore we will now have a look at how exactly the emotional and cognitive state affects a person's attention and learning performance.

One example of such a state is the state of Flow, where the students feel a great sense of motivation and engagement. This state leads to better learning outcomes and a higher level of mental satisfaction [Buil et al., 2019, Csikszentmihalyi and Larson, 2014]. However, Flow is mostly detected indirectly by measuring the perceived task difficulty [Basawapatna et al., 2013]. The idea is that a too easy task bores the student and a too complex task invokes anxiety. Flow occurs when the task's difficulty is just right [Davis, 1977]. Since those indirect measurements are complicated and task-dependent, we will now look at the effects of basic emotions on attentiveness and learning behaviour.

Flow enhances a person's learning capabilities

BASIC EMOTIONS:
The emotions anger, fear, sadness, enjoyment/happiness, disgust and surprise are considered the most basic [Ekman and Oster, 1979, Ekman, 1992]. These emotions can be divided into positive and negative [Watson et al., 1988]. Based on this claim, we classify enjoyment/happiness and surprise as positive emotions and anger, fear, sadness, and disgust as negative emotions. Additionally, if a person is in none of these moods, his emotional state is described as neutral.

Definition:
Basic Emotions

Negative moods like fear and sadness lead to behavioural lapses like mind wandering and task-irrelevant thoughts. Positive emotions, however, decrease those behavioural lapses and enhance the participant's ability to adjust their performance after such a lapse occurs. In addition to the frequency of those lapses, emotions also directly influence attentional commitment. A negative mood reduces a person's attention to a specific task and shifts it to task-irrelevant or personal concerns [Smallwood et al., 2009].

Positive emotions reduce the frequency of behavioural lapses

People in negative moods learn slower and solve problems less efficiently

Furthermore, negative moods can also impair learning success and transfer tasks. In the study conducted by Brand et al. [2007] participants in a negative emotional state needed more repetitions to learn how to solve the three-disk and four-disk Tower of Hanoi problem. Moreover, participants in a negative mood solved the following transfer tasks (e.g. five-disk Tower of Hanoi) less efficiently. In contrast, a positive mindset can increase cognitive flexibility and learning abilities. For example, people learn faster how to categorise stimuli based on rules and they find the optimal classification strategies more quickly when no such rules are given [Nadler et al., 2010].

In conclusion, it can be said that positive emotions like happiness improve a students' attentiveness and ability to learn. Still, negative emotions like sadness lead to behavioural lapses and, therefore, reduce the student's attention.

Chapter 3

Designing a Feedback Application for Online Meetings

For the implementation of our attention measuring tool, we decided to use Zoom as the base application as it is one of the most popular video conference tools in the world[1]. The prototype operates on a simple loop which is executed once every second. This loop can be divided into five steps:

- Finding the participants
- Identifying each participant
- Collecting each participant's data
- Inferring the participant's attention level from the collected data
- Displaying this information to the presenter

The development process is divided into five steps

In the first step, we decided to directly search for the faces of participants since it is difficult to differentiate individual video streams in Zoom by shape alone. Then we match each detected face to the displayed names in the second

[1] `https://zoom.us` Accessed: 12.11.21

step and use them as a unique identifier. This identifier is necessary since we want to keep track of the participant over time. Next we collect data like emotional state and head orientation from each image and save it under the identifier found in the previous step. Lastly, we use this data to infer the participant's attention level, which is then displayed to the lecturer. In the following chapter, each of these steps is described in detail.

3.1 Finding the Participants

First, we must find each participant's video stream. We need to detect the whole video stream and not only the participant's face since we want to use the name in the bottom left of the box to identify each participant. This identification is necessary to keep track of a participant over time as the position of each participant is not fixed in the Zoom window. To find those streams, the application starts by taking a screenshot of the Zoom window once every second. We take a screenshot of the whole window because a special Video-SDK is needed to access a participant's video stream directly. This SDK is provided by *Zoom Video Communications* itself. However, a paid developer account is required in order to access it[2]. Since those accounts are expensive and for our application, the screenshots are sufficient, we decided against using the SDK.

We used screenshots instead of the Video-SDK

This detection task proved to be more complicated than we first anticipated. As shown in Figure 3.1 the standard Zoom-window is divided into multiple boxes, each containing the video stream of a participant. As mentioned before, these boxes can shift around and change in size depending on the number of participants in the Zoom-Meeting. First, we investigated the *VNDetectRectanglesRequest* from *Apple's Vision Framework* to detect each box, but the algorithm was not able to reliably detect the boxes. Likely, the boxes do not provide a constant contrast to the

Individual video stream boxes were not detectable

[2] `https://zoom.us/buy/videosdk` Accessed: 12.11.2021

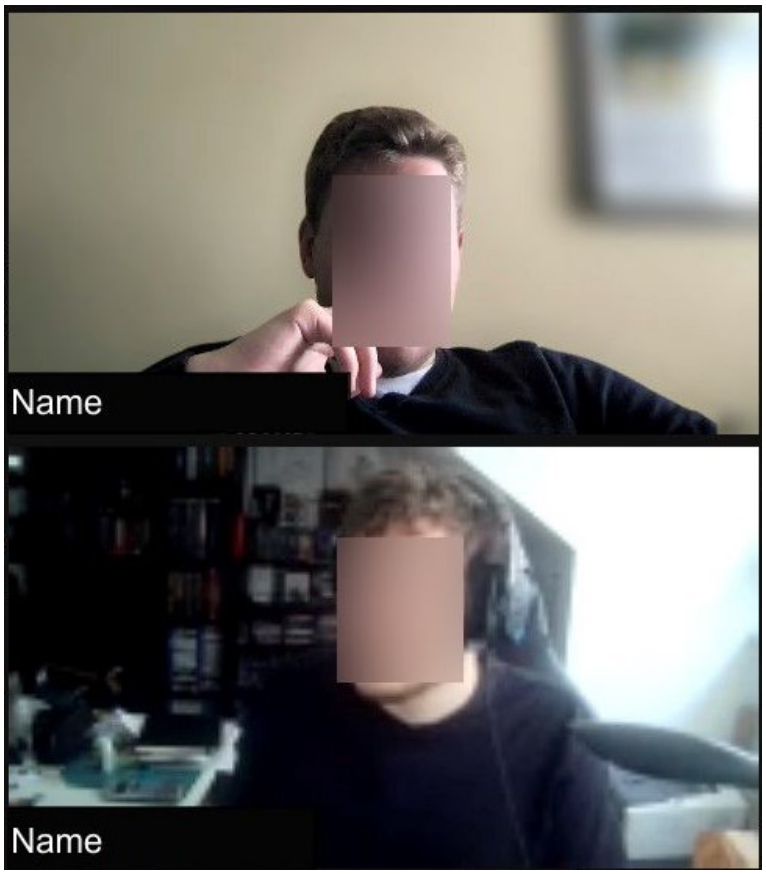

Figure 3.1: The basic Zoom window with two participants. The participant's name is displayed in a transparent box in the lower left of each window. (The names and faces of the participants were covered to protect their identity)

background due to the nature of a video stream, which leads to the framework not recognizing the stream as a rectangle. We tried using an image filter that changes the background-colour to improve the contrast on the edges of each box, but this did not improve the recognition.

VNDetectRectanglesRequest:
This Vision request detects regions of an image with rectangular shapes, like credit cards, business cards, documents, and signs. The request returns the bounding boxes, orientation, and confidence level for each rectangular shape found [Inc., 2021b].

