Applications of Ubiquitous and Interactive Systems in Smart Education

Thesis Submitted by

Pragma Kar

Doctor of Philosophy (Engineering)

Department of Information Technology
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata, India
2023

Applications of Ubiquitous and Interactive Systems in Smart Education

by

Pragma Kar
Registration Number: 1021809002

Thesis submitted for the

Doctor of Philosophy (Engineering)
Degree of Jadavpur University, Kolkata, India

Supervisors:

Prof. Samiran Chattapadhyay
Professor
Dept. of Information Technology
Jadavpur University, Salt Lake Campus
Kolkata-700106
West Bengal
India

Dr. Sandip Chakraborty
Associate Professor
Dept. of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur-721302
West Bengal
India

2023

Jadavpur University
Kolkata 700 032, India

INDEX NO. 227/18/E

Title of the thesis :

Applications of Ubiquitous and Interactive Systems in Smart Education

Name, Designation and Institution of the Supervisors:

Prof. Samiran Chattapadhyay
Professor
Department of Information Technology
Jadavpur University, Salt Lake Campus
Kolkata-700106
West Bengal
India

Dr. Sandip Chakraborty
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur-721302
West Bengal
India

List of Publications

Journal papers

1. **Pragma Kar**, Soumya Banerjee, Sandip Chakraborty, Matangini Chattopadhyay. AutoNotes: A Touch-Free Blink-Based Interactive Model for Generation of Notes from Lecture Videos. Journal of The Institution of Engineers (India): Series B. 2021. DOI:https://doi.org/10.1007/s40031-021-00550-4.

2. **Pragma Kar**, Samiran Chattopadhyay, Sandip Chakraborty. 2020. Gestatten: Estimation of User's Attention in Mobile MOOCs From Eye Gaze and Gaze Gesture Tracking. PACM on Human-Computer Interaction, Volume 4, ACM Press, pp 1–32 DOI: https://doi.org/10.1145/3394974

International conference papers

1. **Pragma Kar**, Shyamvanshikumar Singh, Avijit Mandal, Samiran Chattopadhyay, and Sandip Chakraborty. 2023. ExpresSense: Exploring a Standalone Smartphone to Sense Engagement of Users from Facial Expressions Using Acoustic Sensing. ACM CHI Conference on Human Factors in Computing Systems (CHI 2023).

2. **Pragma Kar**, Samiran Chattopadhyay, and Sandip Chakraborty. 2022. Bifurcating Cognitive Attention from Visual Concentration: Utilizing Cooperative Audiovisual Sensing for Demarcating Inattentive Online Meeting Participants. Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 498 (November 2022), 34 pages. DOI:https://doi.org/10.1145/3555656

3. **Pragma Kar**, Krishna Mishra, Sudipro Ghosh, Sandip Chakraborty, and Samiran Chattopadhyay. 2021. Nosype: A Novel Nose-tip Tracking-based Text Entry System for Smartphone Users with Clinical Disabilities for Touch-based Typing. In Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction (MobileHCI '21). Association for Computing Machinery, New York, NY, USA, Article 26, 1-16. DOI: https://doi.org/10.1145/3447526.3472054

Others

1. **Pragma Kar**, Krishna Mishra, Sudipro Ghosh, Sandip Chakraborty and Samiran Chattopadhyay. 2021. Exploratory Analysis of Nose-gesture for Smartphone Aided Typing for Users with Clinical Conditions. 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Aliated Events (PerCom Workshops), 2021, pp. 380-383, doi: 10.1109/PerComWorkshops51409.2021.9430933.

2. Debasree Das, **Pragma Kar**, Sugandh Pargal and Sandip Chakraborty. FreeSteer: A Smartphone Application for Detecting Anxiety in Novice Drivers through Smart Glasses. 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2023, pp. 427-431,
DOI: 10.1109COMSNETS56262.2023.10041299.

3. Vijay Kumar Singh, **Pragma Kar**, Ayush Madhan Sohini, Madhav Rangaiah, Sandip Chakraborty and Mukulika Maity. Monitoring Engagement in Online Classes Through WiFi CSI. 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2023, pp. 462-465, DOI: 10.1109 COMSNETS56262.2023.10041341.

List of Patents: Nil

List of Presentations in National/International/ Conferences/ Workshops:

1. **Pragma Kar**, Samiran Chattopadhyay, and Sandip Chakraborty. 2022. Bifurcating Cognitive Attention from Visual Concentration: Utilizing Cooperative Audiovisual Sensing for Demarcating Inattentive Online Meeting Participants. Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 498 (November 2022), 34 pages.
   DOI:https://doi.org/10.1145/3555656

2. **Pragma Kar**, Krishna Mishra, Sudipro Ghosh, Sandip Chakraborty, and Samiran Chattopadhyay. 2021. Nosype: A Novel Nose-tip Tracking-based Text Entry System for Smartphone Users with Clinical Disabilities for Touch-based Typing. In Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction (MobileHCI '21). Association for Computing Machinery, New York, NY, USA, Article 26, 1-16.
   DOI: https://doi.org/10.1145/3447526.3472054

3. Debasree Das, **Pragma Kar**, Sugandh Pargal and Sandip Chakraborty. FreeSteer: A Smartphone Application for Detecting Anxiety in Novice Drivers through Smart Glasses. 2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 2023, pp. 427-431.
   DOI: 10.1109COMSNETS56262.2023.10041299.

PROFORMA – I

"Statement of Originality"

I, **Pragma Kar**, registered on **16/04/2018** do hereby declare that this thesis entitled "**Applications of Ubiquitous and Interactive Systems in Smart Education**" contains literature survey and original research work done by the undersigned candidate as part of Doctoral studies.

All information in this thesis have been obtained and presented in accordance with existing academic rules and ethical conduct. I declare that, as required by these rules and conduct, I have fully cited and referred all materials and results that are not original to this work.

I also declare that I have checked this thesis as per the "Policy on Anti Plagiarism, Jadavpur University, 2019", and the level of similarity as checked by iThenticate software is **1%**.

Signature of Candidate:

Pragma Kar

(Pragma Kar)
Date : 15/03/2023

Certified by Supervisors:
(Signature with date, seal)

Samiran Chattopadhyay 15/03/2023

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

1._____
(Samiran Chattapadhyay)

Sandip 15/03/2023

Dr. Sandip Chakraborty
सह. अध्यापक / Associate Professor
कंप्यूटर विज्ञान एवं अभियांत्रिकी विभाग
Computer Sc. & Engg. Deptt.
भा प्रौ सं खड़गपुर/IIT Kharagpur

2._____
(Sandip Chakraborty)

PROFORMA – 2

"CERTIFICATE FROM THE SUPERVISORS"

This is to certify that the thesis entitled "**Applications of Ubiquitous and Interactive Systems in Smart Education**" submitted by **Ms. Pragma Kar**, who got her name registered on **16/04/2018** for the award of Ph.D. (Engg.) degree of Jadavpur University is absolutely based upon her own work under the supervision of **Prof. Samiran Chattapadhyay, Department of Information Technology, Jadavpur University, Kolkata** and **Dr. Sandip Chakraborty, Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur** and that neither her thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

Signatures of the Supervisors with Date and Official Seal

Samiran Chattopadhyay
15/03/2023

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

Prof. Samiran Chattopadhyay
Professor,
Department of Information Technology
Jadavpur University,
Block– LB, Plot– 8, Sector–3,
Salt Lake, Kolkata 700106, India

Sandip
15/03/2023

Dr. Sandip Chakraborty
सह. प्राध्यापक /Associate Professor
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Computer Sc. & Engg. Deptt.
भा. प्रौ. सं. खड़गपुर/IIT Kharagpur

Dr. Sandip Chakraborty
Associate Professor,
Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur,
Kharagpur 721302, India

Acknowledgements

My doctoral journey has offered me an enriching experience that involves persistence, commitment, and determination. It has been nothing short of an adventure that taught me to overcome the fear of “troughs" and stay humble at “crests". The path, indeed, was difficult. However, the company of a few people made it pleasantly memorable. I acknowledge the contribution of everyone who has been a part of my journey.

Firstly, I would like to express my gratitude to my supervisors -Dr. Samiran Chattopadhyay and Dr. Sandip Chakraborty. Not only did they guide me in this journey, but have also believed in me at every step. They helped me select my research domain, allowed me to explore innovative ideas freely, and yet, steered my works in the ideal direction. Dr. Chattopadhyay has ensured that my research progress is never affected by any official or financial dilemma. His expertise and guidance have not only facilitated my academic growth but have also expedited my personal growth. Dr. Chakraborty has taught me the true meaning of diligence. His indubitable efforts towards refining our research, enormous patience with my unceasing queries, and optimistic outlook towards me have inspired me throughout my doctoral journey. He has also provided me access to all the resources required for my work. Both my supervisors have always encouraged me to take up any possible opportunity that is beneficial to my career, and I consider myself fortunate to be able to work under their supervision.

I am grateful to my Research Advisory Committee members for their judicious feedback that helped me shape my work. I thank the former heads of the Department of Information Technology, Dr. Bhaskar Sardar and Dr. Uttam Kumar Roy, for their support. I would thank Dr. Parama Bhaumik, the present Head of the Department, for ensuring an overall smooth research experience. I thank all the departmental faculties and administrative staff for their timely assistance.

I am obliged to the Department of Science & Technology for awarding me the Innovation in Science Pursuit for Inspired Research (INSPIRE) fellowship. I am also thankful to the Gary Marsden Travel Award Committee, Microsoft Travel Grant Committee, COMSNETS Association and LRN Foundation for selecting me for the travel grant awards to support my attendance in International conferences.

I would also like to thank all my colleagues and peers with whom I could discuss my research

ideas and get insightful comments. I would particularly thank Dr. Soumyajit Chatterjee for sharing his expertise in various topics and helping me resolve different technical problems during my work. He has not only been kind enough to provide constructive feedback on my research works but has generously shared his experiences that helped me in research-oriented decision-making. I would also like to thank Dr. Soumya Banerjee, with whom I have shared some quality time and learned a lot during the early days of my doctoral journey. He has always motivated me with appreciative comments and critical questions that helped me improve my work. I am also thankful to Dr. Snigdha Das for all her help and support. I acknowledge the participants and annotators who took part in the human studies. Without their cooperation, this research would not have been possible. I am also grateful to all my collaborators for their valuable input.

I thank my friends and family for their emotional support. Specifically, I thank Antara Bhattacharya, Namrata Ghosh, Debdatta Basu, Gaurab Kumar Dey, Tina Das Chowdhury, and Bidisha Piplai for their constant company during my doctoral journey. I am thankful to Saurav Bhattacharyya, who has been there from the beginning of this journey. He has been a constant source of encouragement for me. His patience and positive nature have always induced strength in me. I am forever indebted to my parents for their unparalleled love and sacrifices in every aspect of my life. I am grateful to my father for being there for me as a friend and supporting me in all the decisions of my life. I am thankful to my mother, whose smile and cheerfulness have provided the strength to face various challenges in life.

Pragma Kar

To
my Father and Mother.

“In the same way that the internet, our mobile phones, medical imaging, satellite navigation and social networks would have been incomprehensible to the society of only a few generations ago, our future world will be equally transformed in ways we are only beginning to conceive. ***Information on its own will not take us there, but its intelligent and creative use will.****"*

– Stephen Hawking, Brief Answers to the Big Questions.

Abstract

The technological advancements in Human-Computer Interaction (HCI) have been facilitated through the evolution of novel solutions promoting smart living. While the contribution of various specialized branches of HCI has been noteworthy, the commensurable increase in the demand for assistive technologies in industry, education, entertainment, communication, and others necessitates further innovations.

Among other domains, academia has experienced an unparalleled transformation during and after the COVID 19 pandemic where the in situ meetings and classrooms changed to virtual meetings and decentralized classrooms respectively. Massive Open Online Courses (MOOC), virtual classrooms and online lectures eventually and partially replaced the traditional classrooms. Besides the numerous advantages of online classrooms like self-paced learning, comfortable environment, resource accessibility, effective visuals, there emerged more convoluted challenges than interrupting connection drops. Online classrooms lacked communication amongst the students. Moreover, in such a setup, manually monitoring the attention of each student, and hence their engagement in the class, became extremely difficult, if not impossible for the faculties. These reasons, amongst others, made both the delivery and absorption of course contents, challenging in virtual classrooms. For pre-recorded MOOC too, attention deficit has led to high dropout rates, thus compromising the overall quality of education. Similarly, for any type of online meetings, lack of attention among the participants adversely affects the meeting's quality. Automated estimation of attention can be used in both personalised assessment or forecasting systems for the learners, and virtual response or feedback system for the educators.

Attention and overall engagement estimation of a person is not only essential in online classes, but is also important in pre-recorded online videos, such as an educational video, or other genres of videos. Here, the challenge lies in obtaining a fair and unbiased estimation of how engaging a media content is. Even though applications like YouTube, provides an option for leaving feedback, the reviews can be highly biased, non-real-time, and sometimes irrelevant, thus providing an unfair estimation of the video quality. In this domain, engagement of a person can be treated as an unbiased feedback to a media content. Estimating how engaging a video, audio, or an article is, can aid in various trailing purposes like improved contents by particular creators, relevant suggestions to an individual audience, fine-grained auto-feedback

and so on. Attention, being a mental process, should be estimated through ubiquitous solutions in a non-intrusive manner.

While “ambient" assistive systems can analyze cognitive processes, HCI also suffers from open challenges surrounding the explicit“interactivity" of a user and the device. On identifying these intricate yet crucial problems, inclusive and globally feasible solutions for all users, including those with clinical disabilities, needs to be designed. Pertaining to the social distancing protocol induced by the recent pandemic, the exigency of touch-free interactive systems came into being. On one hand, in educational domain, MOOC videos frequently needs to be controlled through manual interaction with the device like mouse or button clicks. This becomes proportionately difficult when a learner needs to simultaneously make manual notes of key points taught in the video. On the other hand, for users with clinical conditions like Dactylitis, Sarcopenia, Essential tremor and so on, basic touch-based writing with the device, such as keyboard based text-entry, becomes difficult. These contributes to the exploration of novel modalities in developing HCI models–both ubiquitous and interactive systems, that eliminates the requirement of physical contact between the user and the device.

In this thesis, we first address the challenging problem of automated attention estimation in online meetings, online classes, MOOCs and different YouTube videos through the development of *ubiquitous assistive systems*. In doing so, we classify the various types of attention and explore the corresponding modality of the users that can attribute to each type. While exploring these modalities at each level of attention, we aim at overcoming the challenges of the previous level. **Firstly**, we explore gaze gesture as an indicator of visual attention and high level cognition in MOOC. **Secondly**, we explore video-based facial expressions, vocal emotion, speech intent, head gesture and ambient light reflection as an indicator of cognition and multitasking in online meetings. **Thirdly**, we estimate a user’s engagement to a video, by analyzing their facial expressions, estimated through acoustic sensing.

We then extend the utility of these modalities for developing *interactive assistive systems* facilitating seamless touch-free interactions between different users and devices.
Firstly, we explore blink as the sole modality for controlling MOOC videos and automatic generation of notes from them. **Secondly**, we design and develop a nose-tip gesture-based writing system that operates on smartphones and can help the learners with clinical conditions.

To summarize, we identify and address two key challenges of smart education–(i) The requirement for automated attention estimation during online classes, videos or meetings. The group of lightweight solutions addressing this problem involve extensive studies on attention and its variations, exploration of modalities like gaze, expressions and others for the estimation of attention and the development of novel systems involving video procession and acoustic sensing. (ii) The existing challenges of traditional touch-based interactive systems and the requirement for touch-free HCI models. In this group of solutions, we explore blink and nose-tip gesture for facilitating MOOC video controls along with automatic notes generation and

touch-free writing in smartphones for the users with clinical conditions respectively.

Keywords— MOOC, HCI, Smartphones, Gaze Gesture, Region of Gaze, Video Conferences, Attention Estimation, Multitasking, Acoustic Sensing, Expression Detection, Blink based Interaction, Notes Generation, Contact-less Text-entry, Nose Tracking, Nose-tip Projection; Sensor-based Editing; Auto-complete Suggestion, In-device Computation.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Education, being one of the most essential and basic rights of human beings, has been well defined in terms of its purposes, forms, sources and significance. While the formal definition of education has been a debatable topic among Philosophers, they all converge to the common notion of *"knowledge"*; a concept that has been widely studied in Epistemology[1], and *"learning"*; a process that focuses on gaining knowledge. In the sphere of education, the transfer of knowledge between two or more people can be uni- or bi-directional and can be mediated in a formal classroom setup through structured courses, or in an informal setup through unstructured activities like reading books, watching videos and so on. The mode of education has evolved from centralised physical classrooms to decentralized global classrooms with the assistance of advanced technologies. Correspondingly, the source of education has expanded with the inclusion of e-books, articles, videos and other multi-media contents on websites and other online platforms. Overall, education induces the learning of necessary knowledge and practical skills that contributes to ones pragmatic approaches, emotional balance and logical decision-making skills in life.

Traditionally, under formal setup, in-situ education has always been prevalent globally where learners have visited classrooms during a lecture and read hard-copies of a books to gather knowledge. However, with the unforeseen advent of the COVID-19 pandemic, Academia experienced a major transformation when a significant proportion of educators and learners started opting for the online mode of education. Moreover, most of the in-person academic

[1] https://en.wikipedia.org/wiki/Epistemology (Accessed: Friday 11th August, 2023)

conferences, seminars, meetings, etc., were declared online worldwide during this time. The social distancing protocol indeed played a key-role in initiating this shift, but what further catalyzed it was the number of associated benefits. These benefits included, but were not limited to, a reduction in travel cost, extensive choices in terms of timing and courses, self-paced learning, and so on. As a result, this transformation majorly paced up the adaptation of *smart education* [131], which was already emerging gradually and steadily.

Since smart education involves assistive systems that promote the teaching-learning experience, its domain [226] is highly faceted. This technology revolves around *environment* [223][75] (cloud-based, local, hybrid), *portability* (mobile or web applications for smartphones, laptops, etc.), *data for predictions and recommendations* (Big data, public data sets, in-house data sets), *multimedia content* (video, audio, text, image), privacy issues (information tracking, contextual sensing etc.), cognitive impacts (distractions, engagement, etc.), hardware (sensors, trackers, etc.) and so on. In the light of this discussion, it is essential to understand that smart classroom does not merely refer to online classes. Rather, it can also include a closed physical classroom [277], augmented by sensors and other hardware like camera, trackers, controllers etc. that can facilitate digital content creation, delivery and distribution along with the enhancement of learner's cognition through interaction and attention estimation. Furthermore, a smart classroom can also have features for automating the process of feedback generation through sensing, smart evaluation techniques and so on. However, one of the principal features of smart education is its inclusiveness. Smart education aims at maximizing the number of learners across the globe, by making it reachable to everyone [204]. This is infeasible in a physical classroom setup due to the limited capacity of the room to accommodate a very large number of students. Thus, digital classrooms, that can reach beyond geographical, racial, and cultural barriers, comprise a major part of smart education. Pertaining to the global shift in the mode of education, along with the aim of accessibility for all, we focus on the remote mode of smart education in this thesis and will use the term smart education to imply the online mode of education in the rest of the thesis. Moreover, we essentially focus on the theoretical aspect of smart education as considering both theoretical and practical (e.g. skill development through practical training, laboratory-based experiments etc.) learning is beyond the scope of this thesis.

Based on the difference between the time of content generation and access, smart education can be divided into two broad categories [204]: *synchronous* (also known as Distance Learning) and *asynchronous* (also termed as e-Learning). While earlier researches [166][242][46] presented divergent explanations of these concepts, and either ambiguously or unambiguously used them with terminologies like online learning, web-based learning, and virtual learning, we adhere to a more recent and simplified categorization of these terms. In Distance learning, the two remote parties interact in real time i.e., the time at which a teacher conducts a course is same as the time at which the learners attend the course. A good example of such a system is an online video conference in which a presenter (teachers/employees) presents a topic and the

attendees (learners/employees) attend to it. The presenter can use additional smart devices like smart cameras, interactive whiteboards, and so on. On the other hand, in e-learning the content delivery is not real time. In this mode, the presenter creates a digital content, such as a recorded video, notes, etc. which is stored and distributed later, according to the learner's convenience. Massive Open Online Courses (MOOC)s, where the lectures are pre-recorded, is an example of e-learning. Both these types of smart education have their associated shortcomings which are divergent in nature and requires novel solutions. For example, in Distance learning, improvements in video quality, augmented reality techniques for promoting the virtual experience are some of the areas in which research contributions are being made. Whereas, in e-Learning, features like summarizing lectures automatically, relevant recommendation generation, etc. are being developed. In this thesis, we identified two common challenges associated with both these modes– (1) ***Ubiquitous Assessment of Attention*** and (2) ***Improved interactivity between users and devices*** for seamless teaching and learning. In the following sub-sections, the background of and motivation behind recognising these challenges are discussed, followed by a detailed discussion of the research objectives and contributions.

1.1 Background and Motivation

This section presents the fundamentals of human attention and human-computer interactivity in the context of smart education. The basic concepts then further leads to the discussion of existing challenges in the corresponding domains and how they motivate the development of novel assistive systems.

1.1.1 Ubiquitous Assessment of Learner's Attention

Since learning is the constitutive core of any form and mode of education, identifying the parameters that promote learning is necessary. Cognitive involvement is indeed the prime factor in learning. While such involvement is often generalised as *"Attention"* or *"Engagement"*, the concept is much elaborate. Although psychological studies have extensively explained the aspects and classes of attention, the technique of understating whether a learner is attentive is a demanding task in itself. In case of online mode of education where continuous monitoring is impracticable, understanding the learner's cognitive response is even more difficult. To resolve this challenge, a system, must detect the learner's attention automatically and transparently with the sparse availability of learner's information. However, the development of such systems first requires a thorough exploration of human attention which is presented in the following section.

Dissecting Human Attention

"Everyone knows what attention is. It is the taking possession of the mind in clear and vivid form of one out of what seem several simultaneous objects or trains of thought."– William James, The Principles of Psychology [91].

In real world, the environment contains various forms of information at any given point of time. Attention is the mental process by which a person focuses on a particular information from this information pool at a given instance. Posner [194] defined attention as the ability of a person to be alert, select and process a certain information. Evidently, attention involves the working memory of human beings and thus, is characterised by the limited availability of resources for storing and processing information [177]. The information that we attend to, at a give point, is a function of the top-down control and bottom-up salience [106]. While top-down control involves the voluntary allotment of attention based on the current context and objective of the person, the high salience of a certain information can automatically, give it access to the working memory by using the bottom-up salience filters. For example, ideally during an online class, at any given instance, a learner's attention should be fixated to the course content, i.e. the information being delivered by the teacher. While the objective of the student should be to attend to this information explicitly, in practice, distractions like someone calling their name in the proximity will cause a shift of their attention due to its higher salience, causing a higher neural response. Pertaining to this understanding, any automated attention detection system should allow and account for temporal context switching.

While inferring upon attentiveness through an automated system, we first need to analyze the meaning of inattentiveness. Since human brain allows the shift of attention from one information to another in a continuous process, inattentiveness is a relative and contextual concept. In terms of the objects (or information) we attend to, attention can be classified into several branches. To understand the definitive difference between attentiveness and inattentiveness, we consider the classifications of attention only in the context of education [102]. Next, we discuss these sub-categories of attention and explore how, an automated system can infer upon them using various data modalities.

- Visibility-based: In the first categorization of attention, it is segmented into two classes, based on its visibility. In an online classroom, a learner can perform activities like reading a slide, ask a question, involve themselves in active problem solving, and so on, to promote attentiveness. These visible activities refer to *external attention*. Moreover, if a learner thinks about the topic being taught, correlated it with some previously learnt concepts, think about an answer to a question asked in the course, etc., these mental activities can also promote the overall attention of the learner to the course. However, these processes will be termed as *internal attention*, as they are not immediately visible to the observer. Similarly, *external inattentiveness* would refer to the instances when the learner plays with

their smartphone, eats during the class, opens another tab and reads irrelevant articles, and so on. *Internal inattentiveness* would refer to processes like mind wandering [21] where the learner thinks of some topic, other than the one being taught. In this regard, it is essential to note that complete absence of any activity like "staring blankly at the screen without any active communication" can also indicate internal inattentiveness.

In case of external (in)attentiveness caused by activities like "visually following a content" or "looking at the phone", gaze direction and patterns can reveal the insights of a person's cognitive direction. Hence, visual attention tracking has been widely employed in various works through the use of commercial eye trackers [93][139], and head orientation trackers. Eye-tracking techniques can work significantly well in case of *overt visual attention* [193] which is characterised by precise yet observable shift of gaze towards the object of focus. However, for the mental process of *covert attention* [83], where attention is not indicated by gaze direction, eye tracking methods cannot work with significant accuracy. Moreover, merely relying on gaze region for inferring the location of attention can be misleading due to the instances of inattentional blindness [145] caused by processes like blankly staring at the screen without actually being *aware* [119] of the content. In the lights of this discussion, it is evident that the automated estimation of internal (in)attentiveness would involve other physiological and physical features like EEG [163], Pupillometry [278] and so on. Such techniques are widely discussed in Chapter 2.