Definition: *VNDetectRectanglesRequest*

Only the participant's face is detected

We chose to use the *VNDetectFaceLandmarksRequest* which reliably detects all faces in a given screenshot of a Zoom-Meeting. Since we now only have the faces and not the whole box, we need other ways to identify each participant.

Definition: *VNDetectFaceLandmarksRequest*

VNDETECTFACELANDMARKSREQUEST:
This Vision request finds all faces in the input picture then each face is analysed to detect its facial features. These features include but are not limited to the positions of the eyes, the faces roll or yaw, and the faces bounding box [Inc., 2021a].

3.2 Identifying the Participants

The participant is identified by their name in the lower left corner

At this point, we only have the position of the face of each participant. This position is specified via a bounding box which is shown in Figure 3.2. Identifying people by their faces alone is quite difficult for a computer program. Therefore, we chose to use the participant's name in the lower left of each box to identify the participant. To do so, we must decide which name belongs to which face. Since we know the layout of the box is always the same, we can infer that the correct name is always below and left of each face's bounding box. Since the individual boxes are arranged in a grid, we can extend each bounding box to the left and bottom till we hit the next bounding box or the edge of the Zoom window. The result of this process can be seen in Figure 3.3.

The name is converted from an image to a processable string

There are some cases when this method does not work, but those are very rare and not taken into account in the scope of our work. The extended bounding box now contains the name of the participant. To detect and "read" the name we used the *VNRecognizeTextRequest* of *Apple's Vision Framework* . In this case, reading means the process of converting an image of text to a processable string. To do so, the content of the extended bounding box is cropped from the screenshot. Then we enlarge the size of this cutout and

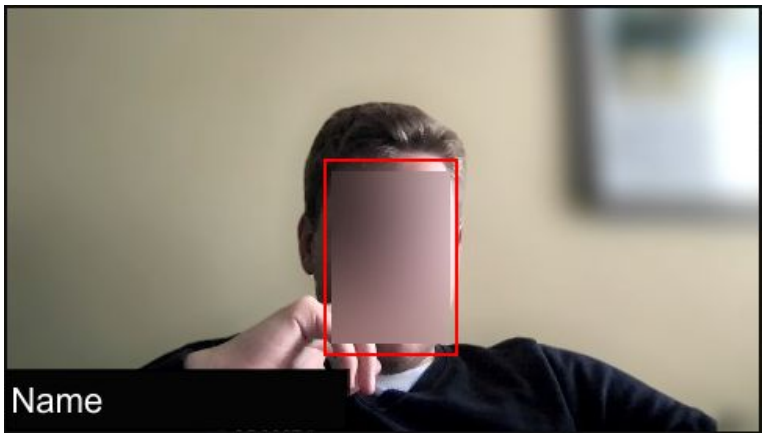

Figure 3.2: Zoom participant with face bounding box (red square). (The names and faces of the participants were covered to protect their identity)

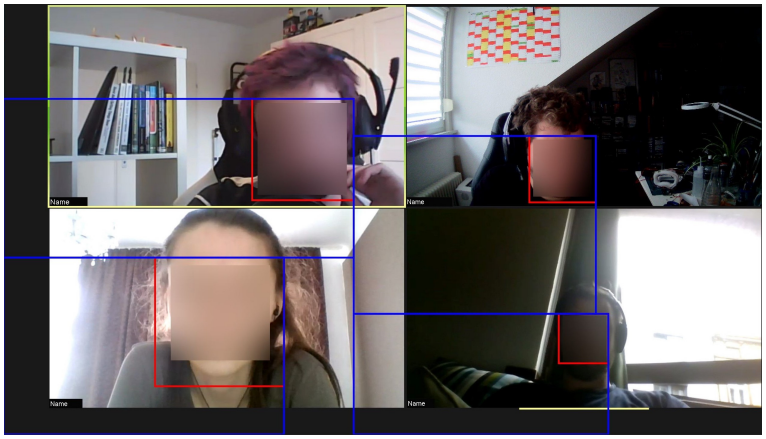

Figure 3.3: Zoom participants with face bounding box (red square) and the extended bounding box (blue square). (The names and faces of the participants were covered to protect their identity)

use a filter to increase the contrast between the white text to the black transparent background. These two steps make the text recognition process more reliable. The *VNRecognizeTextRequest* is now performed on this modified image.

Definition: *VNRecognizeTextRequest*

VNRecognizeTextRequest:
This Vision request locates all characters in a given image. Those characters are analysed and converted into a string. The result of this request contains each analysed string and the bounding boxes of the located text. [Inc., 2021c].

The detected text is processed to remove noise

However, the result is not completely correct at this point and needs further processing. The results may contain incorrectly recognised characters as well as unintentionally detected lines of text from an image's background. We used some simple filters to improve the results: First, we remove every character that is not alphanumeric. Then we remove every text that could be part of the Zoom window itself, like "Mute" or "Stop Video". Lastly, we remove the detected texts that contain less than three characters. If there are still multiple lines of text or no text was detected, the participant's image is not further processed. We use the Levenshtein distance to accommodate the mentioned problem of incorrectly recognised characters when comparing these strings. We consider strings that have a Levenshtein distance smaller than three as equal.

Definition: *Levenshtein distance*

Levenshtein distance:
The Levenshtein distance is the editing distance between two words. It denotes the minimal number of single-character edits (insertions, deletions or substitutions) that is required to convert the first word into the second [Levenshtein et al., 1966].

3.3 Collecting Data

Since we know the positions and names of each participant, it is now time for the actual data collection. To narrow down which features we could extract from a video stream and which of these features would be the most useful, we had a look at similar applications. As a basis we used *AffectiveSpotlight* developed by Murali et al. [2021] which we dissected in Chapter 2.1.3. In their app Murali et al. [2021] used a Convolutional Neural Network (CNN) to infer the emotional state of an audience member and a Hidden Markov Model (HMM) to recognise head gestures. The HMM, however, imposes many constraints regarding the camera angle, lighting, and background of the video stream to work reliably. Therefore we decided to only use a CNN in our analysis.

We want to extract the emotional state of the participants

We assumed that we can estimate the attentiveness of a person from their emotional state alone because positive emotions improve learning rates, and negative emotions lead to mind wandering and behavioural lapses that reduce a person's attentiveness (see Chaper 2.2). We used the CNN presented by Newnham [2018] for our analysis since it was already integrated into *Swift*, the programming language we used for this project, and because it is based on a similar data set as the one used in *AffectiveSpotlight*. The data set of Newnham's [2018] CNN was taken from the 2013 competition of the International Conference on Machine Learning (ICML). It contains 28.000 grayscale images (48 x 48 pixels) labelled with the seven basic emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral (see Figure 3.4).

A CNN can infer a person's emotional state

Due to the data set, this CNN expects centered grayscale images with a size of 48 x 48 pixels. The face must not be obscured, rotated, or turned away from the camera to fit the input requirements of the CNN. All three of those requirements can be checked in the result of the already processed *VNDetectFaceLandmarksRequest*. Since the framework detected the face, it should be visible enough to be processed. The rotation and yaw of the face are also contained in the result. Both are measured in predefined discrete in radian

The images are prepossessed to fit the input requirements of the CNN

Figure 3.4: Excerpt from the 2013 ICML competion dataset. The size of each picture is 48 x 48 pixels and it contains exactly one centred face. Each picture is labelled with one or more basic emotions [Newnham, 2018].

We only processed correctly oriented faces

intervals[3] which are identified by their midpoint. We decided to only process faces whose yaw and rotation angle laid within the interval labelled with *0 rad*. These intervals correspond to a maximum rotation angle of 15° to each side and a maximum yaw angle of 22.5° to each side. Furthermore, we did not adjust the other faces manually because we could not rotate the faces exactly due to the coarse intervals, and even minor differences in a face's rotation may change the CNN's output. Unfortunately, pitch was not supported when we developed this application.

The CNN calculates confidence values for each of the seven basic emotion

After we made sure that each face is oriented correctly, we proceeded by grayscaling and resizing them. These pictures now can be processed by the CNN, which returns a prediction that contains seven tuples. Each tuple contains a string that refers to one of the seven base emotions (anger, fear, sadness, happiness, disgust, surprise and neutral) and a double value representing how sure the CNN is that this emotion is displayed. Those values are given in percent, and all seven add up to 100%.

[3] 30° intervals for roll and 45° intervals for yaw

The collected data is then saved into a participant object. Each of these objects is identified by the participant's name and holds up to five separate measurements. If a participant was already detected and processed in a previous step, this object is updated with the new data. If not, a new object is created. We compared the new participant to each already existing participant by measuring the Levenshtein distance. The new one is matched to the existing one with the smallest distance when the distance is smaller than three.

The new data is saved with the participant's name as identifier

3.4 How to Infer the Attention-Level

After collecting and saving each participant's data, we now needed a way to infer a participant's attentiveness from it. Since there was no proven formula to derive a person's attention level from his emotional state, we developed our own formula which is described within this section.

To design such a formula, we needed sample data from which we could derive its parameters. To get such samples, we conducted a small exploratory study. We performed this experiment with three participants from our computer science department. All three participants attended similar 30-minute online meetings. These meetings started with ten-minute presentation followed by a 20 minute long discussion of its topics. In this experiment, we recorded the participant's face while he used an application to self-report his attention level. The application (shown in Figure 3.5) displayed a bar with six emojis, each representing a certain level of attentiveness. The emoji corresponding to the current attention level was highlighted, and the participant could change the level by clicking on an emoji. The tool was designed so that the level of attention would slowly decay if the application was left alone. To keep a high attention value, the participant has to confirm or readjust his attention level from time to time. Additionally, a higher attention level will decay faster than a lower one. A graphical representation of this behaviour is shown in Figure 3.6.

We conducted an exploratory study to design our attention formula

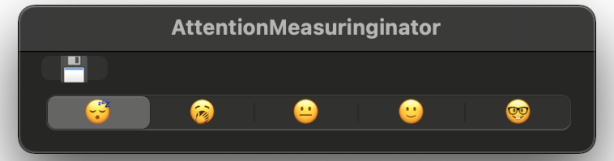

Figure 3.5: The *AttentionMeasuringinator* enables a participant to self report his attention level. This application has six emoji labelled buttons each representing a different level of attentiveness. Currently, the application is set to a low level, since the leftmost button is highlighted. The button in the top left with the Floppy disk saves the recorded values and terminates the application.

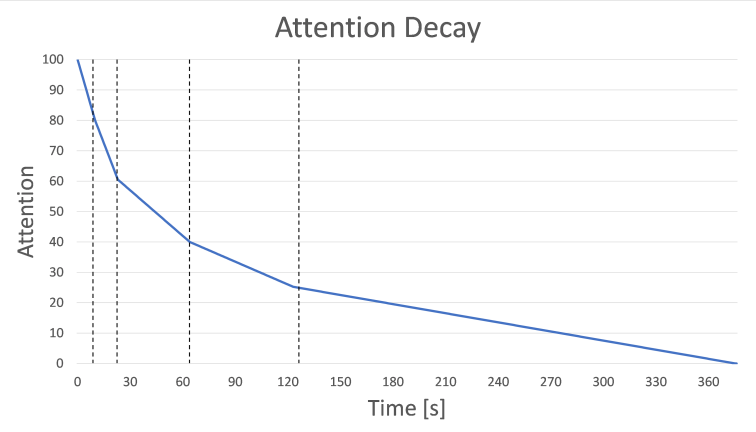

Figure 3.6: Graphical representation of the attention decay formula used in our exploratory study. The dotted lines mark the transition from one attention level to the next one.

We collected and analysed more than 5000 pictures and the corresponding attention values. However, the results of this analysis did not show any correlation between the detected emotions and the self-reported level of attention. Multiple factors can explain this result: One reason could be that our conversion pipeline or the CNN itself is flawed. Another one could be that the self-reported values are inaccurate. We conducted a series of tests trying to fix those possible errors:

Our data showed no correlation between the detected emotions and the reported attention level

Since we could not prove or disprove if the self-reported values were inaccurate, we decided to take a closer look at the CNN and its conversion pipeline. Newnham initially used the CNN in an *iOS* application [Newnham, 2018]. Therefore all conversions like resizing and grayscaling had to be rewritten for a *macOS* application. This conversion was necessary because *iOS* and *macOS* do not use the same data structures to represent images. Some features of the *UIImage* in *iOS* applications are not implemented for *NSImage* in *macOS* applications. Thus we had to implement some of those features ourselves.

We had to port the image processing pipeline from *iOS* to *macOS*

After comparing the results of our application with the results of the iOS implementation presented by Newnham [2018], we confirmed that the results were roughly equal. During the comparison of results, we detected the phenomenon that both implementations produced changing outputs on still images. This noise can result from hardly noticeable differences in the inputs, depending on the bounding box calculated by the *Vision Framework*. Since the face's bounding box is detected automatically, it can move slightly from detection to detection. Those differences can significantly impact the result calculated by a CNN due to the way those networks operate.

The CNN produced noise on still images

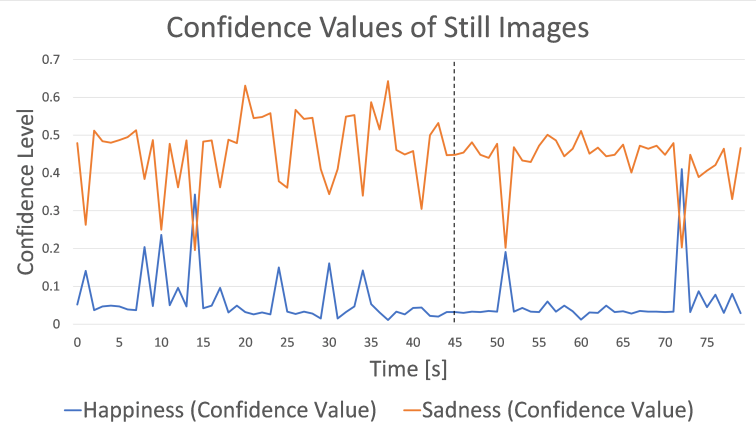

Figure 3.7: This graph shows the confidence values for sadness and happiness when the CNN analyses two different still images for an extended period of time. From second 0 to 45 a picture of a happy person is analysed, from 45 to 80 a picture of a sad man. This change is marked by the dotted line.

The CNN could not detect disgust, surprise, and anger reliably

Another thing we noticed was that the CNN detects some emotions better than other ones. The CNN could not recognise disgust, surprise, and anger reliably even if pictures showed those emotions explicitly. Happiness and sadness were detected more reliably, but the CNN's output was always skewed towards sadness. The confidence value of sadness was almost always between 0.3 and 0.5 irregardless of the analysed picture. The confidence values for happiness moved between 0 and 0.2.

Only happiness and sadness were detected reliably

Both phenomena can be seen in Figure 3.7. From second 0 to second 45, a picture of a happy person was fed into the CNN. From second 45 to 80, the image was exchanged for that of a sad person. Although the confidence value of sadness is almost always higher than the confidence value of happiness, the happiness value spikes more often while the happy face is shown. Furthermore, a high spike in the confidence value for happiness resulted in a low spike in the confidence value of sadness. Therefore, it can be inferred that one emotion excludes the other.