- Time-based: Attention can also be perceived as *sustained* [57], *alternating* [39] or *divided* [232]. In sustained attention, the person pays attention to a particular task for a prolonged duration of time. This type of attention is often characterised by a decreasing vigilance[2] with increasing time span. This is trivial as continued focus is likely to be affected by mind wandering. Nevertheless, it is highly dependant on the required effort to stay focused on the stimuli, the reward related to the attended task, individual's motivation, presentation of the attended information and so on. For example, an online learner will show higher level of sustained attention if the course content is easy to follow or is followed by a certification process. Thus, in online mode of education, sustained attention becomes crucial to estimated from both the student's perspective for personalised recommendations and faculty's perspective for automated feedback. Furthermore, sustained attention can also ensure that the learner is not only paying attention to the content but is also *engaged*[3] to the content, such that they are absorbing the information being delivered. In case of alternating attention, a subject switches their attention between two or more stimuli alternatively. While some studies validated divided attention [232] as a distinct type of attention where a person pays attention to more than one information simultaneously, its

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2865224/ (accessed: Friday 11th August, 2023)

[3] https://speakingaboutpresenting.com/content/attention-to-engagement/ (accessed: Friday 11th August, 2023)

existence is well debated with the argument that divided attention is only a rapid case of alternating attention. Commonly, this type of attention is termed as multitasking.

- Relevance-based: Attention is a continuous process i.e., a learner can either pay attention to the content or be inattentive. However, the term “inattentive" is only relative as, when the learner is not paying immediate attention to the course, they are paying attention to some other information (audio/visual/mental) which is irrelevant to the course. Thus, a more detailed analysis of attention would rather classify it as either *on-topic* attention or *off-topic* attention. However, for simplicity, we will use the term “inattentive" instead of off-topic attention.

Applications of Automated Estimation of Human Attention

On exploring the categories of human attention, it is now essential to realise the necessity of automating its estimation in online classes. The development of such systems are necessitated by and can be applicable to the following domains.

1. Several reports have established the fact that online courses experience high drop out rates and have explored the root causes [4]. Other than difficulties related to technology, connectivity and experience, the most evident cause has been found to be the lack of motivation. This, in turn, is caused by the scope of distractions and inattentiveness in online courses, due to factors like lack of monitoring, feeling of seclusion, low interaction and unfair feedback received by the faculties. Such high dropout rates can be controlled by facilitating automated attention detection in online education.

2. Automated attention estimation can also lead to personalised recommendations for the participants of the online courses or meetings. For example, a learner who is detected to suffer from attention deficiency, can receive more relevant and basic courses than those wit higher attention. Moreover, such systems can be used to forecast grades of learners and generate helpful warnings and schedules. Real time detection of attention can not only prompt the educators to direct their focus to and address the inattentive learners, these personalised predictions can also be transmitted to the educators to arrange the course in an optimal way so that the pace of content delivery can be matched to the learner’s capabilities.

3. Apart from the several branching applications of automated attention estimation, the core necessity lies in terms of promoting the quality of online conferences, lectures, presentations and other related activities. By understanding the attention distribution

[4] https://www.linkedin.com/pulse/so-hidden-problems-dropout-rates-online-learning-borg%C3%BE%C3%B3r-%C3%A1sgeirsson/ (accessed: Friday 11th August, 2023)

among the participants, influential speakers and sought-after topics can be detected. By focusing on such aspects, the overall quality of the online activity can be improved.

Even though Psychological studies have identified the extensive categories of human attention, the early and automated detection of attention through with the help of modern technology is still nascent. Automated estimation of attention, irrespective of its sub types, experiences some inherent and non-trivial challenges that needs to be addressed while selecting the data modality and developing the system. These challenges are discussed in the following subsection.

Challenges of Attention Estimation

To understand whether a learner is paying attention in a physical classroom is a challenging task as it requires manual monitoring of individuals, often infeasible in a classroom of large size. However, the body gestures and activities of the students can reveal much about their attentiveness. This challenge gets exponentially more difficult to address in case of online education due to the following reasons:

1. Ubiquity : The fundamental, indirect and final objective of attention estimation is to ensure that the online participants are as attentive as they would be in any physical classroom setup. Thus, its automation should be designed in a way that it does not cause distractions to the participants. Not only does the system needs to be pervasive or non-intrusive, it also needs to be ubiquitous so that the users do not have to explicitly generate commands to the system. It is particularly challenging to develop a completely transparent system that runs in the background, while the participant attends the online event, as any minor disruption would cause a shift of cognitive focus, thus violating the very purpose of the whole model.

2. Usability : The next challenge lies in developing a system that is globally usable. The usability can depend on two aspects: the ease of adopting the technology and device / hardware availability. While the first can be addressed through sufficient training, the later needs to be addressed by minimising the requirements of additional hardware. The devices available globally should be used as the target platforms for attention estimation models.

3. Security and Privacy : In contrast to the challenge of ensuring ubiquity, it is also essential that the users feel comfortable using the system. Since the system is supposed to work in he background, the users might feel dubious about sharing data. To ensure user's security, the system must not capture contextual information that could pose thread to the user. Moreover, remote processing should be limited to prevent the scope of data leakage.

4. Processing : In continuation to the above challenge, the correct balance between processing speed and accuracy needs to selected. With the limited processing capability of terminal devices at the client's side, especially if its a smartphone, in-device processing becomes difficult and requires lightweight, yet accurate systems.

5. Data : Other than the generic challenge of data availability, selection of data modality is crucial and arduous. In case of online education, only a limited region (facial region) can be captured through the device (assuming the general tendency of sitting in front of the device to attend a course or holding the smartphone near the facial region to view the screen). Due to the occlusion of body gestures, the estimation of attention fully depends on the analysis of facial attributes, if additional hardware/ sensors/ devices are to be eliminated. This, not only makes the selection of the relevant data modalities limited and challenging, but also demands a thorough analysis of the same. Only through the exploration of such modalities, the research can be directed to the difficult task of quantifying mental processes like attention and engagement.

1.1.2 Improved HCI in Smart Education

Now, that the background and details of the first research direction; automated attention estimation in online education; has been discussed, we discuss the other distinct, yet relatable research direction for promoting smart education through novel interactive applications. Majority of the recent works aim at developing smart devices like interactive tables and whiteboards, smart chairs, etc. for in-situ mode of education or interactive smartphone applications for stress management, finger pose detection, touch-based transfer of contents between laptop and smartphones, etc. that can be used in the context of online education. However, we identified a considerable research gap between online education and touch-free interactivity between the users with the devices. To justify the selection of this particular problem, we need to understand why touch-free assistive systems need to be developed in the context of online education. The necessity can be analyzed from two different angles:

- Firstly, the recent COVID-19 pandemic caused the emergence of social distancing protocols and placed a restriction on touching surfaces in public places. Pertaining to the scope of mobility (portability) of online classes, a learner might decide to take a course in an outdoor environment, while travelling. This necessitates novel touch-free interactive methods that would limit the requirement to touch the device.

- Secondly, the development of interactive systems should account for inclusivity of target users. Ubiquitous systems for automated attention estimation does not require explicit commands and hence, can be used by almost everyone. However, HCI models that require physical interaction with the device, like mouse clicks to play-pause a lecture

video on a laptop/desktop, finger-touch for writing a note on smartphones, are difficult to use for users with clinical disabilities like Dactylitis, Sarcopenia, Essential Tremor, Quadriplegia, etc. To expand the usability of the interactive assistive systems in any field, especially in Academia, touch-free systems should be modelled so that they can be used by all users, including those with clinical challenges.

Challenges and Scope of Innovative Interactive System

We now discuss some of the generic challenges associated with traditional touch-based HCI, scope of visualizing novel solutions and the corresponding points of concern to be addressed. The traditional methods of touch-based user-device interaction comes with some inherent challenges like mobility issues. For example, a person travelling in a public transport, like a bus, might find it difficult to type a text with one hand. Similarly, a pedestrian holding a bag in one hand will not be comfortable using a phone to write something, even if it is urgent. Voice commands can be an alternate solution, however,it is not recommendable in public places where the texts can be overheard by others. In an indoor scenario, a learner might be busy taking a note or solving some in-course exercise during which, they might find it difficult to pause and play the videos frequently through mouse clicks. Such shortcomings can be addressed by developing touch-free novel interactive methods.

While modelling such novel systems, feasibility analysis becomes the prime challenge. The choice of input medium should be carefully selected so that the system can be used by a large community. However, it is almost impossible to select one such medium that would be suitable for all user types. For example, a text entry system using gaze would not be usable for people with visual impairments. A system using hand motion for controlling the User Interface will not be usable for users with paralysis and so on. However, the choice of modality should be such that the system can be used by a significant proportion of the global population. Feasibility of the system can also be analysed from the hardware's perspective. As discussed earlier, we eliminate the requirement of any additional hardware, to address this challenge. While the feasibility of a system can be analysed by its usability, another crucial component is its adaptability. Although traditional systems, e are not free from their corresponding drawback, they prevail globally. New interactive methods can be difficult to learn for some users. To address this challenge, significant user training is required to understand the true contributions of the novel systems.

In the view of these discussions, the research objectives of the thesis are enlisted next, followed by the corresponding contributions.

1.2 Objectives

In regards of the above discussions regarding the background, motivation and challenges associated with the development of the assistive systems for smart education, we now discuss the overall research objectives of the thesis. To summarize, the objectives of this thesis are enlisted below.

1.2.1 Ubiquitous Assessment of Attention in Smart education

On identifying the research gap in the domain of ubiquitous estimation of attention of learners in online courses, the first objective of this thesis is to present different applications- both smartphone-based and laptop/desktop-based, that can detect overall attention of the learners in an non-intrusive manner. In particular, we aim to propose different types of systems by considering the nature of the course, modality used, and type of attention. Finally, the objective is to understand the accuracy. robustness and usability of the systems through large and lab-scaled evaluations.

1.2.2 Improved Interactivity between Users and Devices for Smart Education

In the context of interactivity, the current challenges motivates the development of touch-free applications that can be used by the learners with clinical disabilities. In doing so, we aim to limit the inclusion of hardware that are only present in the devices, for the purpose of sensing. Similar to the ubiquitous systems, the objective is evaluate the interactive systems with real-world users under different setups for understanding the system's performance.

1.3 Contributions

In lights of the above objectives, now the major contributions of this thesis are highlighted next.

1.3.1 On-topic External Attention through Gaze Gesture Tracking

As discussed earlier, on-topic visual attention (overt) is external in nature as it can be estimated through observable eye movements. In this work, firstly the three basic principles of visual attention are proposed as: *Observation, Tracing*, and *Focus*, through a large-scaled human study. Along with this, the concept of *prime objects* in MOOC videos are also revealed. Based on these principles, a ubiquitous smartphone application that utilizes gaze gesture, measured through the in-device front-camera, for understanding visual attention of a learner is modelled and proposed. Finally, through real-world human studies, we not only establish the accuracy of the system but also establish a significant correlation between visual attention and high level cognition of a learner.

1.3.2 On-topic Internal Attention & Off-topic External Attention through Expressions & Speech Tracking

For on-topic internal attention, estimating the cognition of a person becomes essential. To overcome the shortcomings of visual attention, in this work, the major contributions encompass the development of a system that can infer upon the cognitive attention of an online meeting (lecture, presentations, etc.) participant. Further, it identifies the instances of off-topic external attention (visual multitasking) and classifies them as either relevant (person reading/watching relevant articles during the meeting) or irrelevant (reading/watching irrelevant articles during the meeting). The system development is related to our concurrent theoretical contribution of devising the association between expressions, communication and cognition of human beings. The system uses modalities like facial expressions of an attendee, speech intent of a speaker, active communication rate of the participants, vocal emotion and ambient light of the participants to understand their attentive involvement and instance of multitasking.

1.3.3 Estimation of Engagement using Facial Expressions through Acoustic Sensing

Envisioning an utopian educational system, the following objective was to remove the challenges of camera-based expression detection systems. With this objective, the major contribution of this work is the development of a smartphone application that leverages acoustic sensing for assessing the on-topic sustained attention of viewers (of online videos), and hence their overall engagement to the content.Through detailed and methodical evaluations, we prove how the system can be used under robust conditions and distinguish well between an engaged and a disengaged user.

1.3.4 Touch-free Interactivity between Users and Devices using Eye Blinks

In this first work of the second line of approaches (systems promoting interactivity), the contribution lies in terms of utilizing visual cues like blinks as the interactive medium, thus eliminating the requirement of touch and clicks. The application allows the user generate commands for playing or pausing a video (MOOC) with eye gestures. Moreover, a learner can take screenshots, select video segments through blinks. The system eliminates the requirement of taking manual notes by automatically generating textual notes with selected keywords, linked to Wikipedia for reference. While it facilitates seamless interactivity for users with clinical conditions, it also promotes the scope of sustained attention by eliminating the necessity of visual context switching.

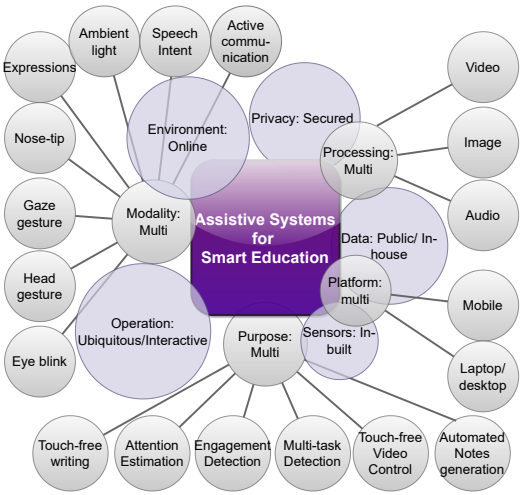


Figure 1.1: Representation of the Facets of Assistive Systems for Smart Education

1.3.5 Touch-free Text Entry in Smartphones using Nose-tip Gestures

The final contribution of this thesis is the development of a smartphone application that uses the in-device camera and sensors to track users' facial orientations and track nose-tip movements for writing on the device in a touch-free manner. The nose movements are used to draw alphabets and numbers in air which are mapped to the device's screen. Further orientation of the phone, tracked through sensors, can be used for editing the entered texts. The usability of the system has been extensively tested with real world users with clinical conditions. Figure 1.1 summarises the contribution of the works presented in this thesis, through the visualization of their multi-dimensional aspects.

1.4 Organization of the Thesis

In this section, the organisation and content of the following chapters are described briefly.

Chapter 2 presents the extensive discussion of the existing literature in the field of automated attention estimation and novel approaches in the fields of HCI. In doing so, we discuss the different dataset, approaches and modalities proposed in such works and identify the limitations and future scopes.

Chapter 3 presents *GestAtten*, an ubiquitous smartphone application that utilises gaze gesture for automatically inferring upon the visual attention and high level cognition of learners attending MOOCs.

Chapter 4 describes *EmotiConf*, an automated ubiquitous assistive system that uses participant's facial expressions, head movement, frequency of active communication and vocal intent to understand their overall attention level in online meetings and classes.

Chapter 5 shows the utility of near-ultrasound signals in classifying user's facial expressions in a non-intrusive, camera-free approach. Through the development of *ExpresSense*, a lightweight smartphone application that utilises such acoustic signals for expression detection, we show how expressions can correlate to the user's overall engagement to online videos.

Chapter 6 discusses eye-blink as a unconventional yet suitable modality for touch-free interaction with devices. Such modality has been used to control online MOOC video playbacks and generating notes automatically for the learners. This Human-computer interactive model has been explained through the development and evaluation of *AutoNotes*.

Chapter 7 addresses the problem of inclusivity through the development and analysis of *Nosype*, a smartphone application that utilises head orientation; hence nose-tip tracking for touch-free writing. The system aims at benefiting the users with clinical disabilities like Dactylitis, Sarcopenia, essential tremor and other problems. The system can be used for taking quick notes on smartphones, in a touch-free manner.

Chapter 8 finally concludes the thesis by summarizing the previous chapters and envisioning the open scopes of future work in these domains.

2
Related Work

This chapter presents an in-depth exploration of start-of-the-art literature in the field of Human-Computer Interaction (HCI), especially focused on ubiquitous and interactive techniques for promoting smart living. Analyzing the research developments around the concepts discussed in Chapter 1, and eventually identifying the scopes, also serves as a foundation on which, the objectives and contributions of this thesis are stacked.

While the primary objective of this chapter is to provide an overview of the formative stages of the following chapters, we also try to answer the following Research Questions (RQ) by studying the *state-of-the-art* literature:

RQ1 : Is human-attention sensing different under physical and virtual setups?
RQ2 : What are the feasible modalities for online attention estimation?
RQ3 : Which forms of attention have been addressed through system development?
RQ4 : Are touch-free interactivity really necessary for novel HCI?
RQ5 : Can we promote such interactivity globally?

Sections 2.1 and 2.2, will aim at deriving the answers to **RQ1–RQ3**. Sections 2.4 and 2.5 will provide answers to **RQ4–RQ5**.

2.1 Attention Estimation in Physical Scenario

Human attention itself has raised research interests in recent times. The research ranges from defining, classifying attention [250], estimating the degree of attention [120], their causes and

effects [187], developing different datasets [13], technologies and tools to identify attention [85] and promote it [214]. In a physical setup, such as a traditional classroom, the requirement of estimating learners' attention automatically, can be argued to be inessential. This is due to the fact that manual inspections are possible in such setups. However, imagine a classroom with 60-70 students and a single lecturer who is responsible for identifying individual's attention level, coordinate the classroom and deliver a lecture [196]. It instantly becomes evident that the classroom requires an assistive system that can capture individual or the overall attention level of the class. Further, based on the proposition that attention promotes comprehension, teacher-learner interrelationship and the overall quality of the course, several works [72, 234, 173, 247, 4] on classroom-based attention estimation has emerged. The works can be divided on the basis of the type of data used to estimate human-attention. For example, while the most commonly used modality is gaze, other modalities include facial expressions, body postures, and other physiological data [26, 233]. Some of these modalities are described below.

2.1.1 Gaze-based Estimation

A number of works in the literature have used gaze to extract the attentiveness of a learners under different educational setup including one-to-one tutorials [92], multi-learner setup [247], robot-based learning [170], educational toys [222] for learners with physical challenges and so on. In some works [86, 85], the authors have analyzed the involvement of gaze in assessing the mind wandering instances of learners in intelligent tutoring systems. The works involve gaze features like saccades and gaze fixations, estimated from the gaze recordings of learners in a classroom, using commercial eye trackers. Under natural setup, gaze hierarchy [79, 248] has also been captured through commercial trackers to understand the common patterns of attention.

2.1.2 Facial Feature-based Estimation

Universal facial expressions like *anger, surprise, disgust, enjoyment(happiness), fear, sadness* [54] have been long studied as an indicator of mental processes like attention [127]. To better understand the mental involvement of learners, several works have relied upon expressions, captured through camera [227, 52]. By analysing the dominant and contextual expressions during lectures, these works have achieved significant accuracy in detecting the learner's attention-related aspects like boredom, drowsiness, engagement etc. Moreover, existing studies [241] have aimed at developing expression databases, specific to classroom scenarios, that can be used to train the automated models for identifying affective states. Apart form macro expressions, some of the works focus on detecting micro expressions [184] caused by fine-grained movement of the facial Action Unit (AU)s.

2.1.3 Physiological data-based Estimation

The three major physiological signals considered to measure attention of a learner are: Electrodermal activity (EDA), Electrocardiogram (ECG) and Photoplethysmography (PPG). In [47], the authors have captured EDA of both the students and the teachers through the Empatica E4 wristband [1]. By deriving the arousal of the learners, their emotional response, and interest at any particular instance of the lecture, the authors classify their engagement. A similar approach [67] has utilised EDA to investigate synchrony between a presenter and attendees in a conference to facilitate automated feedback. EDA has also been used to correlate seating positions of learners in a classroom [62], as study groups, with their engagement level to the academic courses. However, a more recent work [49] shows that there is no significant correlation between the instantaneous EDA and the observed engagement. This questions the applicability of EDA in estimating real-time engagement. On the other hand, bio-signals like ECG and Electroencephalogram (EEG) have been used to detect attention of students during different cognitive tasks like mental calculations, programming tasks etc. Other custom-built devices like smart chairs with pressure mats [175] and head bands with EEG sensors [110] have also been used to estimate learners' attention in a classroom scenario.

2.1.4 Audio-based Estimation

In the context of cognition detection through interactivity, audio has been used to identify students' groups in a classroom [237, 135]. Other works involve the use of web-based applications to detect engagement through interactivity in fixed classroom setups [9]. Despite these approaches, the usage of acoustic signals for understanding human attention directly or indirectly, has remained limited. One possible reason behind this, could be the susceptibility of such systems to privacy breaches.

2.1.5 Multi-modal data-based Estimation

Gao *et al.* proposed a multi-modal engagement detection technique called n-gage [63] that utilises multiple data modalities for assessing student's engagement in a classroom. In this work, the authors have simultaneously recorded the physiological (EDA, PPG, etc.) and environmental (CO2 level, temperature, humidity, noise) data from devices like Empatica E4 writbands, and Netatmo stations respectively. n-gage then predicts the engagement score of the learners in a multi-dimensional scale comprising of the emotional, cognitive and behavioral aspects. Similarly, in [270], facial cues and body postures are combined to identify the behavioral patterns of the learners in classrooms, including activities like yawning, stretching, writing etc. to finally infer upon their attention states. Monkaresi *et al.* [165] used a Microsoft Kinect Face

[1] https://www.empatica.com/e4-wristband (accessed: Friday 11th August, 2023)

Tracker to capture the facial action units and heart rate of the learners, to infer upon their state of mental engagement. Most of these approaches use self-reports as a ground truth estimator.

As observed from the *start-of-the-art* literature, attention estimation in physical classrooms involve effective techniques and can be quite accurate in terms of assessment. However, these approaches mostly involve the use of hardware- either commercial or custom-built. While commercial systems are readily available to the common mass, their prices can be a barrier to the globalization of such systems. Moreover, some wearables are inherently intrusive in nature and causes the learner to get conscious about the setup. In spite of these shortcomings, physical classroom sensing can include various aspects like posture tracking, communication-based activity tracking, group detection, and so on. These expands the scope and design possibilities to simplify the task of automated attention estimation. However, for virtual setups, such techniques are not possible in terms of deployment, hardware availability and physical location of the participants. This answers **RQ1** with the indication that *human attention tracking is indeed different under physical and virtual setup. While deriving this answer, we also establish the requirement of novel solutions, that would largely deviate from the physical-location-based approaches, for automating the process of attention tracking in online courses.*

2.2 Attention Estimation in Virtual Scenario

In order to answer **RQ2**, we need a thorough exploration of the modalities, techniques and applications proposed in the existing literature. We divide this section into the various data modalities that have been popularly used for estimating attention in virtual setups. While discussing these approaches, we also explore the underlying and related techniques and algorithms that can be used to develop the models.

2.2.1 Gaze-based Estimation

Gaze has been a popular modality for attention estimation during online lectures. Although the usage of commercial gaze trackers is not feasible in virtual setup, the device's camera can be used to track the learner's upper body, specifically the face. Hence, the eye region can be extracted and processed for estimating the learner's gaze. These techniques, though not as accurate as the ones with commercial tackers, are sufficiently capable of tracking human gaze locations and gestures. Krithika *et al.* [113] proposed a vision-based technique where learner's eyes and head orientation are detected to infer upon their concentration level as either high, medium or low. Based on the proposition that eye-contacts between an instructor and learner encourages attentiveness, a recent work [5] has proposed the idea of a "virtual digital twin"- a digital representation of a physical classroom where the gaze angle of each learner can be represented in real time. In 2015, Mariakakis *et al.* [152] explored the field of estimating user's

attention, based on their viewing implications on the mobile device. The approach described in the work is lightweight and can be conducted in a stand-alone mobile device. The approach finds its utility in a text based sequential reading application. However, the approach does not take into account the environmental factors like noise, ambient light change etc. which can affect the cognitive aspect of the reader.