Since happiness and sadness were the only emotions that the CNN could detect reliably, we focused on these two emotions to develop our attention formula. However, the results of our exploratory study showed no correlation between these two emotions and the self-reported attentiveness. Therefore, we relied on the results of the related work presented in Chapter 2.2. Those results stated that attentiveness and learning behaviour is improved by positive emotions and decreased by negative emotions. We concluded that a high happiness confidence value would lead to an increased attention level, and a high sadness confidence value would lead to a low attention level.

Happiness increases attention, sadness decreases it

For the final calculation, we used a weighted sum of the two confidence values. The confidence value for happiness was weighted higher than the value for sadness. This weighting was done to accommodate the general difference in the range of both values. The resulting value was then mapped uniformly to a value between 0 and 100. Where 0 is the lowest possible attention level and 100 is the highest. This mapping makes it easier to display the resulting value in the following step. This estimation should be accurate enough to give presenters an idea how attentive their audience is.

We used a weighted sum of happiness and sadness confidence values

3.5 Displaying the Attention Value

We designed three new interfaces

Now that we have an attention value, we need to find out the best way to display it to the presenter. However, we decided against Murali et al.'s [2021] approach, which only shows the most attentive and active members to the lecturer because, with this method there is a risk that the lecturer sees only a tiny part of the audience. This method would have the consequence that the presenter would not be able to estimate the attention level of the whole audience correctly. Therefore, we designed three new interfaces, each emphasising another aspect: A basic single emoji interface, a multi emoji interface, and a graph. The single emoji interface only displays the current attention level. The multi-emoji interface provides additional information about the audience's emotional state, and the graph shows changes in the attention level over time. The user study during which all three interfaces were tested is described in Chapter 4.

3.5.1 Basic Emoji Interface

The *Basic Emoji Interface* is designed to be as simple as possible

First, we developed a very basic interface, which only showed a single emoji. This emoji can have three possible states (see Figure 3.8), each representing a different attention level. If the audience's attention in the Zoom meeting is low - or, more precisely, if the attention value is in the field of 0 to 33 - a sleeping emoji is shown. An attention value between 34 and 66 will result in a smiling emoji representing an average attention level. A nerdy smiley corresponds to a high attention level with an attention value between 67 and 100. We designed this interface to be as simple as possible, so interpreting it introduces only a small cognitive load onto the presenter. Additionally, the sudden changes from one emoji to another may catch the attention of the lecturer, like the alarm presented by Sun et al. [2019]. However, due to its simplicity, this interface provides only

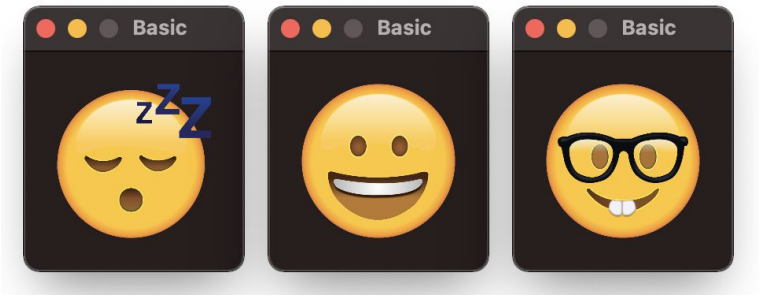

Figure 3.8: Picture of the *Single Emoji Interface*. The displayed emoji changes depending on the attention level of the audience. There are three possible levels: low, average, and high attention (left to right).

a rough estimate of the audience's attention level and no information on why the audience might be attentive or inattentive. Therefore, the lecturer can only react very late to decreasing attention level and may not know how to improve it.

3.5.2 Multi Emoji Interface

This interface which is shown in Figure 3.9 is an extension of the *Basic Emoji Interface*. In addition to the emoji that displays the general attention level, this interface features six other emojis, each corresponding to a primary emotion. Those emotions are happiness, sadness, anger, surprise, fear, and neutrality. We left out disgust to keep a better balance between positive, negative, and neutral emotions and because it was the emotion that was recognised worst by the CNN. The corresponding emojis will change in size to show which emotions are dominant in the audience. The more prevalent a particular emotion is, the bigger the emoji gets. This behaviour can also be seen in Figure 3.9. We developed this more complex interface to give the presenter more detailed feedback on the mood of his audience (for example, if the audience well received a funny story or joke he told). On the one hand, this could help the lecturer react to drops in attentiveness but, on the other hand, makes this interface more difficult to interpret.

The *Multi Emoji Interface* provides more information about why someone might be attentive or inattentive

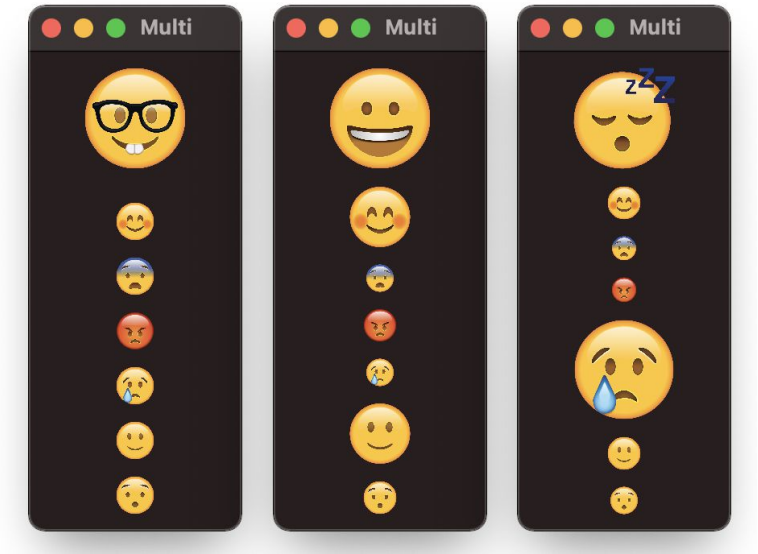

Figure 3.9: Picture of the *Multi Emoji Interface*. The topmost emoji displays the general attention label like in the *Basic Emoji Interface*. The other emojis shows one basic emotion, from top to bottom: happiness, fear, anger, sadness, neutral and surprise. Those emojis can get change in size and get bigger when corresponding emotion is more common in the audience.

3.5.3 The Graph Interface

The *Graph Interface* shows the changes in the attention level of the past 2.5 minutes

The last interface we designed consists of a simple graph which is shown in Figure 3.10. This graph shows the changes in the attention level of the past 2.5 minutes. The graph is updated every five seconds to make the line less turbulent. The most recent point is always on the right edge of the graph. This behaviour is achieved by moving the graph to the left with every update. We designed this interface, so the presenter is not obliged to monitor the interface permanently to catch changes in the attention level. Since the last 2.5 minutes are displayed, he can focus on his talk and check the audience's attention after finishing a particular section of his presentation. Furthermore, he can directly see when the attention level begins to decrease and can react before it gets to low. However, the additional interpretation effort may increase his cognitive load.

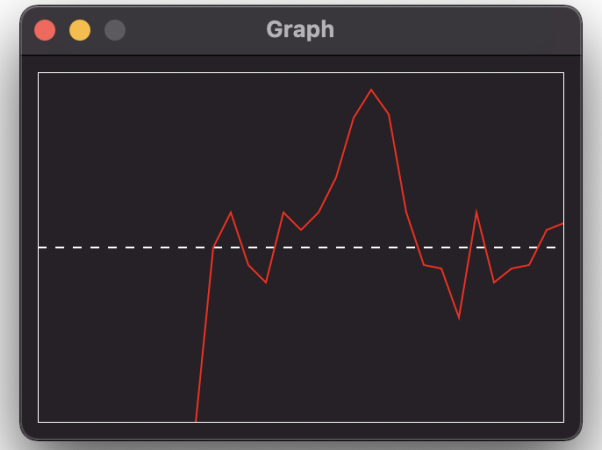

Figure 3.10: Picture of the *Graph Interface*. The dotted line in the middle corresponds to an attention value of 50%. The value on the far right edge is the most recent one.

Chapter 4

User Study and Evaluation

In this chapter, we want to explore how the presenter is impacted by the three interfaces presented in the previous chapter. Therefore, we conducted a user study as well as a small field test to evaluate these interfaces. In the user study, we primarily focused on if and how the presenter recognises and interprets changes in each interface since we wanted to see how the individual interfaces impact the presenters. We used the results of this study to develop a refined interface that combined the *Basic Emoji* and the *Graph*. This interface was then tested in an real lecture to see how it impacts the lecturer in an actual usage scenario.

We conducted a user study and a field test

4.1 User Study

The participants of this user study had to hold three short talks while using a different interface for each presentation. We conducted this study in our lab with premade slides and without a real audience. To simulate the audience, we provided an attention level curve with predefined high and low points. This fixed curve ensures that the peaks have

The user study's audience was simulated

the same height and frequency for each participant, which leads to similar study conditions. After each presentation, we examined if the presenter noticed these distinct peaks and how he reacted to them via a questionnaire and a semi-structured interview.

4.1.1 Hypotheses

For this study, we formulated the following hypotheses:

H1: The three interfaces introduce different cognitive loads onto the presenter.

H2: The recognisability of changes in the attention level varies depending on the interface.

H3: The interfaces allow the presenter to evaluate and adapt his talk swiftly.

4.1.2 Experimental Design

We conducted a pilot study before the actual user study

Before we conducted this study, we ran a pilot study with two participants to test its design. The two participants who took part in the pilot study were excluded from the main study. This pilot study helped us identify display bugs in the interfaces and gave us an estimate of how long the actual study would take (approx. 40 min). The detailed experimental design is described in the following sections.

Environment

The interfaces were displayed on a diffident *MacBook* than the presentation slides

The whole study took place in a separate room with only the participant and the investigator inside. For the entire duration of the study, the participant sat at a table with two laptops on it: One 13-inch *Apple MacBook Air* (2018) displaying the presentation slides in full screen and one 15-inch *Apple MacBook Pro* (2016) showing the respective interface. A still image of a Zoom Window was placed next to the

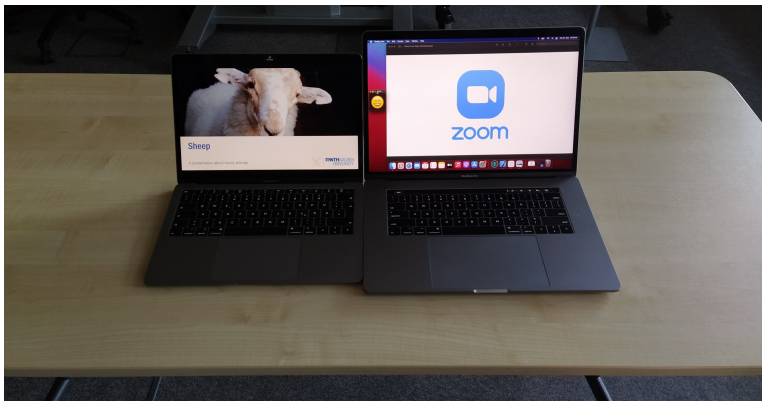

Figure 4.1: Setup of the user study: A *MacBook Air* displaying the presentation slides on the left and a *MacBook Pro* showing the respective interface and a still image of the audience on the right (The Zoom image was exchanged with the still image during the study)

interface to cover up the rest of the screen and remove possible distractions. The *Apple MacBook Pro* was placed to the right of the *Apple MacBook Air* (see Figure 4.1). We chose to use two separate laptops in order to provide a similar setup for the user study and the field study. When working with a real audience, the second screen is necessary for taking screenshots of the Zoom window with our application without interfering with the presentation (see Chapter 3.1).

Interface Types

The impact of these three interface types was measured in the study:

- *Basic Emoji*: Simple interface featuring one emoji with three different states representing a high, medium and low attention level.
- *Multi Emoji*: Extension of the *Basic Emoji* interface. In addition to the attention level it displays the audience's predominant emotions (sadness, happiness, fear, surprise, anger and neutral) with six size-changing emojis.
- *Graph*: Line graph that shows the attention level of the last 2.5 minutes.

We used the three interface presented in Chapter 3.5

For more information on each interface please refer to Chapter 3.5.

Predefined Attention Level

We predefined the attention level for five minutes since we found in our pilot study that this time-frame is long enough for each talk. Since each interface was updated every five seconds, this corresponds to 60 individual values between 0 and 100. We divided these values into three categories:

We used a predefined attention level

- 0 - 33: Low Attention Level
- 34 - 66: Average Attention Level
- 67 - 100: High Attention Level

Each curve had two predefined peaks with a high attention level as well as two peaks with a low attention level. These peaks lasted 20 seconds each, and were distributed with

Each peak lasted 20 seconds

different intervals in between. The rest of the time, the attention level was set to an average value. All participants had the same curves for each of their presentations. Since the *Multi Emoji Interface* shows several emotions at once, a predominant emotion was chosen to be displayed as follows: For an average attention level, the dominant emotion was neutral. The two high peaks exhibited either happiness or surprise. The two low peaks were linked to either sadness or anger.

Procedure

To ensure that any prior knowledge about the lecture content influences the study as little as possible, we selected three commonly known animals as the topics for the predefined talks: Sheep, Fish and Penguins. Each talk consisted of nine slides with pictures and trivial facts for the specific animal group. The order of the talks each participant had to hold was always the same. Still, we randomised the sequence of the interfaces with a 3x6 Latin Square (see Table 4.1) to counterbalance any side and learning effects during the study.

We used three predefined talks

Basic Emoji	Multi Emoji	Graph
Basic Emoji	Graph	Multi Emoji
Graph	Basic Emoji	Multi Emoji
Graph	Multi Emoji	Basic Emoji
Multi Emoji	Graph	Basic Emoji
Multi Emoji	Basic Emoji	Graph

Table 4.1: Latin Square which determines the order of the interfaces for each participant.

At the beginning of the study, we showed the study's setup to the participant. Then, the participant had time to read the informed consent form, and afterwards, the procedure was explained to him. Before every presentation, we explained the interface to the participant and then asked him to memorise the high/low points of the audience's attention during the talk and showed him the questionnaire. We also told the participant that the attention curves

We asked the participants to memorise the high/low points of the audience's attention

were predefined, and no real audience was involved. The participant had five minutes to familiarise himself with the presentation slides. After he was done, the participant had another five minutes to hold the respective talk. Following this, the participant had to fill out the questionnaire, and we conducted a short semi-structured interview. In this questionnaire, the participant had to rate the following statements on a 5-Point Likert Scale:

- The changes in attention were easily noticeable.
- The display method was easy to interpret.
- The display method was not distracting.
- The given information helped to infer the attention of the audience.