For facilitating gaze tracking, several works have aimed at developing lightweight and accurate algorithms. Timm *et al.* [239] presented a cost effective, feature based eye center localization technique that takes into account, the properties of image gradient vectors, intersecting at the center of a circular object, hence the center of detected iris region. Zhu *et al.* [282] presents a novel eye gaze tracking approach that maps the gaze of an user into one of the 8 segments of a computer screen. An infrared illumination technique for glint formation and detection is adopted to track the gaze of the user under free head movement. Gaze estimator using webcams [182] for website visitors adds a new dimension towards generalizing the process of gaze tracking which can elevate its utility in an unrestricted manner. Various gaze-based datasets [111, 274] have also been built by the process of crowd-sourcing or controlled-sessions for the purpose of eye tracking. Approaches have been proposed in [171, 58, 240], that aim at locating the gaze points in continuous locations, rather than blocks of screen segments. A lightweight gaze tracker, that is based on the idea of mouse gesture plugin for Firefox, has been presented in [50]. Although ocular cues like gaze and blinks are generally detected through vision-based techniques, Liu *et al.* [136] proposed a system called BlinkListener that operates on the reflection acoustic chirps to detect blinking instances of an individual. The authors have analyzed the correlation between blink-induced motion and the phase and amplitude of the signal, as captured through a smartphone's microphone. However, the applicability of this system in attention estimation can be explored in future.

Several works have been conducted in gaze tracking, based on the eye model and hence the optical axis [253]. Although these techniques are reliable, their applicability in mobile devices with restricted screen size is a subject of further research. On the other hand, the learning models that uses the feature information of eye images might require large scale databases to be trained accurately. The optimal solution to an eye and gaze tracking problem in mobile environment is hence, a scope of further research. Similarly, approaches based on eye feature fusion [14] have been proposed for detecting gaze points on standalone mobile devices and have been tested on image data sets like GazeCapture [112], MPIIFaceGaze [276], etc. Similar other works have been presented in the literature [30, 260, 100, 98], although the approaches have not been applied for correlating the gestures with the video being played. Moreover, their performance for real time videos is a matter of further experimentation.

2.2.2 Facial Feature-based Estimation

Facial expression has been a rich source of information, revealing insights about a person's attention span. However, the recent approaches also aim at maintaining users' privacy. Pertaining to this, applications that can run on a client's device, without the requirement of contextual data transfer to remote servers, are being developed. Thus, applications that use efficient neural networks models and performs seamlessly on devices in real time are preferred for attention related tasks. In [209], the authors propose such a system that can assess learners' engagement, affect and expressions.

Apart from works proposing novel attention estimation approaches involving expressions, authors have also aimed at developing public dataset where facial expressions are correlated to user's engagement [158]. While most of the expression detection mechanisms rely on learner's facial video capture [31], near-ultrasound acoustic chirps have been used in [64] for capturing various facial expressions and a person's hand-to-face gestures, using commercial microphone arrays. Apart from the microphone arrays, earphones have also been used for monitoring facial muscle movements [125] pertaining to the flexible positioning of the speakers and microphones. Natural facial cues like gaze activity, lip movement, eyelid tracing can also reveal the underlying level of attention [12].

2.2.3 Physiological Data-based Estimation

When it comes to online lectures, the options for physiological data trackers become rather limited. The authors in [263] have used a mobile phone based divided attention monitoring for MOOC videos using photoplethysmography (PPG) signals. These signals are sensed using a commodity based camera that eliminates the requirement of an additional hardware. While, in [262], a PPG based personal event monitoring system has been proposed for indicating disengaged learning and reviving attention by immediate alerts, [188] uses the PPG signals for deciding the difficulty levels faced by a learner and also recommends the best videos based on the perceived difficulty level.

2.2.4 Speech, Mouth Tracking and Acoustic Feature-based Estimations

The utility of acoustic sensing for attention estimation in online classes has remained unexplored, to the best of our knowledge. However, there are some related and interesting applications of acoustic sensing that promises its scope in the domain of automated attention estimation. Joon *et al.* [36] has presented a system that aims at estimating the synchronization between the voice and video of a speaker in the video. In doing so, they take advantage of the audio data and the mouth images to track the mouth movement in the video. Smartphone-based acoustic systems have also enhanced communication clarity by inducing acoustic signal-based

lip movement tracking [272]. Acoustic signals have been used extensively in coarse-grained motion and location tracking of objects [124, 66, 25], as well as tracking subtle movements like respiratory patterns of individuals [87, 252]. In EchoSpot [129], target localization is performed using Frequency Modulated Continuous Wave (FMCW) signals that get reflected in the microphone from different paths. The system can find an individual's location in an indoor scenario by processing the received echo. Apart from chirps, ultrasonic tones [130] has also been used in course-grained motion detection, e.g., sudden falls, using techniques like Doppler Shift [215]. Fine-grained motion tracking facilitates several healthcare applications, such as monitoring an individual's chest wall [229]. Motion tracking can account for learners' engagement to a content by indicating whether or not they are distracted by secondary tasks. Whereas, fine-grained motion like respiration can indicate stress level of the learner during online evaluations.

2.2.5 Multi-modal data-based Estimation

Apart from the single modalities, models have also utilised multi-modal data. Video conferencing systems generally consider individuals to be connected in a virtual conference room. In most cases, one user is present at each end in a static condition. However, research has also been extended towards considering scenarios where multiple users can be present at one end of the conference and are connected to other participants over the video conferencing platform. In [60], the audio and video-based features are enforced to develop a speaker identifier so that the camera's focus can be auto-shifted towards the active speaker around the table. In [144], the authors have designed a platform for online lectures, where the instructors can view behavioral summary of the learners, even when their videos are turned off. The system displays their engagement level, based on their expressions, their emotions, head gestures and on/off-screen gaze gestures. Although the system works well for formal presentation-oriented courses, its applicability in interactive sessions can be explored through further studies.

The discussion of the modalities leads to the answer to **RQ2**. Although virtual mode of education restricts the utilization of body sensors and specialised devices, modalities like gaze, facial expressions, oral features (mouth movement/speech analysis) can be employed to detect users' attention. This can be facilitated through vision-based techniques, as well as acoustic signal processing. Next, we discuss the systems that explore individual attention types and aims at automating the estimation of various forms of attention.

2.3 Systems for Various Attention types

As discussed previously, early research in the field of Psychology revealed different stages of attention: focused attention, sustained attention, selective attention, alternating attention,

and divided attention [161]. In [1], the authors present a novel approach that correlates facial temperature, captured through thermal imaging, with their cognitive state. Based on the hypothesis that sustained, alternating, selective and divided attention are associated with different levels of cognitive load, they employ thermal imaging and gaze tracking to estimate and classify user's attention. The methodical selection of audio-visual and Stroop tests for data collection, along with the accuracy of the system proves the applicability of thermal scans as one possible modality for human attention classification. The study, presented in [118], facilitates selecting the optimal degree of cognitive or observable attention to be allowed in the meeting by the meeting authorities. Moreover, the modular subdomain of attention categories is depicted as attention by direction, attention by action, and attention by state. Attention by direction is a direct derivative of the participants' visual direction in the meeting. In contrast, attention by action is more inclined towards verbal addressing, giving a purposeful insight of attention.

On the other hand, the impacts of multitasking (causing divided attention) at workplaces and during remote meetings have been well studied [27]. The study presented in [143] shows an essential comparison between the level and effect of multitasking during formal meetings, teleconferences, and virtual meetings. It reveals that even though multitasking is not prevalent in face-to-face meetings, they are almost inevitable in teleconferences and virtual meetings. It further explains the statistical results regarding the positive or negative impact of multitasking and its effects on attention span. Avrahami *et al.* [11] proposed a system that reduces the effect of multi-tasking based impoliteness in video meetings by automatically selecting the camera showing the best view of the participant's face in a dual monitor setup. While this model aims to support the necessary simultaneous tasks during a meeting, it is often essential from the organization's perspective to identify whether or not a participant is losing attention due to multi-tasking. In [155], the authors have conducted a series of extensive interviews to reveal that multitasking involving the same device/screen as that of the online meeting is better accepted than that involving other devices like smartphones or another monitor. The study also suggests some design ideas like redirecting notifications to the meeting's screen, auto-switching of camera angles and optimized meeting layouts. Moreover, the positive and negative effects of parallel chats during a virtual meeting have been studied extensively in [208]. In [264], the authors have proposed a smartphone-based divided attention detection system that records the user's PPG, using the back camera of the device. However, this imposes a severe restriction of placing the finger-tip on the back camera for continuous sensing.

From this discussion, we can answer **RQ3**. We see that works have been conducted around sustained, divided, alternating, and selective attention. However, majority of these works are suitable for physical setups as they require commercial trackers and specialised sensors. Others impose severe restrictions on the users, thus being non-pervasive in nature. This necessitates the development of ubiquitous techniques, dedicated to these attention types.

2.4 Interactive Systems for Content Control

In this section, we discuss some of the common parameters that have been used for touch-free control of online media contents, mouse movements, and other interactive purposes. Interaction has been identified as a key component for determining user experience in virtual scenarios. Although analysed in terms of Virtual Reality (VR), [45] summarises different categorical analysis of hand gestures as an interactive modality. While some systems allow static poses, others work with moving gestures for controlling the elements of the virtual environment, or establishing communication with the system. These interactive gestures can also be classified based on their functions, like, directional indications, moving / translating virtual objects, etc. By discovering the current limitations of electromyography-based gestures (as compared to touch-less controller-based interactions) in, the study [45] presents an open scope of improvement in this domain of research. Facial gestures have been largely contributing to the research domain involving almost every aspect of life. While facial gestures largely involve facial emotion detection [142] for a wide range of applications starting from medical condition detection [202] to cognition estimation [169], eyes solely can contribute majorly to these purposes. In [17], Zhen *et al.* uses a commercial depth camera that tracks the nose position along with the subject's status of mouth. While the facial location is mapped to the cursor locations, mouth-open gesture is used to re-adjust the position of the cursor. The system has been found to be beneficial for people suffering from Tetraplegia.

Gaze and gaze gestures are widely used in ubiquitous computing applications but require extensive continuous tracking for significant performance. Eye blinks [114] are much simpler to identify through different approaches like color appearance [181], feature based [167] and neural network based models [128] to contribute in fields like transport safety by identifying fatigue level of drivers [41], contact free mobile/computer interactions [179][164] primarily to help the physically challenged individuals, security domain [128] and so on. Similar tro [17], [162] depicts a system that uses blink to generate mouse click commands in desktops. However, the application of blink based human computer interaction in the domain of academia is still nascent.

Audio commands are commonly used for touch-free interactions with the device. In this context, voice recognition is a prerequisite for tasks like voice search. Google developed a very rich language recognition model [84]. Schalkwyk et. al. presented a study [212] on Google Search by Voice and demonstrated its accuracy. However, voice commands in mobile environments might get distorted due to facial obstructions like masks. Moreover, vocal commands are prone to eavesdropping in outdoor scenarios, thus limiting its applicability.

In the context of educational videos and slides, auto-summarization is a well researched area [224]. Text processing has been widely been used for this purpose. Text processing can be executed from images through using optical character recognition [35]. However, identifying

the salient concepts from a group of texts has been an open challenge for many years. The classical approach had been to identify the significant keywords in the text utilizing approaches like POS tags, n-grams etc [138][244][183]. However the limitation of this approach is that oftentimes the actual salient concept is not mentioned in the text but must be inferred from the contextual information. To address this limitation, it is common to utilize an external knowledge base to infer the contextual information. Tagme [59] is a very widely accepted entity linker identifies and links text topics to wikipedia entries. [189] is an improvement to the Tagme pipeline that significantly improves its performance. However, neither tagme nor WAT is designed for the entity salience task. SWAT, extends on the WAT approach to propose state of the art salience detection [191]. In the lights of these discussions, we now focus on some HCI models that can be used for touch-free writing (of notes, short texts, quick points, etc.) on portable and static devices. These systems are discussed in the next section.

2.5 Interactive Systems for Text Entry

The concept of hands-free typing assists people with physical and motor impairments. The incorporation of camera and tracking facility in mobiles has led to the use of gaze cues [256, 70] and head orientation of the user to select particular characters from soft keyboards [267, 265]. Even though different approaches can use these tracked features, mapping it to the exact location of a soft key is particularly difficult in terms of accuracy. As soft keyboard layout [148] places the characters in close proximity, even a minor error or shift in the mapped gaze point on screen can lead to selecting a wrong key, thus requiring the user to retype it several times. The characters typed per second can be decreased if the error rate in character selection is high. Conversely, if a sequence of pose estimation stages has to be performed to detect a single key, the typing speed decreases as well. Some other approaches use voice dictations for voice-text typing [115, 200]. While this approach is accurate for a noise-free environment, it can be significantly affected by background noise or the presence of barriers like face masks. The recent pandemic caused by the COVID 19 virus requires touch-free interfaces that are not affected by the use of facial masks.

Systems like [243] prove that for unconventional systems like gaze-based typing, which largely deviate from touch-based text entry, training is essential to achieve equivalent typing speed and acceptance. In this domain, successful approaches have been proposed for desktop computers [213], and mobile devices' development is still nascent. In [146], the authors propose a gaze-blink based text entry system for disabled users that show promising results. Räihä and Ovaska[197] presents a study of metrics for evaluating such systems using eye tracker and computers on different participants. Apart from this, the major focus has also been placed on diminishing the dwell time for gaze-based text entry systems, even for medium portable devices [207, 246]. Novel fast systems for performance enhancement have been proposed in

doing so [48]. Gaze localization has been achieved involving large scale dataset based training [105, 123] or iris localization and calibration based mapping functions [68]. Other text entry systems also employ hand gesture tracking using other equipment [76], muscle movements [141], touch-based gesture [281], silent speech commands [235], peripheral vision [140], head orientation [71] and key compaction, and even tongue orientation concerning teeth locations [174]. Khan *et al.* [104] presents a desktop based system that employs nose tracking to control the cursor locations by clinically challenged users.

Face, hence nose tracking, has been a prevalent research problem for years as it is one of the most contributing facial features [73] in image processing applications. With the increasing demand for technology and its involvement in daily life, this research domain's approaches and applications have expanded and are still evolving. This section discusses some of the nose tracking approaches and their existing applications. Like gaze tracking, nose tracking can also be incorporated using deep learning frameworks [257] and used in various applications, including driver's fatigue detection. Nose shape [268] and feature construction can not only aid in interface control but also help in facial projection, graphical reconstruction, and expression detection. Another unconventional application of nose tracking has been proposed by Yasuyuki *et al.* [266] where the olfactory display is considered for virtual reality applications. The applicability of nose tracking mainly includes interfacing control in desktop computers due to the easy accessibility of webcam images, static environment, and high processing power [74]. The approaches for such an application can vary from HAAR based feature extraction for nose localization to Multi-Domain Networks and other neural network models [271, 19, 225]. However, for mobile devices, the tracking nose should be executed using lightweight approaches, thus eliminating the chances of frame lag.

We can answer **RQ4** by considering the utility of the existing interactive systems, as discussed above, as well as their future scopes and limitations that require novel solutions. Indeed, such interactive systems are extremely useful for learners with clinical issues. However, whether the systems can be globally promoted (**RQ5**) depends on the ease of their usability. We believe that with considerable training, novel systems can be incorporated seamlessly in daily lives.

2.6 Summary

By analyzing the existing literature, we not only perceive their contributions in the respected domains, but also identify their limitations. Firstly, with respect to RQ1 and RQ2, we understand that assessment of attention in virtual educational setup is not only different, but also difficult. The current works can efficiently work in physical classrooms where specialized hardware can be deployed in the environment for multi modal sensing. However, in a decentralised setup, commercial or custom-built trackers cannot be considered. Thus, some of the current works have

utilised the commodity smartphones and laptops for tracking an individual's behavioral cues and bio-signals. However, these methods are often intrusive in nature and causes distractions. This leads to the motivation behind (1) developing systems built specifically for virtual educational setups like MOOCs, Online Classes, etc. that can optimally work with a restricted set of data modality, (2) innovating assistive systems that can ubiquitously estimate subjects' attention, and (3) considering in-built sensors and hardware components of smartphones/laptops. With respect to RQ3, we observe that the categorical analysis of attention is quite limited in virtual setup. While a few of the works individually address a single form of attention, say divided attention, there is no common direction along which these works can be linked and evolved. Moreover, inevitable activities like visual multitasking needs automated identification. Although some works have aimed at identifying micro and macro activities in different other contexts, visual multitasking in online meetings has not been addressed in any of the prior works, to the best of our knowledge. This limitation motivates (1) The development of systems that are dedicated to individual attention types and (2) building such models in a stratified manner to identify and eliminate the shortcomings of the previous model. While answering RQ4 and RQ5, we explored the different existing works that propose novel approaches towards contact-less communication with devices. However, most of them use custom-built hardware with limited availability to the common mass. Moreover, some of these systems are suitable for desktops and require commercial devices like a webcam. Similarly, with voice-based approaches, the utility gets restricted due to the systems being prone to eavesdropping. The existing literature also lacks the correlation of these systems in educational domains. These limitations motivate the development of touch-free systems for learners with clinical issues. Using these motivations, derived from existing the research gaps, we design and develop assistive systems that can be used for automated cognitive assessments and seamless interaction with the devices in the context of smart online education. The following chapters discuss there systems in details.

3

Ubiquitous System for Estimating Visual Attention of Learners in MOOC

In Massive Open Online Courses (MOOC), the cause of high dropout rates can be related to the lack of learner's attentive engagement with the online videos [258, 236, 37, 263]. Moreover, even if a learner completes the certificate course online, manual evaluation neither guarantees the full comprehension of subject by the learner, nor the attentiveness of the learner throughout the course [28]. Automatic attention estimation and feedback, not only estimates the learner's involvement, but also appraises the efficacy of the video tutorial. A number of recent studies [279, 116, 188, 97] has shown that mobile based MOOC learning has been emerged as a widely-acceptable platform; consequently, almost all the popular MOOC platforms come with an associated mobile application. The existing approaches for learner attention estimation on mobile MOOC learning platforms primarily rely on physiological signatures, such as heart rate sensing and PPG signals obtained through the back camera of the mobile phone [261, 263, 262, 188], finger tracking [56], and so on. However, the accuracy of such approaches depends on the assumption that at least one of the fingers will be in the vicinity of the back camera when the learner holds the mobile on her hand; otherwise the accuracy drops down significantly.

In contrast to the above approaches, the front camera of the mobile can be effectively used to capture the visual expression of the learner. Visual aspects like gaze and gaze gesture can furnish sufficient relevant information regarding the attention level of a learner. Based on an online

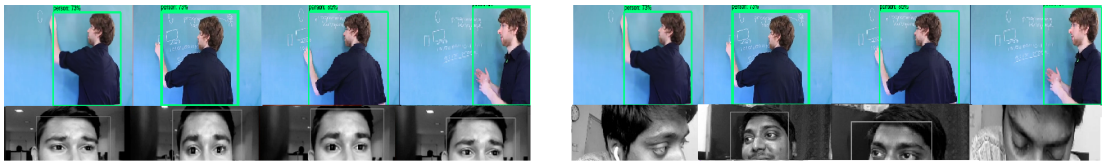


FIGURE 3.1: Correlation of visual cues and object movement trajectory in MOOC videos. The set of MOOC frames to the left depicts the movement of the lecturer which is highly correlated to the eye movement of the user, showing high level of attention. The visual gesture of the user on the right shows no similarity to the movement trace of the lecturer, implying low attention level.

survey conducted over more than 1200 participants across the globe (details in Section 3.1), we observe that the gaze gestures of a MOOC video user depend on the content of the video, like the movement patterns of the instructor during the lecture delivery, the textual contents in the presentation as they appear during the lecture, the pattern of the pen or hand movement as the instructor illustrates something by writing it over the display area, and so on. Therefore, a sufficient estimate of visual cues can adequately indicate the learner's attention level and even enhance the learning process [219, 220]. Figure 3.1 shows an example, where the eye gaze gesture of the user should follow the hand movement pattern of the course instructor, when the instructor explains the concepts by writing texts on a board.

However, there are multiple challenges for inferring cognitive attentiveness of a user from gaze and gaze gesture patterns, as mentioned below.

1. Multi-tasking is common while observing a MOOC video over the mobile platform. Our online survey indicates that a MOOC user may perform different other related (or even sometime unrelated) tasks while going through a MOOC video, primarily, taking notes, solving related problems as being discussed in the lecture, searching for the reference materials, checking emails or Internet, looking into the messages in mobile, and many more. Because of this reason, it is not expected that the user will continuously gaze into the video, and the gaze pattern may shift from the video to the outsize objects frequently. This needs to be accounted for while developing a system from the eye gaze patterns.

2. To understand whether the user is following the video, the eye gesture pattern of the user needs to be correlated with the object movement patterns in the video. These objects can be multi-fold – the teaser of the instructor, the hand and the pan of the instructor while the instructor writing something in the board, the text that is being illustrated by the instructor, and so on. A proper methodology is required to understand each of these important objects (we call them as the prime objects) in the video and then to check which object the user is following or whether the user is following the correct object. This can be challenging considering the diverse types of videos available in the MOOC platforms.

3. Existing approaches for eye gaze and gaze gesture tracking [275, 239, 180] rely on computationally heavy methods for video processing at the pixel level as well as incorporates complex supervised machine learning techniques, for which server-side computation is required. However, it is not a good idea to transfer the video captured through the front camera of the mobile to a different server primarily because of two reasons – first, it invades the privacy of the learner as the learner may watch the video during her personal free time and may not like to get captured in a video, and second, transferring the video from the mobile to the server will also incur an additional non-negligible cost.

4. As we consider the scenarios of observing MOOC videos over handheld devices like smartphones or personal laptops, a user can use the platform under different environment; even the environment may change while the user is observing the video. For instance, the user may observe the video while traveling in a cab – thus the ambient light of the environment may change continuously. This affects the lighting conditions of the video that is being recorded by the front camera of the smartphone or laptop for the purpose of analyzing the eye gaze and gaze gesture. The existing light-weight eye detection techniques [180] rely on threshold based approaches. However, a single threshold does not work under varying ambient lighting condition. This poses a major challenge in the development of an on-device processing methodology for eye gesture tracking.

Owing to the above challenges, in this paper, we develop an ubiquitous platform for automated detection of user attentiveness as an ad-on service over a MOOC video platform. Based on the eye gaze and gaze gestures of the user as processed from the video captured through the front camera of the hand-held device, the add-on service over the MOOC video platform assigns a score against the perceived attentiveness of the user, which can be used to develop multiple recommendation services for the MOOC service providers, course developers as well as the learners. To mitigate the challenge of preserving privacy as we capture the user's video, in this paper, we rely on in-device computing to extract the eye gaze and gaze gesture from the video captured through the front camera of the hand-held device. Consequently, we develop *Gestatten* where the cognitive aspect of a person is assessed by correlating her tracked gesture of the eye with the movement of an object of interest in the MOOC video, as the frames progress. The initial offline module of our approach extracts the prime object (an object where the user should focus) movement trajectory from the MOOC video. The real time module extracts the user's eye movement trajectory from the live capture through the mobile's front camera. These two trajectories are correlated together based on three criteria – (i) gaze tracking, (ii) gaze gesture patterns and (iii) importance of the MOOC video object as captured through the gaze.

The model developed in this paper serves a two way purpose – the technique finds its way in estimating whether or not, the learner is attentive in a course and unlocks a scope for personalized recommendation system. Moreover, the model determines the effectiveness of the course, based

on the average attention level of the students. Secondly, the proposed technique uses efficient in-device low-overhead unsupervised machine learning techniques for classification, simple pixel level calculations for processing the video streams and pattern mapping procedures for inference. The absence of high computational overhead makes the model suitable for mobile devices. Storage of user specific data is only restricted to files containing coordinates and labels. No real time images or videos are stored for training or testing purposes, thus maintaining user's privacy. In doing so, the model proposes a novel concept of identifying prime objects of focus in a video and tracks the user's visual cues under dynamic ambient light conditions for estimating their degree of attention. The claim that gaze gesture can reveal relevant information regarding the high level of comprehension of the topic, has been proved by correlating the model generated and subjective scores for 48 different participants. We have implemented and tested *Gestatten* over 48 participants using various models of mobile phones, using two different subjective tests. In the first test scenario, the learners have participated in a small cognitive examination that judges their attention level during watching the MOOC video. In the second test scenario, a set of different adjudicators judged the attention level of the learner, and the learners have been ranked accordingly. The results of the subjective tests have been compared with the *Gestatten* generated results, and we observe a high accuracy with 8.68% average absolute error rate in estimating the attention level of the learner. We also benchmark the Android application developed for *Gestatten*; we observe that the application is significantly lightweight compared to existing baseline eye gesture tracking techniques, whereas it consumes minimum battery power while in execution.

3.1 User Study

The advent of ubiquitous learning and advanced mobile frameworks has instigated a convention of smartphone usage for the purpose of learning. This trend is not only limited to young population but has also spread among individuals belonging to all age groups. To establish the necessity of an automated mobile based lightweight approach for the estimation of attention, we have conducted an online anonymous survey over more than 1200 MOOC users. The survey aimed at understanding the popularity of mobile devices for MOOC videos and also to identify the pattern in which learners watch a video as well as to examine whether visual cues can sufficiently contribute for estimating learner's attention level.