The participants were interviewed and had to fill out a questionnaire

Additionally, the participants had to rank the three interfaces based on their usefulness (for the complete questionnaire, see Appendix A). In the interview, we asked the participant on which slides the audience was most/least attentive. We also asked how high the attention level was at specific points of the presentation that were not mentioned during the first question. In total, we checked three to five points per presentation.
This procedure was then repeated for the other two interfaces. The participant had the chance to take a break after each talk. During these breaks, we offered sweets and drinks.

4.1.3 Participants

12 people took part in the user study

Twelve people participated in this study. Their age ranged from 22 to 30 ($M = 25.09$, $SD = 2.39$). Four participants were female, and seven were male. One participant reported neither age nor gender. All participants had a scientific background, most of them being computer scientists. Each participant had at least some experience with public speaking and online presentations.

4.1.4 Results

In the following two sections, we will present the results collected from the questionnaire and during the interviews. The first section will focus on the quantitative results, i.e., the answers to the 5-Point Likert Scale questions and how well the participants remembered the audience's attention levels. In the second part, we will focus on the comments given by the participants.

Quantitative Results

The results of the questionnaire are shown in Table 4.2. There were almost no differences between the interfaces regarding noticeability of changes, helpfulness and rank. Furthermore, the participants perceived all three interfaces as equally distracting (see Figure 4.2). However, the *Multi Emoji Interface* was considered as harder to interpret in contrast to the *Basic Emoji Interface* and the *Graph Interface* (see Figure 4.3). In addition, participants had a poorer recall of the audience's attention while using the *Multi Emoji Interface* compared to the other two. This also applies to the recall of the audiences prevalent emotions (see Table 4.3 and Figure 4.2).

The *Multi Emoji Interface* was harder to interpret than the others

	Noticeability		Interpretability		Distraction		Helpfulness		Rank	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Basic Emoji	1.92	0.64	1.30	0.47	1.92	0.95	1.83	0.69	1.75	0.72
Multi Emoji	2.42	0.95	2.25	0.72	2.33	1.31	2.33	0.75	2.58	0.49
Graph	1.83	0.80	1.50	0.87	2.00	0.91	2.00	0.91	1.67	0.85

Table 4.2: Results of the questionnaire regarding noticeability, interpretability, distraction, helpfulness and ranking.

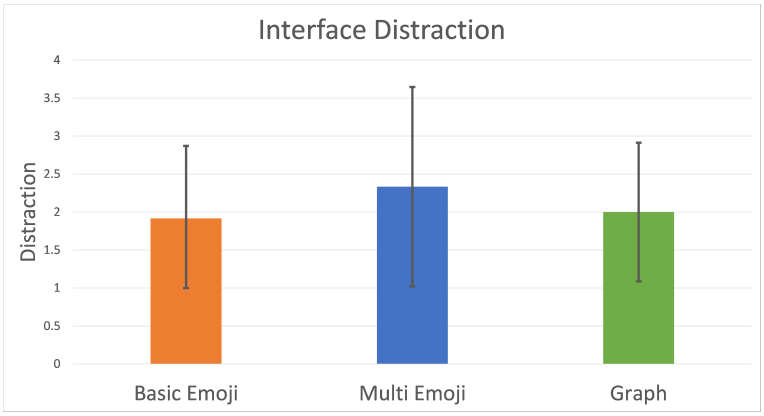

Figure 4.2: Mean and standard deviation of the distraction an interface imposed onto the participant rated on a 5 point Likert Scale. (1 = No distraction 5 = High distraction)

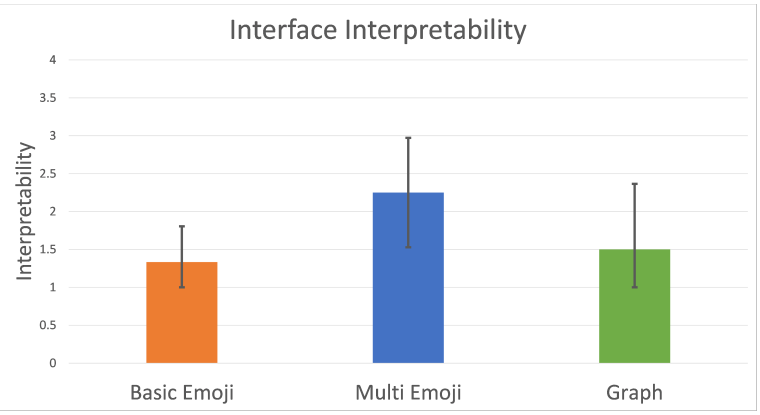

Figure 4.3: Mean and standard deviation of the interpretability of the interfaces rated on a 5 point Likert Scale. (1 = Easy to interpret 5 = Hard to interpret)

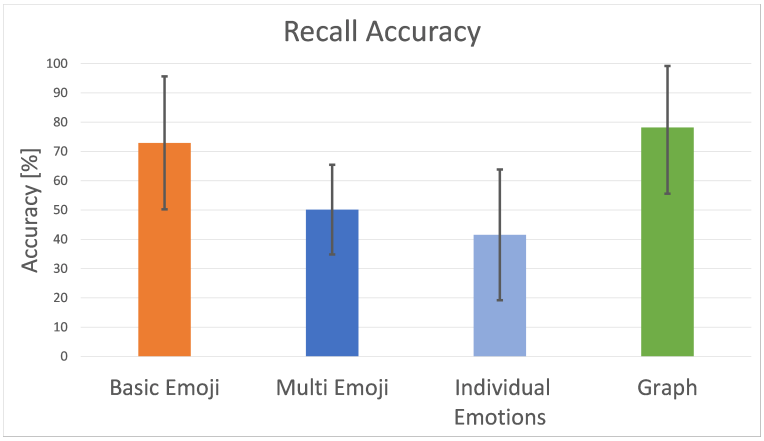

Figure 4.4: Mean and standard deviation regarding how accurately the participants could remember the audiences attention level at specific points in percent.

	Recall Accuracy	
	Mean	SD
Basic Emoji	72.92%	22.69
Multi Emoji	50.14%	15.30
Individual Emotions	41.53%	22.31
Graph	78.19%	21.00

Table 4.3: Interview results regarding how accurately the participants could remember the audiences attention level at specific points.

Comments

During the interviews, the participants had the opportunity to comment on problems they had with the interfaces and what changes they would suggest.

The participants did not know how to react to low attention levels

After using the Basic Interface, two participants mentioned that this interface is easy to interpret since it had only three distinct states. Furthermore, five participants stated that the switch from one attention level to another was clearly noticeable. However, three of them said that they did not

know how to react to low attention levels and therefore deemed this information as unnecessary during the presentation. According to the participants, it was difficult to adequately respond to the attention drop because the presentation required too much of their concentration.

The participants could not process the additional information provided by the *Multi Emoji Interface*

Concerning the *Multi Emoji Interface*, eight of the twelve participants said that it was hard to read and the amount of information this interface provided was too much for them to process during the talk. Three participants stated that they noticed that something had changed. Still, they could not always identify what exactly it was. They said they had these difficulties because they could not remember the previous state of the interface. The opinions of the participants were split on whether a single emoji would be better to reflect the prevailing mood of the audience: Four participants agreed on this, while three said that interpreting even this reduced interface would be too hard. However, all seven stated that the most interesting information by far was the general attentiveness displayed by the topmost emoji.

The *Graph Interface* helped the participants to see the changes in the attention level even when they concentrated on their talk

Three participants reported that the history the *Graph Interface* provided helped them to detect parts of their talk where the audience was (not) attentive. They said they noticed the peaks and dips even when they were entirely focused on their talk and only looked at the graph after finishing a section. Four participants said that the *Graph Interface* provided a more detailed representation of the attention value since it showed even small changes. However, three of them said that this makes the *Graph Interface* more challenging to interpret. They suggested that it might help to combine the *Basic Emoji Interface* with the *Graph Interface*. The emoji could then improve the readability and interpretability of the interface during the talk. Additionally, three participants stated that the graph could be more helpful to analyse the talk in retrospect since they had no idea how to react to a low attention level.

4.1.5 Discussion

During this study, we found that although all interfaces were equally distracting, the difficulty to interpret them varied. The *Basic Emoji* and the *Graph* were easy to interpret, but most presenters could not use or process the additional information that the *Multi Emoji Interface* provided. This led to the fact that fluctuations in attention were detected less or not at all. Therefore we argue in favour of H1 and H2. Since the participants stated multiple times that they have no idea how to respond to low attention levels because they were too occupied with the presentation, we argue against H3.

We argue in favor of H1 and H2 but against H3

To reduce the mental load the interface introduces onto the presenter, we reduced the interface for our field study to its essentials. This meant that this interface would only show the attention level and not the prevalent emotions. In addition to that, we chose to combine the biggest advantages of the *Basic Emoji* (readability) and the *Graph* (history) in the new interface as some participants suggested. The new combined interface is shown in Figure 4.5. This interface consists of two parts: The *Basic Emoji* (top) to assess the current audience's attention level at one glance and the *Graph* (bottom) for a more detailed view on the attention level in retrospect.

Combined the *Simple Emoji* with the *Graph* for the field study

4.2 Field Study

We conducted a field study to test our system in an actual lecture and to increase the external validity of the laboratory study's results. The study took place during a 90-minute computer science lecture with around 40 participants. We used the setup described in Chapter 4.1.2. However, we exchanged the still image with the real audience's video feeds and used the refined interface shown in Figure 4.5. We furthermore used the actual attention level of the audience instead of the predefined curves. After the lesson, we interviewed the lecturer on how practical the application was during his presentation:

We conducted a field study to increase the external validity of our findings

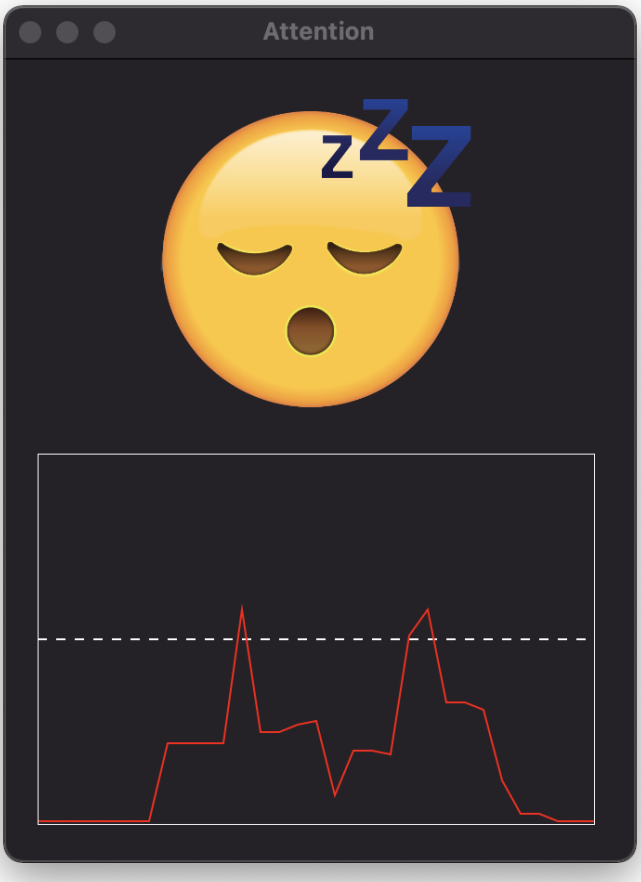

Figure 4.5: This interface combines the fast readability of the *Basic Emoji* (top) and the detailed attention level history of the *Graph* (bottom).

The field study agreed with the results of the lab study

The lecturer stated that in most situations, he had no idea how to raise a low attention level, so he paid little to no attention to the emoji. However, he used the graph to check how certain sections of the lecture were received by the audience. He wanted to use this information to analyse which parts he needs to repeat or change in future lectures. In his opinion, reviewing the graph while preparing the following classes would be much better than just checking it during the lecture. These comments agree with the results of the laboratory study.

Chapter 5

Summary and Future Work

To conclude this thesis, we will summarise our work and our contributions. Then we will present an outlook on possible future work in this field.

5.1 Summary and Contributions

In this thesis, we developed an application that implicitly assesses the audience's attention level during online meetings. To do so, we examined how other researchers measured attention and engagement in lectures or meetings. We found two main approaches: The first one uses special devices like EEGs, the second approach only needs video feeds. The devices that are required by the first approach are often expensive and not widely accessible. Therefore, we decided to work with the second approach. Since the recognition of gestures like head shaking is highly dependent on camera angles and the orientation of the observed person, we used a CNN to measure the person's emotional state. Our application can infer a person's attentiveness from this state as positive emotions improve attention and learning capabilities.

We used a CNN to measure a person's emotional state

We designed three interfaces: Basic Emoji, Multi Emoji and the Graph

We then developed three interfaces to display the gathered information to a presenter: The *Basic Emoji Interface* only shows if the audience's attention level is either high, average or low. The *Multi Emoji Interface* provides additional details on the audience's emotional state and the *Graph Interface* displays the attention level changes of the last 2.5 minutes.

We found that participants could not utilise most information while holding a presentation

These interfaces were then evaluated in a user study to see how they impacted the presenters. We found that the *Multi Emoji Interface* was more challenging to interpret than the other interfaces which made it hard to use the additional information on the predominant emotions of the audience. Furthermore, some participants suggested that the history feature of the *Graph Interface* helps to assess the audience's attention in retrospect. They also said that this retrospective view could help them to plan and improve future lectures, since they had no idea how to react to low attention levels on the spot. Using the findings of this study, we developed an interface that combined the easy readability of the Simple Emoji and the history feature of the Graph.

Our application may be more useful to prepare the next lectures

We tested this interface in an actual lecture. In this small field study, the participant of the field study agreed with a tendency from the user study that lecturers are actually more interested in the retrospective view than in the current attention levels. The lecturer stated that reviewing the Graph while preparing the following lecture would be more helpful than seeing the attention level during the lesson.

5.2 Future Work

In our user study and our field study, we found that most presenters can not change their talk or lecture on the spot to accommodate a low attention level. Therefore, future research should focus on the post-processing of a class. These review applications can show more detailed information without the risk of distracting the presenter since analyzing the lecture is his main task. Future studies could explore which information the presenter needs to improve his classes and the best way to display them. For example, the application could highlight the most confusing slides or which parts of the lecture were perceived as the most interesting.

Explore applications for reviewing lectures retrospectively

Another part of this field is the technical improvement of these applications. More advanced programs could, for example, capture more complex emotions such as confusion. In the best case, these algorithms would be able to process any video stream and no longer depend on camera angles, body position, and lighting conditions.

Make the algorithms more reliable

Appendix A

Study Questionnaire

The following pages contain the complete questionnaire which was used in the user study.