The survey, apart from focusing only on MOOC usage patterns, also collected some basic information like age group, gender, demography and eye color of the participants. The survey was attended by a total of 1256 participants from different locations across the globe including India, Chile, Canada, United States, United Kingdom, Bangladesh, United Arab Emirates, Switzerland, Germany, Singapore, Nigeria, Brazil, Bangladesh, Switzerland, Mauritius, Singapore and so on. The variation of eye colors was limited to black (73.2%) and brown (19%) and

blue(.3%), as per the responses. The rest 7.5% did not reveal their eye color. We use "black eye" for "dark brown" eye.

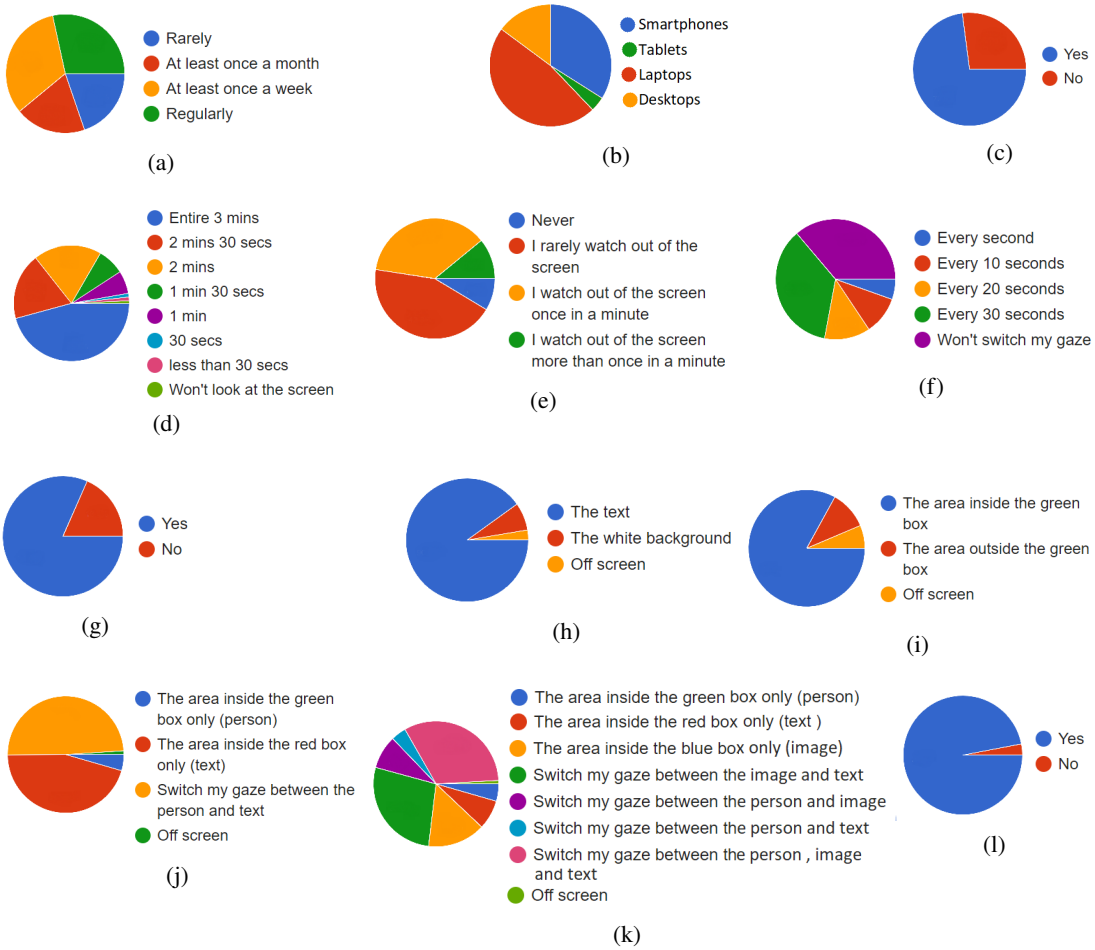


Figure 3.2: (a) Popularity of MOOCs, (b) Choice of device, (c) Multitasking (d) Constant on screen gaze time for a 3 mins video (e) Frequency of watching off screen (f) Context switch for a 3 mins video (g)Preference of video over audio (h) Single Object of focus (text) (i) Single Object of focus (person) (j) Two Objects of focus (Text and person) (k) Three Objects of focus (Text, image and person) (l) Visual preference for better comprehension

3.1.1 Observations from the Survey

Figure 3.2 summarizes the outcomes from this survey. The survey unveiled some of the key points involving MOOCs, their usage pattern and the role of visual cues for attentiveness and comprehension. The details are discussed next.

Popularity of MOOCs: The survey indicated a high popularity of MOOCs among people belonging to different age groups 71.6% of the people watches these videos more that or at least

once a month (Figure 3.2(a)). The high acceptance rate of these videos demands an automated evaluation that is free from manual bias.

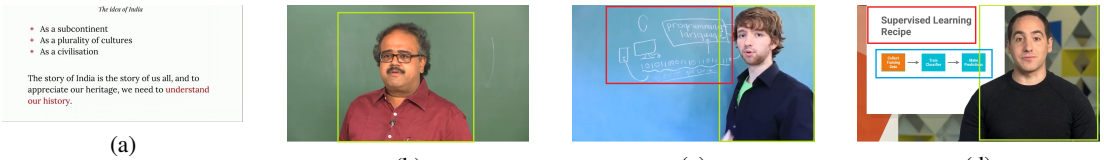
Choice of devices: The usage of mobile devices like smartphones and tablets is strongly portrayed by the responses that showed a spiked percentage of 52.8% of smartphone use plus 6.1% of tablets (Figure 3.2(b)). Along with this, 73.5% of the users use personal laptops for watching MOOC videos. For device usage, participants were allowed to select multiple options.This motivates us to develop a lightweight attention estimator for MOOCs that can seamlessly run on smartphones and other mobile devices.

Multitasking: The survey revealed that majority of the MOOC users perform simultaneous activities while watching a video lecture (Figure 3.2(c)). 72.9% of the participants opted for multitasking that includes reading books, taking notes, solving problems, looking for alternative sandbox sites or additional examples, playing games, eating etc. This factor proves that performing relative activities require shift of visual gaze, while still indicating discrete attention of the learner that does not hamper comprehension. The provision of multitasking is hence incorporated in our approach.

Constant on screen gaze time: As a contrary to the previous fact of context switching, it is also established that a certain amount of constant gazing is required to understand the topic. 99.3% of the responses validate this factor by choosing different rates of constant on screen gaze times (Figure 3.2(d)). Thus it can be inferred that a certain level of constant viewing can reveal significant visual information. This fact further motivated the design of *Gestatten*.

Context Switching: The absence of multitasking cannot prevent the occurrence of periodic shifts of the learner's gaze. Even though, looking at the video promotes comprehension, it is quite infeasible for the learners to continuously gaze at the video, thus leading to occasional off screen gazing. 91.4% of the participants stated that they would switch their gaze to some off screen object, once in a while. 63.8% of the participants supported periodic visual context switching at different rates varying from 1sec to 30 secs, even for a short 3mins video lecture (Figure 3.2(e) and Figure 3.2(f)). This constraint is considered and addressed while developing *Gestatten*.

Visual Emphasis: 81.6% of the participants preferred to watch a video lecture, rather than just listening to the audio and 97% supported the claim that observing a video lecture assists comprehension (Figure 3.2(g) and Figure 3.2(l)). This forms a major motivation for the design of an automated attention estimator that can rely on visual sequence.



(a) (b) (c) (d)

Figure 3.3: (a) Frame with single object of focus (text), (b) Frame with single object of focus (person), (c) Frame with two objects of focus (text and person) (d) Frame with three objects of focus (text, image and person)

Object of Focus: The participants were asked to select one or more of the marked key objects from the frames they are more likely to observe to understand the content of the frame. To understand a basic visual preference, we selected four video frames containing objects like a lecturer, explanatory texts and images. Figure 3.3 presents the four different frames shown in the survey. The four different frames are – (a) a text only frame (Figure 3.3a), (b) a frame with the instructor as the primary object (Figure 3.3b), (c) a frame where the instructor explains by writing on a board (Figure 3.3c), and (d) a frame where the instructor explains using a presentation slide (Figure 3.3d). In the frames, we have highlighted the major objects using different colored boxes. For the frames with multiple objects, the object at which the learner focuses, becomes dependent on the learner. To learn, which of these objects stands out to be most relevant, all possible combinations of informative objects have been marked in these frames from different video lectures. We limited our study to a maximum of 3 such objects due to space constraint in each video frame and the fact that too many objects in a frame can automatically lead to diversion of attention.

For a frame containing a single object of focus (Figure 3.3a and Figure 3.3b), majority of the people (90.1% for texts as shown in Figure 3.2(h) and 83% for the instructor as shown in Figure 3.2(i)) chose to look at the object instead of the background or off screen. For more than one of such objects, majority of the participants decided to switch their gazes between the objects (Figure 3.2(j) and Figure 3.2(k)). This factor is attributed by a concept of *prime objects*, their multiplicity and variations in the design of *Gestatten*.

3.1.2 Challenges and Opportunities

The user study provides as various insights which highlight the challenges in developing an ubiquitous mechanism for user attention prediction based on eye gaze and gaze gestures. This is summarized as below.

1. Eye gaze gives a good indication for learner's attention estimation. The learner prefers to look into the video objects, more particularly if it is a text object, some presentation or the instructor is writing something on the board. Although the learners may not continuously gaze at the video while performing multitasking like taking notes or searching for the

references, they definitely gaze at the video within a time duration depending on the type of the content shown in the video. This gives us the opportunity to use eye gaze for an estimation of the learner's attentiveness.

2. The survey indicates that the learner prefers to change the gaze among different objects, when multiple objects are shown simultaneously in the video. This pattern of gaze changing is also influenced by the priority of the objects in the video shown. Therefore, gaze gesture also gives an indication for the estimation of a learner's attentiveness. Further, there is expected to be a high correlation among the gaze gesture of the learner and the object shifting pattern in the video. We capture this correlation in *Gestatten* to estimate the attentiveness of the learner.

Based on this, we develop the detailed framework for *Gestatten* as discussed in the next section.

3.2 The Design of Gestatten

This section gives the overall architecture and design details of *Gestatten*.

3.2.1 Architectural Overview

The overall architecture of *Gestatten* can broadly be classified into two sub modules – (1) The video tracker that generates the tracked locations and shifts of the objects displayed in the video. (2) The attention estimator that performs a three fold evaluation of user's attention, based on eye detection, gaze region and gaze gesture. A mapping function correlates the tracked video object to the tracked gaze of the user for this purpose. Figure 3.4 presents the overview of the proposed model. The video tracking module accepts the MOOC video as input. A frame-wise processing of the video is performed by the Single Shot Multibox Detector (SSD) [137] model using MobileNet [81] as a feature extractor which performs classification and tracking of video objects and locates them using a bounding box (anchor box), as shown in the figure. This model is trained using the *Common Object in Context* (COCO) dataset [133]. The box coordinates provide the location of the object and its relative shifts in each frame, which eventually forms the output of this module.

The *Attention Estimator* module receives the frames from the device's front camera, previewing the user's face. The user can optionally create personalized templates or use the default ones in the *Template Creation* sub-module – the templates help in mapping the eye center with the device coordinates for various different sizes of mobile front screen. The *Gaze Gesture Tracking* sub-module starts once the user opens the MOOC video, tracked by the video tracking module. This sub-module accepts the variable threshold value generated by the ambient light

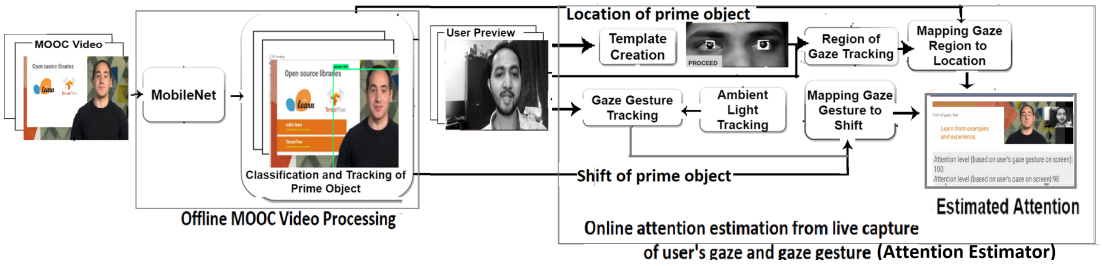


Figure 3.4: *Gestatten* Architectural Components

tracking phase. The *Gaze Region Tracking* module simultaneously accepts each of these user preview frame and estimates the region of gaze using the templates. A mapping function is required to match the shift sequence of the video object with that of the user's eye. This mapping results in statistical estimation of the user's attention and is performed at the end of the course, thus estimating the overall attention level of the user. However, the other mapping function estimates the object of gaze by mapping the gaze region of the user with the location of the video object from the previous module. This is a continuous frame wise mapping, generating the object of gaze throughout the course. The estimated attention is, hence, both a statistical average evaluation as well as a continuous monitoring process. Next, we discuss the MOOC video pre-processing module to extract prime objects from video frames.

3.2.2 MOOC Video Pre-Processing: Tracking Prime Objects

The offline video tracking module focuses on classifying and tracking different video objects. It is assumed that a course video will mostly contain 1-3 different objects that might include person (the lecturer), boards, pen/pencil etc. The range of various objects to be tracked can be adjusted based on the video being used. In the proposed model, the concept of a *prime object* is used to estimate the attention level. Each video is assumed to include at least one and at most three main objects that should be focused on. In most of the video lectures, the lecturer is the main object in the video, whose movement and activities should be followed by the learner. This object will be referred to as the *prime* object in the video. It is to be noted that a prime objects can either be one of the following or a combination of the following types:

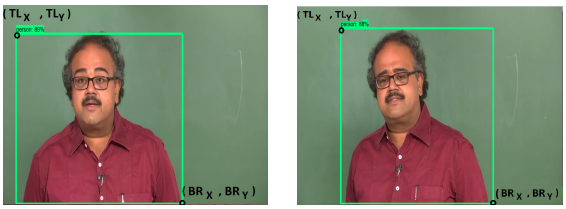
- Type 1 : These objects are content-wise dynamic and location-wise static. The example of type 1 prime objects are static texts like the ones shown in simple presentations.
- Type 2 : These objects are content-wise static and location-wise dynamic. Type 2 prime objects are lecturers whose location changes inside the frame as they walk around during the lectures.
- Type 3 : These objects are content-wise dynamic and location-wise dynamic. Example

of type 3 prime objects are dynamic texts flashing at different regions of the screen. This type of prime objects are rare for video lectures.

- Type 4 : These objects are location-wise and content-wise static. Type 4 prime objects are lecturers standing at a particular location while delivering the lecture.

In our approach, type 1, 2 and 4 prime objects are considered due to the restricted screen size and limited gaze area. A video lecture may contain any combination of these types of prime objects. We also consider the fact that the importance of prime object can vary from user to user. For eg. in a lecture where both the lecturer and a textual data is displayed simultaneously (as shown in Fig. 3.4), one user might visually follow the lecturer, while the other might find it convenient to gaze at the textual part. In such cases, both the objects are prime and if the user gazes at either of the two objects or even at the intersection of the two objects (considering overlapped/proximal objects), he/she is considered as attentive. Also, for multiple prime objects, trajectory correlation can be maintained respectively. The prime objects can be one of the 100 different objects (like person, board, pain, text, etc.) that can be detected using MobileNet.

Figure 3.5 depicts an instance of a type 2 prime object. By comparing the positions of the prime object in frames given in Figure 3.5a and Figure 3.5b, it can be seen that object has shifted to the right. This positional shift is tracked. For classification of these objects, a *SSD MobileNet* model in used due to its high speed and significantly reduced computational cost [82]. The above model can classify and track up to 100 different objects in a video. However only up to 3 prime objects are considered in *Gestatten*.



(a) Frame 1 (b) Frame 2

Figure 3.5: Prime Object Tracking : Shift of prime object in different video frames. The prime object in (b) shows a right shift (increase in the value of TL_X AND BR_X in (b)), relative to the prime object in (a).

As an outcome of this phase, two different record sets are generated for each object in the video – (a) the location and (b) the shift of the object in each frame. The records in location set provides an account for the top left and bottom right coordinates of the bounding box, along with the frame dimensions. The frame dimensions are recorded to find the relative region of location of the object, so that it can be flexibly adapted to a wide range of devices with varied screen sizes. The location records are required by the mapping module,

dedicated for the third level of evaluation, where region of gaze is estimated, thus inferring if the user is observing the prime object in the video. Each record is of the following format: $< frame, TL_x, TL_y, BR_x, BR_y, length, breadth, class >$, where $frame$ is the frame number, TL_x, TL_y are the X and Y coordinates of the top left coordinates of the object in $frame$ (Figure 3.5), BR_x, BR_y are the X and Y coordinates of the bottom right coordinates of the object in $frame$ (Figure 3.5), $length, breadth$ are the length and breadth of the video frame on the device's screen, $class$ defines the class of the object being classified using SSD MobileNet. The classes are according to the COCO dataset labels that include person, desk, TV, laptops, cell phone, book and other common objects. From the given coordinates and the frame dimensions, the relative region of the object can be determined. The shift information of each classified object in the video is tracked independently based on the class label, the confidence score and the location information. The location and shift information, as discussed above, are stored along with the video in the video database and sent to the mobile along with the video content, when a learner views the video. These information are used for attention estimation, as discussed in the next section.

3.3 Attention Estimation

This is a real-time module which runs in the mobile while a learner views a MOOC video. In this module, the prime object movement trajectory is correlated to the eye gaze and gaze gesture of the user. The module begins with creation of a personalized template by the user, followed by a dynamic binarization of live captured user's eye regions, aided by the inbuilt ambient light sensors in mobile devices for evaluating the iris center of the user's eye. These centers are tracked in subsequent frames to infer upon the gaze directions and region of interest in the MOOC videos. The gaze is detected and the gaze directions (gesture) are mapped to the direction of the prime object movements to estimate whether the user has visually followed it, thus estimating her level of attention. Each of these phases are discussed in the following subsections.

3.3.1 Template Creation

The cardinal phase is directed by a calibration process, based on a one-time template creation. The templates are user specific facial record for estimation of gaze region or object. This calibration process is required to compensate the variations in dimensions of the device's screen and user's position. The user is asked to gaze at 5 different sections of the mobile's screen as shown in Figure 3.6 where TL, TR, BL, BR and C represents the top left, top right, bottom left, bottom right and center of the mobile's screen, respectively. At these instances, the user's eye locations are captured, the iris centers are calculated and stored as calibrated points.

Firstly, the device's front camera is accessed to preview the user's face. Each frame is converted

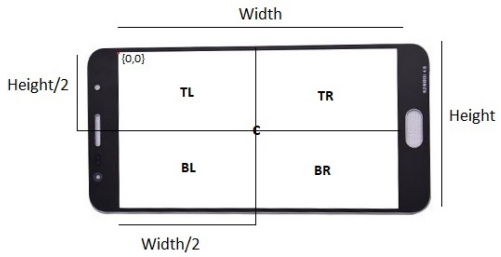


Figure 3.6: The segments of the mobile screen showing Top Left (TL), Top Right (TR), Bottom Left (BL) and Bottom Right (BR) regions.

from RGB to gray-scale for further processing. Based on the tracked gray-scale frames, the user's face location and eye locations (Region of Interest) are identified using a trained *Cascade AdaBoost Classifier* [249, 259] model. It can be noted that the training is done offline, and the trained model is loaded to the mobile; hence, it does not incur significant computation cost. Further, classification using a Cascade AdaBoost Classifier is lightweight, therefore can be performed effectively on a mobile [77]. Now, the user's facial images and hence the *Region Of Interests* (ROI) i.e. the left and right eye centers (iris centers) are tracked while associating the iris centers with the gaze locations on the screen in a predefined order of top-left corner, top-right corner, bottom-right corner, bottom-left corner and center (calibration points).

The criteria based binarization of ROI relies on the fact that the intensity of pixels in a grayscale ROI varies from 0-255, 0 representing the darkest pixels while 255 representing the brightest ones. Experimentally, it can be seen that the pixel intensities, in the iris region of the ROI, are the lowest, representing the darkest region i.e. below a threshold. The proposed technique uses a simple threshold comparison and scalar addition or substitution of pixel values. This results into the binarization of the ROI, where the pixels in iris region (along with some noise due to the presence of eyelashes) are represented by black pixels (0) and the rest of the area by white pixels (255). It is assumed that most of the dark pixels are accumulated in the iris region. Estimating the centroid of the black pixel locations provides the approximate iris center. The eye centers are tracked for the subsequent frames, using the eye center localization algorithm described in Algorithm 1. In this algorithm, each ROI is considered as a group of two sub ROIs, each referring to the left or the right eye. The algorithm only considers the frames where both the eyes are detected. For refined iris tracking, the first fifty rows and columns are pruned out from the intensity matrix, as they mostly cover parts of eye brows, and adjacent eye regions. During template creation, the user can select proper frames where the eye centers are tracked accurately. Only these selected frames are captured and hence five records corresponding to the five calibration points are recorded for the mapping of user's gaze to the screen coordinates. Each record is associated with the following

Algorithm 1: Eye center localization in a single frame

Input: Candidate ROI including left and right eye regions.
Output: The estimated iris centers for left and right eye (C_x, C_y).
if *count(ROI)==2* **then**
  foreach *ROI* **do**
    Set $C_x \leftarrow 0, C_y \leftarrow 0, count \leftarrow 0$
    Extract pixel intensity matrix (M_{px}).
    for $k \leftarrow 50$ *to* $(rows(M_{px}) - 50)$ **do**
      for $l \leftarrow 50$ *to* $(columns(M_{px}) - 50)$ **do**
        $val_i \leftarrow M_{px}$[k][l]
        if $val_i > threshold$ **then**
          $val_i \leftarrow val_i$+255 ; // set pixel to white
        else
          $val_i \leftarrow 0$; // set pixel to black
          $C_x \leftarrow C_x$+k ; // add X locations
          $C_y \leftarrow C_y$+l ; // add Y location
          $count \leftarrow count$ +1 ; // count black pixels
    $C_x \leftarrow C_x/count$; // mean of X
    $C_y \leftarrow C_y/count$; // mean of Y
    if $C_x < FrameWidth$ **then**
      Record C_x, C_y as eye center for left eye.
    else
      Record C_x, C_y as eye center for right eye.

structure: $< C_x, C_y, TL_x, TL_y, BR_x, BR_y, E_{index}, Label >$, where C_x, C_y are the estimated X and Y coordinates of the iris center inside the eye frame, respectively, TL_x, TL_y are the X and Y coordinates of the top left coordinates of the detected eye frame (ROI) inside the facial frame on screen, BR_x, BR_y are the X and Y coordinates of the bottom right coordinates of the detected eye frame (ROI) inside the facial frame on screen, E_{index} is the eye index and can take the value of either "Left" or "Right" i.e. $E_{index} \in \{Left, Right\}$, Label represents the corresponding gaze location on screen | $Label \in \{TL, TR, BR, BL, C\}$, indicating whether the user was looking at the top left, top right, bottom right, bottom left or center of the screen, respectively, at that instant.

3.3.2 Ambient Light Tracking

The ambient light level plays an important role while analyzing video frames. Figure 3.7 presents the overview of how the ambient light sensing module works in parallel with the live video processing. A threshold that decides the probability of a pixel to be considered as an iris pixel is affected by the ambient luminance of the user, measured in lux (lx) units. The fixation of manual thresholding can incorporate considerable misapprehension in the desired output. This is due to the fact that ambient light affects the intensity values of the frame pixels. An iris pixel x_i having intensity y_i under ambient light of L lx, will have an intensity of y_i' under the ambient light of $(L + \delta)lx | y_i' > y_i$. In this scenario, if the threshold is fixed, the same pixel x_i, might not be identified as an iris pixel under a higher luminous. This scenario is handled by sensing the ambient light of the user using the mobile's inbuilt light sensor. Based on the

change in the sensed luminous level, the threshold for binarization is adjusted. Figure 3.8 shows the change of pixel intensities of iris region under different ambient light conditions and natural face orientations. The gray cross depicts the estimated iris centers.

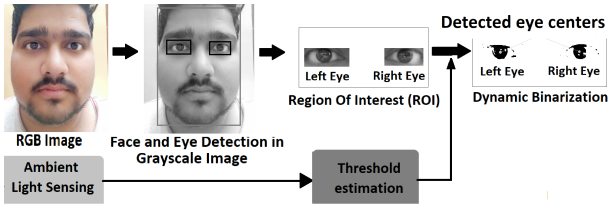


Figure 3.7: Ambient light sensing and eye center localization : The camera preview showing the user's face is captured and converted to grayscale image, from which, the eye region is extracted. Parallel to the eye region extraction, the ambient light is sensed and dynamic threshold is estimated for binarization. The extracted eye regions and dynamic threshold value are integrated and processed for eye center localization.

(a) (b) (c)

Figure 3.8: Eye center localization at different levels of ambient light and face orientations: (a) shows eye region of the user, looking straight in a bright room (b) shows the eye region of the user, with the face oriented towards right, in a moderately illuminated room, (c) shows the eye region of the user, looking straight, in a comparatively dark room. The eye centers are correctly identified in all 3 cases.

The estimation of varying thresholds is conducted by evaluating the luminance level of the BioID dataset[1] images as shown in Algorithm 2. The dataset contains 1521 images under a large variety of illumination, background, and face size; therefore, it can be considered to pre-learn the thresholds need to be used for a specific illumination level. In our approach, the ambient light sensor of the mobile phone captures the illumination level of the environment. These captured levels are compared to the selected ranges of luminance, and the corresponding estimated threshold is used for the eye center localization in the mobile application.

3.3.3 Gaze Gesture Tracking

The gaze gestures are tracked using a continuous and simple string pattern creation. This process is similar to the approach presented in [50]. However, unlike [50], there is no start or stop criteria once the gesture tracking is started. The gestures are tracked continuously throughout the length of the usage of the application. Firstly, each frame is processed to identify a valid pair of ROIs. If two ROIs, separated by a minimum distance (non-overlapping pair) in the X-direction of screen coordinates, are identified in the frame, the frame is considered as a valid frame. Only these frames are considered, eliminating any further processing of the

[1] https://www.bioid.com/facedb/ (accessed: Friday 11th August, 2023)

Algorithm 2: Estimation of threshold based on ambient light

Input: BioID images
Output: Thresholds for given luminous range ($AvgThresh$)
foreach $Image \in BioID$ **do**
  $C_{xlg}, C_{ylg} \leftarrow$ Marked X,Y coordinate for left eye
  $C_{xrg}, C_{yrg} \leftarrow$ Marked X,Y coordinate for right eye
  $ImgAvg \leftarrow$ Average pixel Intensity of entire frame
  $FaceAvg \leftarrow$ Average pixel Intensity of face region
  $EyeAvg \leftarrow$ Average pixel Intensity of left/right eye region
  $lx_{Image} \leftarrow .2 * ImgAvg + .5 * FaceAvg + .3 * EyeSAvg$
  $min \leftarrow 9999$
  for $i \leftarrow 0$ *to* 255 **do**
    $threshold \leftarrow i$
    $C_{xl}, C_{yl} \leftarrow$ Calculated X,Y coordinate for left eye using Algorithm 1
    $C_{xr}, C_{yr} \leftarrow$ Calculated X,Y coordinate for right eye using Algorithm 1
    $d1 \leftarrow (C_{xlg} - C_{xl})^2 + (C_{ylg} - C_{yl})^2$
    $d1 \leftarrow \sqrt{d1}$
    $d2 \leftarrow (C_{xlg} - C_{xl})^2 + (C_{ylg} - C_{yl})^2$
    $d2 \leftarrow \sqrt{d2}$
    if $d1 < min$ **then**
      $min \leftarrow d1$
      $thresh_{Image} \leftarrow threshold$; // minimum error
    if $d2 < min$ **then**
      $min \leftarrow d2$
      $thresh_{Image} \leftarrow threshold$; // minimum error
  Record $lx_{Image}, thresh_{Image}$ in $LxRep$
$R_u \leftarrow$ Defined upper limit of light range
$R_l \leftarrow$ Defined lower limit of light range
/* Ranges are set as: 0, 1-30, 31-60, 61-91, 92-122, above 122 */
$ctr \leftarrow 0, avg \leftarrow 0$
foreach $lx_{Image} \in LxRep$ **do**
  if $lx_{Image} <= R_u$ && $lx_{Image} >= R_l$ **then**
    $ctr \leftarrow ctr+1$
    $avg \leftarrow avg+ thresh_{Image}$
$AvgThresh \leftarrow avg/ctr$

rest of the invalid frames. Secondly, for the first valid frame after opening the camera, the $(X_{start}, Y_{start})_{left}$ and $(X_{start}, Y_{start})_{right}$ are estimated where $(X_{start}, Y_{start})_{left}$ is the start location of the iris center for the left eye and $(X_{start}, Y_{start})_{right}$ is the start location of the iris center for the right eye according to the screen coordinates relative to the detected ROIs respectively. Next, for each consecutive valid frame, $(X_{now}, Y_{now})_{left}$ and $(X_{now}, Y_{now})_{right}$ are estimated. The X_{now} of the detected iris location is subtracted from the X_{start}, and the Y_{now} coordinate of the detected iris location is subtracted from the Y_{start} for both the left and right eyes, to estimate the magnitude of the shift. The shift direction and associated string pattern is estimated according to Table 3.1. Lastly, the new $(X_{start}, Y_{start})_{left}$ and $(X_{start}, Y_{start})_{right}$ are updated with the current $(X_{now}, Y_{now})_{left}$ and $(X_{now}, Y_{now})_{right}$ after processing each frame.

This technique of gaze gesture tracking provides a better and more accurate estimation of user's gaze over tracking the point of gaze. The main reason behind this is its simplicity in calculation and allowance of free head and device movement. It can be stated that even if a false gesture is estimated due to the sudden movement of user's head position, face orientation or

Table 3.1: Shift Direction and Associated Symbols

Shift in X direction	Shift in Y direction	Shift direction	Symbolic representation
=0	=0	No shift	X
>0	=0	Left	L
<0	=0	Right	R
=0	>0	Top	T
=0	<0	Bottom	B
>0	>0	Top Left	M
>0	<0	Bottom Left	P
<0	>0	Top Right	N
<0	<0	Bottom Right	O

device's frame, the subsequent gesture strings will be related on the preceding coordinates only. Hence the following patterns should match the intended sequence. The obtained sequence of tracked gaze gestures is recorded and each record contains the following structure: $< frame, Symbol, E_{index} >$, where $frame$ corresponds to the frame number and $Symbol$ is the gaze shift direction depicted in Table 3.1.

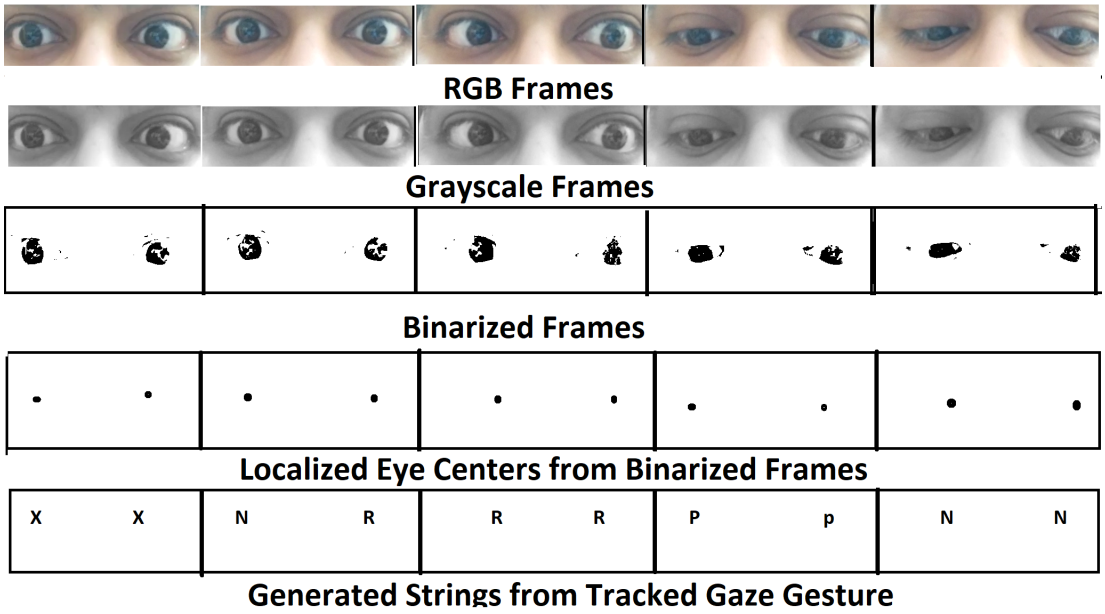


Figure 3.9: Eye gesture tracking and string generation: The first row of RGB frames show a sequence of eye movements by the user. These regions are converted to grayscale in the second row. The third row shows the dynamically binarized representation of the grayscale frames. Eye centers are generated from the binarized frames in the fourth row. The initial location of the centers are represented by 'X' and the consecutive locations are tracked by their relative positions to their previous frames and assigned to the appropriate string symbols in the fifth row.

Figure 3.9 shows a sequence of captured frames depicting user's eye movement and the process of generating strings based on tracked gaze gesture. Here, the left eye movement trajectory generates a sequence of XNRPN , while the right eye generates a string of XRRPN.

3.3.4 Region of Gaze Tracking

The region of gaze tracking is executed by mapping the approximate gaze point on one of the four regions of the mobile screen, as shown in Figure 3.6. $\{0, 0\}$ represents the origin of screen coordinates. The orientation of the device is set according to the given figure for wide area display of video frames. The region of gaze tracking is performed according to Algorithm 3. The tracked region of gaze for both left and right eyes, obtained as $leftS, rightS$ are further

Algorithm 3: Region of Gaze Tracking

Input: The records obtained from calibration stored in Template file (section 3.3.1) and localized eye center from each frame (Algorithm 1).
Output: The estimated region of gaze ($leftS, rightS$).

$l \leftarrow 0, r \leftarrow 0, minL \leftarrow \inf, minR \leftarrow \inf, leftS \leftarrow \phi, rightS \leftarrow \phi$
foreach $record \in Template$ **do**
  /* left eye coordinates & labels */
  if E_{index}=="$Left$" **then**
    $leftCoord$[l][0]$\leftarrow C_x$
    $leftCoord$[l][1]$\leftarrow C_y$
    $leftLabel$[l] $\leftarrow Label$
    $l \leftarrow l$+1
  else
    /* right eye coordinates & labels */
    $rightCoord$[r][0]$\leftarrow C_x$
    $rightCoord$[r][1]$\leftarrow C_y$
    $rightLabel$[r] $\leftarrow Label$
    $r \leftarrow r$+1
foreach $frame_i$ **do**
  foreach E_{index} **do**
    if E_{index}=="$Left$" **then**
      for $j \leftarrow 0$ *to* $length(leftCoord)$ **do**
        $d \leftarrow (C_{xi} - leftCoord[j][0])^2 + (C_{yi} - leftCoord[j][1])^2$
        $d \leftarrow \sqrt{d}$
        if $d < minL$ **then**
          $minL \leftarrow d$
          $leftS \leftarrow leftLabel[j]$
    else
      for $j \leftarrow 0$ *to* $length(rightCoord)$ **do**
        $d \leftarrow (C_{xi} - rightCoord[j][0])^2 + (C_{yi} - rightCoord[j][1])^2$
        $d \leftarrow \sqrt{d}$
        if $d < minL$ **then**
          $minR \leftarrow d$
          $rightS \leftarrow rightLabel[j]$
  Record $leftS, rightS$

analyzed to draw a rectangle on the focused segment of the screen.The top left (R_{TL}) coordinates and bottom right (R_{BR}) coordinates of the rectangle to be drawn are evaluated based on Table 3.2, where *Width* and *Height* refers to the frame width and frame height. The $leftS$ and $rightS$ in a single frame are expected to be the same as both the iris movement should have nearest points belonging to the same label. However, in some cases, computational inaccuracy might lead to the difference in values of $leftS$ and $rightS$. In these cases, either one of the tracked location is considered for a frame.

Table 3.2: Coordinates of Gaze Region

$leftS$ or $rightS$	R_{TL}(x)	R_{TL}(y)	R_{BR}(x)	R_{BR}(y)
TL	0	0	Width/2	Height/2
TR	Width/2	0	Width	Height/2
BR	Width/2	Height/2	Width	Height
BL	0	Height/2	Width/2	Height
C	Width/2	Height/2	Width/2	Height/2

3.4 Mapping Gaze Gesture and Region to MOOC Video Object

The final phase, for estimating the user's attention, comprises of mapping the gaze gesture and the region to the tracked video object. We assume, each video contains one to three prime objects which require to be focused on, by the user. For simplicity, we restrict our mapping to one prime object of the video. In most MOOC courses, the prime object can be assumed to be the lecturer (person). However, the system can be adapted for annotating the prime object while uploading the video, in case it differs from a lecturer.

The estimation of user's attention level is a 3-fold process that combines the output from the previous phases of the proposed model. Figure 3.10 presents the overview of the 3-fold evaluation technique. The three levels of evaluations rely on three basic assumptions in this work which encircles the core notion about the correlation between high level cognitive attention and low level visual cues [211, 55]. The fact that an automatic pruning of visual data is performed by visual systems, ensures the visual attention on only a relevant segment of the visible image, and in turn, promotes learning [251]. The three principles considered in this approach are : Observation, Tracing and Focus.

3.4.1 Observation

The **first** level of estimation relies on the assumption that *the user has to observe the video, either constantly or periodically, to be considered as attentive*. The objective of this level is to identify whether the learner is watching the video at least occasionally. Based on this criteria, each frame is analyzed to obtain valid ROIs. If and only if a valid pair of eyes are found, it is inferred that the user is looking at the screen and is "Attentive". In cases where eye regions are not detected, the user is treated as "Inattentive". However, in some frames, even though the user is looking at the screen, valid eye pair might not be detected due to the improper lighting condition, calculation errors or inaccurate classification of faces. To avoid such scenario and allow periodic glancing, a window interval is maintained. This estimates the level of attention based on whether or not, the user is looking at the screen for at least once during the window. For videos, where plain texts are the prime objects (or no specific prime object), this estimation technique is helpful, as text movement on screen is minimal.

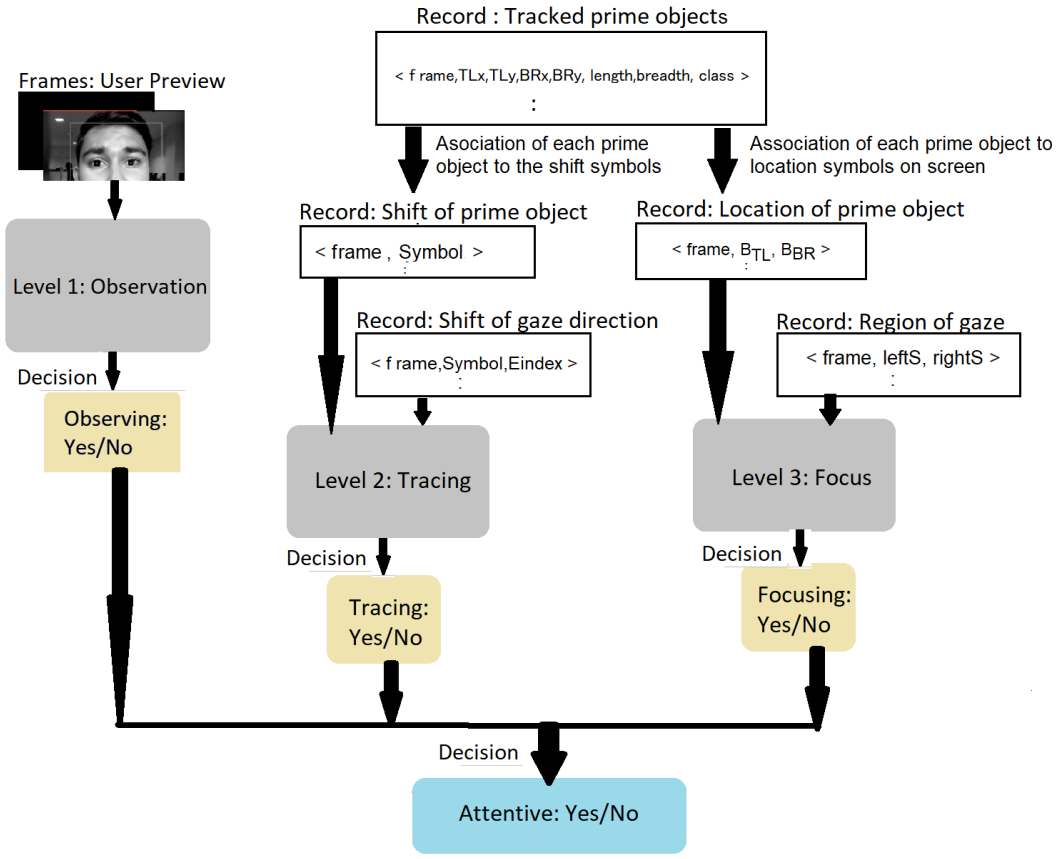


Figure 3.10: Attention estimation based on the 3-fold evaluation : Level 1 evaluation utilises the rendered frames with valid eye pairs. Level 2 and level 3 evaluations use textual data extracted during prime object, gaze gesture and region of interest tracking. Each level results to a certain binary decision regarding attentiveness depicted by scores or on screen alerts, which jointly decides the final attention level of the user (scores).

3.4.2 Tracing

The **second** level of estimation is based on the assumption that *the user must visually follow the prime object in the video to be considered as attentive.* This process aims at tracking the visual gesture of the learner to identify whether the learner is observing the relevant object in the video.To satisfy this criteria, the gaze gesture tracking is considered. From the tracked locations of the prime objects in the MOOC video, the shift symbols for the prime objects are fetched for each frame. The corresponding shift symbols derived by tracking the gaze gesture of the user's eyes are derived. If, either the left eye shift symbol or the right eye shift symbol for a frame matches with that of the shift direction of any of the prime object in that frame, the user is marked as attentive. However, due to free head and device movement, not all shift symbols match with the corresponding video object shift symbols.

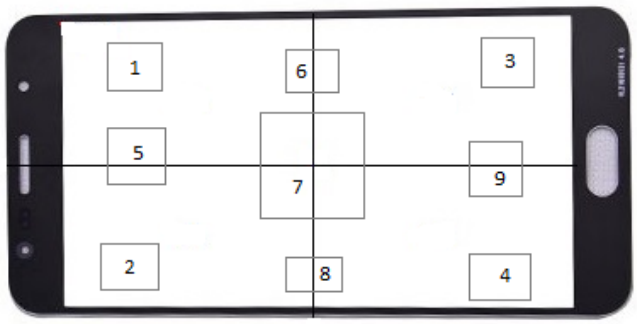


Figure 3.11: The locations of video objects on the segmented regions of the mobile screen showing all possible locations any video object (numbered) can hold on the mobile screen.

Table 3.3: Mapping Objects to Gaze Region

Object no. (Figure 3.11)	B_{TL}	B_{BR}	$leftS$ or $rightS$
1	TL	TL	TL
2	BL	BL	BL
3	TR	TR	TR
4	BR	BR	BR
5	TL	BL	TL/BL
6	TL	TR	TL/TR
7	TL	BR	TL/BR/TR/BL/C
8	BL	BR	BL/BR
9	TR	BR	TR/BR

3.4.3 Focus

The **third** level of evaluation considers the assumption that *the user might look at the screen and randomly gaze at different points, leading to some level of gesture matching. For the user to be attentive, he/she must focus on the prime object in the video.* This evaluation is an extension of the previous evaluation stage that further strengthens the estimation of gaze gesture and accounts for the gaze location which is important in the presence of multiple overlapped or discrete prime objects. To further ensure that the user's gaze is focused on the prime video object, a gaze region mapping is performed. Figure 3.11 displays the possible location of the video object on the device's screen. The figure provides an estimate on how the object orientation can be mapped to the gaze region. For object 5, since the top left point of the bounding box lies in top left region of the screen and the bottom right corner lies in bottom left region of the screen, the user's gaze region should be either top left or bottom left. Table 3.3 tabulates the possible object locations displayed in Figure 3.11, the coverage of their bounding boxes (top left and bottom right regions as B_{TL} and B_{BR}) and the corresponding allowed gaze regions ($leftS, rightS$) for the user to be marked as attentive. If for a particular frame, the $leftS$ or $rightS$ falls in the permitted set of values, according to the table, it is inferred that the user is focusing on the prime object in the video.

Figure 3.12 shows the visual representation of mapping prime objects to gaze location,

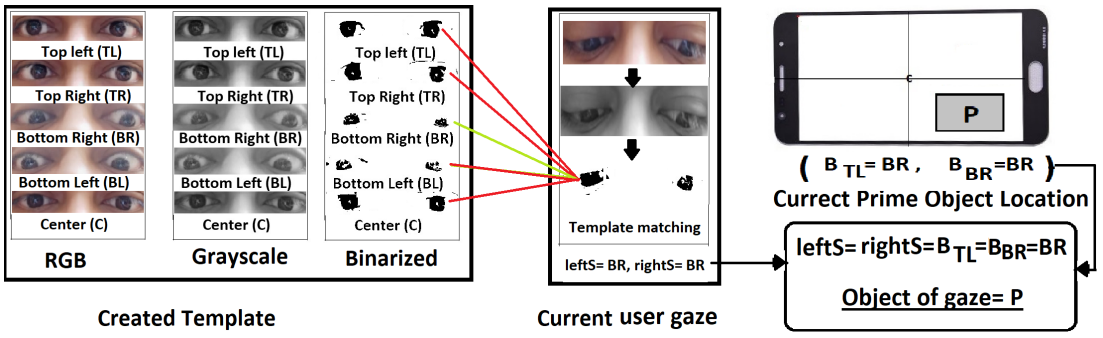


FIGURE 3.12: Mapping prime object location to gaze location : In this figure, the user is looking at the bottom right corner of a Video frame and the prime object in that video frame is also present at the bottom right corner(i.e. Top left and bottom right locations of the prime object's bounding box falls inside the bottom right region). Hence B_{TL}=B_{BR}=BR. The current eye center is mapped to obtain the nearest known eye center from the templates. In this case, the current eye center is nearest to the Bottom right eye center in the template list. Hence the current centers are associated with the BR label (leftS=rightS=BR). Since the $leftS$ or/and $rightS$ matches with B_{TL} or/and B_{BR}, it is concluded that the user is focusing at the prime object in this frame.

according to the third level of evaluation. The template matching shown in the figure, is in terms of obtaining the eye center coordinates from the template records, that is nearest to the eye center of the current captured frame. In this case, the location of the prime object matched with that of the gaze location of the user, hence inferring that the user is looking at the prime object.

3.4.4 Design of Dynamic Window for Handling Context Switches and Multitasking

The usage of window frames alleviates the requirement of continuous gazing as an indicator of attentiveness. In case of mobile platforms, continuous gazing is rather unrealistic, unlike in traditional classrooms. Mobile based lectures can be attended with periodic multitasking like taking notes, simultaneous referencing to books or reading relevant materials and occasionally listening to lecture audio only. While temporary shift of visual focus from the video can still indicate attentiveness, in most cases, a learner prefers to pause the lecture while he/she is focusing on a note or book. Hence, it is expected that even if the visual focus is temporarily shifted from the video, the learner will continue looking at the video after a certain period. This occasional glancing is allowed and facilitated by the use of window frames.In our implementation, the first level of evaluation uses a window of 50 frames. If the user's eyes are not detected within this interval, the first attention level is depleted by one from the current gaze attention level. For the second level of evaluation window size of 20 is maintained. If the shift symbols of tracked eye gesture in 20 frames, do not match with the corresponding shift symbols of the video object, the second attention level is depleted by one from the current gaze gesture based attention

level. Since, short lecture videos are considered in the experiments, the window frames are proportionally less. For longer videos, these videos can be expanded accordingly to facilitate appropriate occasional glancing.

The maximum gaze and gaze gesture attention levels are set to Max_{gz} and Max_{gs} at the beginning, where Max_{gz}= (Number of frames in the video)/(Window size for gaze estimation) and Max_{gs} =(Number of frames in the video, where eye centers were detected)/(Window size for gesture estimation). The third level of evaluation is not mapped to a statistical output, rather produces a continuous alert to the user during the course. The overall estimation of attention is hence based on the following equation: ET=$Max_m - (Max_{gz} - ET_{gz}) - (Max_{gs} - ET_{gs})$

where, ET, ET_{gz} and ET_{gs} are the overall estimated attention, estimated attention based on gaze and estimated attention based on gaze gesture, respectively. ET_{gz} and ET_{gs} are obtained from first and second levels of evaluations. For our experiment, Max_m=Max_{gz}+Max_{gs} and finally the estimated ET has been scaled down within the range of 0–100. The differences $(Max_{gz} - ET_{gz})$ and $(Max_{gs} - ET_{gs})$ signifies the points lost due to inattentiveness based on gaze and gesture. Subtracting these differences from Max_m evaluates the total score, based on the points lost in gaze and gesture tracking.