ID:
Gender:
Age:

Basic Emoji Interface

The changes in attention were easily noticeable:

Totally Agree				Totally Disagree
1	2	3	4	5

The display method was easy to interpret:

Totally Agree				Totally Disagree
1	2	3	4	5

The display method was not distracting:

Totally Agree				Totally Disagree
1	2	3	4	5

The given information helped me to infer the attention of the audience

Totally Agree				Totally Disagree
1	2	3	4	5

Multi Emoji Interface

The changes in attention were easily noticeable:

Totally Agree				Totally Disagree
1	2	3	4	5

The display method was easy to interpret:

Totally Agree				Totally Disagree
1	2	3	4	5

The display method was not distracting:

Totally Agree				Totally Disagree
1	2	3	4	5

The given information helped me to infer the attention of the audience

Totally Agree				Totally Disagree
1	2	3	4	5

Graph Interface

The changes in attention were easily noticeable:

Totally Agree				Totally Disagree
1	2	3	4	5

The display method was easy to interpret:

Totally Agree				Totally Disagree
1	2	3	4	5

The display method was not distracting:

Totally Agree				Totally Disagree
1	2	3	4	5

The given information helped me to infer the attention of the audience

Totally Agree				Totally Disagree
1	2	3	4	5

Please rank each method from 1(best) to 2(worst)
(Basic Emoji/Multi Emoji/Graph)

1	2	3

Additional Remarks:

Bibliography

Mostafa Akbari, Georg Böhm, and Ulrik Schroeder. Enabling Communication and Feedback in Mass Lectures. In *2010 10th IEEE International Conference on Advanced Learning Technologies*, pages 254–258. IEEE, 2010.

Mostafa Ataei, Saeid Saffarian Hamedani, and Farshideh Zameni. Effective methods in medical education: from giving lectures to simulation. *Journal of Advanced Pharmacy Education & Research*, 10:37, 2020.

Maryfran Barber and David Njus. Clicker Evolution: Seeking Intelligent Design. *CBE—Life Sciences Education*, 6(1): 1–8, 2007.

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283, 2016.

Ashok R Basawapatna, Alexander Repenning, Kyu Han Koh, and Hilarie Nickerson. The Zones of Proximal Flow: Guiding Students Through a Space of Computational Thinking Skills and Challenges. In *Proceedings of the ninth annual international ACM conference on International computing education research*, pages 67–74, 2013.

Ronald Bassett, Ralph R Behnke, Larry W Carlile, and Jimmie Rogers. The effects of positive and negative audience responses on the autonomic arousal of student speakers. *Southern Journal of Communication*, 38(3):255–261, 1973.

Andrew A Bennett, Emily D Campion, Kathleen R Keeler, and Sheila K Keener. Videoconference fatigue? exploring

changes in fatigue after videoconference meetings during covid-19. *Journal of Applied Psychology*, 106(3):330, 2021.

Serge Brand, Torsten Reimer, and Klaus Opwis. How do we learn in a negative mood? Effects of a negative mood on transfer and learning. *Learning and instruction*, 17(1): 1–16, 2007.

Isabel Buil, Sara Catalán, and Eva Martínez. The influence of flow on learning outcomes: An empirical study on the use of clickers. *British Journal of Educational Technology*, 50 (1):428–439, 2019.

Jane E Caldwell. Clickers in the Large Classroom: Current Research and Best-Practice Tips. *CBE—Life Sciences Education*, 6(1):9–20, 2007.

AT Chamillard. Using a Student Response system in CS1 and CS2. In *Proceedings of the 42nd ACM technical symposium on Computer science education*, pages 299–304, 2011.

Milton Chen. Visualizing the Pulse of a Classroom. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 555–561, 2003.

Xinyue Chen, Si Chen, Xu Wang, and Yun Huang. "I was afraid, but now I enjoy being a streamer!": Understanding the Challenges and Prospects of Using Live Video Streaming for Online Education. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–32, 2021.

Mihaly Csikszentmihalyi and Reed Larson. *Flow and the Foundations of Positive Psychology*, volume 10. Springer, 2014.

Murray S Davis. Beyond Boredom and Anxiety: The Experience of Play in Work and Games, 1977.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

Paul Ekman and Harriet Oster. FACIAL EXPRESSIONS of EMOTION. *Annual review of psychology*, 30(1):527–554, 1979.

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*, pages 5114–5119, 2017.

Mariam Hassib, Stefan Schneegass, Niels Henze, Albrecht Schmidt, and Florian Alt. A Design Space for Audience Sensing and Feedback Systems. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

Apple Inc. Vndetectfacelandmarksrequest Apple developer documentation. `https://developer.apple.com/documentation/vision/vndetectfacelandmarksrequest`, 2021a. Accessed: 25.11.2021.

Apple Inc. Vndetectrectanglesrequest Apple developer documentation. `https://developer.apple.com/documentation/vision/vndetectrectanglesrequest`, 2021b. Accessed: 25.11.2021.

Apple Inc. Vnrecognizetextrequest Apple developer documentation. `https://developer.apple.com/documentation/vision/vnrecognizetextrequest`, 2021c. Accessed: 25.11.2021.

Ashish Kapoor and Rosalind W Picard. A Real-Time Head Nod and Shake Detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5, 2001.

Vladimir I Levenshtein et al. BINARY CODES CAPABLE OF CORRECTING DELETIONS, INSERTIONS, AND REVERSALS. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

Peter D MacIntyre, Kimly A Thivierge, and J Renée MacDonald. The effects of audience interest, responsiveness, and evaluation on public speaking anxiety and related variables. *Communication research reports*, 14(2):157–168, 1997.

Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. AffectiveSpotlight: Facilitating the Communication of Affective Responses from Audience Members during Online Presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

Ruby T Nadler, Rahel Rabi, and John Paul Minda. Better Mood and Better Performance: Learning Rule-Described Categories Is Enhanced by Positive Mood. *Psychological Science*, 21(12):1770–1776, 2010.

Joshua Newnham. *Machine Learning with Core ML*. Packt Publishing Ltd., 2018.

Rosalind W Picard and Jocelyn Scheirer. The Galvactivator: A glove that senses and communicates skin conductivity. In *Proceedings 9th Int. Conf. on HCI*, 2001.

Alan T Pope, Edward H Bogart, and Debbie S Bartolome. Biocybernetic System Evaluates Indices of Operator Engagement in Automated Task. *Biological psychology*, 40 (1-2):187–195, 1995.

J Poulis, C Massen, E Robens, and M Gilbert. Physics lecturing with audience paced feedback. *American Journal of Physics*, 66(5):439–441, 1998.

Verónica Rivera-Pelayo, Johannes Munk, Valentin Zacharias, and Simone Braun. Live Interest Meter: Learning from Quantified Feedback in Mass Lectures. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pages 23–27, 2013.

Jonathan Smallwood, Annamay Fitzgerald, Lynden K Miles, and Louise H Phillips. Shifting Moods, Wandering Minds: Negative Moods Lead the Mind to Wander. *Emotion*, 9(2):271, 2009.

Wei Sun, Yunzhi Li, Feng Tian, Xiangmin Fan, and Hongan Wang. How Presenters Perceive and React to Audience Flow Prediction In-situ: An Explorative Study of Live Online Lectures. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–19, 2019.

Jaime Teevan, Daniel Liebling, Ann Paradiso, Carlos Garcia Jurado Suarez, Curtis von Veh, and Darren Gehring.

Displaying Mobile Feedback during a Presentation. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 379–382, 2012.

David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

Matin Yarmand, Jaemarie Solyst, Scott Klemmer, and Nadir Weibel. "It Feels Like I am Talking into a Void": Understanding Interaction Gaps in Synchronous Online Classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2021.

Index

Typeset January 21, 2022