3.4.5 Dissecting Gestatten

The accuracy of the eye center localization algorithm used in the proposed approach greatly relies on the variation in sensed ambient light level, using the inbuilt sensor. We examined the binarization technique with a static threshold value on the BioID dataset, containing 1521 different faces under various lightning conditions. The evaluation was based on the error levels of the better, worse and average eye detection, according to equations : $error \leq min(d_l, d_r)/d_m$, $error \leq max(d_l, d_r)/d_m$ and $error \leq (d_l+d_r)/2d_m$, where d_l, dr, d_m are the Euclidean distances between measured and marked left eye centers, right eye centers and the distance between the marked left eye and right eye centers respectively. For fixed thresholds, the worse, better and average eye hits rates for error less than or equals to 0.25 are 80.21%, 99.45% and 81.59% respectively. However, the respective accuracy for error less than equals to 0.1 significantly decreases to 56.86%, 88.46% and 64.01%. This leads to the consideration of varying threshold levels based on the ambient light of the user. To address this problem with static threshold, we have used dynamic threshold levels, based on the instantaneous ambient light, as discussed in the previous section. Using these dynamic thresholds, the accuracy of worse, better and average eyes significantly increases to 78.84%, 99.17% and 81.59% for error less than or equals to 0.1 and 82.14%, 99.72% and 82.15% for error less than or equals to 0.25. The respective accuracy shown in [239] are 93.4%, ≈100% and ≈95% for error less than or equals to 0.1 and 98%, ≈100% and ≈100% for error less than or equals to 0.25 which are comparatively higher. However, Gestatten, that depends mostly on relative shifts, rather than

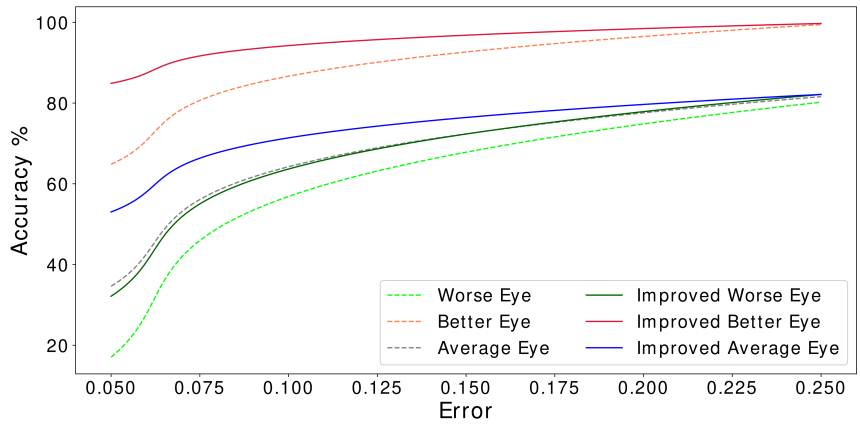


FIGURE 3.13: Analysis of localization approach with static and dynamic thresholds

accurate gaze points, is significantly lightweight. Figure 3.13 compares the performance of localization approach for constant and varying thresholds on the BioID dataset. We observe that the dynamic threshold approach used in *Gestatten* significantly improves the gaze detection accuracy.

3.4.6 Benchmarking Gestatten

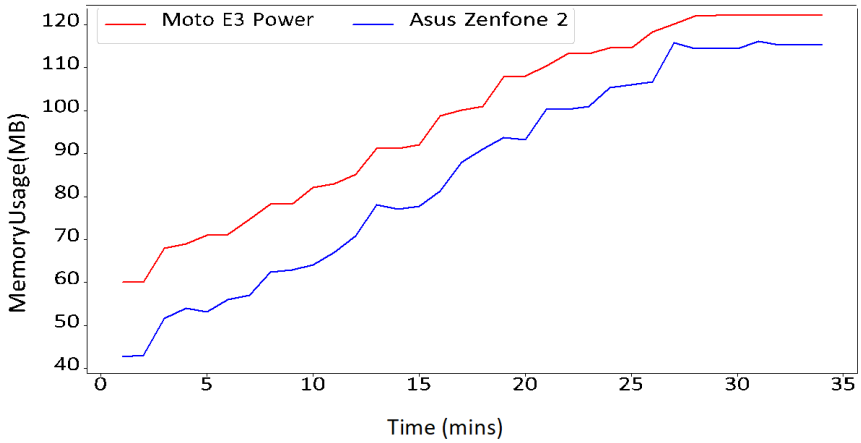


FIGURE 3.14: The android application profiler for Memory usage

Memory and CPU usage: Figure 3.14 shows the memory usage of the Gestatten. The performance of two smartphones has been measured in this approach. It can be seen that the maximum memory usage is 122.2MB. To visualize the maximum memory usage by the application, the internal visual record files are retained throughout the application usage. In practical scenario, the memory usage will be much less as these files can immediately be auto deleted after viewing each video shown by the application. This is due to the fa ct that these

textual data are not used any further, once the application scores are generated immediately after each video ends. The CPU usage by the Moto E3 Power is 18-24% and that of Asus Zenfone 2 is 27-33%, making it significantly lightweight. These parameters are measured by android studio application profiler.

Battery Consumption: For estimating the battery usage by the application, the smartphones were initially fully charged. The applications were used continuously on both the smartphones till the battery charge dropped to 20%. The Asus Zenfone 2 could run the application continuously for 4hrs, while the Moto E3 Power ran the application for more than 5hrs. The smartphones were again charged to 100% and only the videos shown by Gestatten were continuously played on the same devices, till their battery charge dropped to 20%. This experiment was to test the battery usage of the standalone videos only, without their involvement in Gestatten. The Asus Zenfone 2 could display the videos continuously for 6hrs, while the Moto E3 Power displayed the videos continuously for 6hrs 50minutes. This statistics shows the low battery drainage by the operations involved in Gestatten.

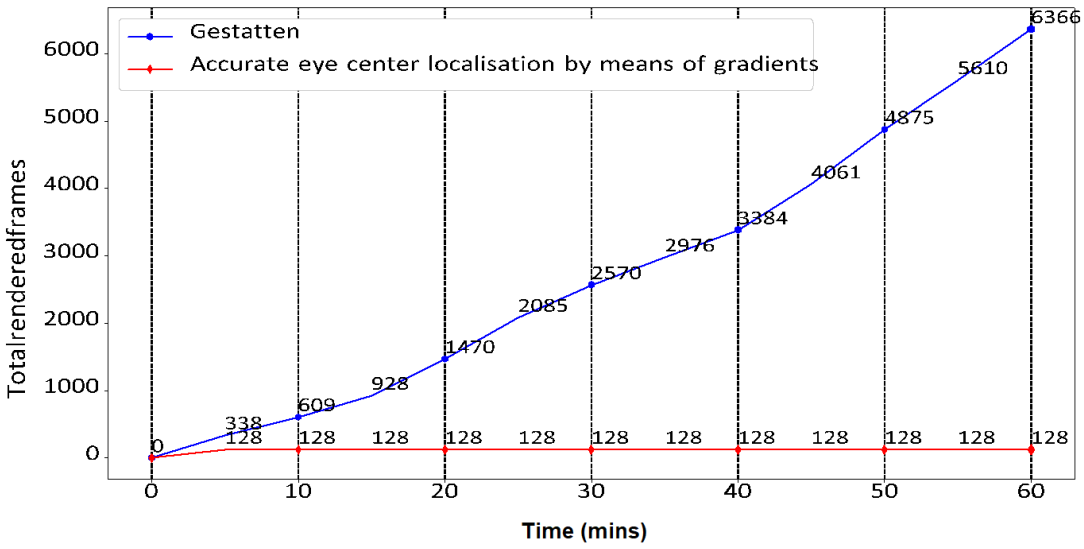
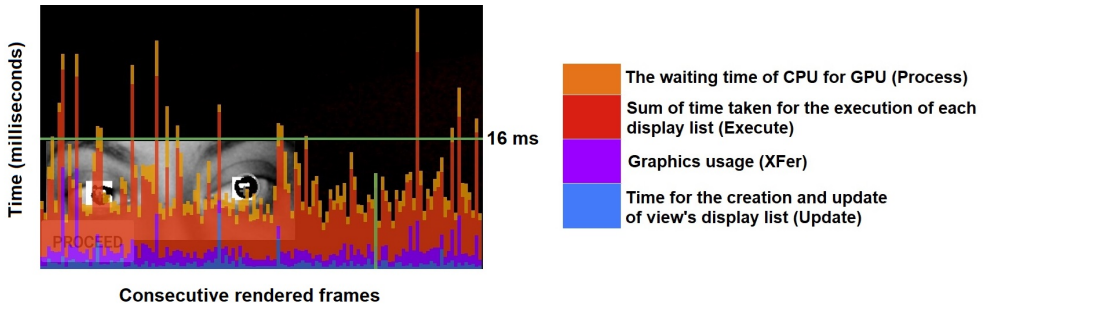


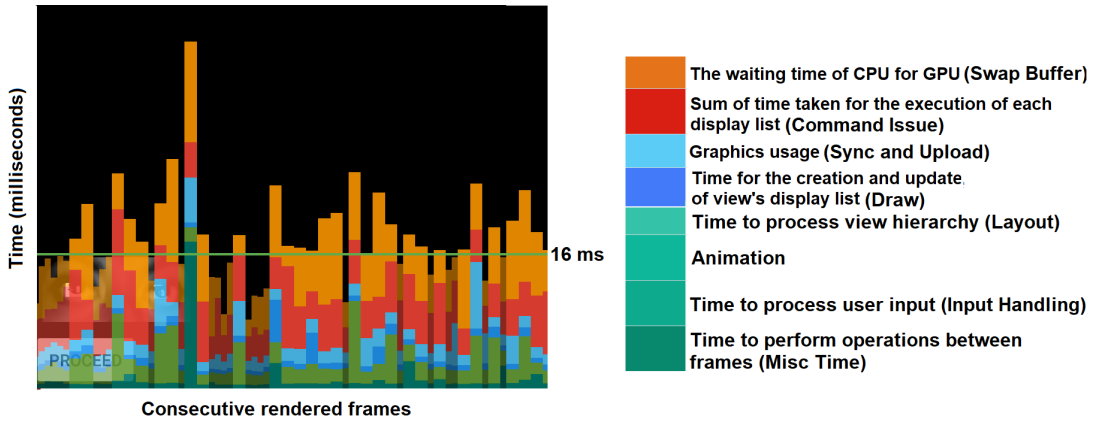
Figure 3.15: Comparison of eye tracking algorithms used in Gestatten and Accurate eye center localisation by means of gradients [239]

3.4.7 Baseline Comparison

Finally, to establish the feasibility of using a lightweight model like Gestatten, for mobile devices, we compare the eye center tracking algorithm proposed in Gestatten, with that of one of the bench marking algorithms presented in [239], which also uses pixel level calculation for locating the eye centers. The comparison is made due to the following similarities between Gestatten and [239]: (1) Both works deal with pixel level calculations on eye region only and



(a)



(b)

Figure 3.16: Profiling GPU rendering of Gestatten in mobile device : The graph depicts the time (milliseconds) for each frame to be rendered by the application within 16milliseconds depicted by the green horizontal line, in (a) an android device with Android version 5.0, 1.8GHz CPU (b) an android device with Android version 6.0, 1000MHz CPU.

claim to be computationally simple. (2) Similar to [239], Gestatten uses BioID dataset and similar metrics for error calculation (accuracy of eye centre estimation) like worse eye, average eye and better eye. Moreover, the comparison shows why, the highly accurate pixel based eye centre estimation that are computationally simple enough for other computing devices, might not work for smartphones.Both the algorithms are implemented for android devices using Android studio. The in-device run-time frame rendering is analyzed using profiler for GPU rendering using bars on screen (Figure 3.16) and the back-end run time frame rendering information is analyzed by using dumpsys. Figure 3.15 presents the comparison graph. The total rendered frames indicate the initial frames before the camera is turned on and only the total number of frames where valid eye pairs were detected and processed to evaluate their centers, after the camera is turned on. The number of total frames shown in the graph is based on the natural viewing pattern of the user, consisting of occasional context switching and multitasking, for 1 hour. The aim of this task was to identify whether the eye center localization module of Gestatten and that of [239] can work in real time, rather on the level of attention paid by the

user. The number of frames is proportional to the constant on screen gaze time of the user. The performance of the eye center localization module was examined for 1hour, using intervals of 5minutes. This lead to the problem of frame congestion for the algorithm used in [239], due to its intensive pixel level calculations. From the graph, it can be seen that throughout the period of 1 hour, the rendered frame has a constant value of 128. This is due to the fact that at the 128th frame, a valid eye pair was detected and the algorithm was initiated by the mobile application for estimating the centers. The frame took more than 1 hour to be processed by the algorithm, hence rendering no further frames. It took 67 minutes, to process this single frame by the algorithm, whereas Gestatten worked in real time. This proves the feasibility of Gestatten in mobile devices, as it requires simple calculations, optimal for devices with restricted powers. Figure 3.16 and 3.15 validates the lightweight nature of our approach and depicts that the eye center localization algorithm of Gestatten can work seamlessly for mobile devices. In Figure 3.16, the green horizontal line depicts the standard benchmark of 16ms (default rendering time of each frame is set to 16ms). This is due to the fact that the modern smartphone's have an inbulit refresh rate of 60 frames per second for seamless display of motions. Even though 24-30fps is sufficient to provide smooth visual effect for basic frames, for frames with high graphical effects, human brains can perceive consistent and fluid video motions at 60fps. Hence, it is the default benchmark in GPU rendering profilers. In the figure 3.16a, we can see that most of the frames are rendered within 16ms. Even for figure3.16b, we can see that the elevated bars are only due to the GPU usage, however, the Misc time indicating the execution time for operations between the frames by the application, in significantly within the limit of 16ms. Also, additional processing is required to display the extracted eye region in each frame, that adds to the height of the bars. Figure 3.16 shows the suitability of our approach for mobile devices, working at 50-60fps. The proposed approach is computationally simple and hence is faster than several other existing mobile based approaches which allows lower frame rates of 10-30 fps [100, 171, 260].The performance can further be optimised by eliminating the display of the eye region in the interface design of the application.

3.5 Experiments and User Study

Gestatten is a standalone android application that requires no separate Internet access, can display videos and use the device's front camera to preview the user's face, while simultaneously processing the facial images. The model has been implemented using Android studio and has been tested in several mobile devices including Asus Zenfone 2 with Android version 5.0, Moto E3 Power with Android version 6.0.

3.5.1 Experimental Methodology

We have tested *Gestatten* over 48 participants divided into two sets of experiments. The participants are primarily university undergraduate and graduate students, in the age group of 19–30 years. The ethical considerations for these experiments have been followed and participants have been given a token incentive of $3 Amazon Gift card for participation in each experiment. In overall, the participants have observed video segments from some pre-selected MOOC courses over a mobile with a player that incorporates *Gestatten* as an add-on module. We use two different evaluation technique as discussed next.

Method-1: Assessing the Correlation between the System Predicted Attentiveness Score and the Learner's Performance after Completing the Video Course

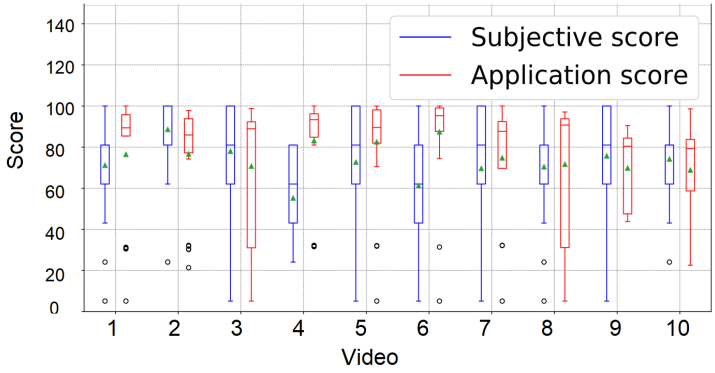
The **first set** of subjective evaluation verifies the results of the application using a Multiple Choice Question (MCQ) based evaluation on 28 participants for the generation of ground truth. This technique also correlates the attention level of a learner to his/her degree of comprehension. It is evident that for a short video lesson, the participants will remember the answers, at least for a short time, if not for a long time, only if they have paid attention to the video, as established in the existing literature [263, 103, 273, 262]. We try to find out, how visual attention can account for such comprehensions, and this provides an indirect measurement of the success of our application. If we can observe a high correlation between the predicted attention level by *Gestatten* and the performance of the learner in the test just after watching the video, we can say that the system does a good job in understanding the attentiveness of the learner. The hypothesis for this experiment is to check how the system predicted attention level correlates with the user performance in an "attentiveness test", pertaining to various user studies from the literature that shows high correlation among the two. From this correlation, we show how good the system can predict the attentiveness of the user. In this technique, each participant was shown a sequence of 10 different MOOC videos of average duration of 3 minutes and belonging to different field of studies, in succession. The videos were selected in a way that each contained at least one prime object, except videos 4-7, which had no fixed prime object. Videos 4 and 6 contained mostly textual and audio based information. The videos used for the experiment are mostly lecture videos containing textual as well as visual presentations. The topics covered in these videos varies from Engineering subjects to Political Science and History lessons. We also believe that, for a 3 minutes demonstrative video, the user should visually follow the lecture material to have good interpretation of the topic being taught. After each video, the participant was provided with 5 questions based on the shown video. Each question of 1 mark belonged to a certain difficulty level in ascending order. Hence, the subjective scores ranged from 0 to 5. The MCQs were directly from the video content and have been answered following each video, so that the participant can answer them only if they have observed it in

the video. Further, they have been discouraged to guess an answer. In addition, it has been ensured that the participants have observed those videos for the first time to eliminate any bias during the MCQ answering. We have used external stimuli (such as loud noise through music player, dropping the sound of the video, etc.) to distract the attention level of 5 participants out of the 28 participants, to emulate an environment when the learner is not attentive. For each of the videos, the mobile application generated the gaze gesture based scores and the gaze based score, thus finally generating the estimated score ET. The subjective scores were scaled up and normalized to the range of the application generate score.

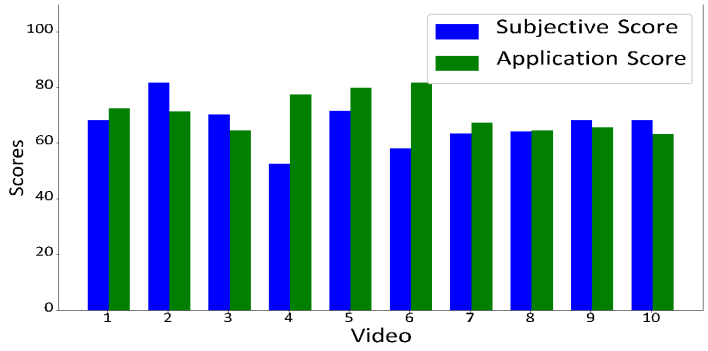
Method-2: Validation Through Independent Observers

The **second set** of subjective evaluation has been conducted with 20 more participants under different environments such as low illumination, at different rooms, etc. The objective of this user study was two-folded: to examine the usability of the application and to determine whether "human perception" matches with the "machine perception". For this second objective, we have matched the agreement among the annotators.In case of manual agreement, the score given by the annotator is compared with the score generated by Gestatten. the hypothesis behind the second experiment is to find out the agreement between the machine generated score and human evaluated score. The machine may fail in various adverse conditions; so we may not be able to draw any conclusion from the first set of experiments (attentiveness test) under those adverse settings. Therefore, from the second experiment, we have shown the impact of such adverse conditions, like low light intensity, absence of prime objects, etc., when Gestatten may fail to provide a reliable score. Moreover, further analysis of the manual evaluation has been performed to explore its difference from the application generated results. The task was performed in groups of 3 to 5 participants. For each group, one participant was initially requested to use the application (user), while the rest manually evaluated the attention level of the user (evaluators). This procedure was repeated for each participant of the group and for all the groups, where each volunteer had to play the role of the user once and evaluator for at least twice, depending on the group size. Each evaluator was assigned the task of observing the user and identifying her point of gaze in the video. For mobile devices, with restricted screen dimensions, this is a difficult task to perform, as the point of gaze can be confused between multiple proximal objects on the screen from the manual point of view. Moreover, it is practically impossible for the evaluators to note the point of gaze for every frame for a video played at about 60 frames per second. To address these problems, videos with one prime object and one to two secondary objects shown in each frame, are selected. Also, an interval of 10 seconds between two consecutive evaluations has been fixed for the evaluators, eliminating the requirement for frame-wise monitoring. This might lead to another problem where, discontinuous monitoring can result in disregard of some instances, where the user might

have lost attention within the time frame of 10 seconds. To avoid this, different evaluators were assigned to different random time intervals. After the observation was made according to the time intervals, the evaluators marked the user from 0-100, based on the amount of time, the user paid visual attention to the video objects. The evaluators performed a strict monitoring on the user by sitting at different locations within a close proximity of the user. To facilitate this process and eliminate incorrect readings, the seating were arranged in such a way that both the user's eyes and the mobile screen could be simultaneously visible. Also, since the number of prime objects were only restricted to 3, it was easier for the evaluators to marks the object of gaze. The evaluators continuously marked whether the user gazes towards the video, or looks in a different direction other than the video objects. We believe that a single evaluator might not be trusted to assess the performance. To ensure reliability, multiple evaluators are employed. For the videos to be displayed on full screen, the landscape mode is used in our experiment.This is due to the reason, that for portrait mode, the screen dimension is further restricted, making the subjective observation nearly impractical. However, since the shift and location of the prime objects are relative to the device, the application result would not be practically affected by the orientation of the screen.



(a) Score distribution



(b) Average score comparison

Figure 3.17: Video-wise Comparison of Subjective and Application Generated Scores

3.5.2 Results

Here, we discuss our salient observations from the two different types of experiments as discussed above.

Correlation Between the Subjective Score (Test Performance) and the Application Generated Score (Experiment-1)

In the **first set** of subjective evaluation, for each video, the subjective scores of 28 participants are compared with the respective application generated estimated scores (ET) and has been depicted in Figure 3.17 by the parameters like range, quartiles, means and medians (Figure 3.17a). The average subjective and application generated scores for 28 participants in each video has been compared in Figure 3.17b. It can be observed, that for all videos except 4 and 6, the subjective scores and application generated scores are comparable and has a minimal deviations. The inaccuracy of the scores for videos 4 and 6 can be attributed to the absence of prime objects and projects the necessity of gesture based attention estimation. For videos where no particular object of interest is present, the participant was free to visually follow any part of the screen. The gaze gesture based estimation is not functional in this case. The application's attention estimation hence, only depended on the gaze of the participant for these cases. For these videos, even though the students stared continuously at the screen, making them score highly in application generated estimation, their actual apprehension of the subject lacked as the textual and audio information were not adequate enough to be comprehended by the participants, which was reflected in their subjective evaluations. This proves, not only the applicability of our approach, but also proves the efficacy of explanatory MOOC lectures over textual videos.

Figure 3.18 presents the comparison of participant wise average estimations. For each participant, the average subjective score across 10 videos has been compared with the average application generated scores. The average absolute error between the subjective and estimated score generated by the application is found to be 8.68. Further, we observe that *Gestatten* can identify both the learners who are attentive as well as no attentive as the average subjective scores matches with that of the application generated scores. Low scores have been observed for the last four users (for both the application generated scores as well as the subjective scores) for whom we have used external stimuli to divert their attentions.

3.5.3 Analysis of Application Generated Score versus the Score Given by Evaluators (Experiment-2)

Table 3.4 presents the manual and application based results for the participants from five different groups in the **second set** of subjective evaluation. Different groups of users have been

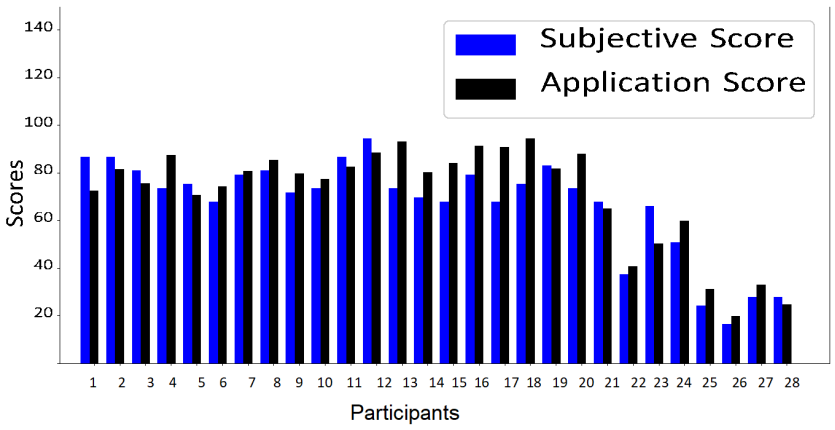


Figure 3.18: Participant-wise Comparison of Subjective and Application Generated Scores

evaluated at different environments as shown in the table. Here, we have grouped the scenarios based on the evaluation environments, and we observe that the disagreements between the human annotator and the system are because of unfriendly environments, such as low lighting, absence of prime objects in the video, etc. It can be observed that, for users 1-5 and user 12, the application failed to produce any estimation of the gaze gesture based attention level. The ambient illumination for all these cases were set to a low value where, the camera of the device and hence the algorithm could not capture and identify the eye center of the user's face for each frame. However, in these cases, the eye regions were still correctly identified and an estimation of attention level, based on whether or not, the user was looking at the screen, was produced by the application. It can be seen, that for group 4 and 5, the average manual evaluations has less deviation from the application generated results. The video displayed for this group had distinct primal object which had comparatively less movement in the video and location wise well separated from the other video objects. Also, angle of observation affected the manual evaluations greatly. The manual evaluation in group 5 produced the closest results. This can be attributed to its small size. as the group size increases, manual evaluations deviate greatly, proving the necessity of personalized and automated models to monitor the attention levels of individual students. Moreover, the huge deviation of subjective evaluations for individual user proves the inefficiency of manual estimations. User 15 was evaluated for 90 seconds that covered 47.36% of the total video. The user preferred to stop the video instead of completing it. The subjective scores were based on the evaluator's perception. To generate the ground truth for proving the correctness of the application generated results, 11 participants mentioned their object of gaze for 20-40 seconds. During this time, their faces and application screens were simultaneously recorded. The application continuously generated the point of gaze (the object at which the user is looking at, based on the region of gaze), while the face of the user proved the correctness of the dictated point of gaze by the user. The dictated point of gaze was

Table 3.4: Subjective Evaluations and Application Results Under different Environments (Eval indicates Evaluator)

User	Env.	Eval 1	Eval 2	Eval 3	Eval 4	Gaze	Gesture	Comments
1	Textual video	65	90	60	X	95	X	The high deviation between subjective and application score can be attributed to the lack of prime objects in textual videos and group size. For larger groups, observation and manual evaluation gets erronous and difficult.
2		60	70	60	X	95	X	
3		60	50	40	X	95	X	
4		50	60	50	X	98	X	
5		80	80	80	X	94	X	
6	Medium low light	54	45	50	79	97	94	The deviation in subjective and application score is due to the poor lighting condition where the subjective evaluators faced difficulty in following the user's eye movements.
7		55	55	75	X	95	100	
8		70	75	70	X	95	84	
9		80	82	85	90	96	98	
10	Low light	40	8	30	X	97	99	The deviation in subjective and application score is due to poor lighting condition where the subjective evaluators faced difficulty in following the user's eye movements.
11		70	90	40	X	99	87	
12		60	X	X	X	95	X	
13		50	45	X	X	98	94	
14	Normal environment Room-1	90	90	95	X	98	100	The similarity between subjective and application score is due to proper lighting condition and presence of prime objects where the subjective evaluators could easily track the point of gaze of the user.
15*		90	90	90	X	99	100	
16		70	80	85	X	97	99	
17		70	50	79	X	97	94	
18	Normal environment Room-2	89	97	X	X	98	93	The similarity between subjective and application score is due to the proper lighting condition, presence of prime objects and smaller group size where the subjective evaluators could easily track the point of gaze of the user.
19		86	98.5	X	X	99	88	
20		95	95	X	X	97	99	

matched with the application generated point of gaze. The results showed that 76.31% match was generated in average, for the dictated and generated values.

3.6 Discussion

In this section, we dicuss some of the limitations of Gestatten and their remedies as future scopes.

1. In Gestatten, we focus on developing a lightweight attention estimator that can be operated seamlessly on a mobile device. While most of the existing works rely on techniques like gradients, glint detection, we employ simple binarization of grayscale eye frames and centroid calculation. Apart from this the novelty of our approach lies in terms of incorporation of ambient light sensors, prime object tracking and trajectory mapping. However, the variation in iris colors has not been considered for our experiment. We

believe that the iris/ pupil pixels are mostly darker than the surrounding sclera pixels and the threshold based binarization will convert them into black pixels for centroid calculations.

2. The application currently needs the videos to be locally added to the device along with their tracked prime object trajectories that are obtained in an offline module. In future, for practical deployment, we aim at eliminating the requirement of local storage and use cloud services, so that different MOOC organizers can simultaneously upload their videos for Gestatten to stream them. Although significant result has been obtained, accuracy of region of gaze algorithm is affected by free head and device movements. This can be addressed by considering the user's face location while estimating the nearest calibration point.

3. The subjective evaluation of the model purposely avoids in-video prompts and quizzes for estimation, and instead relies on quizzes only after the completion of a video. This is due to the fact that in-video prompts leads to the distraction of the learner, not only visually, but also mentally. While prompts and scheduled tasks can estimate the learner's cognition, they practically also diverts the concentration. We tend to make the evaluation of attention more seamless and automated for the learner, which can be achieved using visual cues. To promote attentiveness, the users were periodically shown the instantaneous level of their attention on screen. One disadvantage of relying on visual trajectories is that the learner may choose to rely on the audio. However, this is unlikely to happen in majority of the cases as video contents display information which can mostly be comprehended through simultaneous hearing and viewing. For most of the textual videos, the results shown reveals the fact that visual gesture cannot account for the underlying attention and comprehension level of an individual. Based on the experimental results, mere glancing at such videos reveal little or no relevant information regarding the participant's cognition. Even though, from experimental observations, we infer that visual attention cannot be strongly correlated to cognition in absence of a prime object, the approach is practically scalable for videos with any number of prime objects. The fact that most of the MOOC videos contain limited number of objects, we restrict our case studies to 3 prime objects only.

3.7 Summary

In this chapter, we proposed *Gestatten*, a ubiquitous mobile platform for learner's attention estimation from eye gaze and gaze gesture tracking. The significance of the proposed approach lies in terms of its lightweight nature and applicability. Experimental results reveal its practicality

in shaping one of the major aspects of academics. However, the proposed approach can be improvised to integrate other contextual information like noise level and video specific attention requirement, which we plan to keep as a future direction of this work. From the experiments, we understood that background noise level can be a contributing factor in assessing the learner's attention. Moreover, some video lectures do not require continuous observations and emphasize more on audibility. Hence, involving external and video sound levels will reveal the concealed aspects of learner's attention levels. Parameters like initial attention level, number of prime objects, etc. can be modified to substantiate the scalability of the approach.

4

Ubiquitous System for Estimating Cognition and Multitasking in Online Meetings

For over two decades, video conferencing has been a productive approach for exchanging conversations between multiple participants through a digital online mode [88]. During the COVID-19 pandemic and beyond, it became a necessity rather than an option when almost every meeting, be it a classroom teaching or a business meeting, is being conducted virtually through various online video conferencing platforms. Nevertheless, there has been a serious concern about these meetings' quality due to the lack of engagement from the participants, particularly in the business meetings or the classroom teachings, educational seminars, etc [44]. Many participants tend to be passive during the sessions, mainly when they find other more exciting activities, like reading a storybook or an article over the Internet or browsing through their social networking feeds [160]. Consequently, attending the meeting becomes merely a proof of participation, like giving the class attendance while not following the lectures!

Understanding participants' engagements in an online meeting are essential for the organizers and speakers because it helps plan quality meeting sessions. The Zoom Platform incorporated the attention tracking feature during the onset of COVID-19. Still, they later removed it due to privacy concerns[1]. Various works in the literature [250, 187, 118, 214, 110]

[1] https://support.zoom.us/hc/en-us/articles/115000538083-Attendee-attention-tracking (accessed: Friday 11th August, 2023)

have argued for estimating the users' attention levels during online meetings. However, assessing the attention level in such online meetings is challenging. Notably, the most prominent cue available during such meetings is the video feeds from the participants, which contain their facial expressions and gestures. Although participants prefer to keep the video feed off during general meetings such as academic conferences or seminars, specialized meetings like business discussions or online classrooms typically mandate keeping the video turned on.

Participants' attention during a video meeting is difficult to quantify as it involves a cognitive process. Overt attention [193] involves precise shifting of gaze towards the instantaneous point of interest. However, covert attention is more of a mental rather than a sensory process. Thus, it is not reflected by gaze shifts (they can be related to fine-grained saccades). Psychological studies have also discovered cases where a subject "looks without seeing" [145]. In light of this concept, Lamme [119] has distinguished visual attention from awareness. In a video meeting, which lacks alertness to draw attention, a person might blankly gaze at the screen (which we refer to as visual concentration/attention in this paper) without processing or being aware of any of the information shown/delivered, as their attention is fixed on another complex object/thought. "Cognition", on the other hand, involves memory and processing of information [192]. In this paper, we use the term "cognitive attention" to imply the germane cognitive load [96] or awareness of the subject concerning the online meeting being attended.

In this line, several works [12, 34, 156] correlate visual concentrations with cognitive attention. However, such techniques are difficult to apply for online meetings because of the following reasons. (1) During online meetings, participants may not continuously gaze at the screen; they may look in other directions and still listen to the speaker attentively. Thus a lack of visual concentration may not imply inattentiveness all the time. Eye-tracking has already been applied to track divided attention [8], overt attention [190], selective attention [18] and so on. Techniques like blink detection [149, 43] and Pupillometry [245] have been found to hold a high correlation with sustained attention. While these techniques can be used during online conferences, they generally require devices like commercial eye trackers, electrodes for electrooculograms, etc., which are not usually available to the common mass. Given the current resolution of the webcams or in-built cameras in laptops, it isn't easy to accurately track intricate movements like pupil dilation. Even though commodity cameras can detect eye blink, this technique is likely to fail if the participant is paying attention to an irrelevant article (cognitive disengagement in the meeting) on another tab instead of the meeting. (2) Eye gaze-based attention estimation can fail in the scenarios when the participant opens up a new tab in the browser to browse her social profiles, chat with a friend using text messages on a different tab, take notes, or even look into online materials related to the topic of discussion (we call such activities '*multitasking*') [154, 153, 160]. Interestingly, a recent work [27] has shown that multitasking is unavoidable during remote meetings; indeed, in many instances, multitasking may promote attentiveness and productivity. The existing methods [12, 34, 156] may still find

the participant attentive, whereas her activities are entirely uncorrelated with the online meeting going on. On the other hand, certain multitasking instances, like taking notes during a meeting, can promote engagement of the participants; however, a gaze gesture-based attention estimation method may mark the user as inattentive. Therefore, it is necessary to correctly identify such multitasking instances and correlate them with the participant's cognitive attention.

The fundamental premise behind this paper is that a lack of visual concentration towards the meeting app may not imply inattentiveness; similarly, visual engagement towards the computer screen does not indicate that the participant is engaged with the meeting. Other metrics, like facial expressions and active participation, can account for the alertness of a subject. Such an attention estimation mechanism can only add to the accuracy of gaze-based attention assessment systems. The visual (multi)tasks during such meetings can further enhance the traditional gaze-based methods. Accordingly, we propose *EmotiConf* – a cooperative framework for automated estimation of participants' cognitive attention towards the meeting by leveraging audiovisual sensing over the meeting participants coupled with some interesting observations from human social behavior during virtual interactions. *EmotiConf* develops a software wrapper on top of the meeting platform and is entirely a client-side application that does not send any data outside the user's device.

4.1 Contributions

In the backdrop of the existing works on participants' attentiveness estimation during an online meeting, our novel and salient contributions in this work are as follows.

(1) Bifurcating Cognitive Attention : One of the significant primitives behind the design of *EmotiConf* is that attentiveness toward a meeting's discussion causes similar emotional states as reflected in their facial expressions. A typical facial expression is imperative even if a participant is not visually concentrating on the screen but is cognitively connected to the virtual interactions in the meeting. We conducted a human study to establish this fact and to infer the participants' cognitive attentiveness during a meeting session. Visual multitasking significantly impacts the cognitive attention of a participant. We also demonstrate that certain multitasking instances, like using a notepad to take notes, searching for relevant materials, etc., promote cognitive attentiveness during an online meeting; we call this positive multitasking. On the other hand, activities like browsing social media profiles, chatting with friends, etc. affect attentiveness; we call this negative multitasking. *EmotiConf* uses a cooperative and collective audiovisual analysis of all the meeting participants to classify multitasking instances into positive and negative using three interesting primitives. (a) A change in the vocal expression of active speakers triggers a difference in the facial expression of the participants engaged in the meeting. (b) Positive multitasking instances have a causal relationship with the intent of the

speech; for example, a question asked by the speaker may cause a participant to open a new tab to find out the answer. (c) The degree of facial movements, like head nodding, indicates that the activities are related to the meeting.

(2) System Development and Evaluation: For *distributed expressions and attention estimation*, we analyze the lip movement patterns of the participants to infer the active speakers in an online meeting. As the active speakers are sure to be engaged or attentive towards the meeting, we use their behavior as the ground truth for attentiveness estimation of other speakers. This information, coupled with the facial expression of the participants in a cooperative way, helped us to mark the inattentive participants in a meeting correctly. Moreover, *EmotiConf* uses a novel approach for the *separation of positive and negative multitasking instances* by analyzing the change in the on-screen ambient light reflected on the faces of the participants. We have developed a prototype of *EmotiConf* and evaluated it over a pilot study (30 meetings, 3-12 participants per meeting) and through feedback from 96 in-the-wild participants from a public broadcast about the platform. While we observe a good performance of *EmotiConf* in identifying attentive/inattentive participants and various positive and negative multitasking instances performed by them, the in-the-wild study returned an average usability score of > 80% on the system usability scale (SUS).

4.2 Human Study

The primary motivation and assumptions for the development of *EmotiConf* have been established by two different human-based studies: an anonymous public survey and an annotation task. The detail follows.

4.2.1 Anonymous Public Survey – Realizing the Notion of Attentiveness during Online Meetings

The objective of the public survey was to understand the factors that influence participants' attention during an online meeting. Here we discuss the survey procedure followed by our observations.

Survey Procedure and Participant Details

We floated a Google form containing a set of questions over various social media and newsgroups to obtain the public view about participants' attention and its implications during an online meeting. The form contained 18 questions divided into two sections. The first section contained 4 questions related to the participant's age, gender, demography, and profession. The

second section had two subsections – the first subsection contained 8 questions focusing on the frequency and purpose of the online meetings that the participant typically participates in and the type of different activities that the participant performs during an online meeting. The second subsection contained 5 questions where the participant had to rate the importance of the following factors during an online meeting, on a scale of 1 (Low) to 5 (High) – visual attention, positive multitasking (tasks related to the meeting, like taking notes), negative multitasking (tasks not related to the meeting, like playing a game), cognition (auditory attention on the meeting discussion) and active participation (participating in the discussions). The different types of activities that can positively or negatively impact meeting efficiency have been widely discussed in literature [178, 143].

We received ~ 600 responses from different age groups, locations, and professions. 74.7% of the participants belonged to the age group of 18-30 years, 16.3% to that of 31-40 years, 5.3% to 41-50 years, and the remaining were above 60 years. 75% of the participants were male, and the rest were females. The participants were from different countries like Canada, China, Germany, India, Italy, Nigeria, the UK, and the USA. Most of the participants (~ 75%) were from academics; however, people from other professions, like corporate (12.9%), government service (.3%), freelancers (.3%), etc., also participated in the survey. Among the participants, 65.3% mentioned that they use video conferencing applications at least once a day. The survey also showed that video conferencing systems serve a wide range of purposes, including Corporate meetings (17.1%)[2], academic online classes (79.2%), educational meetings (52.6%), personal conversations (44.5%), Webinars (0.3%), etc. We next discuss the observations made from this survey.

Does Inattentiveness Impact the Meeting's Quality?

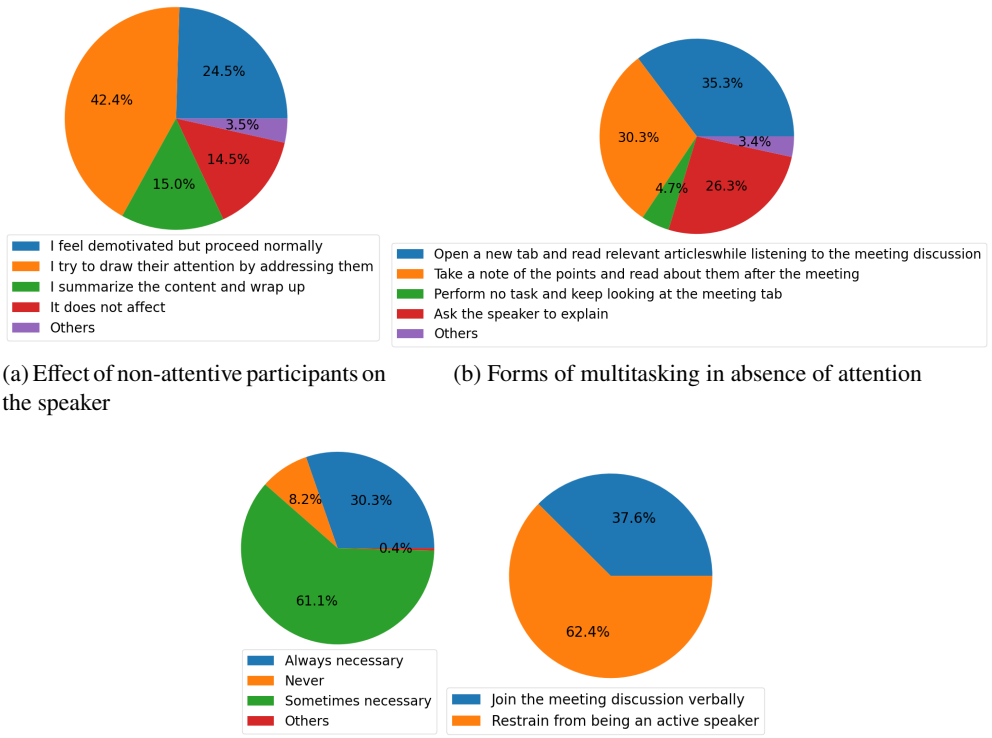
Hypothesis: This question is included to test whether our hypothesis–*Lack of attention impairs the quality of the meeting*. There can be a lot of factors attributing to the disengagement of a person in a meeting – the participant might be attending consecutive sessions of different meetings, preoccupied with other tasks in mind, unable to follow the meeting's discussion, fatigued, and so on. We consider inattentiveness (cognitive) to have a direct impact on the quality of the meeting.

Findings: To assess how the speaker and hence the meeting's quality is affected when the participants are not attentive, we asked the participants how their speech/presentation gets affected in a meeting where they notice that others are not attentive. Even though 42.4% of the responses indicated that they would ignore the inattentive participants or draw their attention, 24.5% mentioned that they would feel demotivated, and 15% would like to shorten the actual content (Figure 4.1a), thus degrading the quality of the meeting. Therefore, inattentiveness is a

[2]The numbers in brackets indicates the percentage of respondents who used video-conferencing for such purposes. The participants could select more than one option.

severe concern during an online meeting.

Recommendation: The survey results prove that inattentiveness indeed affects the meeting's quality. This finding recommends the development of a system that can, not only infer upon the cognitive attentiveness of the participants in online conferences but also perform it in realtime, so that the speaker/ the participant can be alerted. Thus, the real time attention detection method is incorporated in *EmotiConf*.



(a) Effect of non-attentive participants on the speaker

(b) Forms of multitasking in absence of attention

(c) Necessity of positive multitasking

(d) Correlation between active speech and negative multitasking

FIGURE 4.1: Outcomes from the Online Survey

Is Multitasking Common during an Online Meeting?

Hypothesis: The set of questions pertaining to "multitasking" was based on the hypothesis that multitasking, positive or negative, is inevitable in online conferences where manual surveillance is difficult. The objective was also to understand the range and type of tasks people perform during online meetings.

Findings: When asked, 83.2% of the participants mentioned that they perform simultaneous activities during online meetings. This proves that multitasking is often inevitable in online meeting scenarios. However, such activities included a wide range of tasks, including reading books, taking notes, texting, reading emails, reading references, writing notes, eating, playing

games, checking social media profiles, and so on. However, the most popular activities are reading articles in a different browser tab. The articles can be both meeting-related (responded by 68.2% of the participants) or meeting-unrelated (responded by 26.3% of the participants). Additionally, 37.9% of the respondents mentioned that they even use other electronic devices, like smartphones or tabs, during the meetings. It can be noted that a participant could choose multiple options in this question. The survey indicates that people typically engage in various activities during a meeting; many of which are not related to the meeting's context. 63.68% of the young participants, belonging to the age group of 18-30 years, were found to perform simultaneous activities during a meeting, but only 26.5% of these young people mentioned that they read relevant articles during the meeting. The rest of them said that they read meeting-irrelevant online articles or newspapers (10.05%), use electronic devices (6.5%), read books (2.3%), and so on.

Recommendation: This finding indicated that *EmotiConf* should be designed in a way that it can capture the multitasking instances and classify the instances as positive or negative. In doing so, we aim to identify those specific instances where a participant uses the same screen to read relevant/irrelevant articles. This is due to the fact that these types of multitasking involve little to no facial muscle movement or apparent visual context switching and are hence, very difficult to be detected. The categorization of positive and negative multitasking has been explained in the following subsection.

Attention and Multitasking – How do They Go Together?

Hypothesis: This set of questions was circulated in light of three hypotheses–(1) visual focus does not always guarantee cognitive attention and vice versa. (2) Multitasking does not always hinder attention and (3) Active participation is related to attention.

Findings: As we mentioned earlier, attention can be perceived as **visual attention** (or '*concentration*') and **cognitive attention**. To understand whether the former implies the latter, our survey questionnaire included a scenario where the participants were asked to mention the activities they would perform if they could not follow the meeting contents. The answers to this question are related to visual attention and cognition, as explained next. According to Figure 4.1b, 35.3% of the participants said that they would open a different tab to read relevant articles to understand the discussion. Even though it is a form of multitasking, such activity leads to insufficient visual attention towards the meeting yet indicates high cognition as the participant is ultimately focusing on the content of the meeting. We categorized such activities as **positive multitasking** as they promote cognition towards the topic of the meeting.

On the other hand, 30.3% of the participants mentioned that they would note down the points and read about them after the meeting. This will lead to close visual attention in the meeting and average synchronous cognition due to context switching. This form of activity

is also positively attributing to some aspect of attention. However, 4.7% of the participants said that they would perform no action and keep looking at the meeting tab, which will lead to close visual attention but low cognition. Finally, 26.3% of the participants mentioned asking the speaker to explain more details for every point they could not follow. In this case, there will be close visual attention, active participation, and high cognition. Among the other 3.4% responses, activities like playing games were mentioned. This form of activity results in low visual and cognitive attention and can be considered as **negative multitasking**. Hence, it can be inferred that multitasking (positive) can also promote attention depending on the context. This claim is supported by Figure 4.1c where the majority of the participants mentioned that positive multitasking, if not always, sometimes is necessary for an online meeting scenario. Incidentally, most participants (62.4%) prefer to be passive attendees in the online meetings (Figure 4.1d), during disengagement. Therefore, we can infer that multitasking instances are not negligible in such online video conferencing-based meetings. From the third section of the questionnaire which involved questions on how different types of attention and multitasking are important for maintaining the quality of a meeting, it was also observed that 37.6% of the people stated that Visual attention is essential (score=4) , 33.5% mentioned that positive multitasking is essential (score=4), 53.6% mentioned that negative multitasking is not essential (score=1), 47.7% mentioned that cognitive attention is most essential (score=5) and 49% mentioned that active participation is most essential (score=5).

Recommendation: The results infer that *EmotiConf* should involve a measure to identify the cognitive involvement of a participant, rather than just their visual engagements. Active participation (communication), being a good indicator of engagement, has also been used in our model.

4.2.2 Human Experiments to Correlate Attentiveness with Facial Emotions

The design of *EmotiConf* is based on the hypothesis that facial emotions play a pivotal role in defining participants' attentiveness during an online meeting. We argue that facial emotion plays an essential role in determining the participants' attentiveness in an online meeting. For instance, the context or the discussion during the meeting should directly impact the participants' facial emotions. Suppose a participant is attentive towards the meeting. In that case, his facial emotion should reflect such contexts, and ideally, it should tally with most of the participants' facial emotions in the meeting. To validate this argument, we performed an annotation-based evaluation to answer the following two questions. (1) Does the facial emotion of a participant get changed during an online meeting? (2) If the answer to the above question is "yes", then do the facial expressions of the majority agree to a common notion?

Experiment Procedure

This experiment performs an annotation task to map facial emotions with attentiveness in an online meeting. We recruited 9 different annotators with low to high domain knowledge and low to increased experience with video conferencing. This annotation task aimed to validate the claims – (1) facial emotions vary over time in an online meeting, and (2) correlated emotions among participants indicate greater attention. In this task, four annotators were male, and five were female. Two of them had domain knowledge of video processing and were research scholars. One of them had moderate experience with video processing (IT professionals), and the remaining was a novice. Two of the annotators used video conferencing rarely (once in 6 months), three of them used it ordinarily (once in a month), and the remaining four used video conference-based meetings very frequently (once a day).

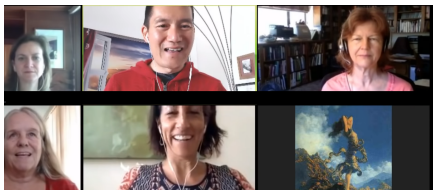


FIGURE 4.2: A sample frame for human annotations of the facial expressions

	1	2	3	4	5	6	7	8	9	Scale
1	1	0.75	0.92	0.81	0.88	0.83	0.94	0.87	0.97	1
2	0.75	1	0.73	0.63	0.7	0.66	0.74	0.68	0.75	0.95
3	0.92	0.73	1	0.85	0.81	0.77	0.95	0.88	0.92	0.9
4	0.81	0.63	0.85	1	0.74	0.69	0.87	0.83	0.8	0.85
5	0.88	0.7	0.81	0.74	1	0.76	0.86	0.77	0.86	0.8
6	0.83	0.66	0.77	0.69	0.76	1	0.79	0.76	0.82	0.75
7	0.94	0.74	0.95	0.87	0.86	0.79	1	0.9	0.94	0.7
8	0.87	0.68	0.88	0.83	0.77	0.76	0.9	1	0.87	0.65
9	0.97	0.75	0.92	0.8	0.86	0.82	0.94	0.87	1	0.6

FIGURE 4.3: Heatmap of inter-annotator agreement

The experiment has been conducted as follows. We considered 5 pre-recorded online meetings, each of ≈ 30 minutes duration, consisting of different types, including formal group discussions, online video conferencing, group discussion with presentations, and online classes. Each of the meetings had a minimum of 3 and a maximum of 8 participants, with a total of 24 participants (P1 - P24) over all the meetings. For these 5 meetings, we selected up to 36 frames in the timescale for each of the meetings (a sample frame from the first meeting having 5 faces is shown in Figure 4.2) and asked 9 independent annotators to label the perceived emotion of the individual faces in those frames, from a set of 8 emotion levels – '*happiness*,' '*anger*,' '*surprise*,' '*disgust*,' '*contempt*,' '*neutral*,' '*sadness*,' and '*fear*.' The frames were selected to ensure that the overall change in expressions (if any) could be captured from the timeline of the meeting. We intended to choose 1-2 frames from each time frame (≈ 1 min) of the video to increase the chance of capturing the overall emotion of the minute. Since macro expressions are known to change within a short span of time, we carefully choose the frame(s) within a time frame, so that the change of expression can be captured before it is faded off. For example, if the speaker is discussing a funny instance, we randomly choose a frame within the tenure of that discussion. If there is no perceivable change in the discussion topic within the consecutive time frames, we choose the frames randomly from those time frames. We use the following expressions: anger, surprise, disgust, enjoyment(happiness), fear, contempt and

sadness Additionally, we have allowed the annotators to mark facial emotion to be "neutral" if no prominent expression was found. The annotators were not the participants of those recorded meetings.

To see the viability of these annotations for further study, we compute the inter-annotator agreement using Cohen's Kappa statistics [157]. We observe an average score of 0.813 that indicates a reasonable degree of agreement among the annotators. Figure 4.3 shows the heatmap of inter-annotator agreement, suggesting that most of the annotators agree on the label of the facial emotions. The heatmap indicates that nearly 60.4% of the Kappa values are more than 0.8 indicating an almost perfect agreement. In contrast, the remaining 40% values are between 0.6 and 0.8, indicating substantial agreement among the annotators. The overall inter-annotator agreement, measure by the Fleiss' Kappa[3] statistics [195] is found to be .809. We next try to find out the answer to the two questions, as mentioned above, from these annotated values.

Observations

The pre-consideration for the first question is that, the facial expression of the subjects in a conference will change gradually or rapidly with respect to the topic of discussion. To answer the first question, we compute the emotion labels' change points as marked by the annotators for individual meeting participants. Then we calculate the agreement among the annotators on those change-points. For example, whenever we move from frame i to frame $i+1$ for a participant P if n out of the 9 annotators mark a change in the facial emotion labels (as inferred from the labels), we compute the agreement as $\frac{n}{9} \times 100\%$. An agreement value of 100% indicates that all the annotators have agreed that the participant's emotion has changed over time. Similarly, a value of 0% suggests that all the annotators have agreed that the participant's emotion has not changed at all. The value has been averaged over all the frames for each of the 24 participants. Figure 4.4 shows the average agreement on the emotion change for all the 24 participants in different video sessions. The figure indicates that for the majority of the participants, the facial emotions change over time. This recommends the utility of facial expression variation in attention estimation. We base the expression detection module of *EmotiConf*, based on this observation.

We next answer the second question. We consider expression to collective in nature during online conferences, i.e. people having cognitive engagement are likely to react/ express in a similar way to the meeting content. We have estimated the percentage of the match for each of the facial expressions found in each frame. For each of the facial images in a frame i, the label selected by most annotators has been selected as the participant's labeled emotion. We then count the number of different emotion labels found in each frame and the percentage of a match (m_e) for each emotion label e found in the frame. In an ideal scenario, for a particular

[3] https://en.wikipedia.org/wiki/Fleiss%27_kappa (accessed: Friday 11th August, 2023)

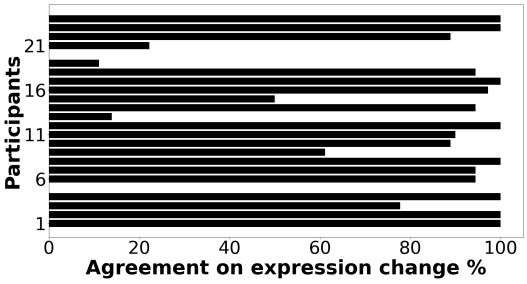


FIGURE 4.4: Agreement on expression change from consecutive frames

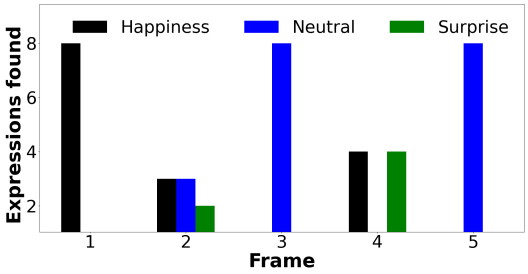


FIGURE 4.5: Matched expressions in frames

frame i, there should be one emotion label for all the faces with a 100% match indicating all the participants have the same emotion label. If there are k participants shown in frame i, and l out of those k participants have the emotion label e then $m_e = \frac{l}{k} \times 100\%$. Figure 4.5 shows a sample scenario with 5 frames having 8 participants. We see 1–3 different emotion labels as expressed by the participants, and each of those emotion labels has been contributed by an almost similar number of participants. On average, 1.13 different expressions have been found on each of the frames. Further, we observe that these emotion labels are closely similar – like 'neutral' and 'happiness.' Therefore, we can say that if there is a drastic difference in the emotion label of one participant from the emotion labels of others, that participant is likely to be inattentive towards the meeting. This leads to the development of the emotion mapping module of *EmotiConf*. It is to be noted that there are multiple streams of literature that support [53, 201] or question [15] the role of discrete expressions in expressing emotions. The intent of this study is to analyze and emphasize the similarity of expressions (positive expressions like happiness or negative expressions like disgust etc.) rather than how well these expressions reveal the emotions. Empirically, we found that in a meeting set up, positive and negative emotions, expressed through these 8 broad categories are indeed similar for attentive participants. However, in the future, overlapping micro-expressions, revealing further insight into a person's emotions will be explored.

4.2.3 Lessons Learnt

From the public survey and human annotation-based study, we have the following takeaways. (1) Multitasking can promote attentiveness if utilized positively. Therefore, 'visual concentration' does not always imply 'cognitive attention.' Consequently, the existing studies based on visual cue and gesture analysis from the participants' video feeds are not very suitable to accurately quantify the attentiveness of a participant towards a meeting. (2) We observe that facial emotions play a pivotal role in defining the participants' cognitive attention towards the meeting. Even if the participant is not directly concentrating on the browser tab or the application running the meeting platform, (s)he may still get cognitively connected to the discussion going on in

the meeting. Such cognitive attentiveness is likely to get reflected in their facial emotions; therefore, the facial emotions of the participants should match. We next proceed with these observations to develop *EmotiConf* to identify the inattentive participants in a meeting.

4.3 Proposed Model – An Overview

EmotiConf infers the following from a cooperative audiovisual analysis of the participants' video feeds – (a) cognitive attentiveness of a participant, (b) multitask instances over the device/computer used for the meeting, and (c) whether the multitask instances are positive or negative. It is designed to work on a client's device that displays the participants' facial previews (video feeds) in a windowed (grid) layout. Based on empirical evidence from our experiments, the grid layout has been chosen so that the facial previews of all participants can be monitored simultaneously and continuously by *EmotiConf*. The overall idea of *EmotiConf* is to analyze six different multi-modal aspects from the facial region of interest (ROI) and audible conversation: (1) mouth movement, (2) facial expression, (3) change in the reflection of light, (4) facial movement, (5) vocal expression, and (6) conversation type. While the two former aspects identify the cognitive attention level, the third one infers visual multitasking. The rest of the parameters are used to classify the detected multitask instances as positive or negative.

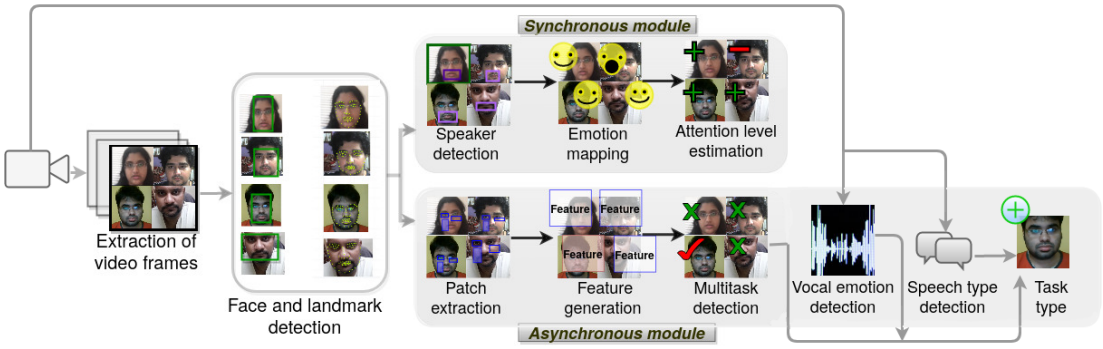


FIGURE 4.6: The overview of *EmotiConf*

Figure 4.6 shows an overview of *EmotiConf*. The pipeline of *EmotiConf* includes a real-time (synchronous) estimation of attention and a non-real-time (asynchronous) detection of visual multitasking categorization of these tasks. *EmotiConf* assumes that the video feeds from a majority of the participants are available and processes these video feeds locally as a software overlay on top of the meeting app. We extract 68 facial landmarks [101] that help us to analyze any specific parts of the face. These facial landmarks are used in both the synchronous and the asynchronous modules.

4.3.1 Synchronous Module for Attention Estimation

The synchronous module works in real-time for a frame rate of up to 60fps. This module aims at presenting a statistical comparison of the attention levels of the different participants. The synchronous module analyzes the facial landmarks over a person's mouth and the entire facial area. *EmotiConf* tracks the lip movements of individual participants using these facial landmarks. Unless the person talks offline outside the meeting, lip movements indicate their engagement as a meeting speaker. We utilize this information to find out the active speakers by correlating the lip movements of all the meeting participants. For the non-active speakers, the synchronous module of *EmotiConf* performs a frame-wise matching of the participants' emotional states (facial expressions) captured through a pre-trained deep neural network model resulting 8 different classes of emotional states. By combining the cooperative observations from active speaker detection and facial expression matching, the synchronous module of *EmotiConf* marks a participant as attentive or inattentive with a confidence value.

4.3.2 Asynchronous Module for Multitasking Analysis

The asynchronous module processes the collection of features derived from all the video frames. One of the significant sources of multitasking during an online meeting is to use different tabs or applications on the computer beyond the meeting app. *EmotiConf* uses a novel idea to detect such multitask instances by monitoring the reflection of the ambient light from the device's screen on the face of the participant. By analyzing the patterns of the reflected light, *EmotiConf* determines whether a participant performs multitasking during the meeting hour.

Tagging Multitasking as Positive and Negative

To classify the multitask instances as positive or negative, *EmotiConf* models the general human behavior by combining the facial expression (emotional states) of the participants, as captured earlier, with the vocal expressions and the intent of the speech by the active speakers. By analyzing the behavioral patterns of individuals while engaging in an online meeting, *EmotiConf* designs a rule-based approach to mark the multitask instances as positive or negative. For this purpose, *EmotiConf* utilizes pre-trained models [176, 238] that are well-established in the literature for extracting the vocal expressions and the intent of the speech from the audio signal captured during the meeting.

4.4 Synchronous Module: Who Are Inattentive in My Meeting?

The synchronous module works in real-time to determine attentiveness of the participants. The details follow.

4.4.1 Detection of Faces and Facial Landmarks

EmotiConf detects the facial region of the participants using the method proposed in [40], following which, the 68 facial landmarks are detected using the method shown in [101]. More details can be found in Appendix A.1.

4.4.2 Extraction of Mouth Region and Speaker Detection

This module considers the facial landmarks detected near the lips to precisely extract a participant's mouth region.

Feature Extraction

From the outer and inner lip regions' landmarks, we extract the following two features for each participant.

(i) Average Pixel Value at the Mouth Region ($\mathcal{F}_1$): The first feature is a pixel-based mean intensity value from the inner region of the mouth. For example, the landmarks 50 and 57 approximately indicate the two corners of the mouth. Accordingly we extract a rectangular patch from $\{X_{50}, Y_{50}\}$ to $\{X_{57}, Y_{57}\}$, where $\{X_\ell, Y_\ell\}$ denotes the x and y coordinates of the landmark ℓ. The first feature ($\mathcal{F}_1$) from this given patch is estimated by taking the average of all the RGB pixel values along with the patch. Since mouth movement occurs in the presence of verbal communication, a large deviation in this value in consecutive frames should be tracked if a person is talking.

(ii) Mouth Height to Face Height Ratio ($\mathcal{F}_2$): The second feature is derived by the formula,

$$\mathcal{F}_2 = \frac{\frac{1}{5}(\mathcal{D}(50,60) + \mathcal{D}(51,59) + \mathcal{D}(52,58) + \mathcal{D}(53,57) + \mathcal{D}(54,56))}{\mathcal{D}(20,9)}$$

where $\mathcal{D}(\ell_1, \ell_2)$ is the Euclidean distance between two landmarks ℓ_1 and ℓ_2. Instead of relying on a single pair of points, the numerator takes an average of mouth heights, measured at five different points on the mouth region. If this ratio is more than a threshold, it can be assumed that the mouth is opened. However, an increase in this value can also indicate non-verbal activities like yawning. The difference between talking and yawning in terms of $\mathcal{F}_2$ is that yawning increases the mouth height for a few consecutive frames. However, talking required more rapid movement of the mouth region; thus, the inter-frame difference will be more for the consecutive frames. Based on empirical evaluations on different video recordings of people attending online conferences (both in-house and open source YouTube videos), we set the inter-frame thresholds for these two features as $T(\mathcal{F}_1) = 1$ and $T(\mathcal{F}_2) = 0.15$, respectively.

Detection of Lip Movements

Let $\mathcal{I}(i, p)$ be an indicator variable that determines whether a lip movement is detected for participant p over the video frame F_i. Let, $\mathcal{F}_1^i$ and $\mathcal{F}_2^i$ be the feature values measured over the video frame F_i. For each participant p, we compute $d_1 = |\mathcal{F}_1^i - \mathcal{F}_1^{i-1}|$ and $d_2 = |\mathcal{F}_2^i - \mathcal{F}_2^{i-1}|$. If $d_1 > T(\mathcal{F}_1)$ and $d_2 > T(\mathcal{F}_2)$, then we set $\mathcal{I}(i, p) = 1$ (indicating a lip movement), else $\mathcal{I}(i, p) = 0$ (indicating no lip movement).

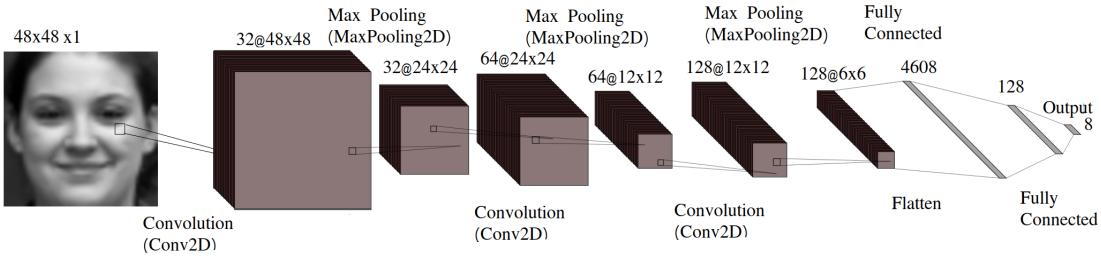


FIGURE 4.7: The CNN architecture: Each Convolution layer (Conv2D layer is used since the input is an image of shape 48x48x1) uses a kernel of size 3X3 and is followed by a LeakyReLu function for activation. The MaxPooling layers use a kernel of 2X2 and are followed by Dropout layers. The final output layer produces a vector of size 8 representing the eight classes of facial emotion from the CK+ dataset.

4.4.3 Emotion Mapping

EmotiConf estimates the emotion of a participant from her facial expression, when no lip movement is detected over a frame, i.e. $\mathcal{I}(i, p) = 0$. For this purpose, we use a Convolution Neural Network (CNN) trained with the extended Cohn-Kanade dataset [142] that consists of eight different classes of facial expressions – *anger*, *sadness*, *happiness*, *surprise*, *contempt*, *fear*, and *neutral*. The structure of the CNN is shown in Figure 4.7.

Quorum on Emotional States

EmotiConf executes a quorum function `Quorum(`$\mathcal{E}_p$`, p)`, shown in Algorithm 4, where p is a participant and $\mathcal{E}_p$ is the integer encoding of the emotional state of participant p. This function is executed for each participant independently, and the function determines whether the participant is cognitively attentive or not.

4.4.4 Marking the Cognitive Attentiveness for Each Participant

EmotiConf maintains a positive frame count $\mathcal{A}_p$ for each participant p to count the frames where the participants are found to be attentive. This frame count is processed as follows. For a frame $\mathcal{F}_i$, $\mathcal{A}_p$ is incremented if one of the following two conditions are satisfied – (i) a lip

Algorithm 4: Emotion Mapping – The Quorum Function

Input: Current Frame ($\mathcal{F}_i$), Number of participants ($\mathbb{N}$), Set of emotion labels (integer encoding of the emotional states) of the participants ($\mathbb{E}$) predicted by the trained CNN
Result: Estimated attention of a particular participant (`True`/ `False`) through mapped emotion
Function *Quorum($\mathcal{E}_p$, p)*

```
lcounter[8] ← 0;
foreach E_x ∈ 𝔼 do
    // E_x denotes the emotion E (mapped to integer value) of participant x among the
       total set of participants.
    if x ≠ p then
        lcounter[E_x] ← lcounter[E_x] + 1;
        // Increases the counter of that emotion which is shown by the participant x.
index ← −1, maxval ← 0, el ← 0;
while el < 8 do
    // Finds the maximum emotion shown by the participants excluding p.
    if max < lcounter[el] then
        max ← lcounter[el];
        index ← el;
    el ← el + 1;
if E_p == index then
    // If the emotion shown by max participants match with participant p's emotion
       E_x.
    return True;
    // Attentive.
else
    return False;
    // Inattentive.
```

movement is detected for the participant p over frame $\mathcal{F}_i$, (ii) `Quorum(`$\mathcal{E}_p$`, p)` returns `true` over frame $\mathcal{F}_i$. Let within the time duration $[0, t]$, f number of frames have been processed; then *EmotiConf* marks the cognitive attentiveness of the participant p as $\frac{\mathcal{A}_p}{f} \times 100\%$.

4.5 Asynchronous Module: What are You Doing, Man?

The asynchronous module works offline after the meeting is over and finds out the occasions of positive and negative multitasking after analyzing all the detected multitask instances for a participant. This module starts with the detected facial regions of each participant along with 68 facial landmarks for each of them. The different steps of execution in the asynchronous module follow.

4.5.1 Detection of Multitasking Instances

The core idea is to analyze the ambient light reflected from the participants' faces to check whether the participant performed multitasking during the meeting session. Whenever a participant changes the foreground application on her device, the ambient light from the device's screen should also change. It can be noted that if the participant starts playing a video, that would also trigger continuous changes in the ambient light; however, the frequency of such changes will be very high. On the other hand, for a typical meeting app, even if the main

display of the meeting app shows some presentation, the intensity of the ambient light is not likely to change much. In order to understand how much the light intensity will vary during presentations, we considered about 10 presentations and calculated the average pixel values from each slide on presentation mode. We calculated the difference between two consecutive slide intensities and found that for 69% of the slides, the difference was only up to 6.67. These presentations are mainly academic lecture materials following formal themes or presentations of research papers. However, we do agree, that other types of animated presentations might affect the ambient light, to some extent. Based on this idea, *EmotiConf* monitors the pattern of the ambient light reflected from the participants' faces to detect a change in that pattern.

Feature Generation

From empirical experiments, we observe that for a typical right-handed person, the reflection of the ambient light is maximum from the right cheek of the participant[4]. Further, we observe that the red and the green channels of the facial image have the maximum impact on the ambient light; therefore, we use the intensity of the ambient light over these two channels, as reflected from the right cheek of the participant. This intensity is used for finding out a "*significant*" change in the intensity distribution, as we discuss next.

4.5.2 Extracting Multitask Instances

To extract the multitask instances based on an observation of the reflected light intensity, we use an online unsupervised learning mechanism – *Hierarchical Temporal Memory* (HTM) [2]. An HTM is suitable when we want to find out a deviation, called *anomalies*, over time-series data. HTM is lightweight and fast in processing time-series data to identify anomalies in its pattern; therefore, an HTM perfectly suits our case.

Architecturally, an HTM consists of an Encoder that takes the feature at time t and encodes it, then passes it to a spatial pooler layer that generates a sparse vector. The active columns of a spatial pooler are based on permanence values, indicating how confidently the column can represent a particular input data feature. The model views the input steam series; it gradually learns to predict the next timestamp's value using a temporal memory. New patterns are also learned as the columns responsible for predicting the data are continuously updated.

In our approach, we use the time-series average light intensity over the red and the green channels as the input/output distribution for the HTM. Based on the predicted intensity values, *EmotiConf* computes the prediction error for each timestamp t by taking the absolute difference between the actual and predicted light intensity values. Since the data is processed in asynchronous mode, the mean error is estimated after generating the errors at each timestamp, up to t_n, where t_n is the video's duration. It is observed that task switching on the screen is

[4]For a left-handed person, we can similarly use the left cheek

indicated by a sharp increase in error, preceded and followed by slightly increased errors for a few consecutive timestamps.

Taking advantage of this pattern, we estimate the threshold to be twice the average error and identify the continuous blocks of timestamps where the errors are more than this threshold. These blocks capture multitask instances.

4.5.3 Classification of Visual Multitasking

For classification of visual multitasking into positive and negative instances, *EmotiConf* uses the idea that for positive multitasking, a participant typically switches the task in sync with the discussions going on in the meeting. Consequently, the vocal emotion of the speaker and the intent of the speech, cooperatively, plays a pivotal role here.

Audio Processing for Vocal Emotion Detection

We extract the vocal emotion of a speech signal using the well-adopted approach proposed in [186]. *EmotiConf* captures the speech signal from the meeting and divides it into audio segments of 5 seconds duration. The duration of individual and continuous audio signals has been empirically estimated through experiments by considering two aspects, discussed in the literature. In [185], it has been shown that different vocal emotions require varying amount (in duration) of auditory data to be identified. The work shows that the duration varied from .5-1 second for 6 different emotions. Conversely, a complete sentence with $10-15$ words on average, takes around $4-6$[5] seconds to be uttered. While segmenting the audio, the intention was to encompass as many complete sentences as possible, from individual audio segments. Hence, an optimal duration of 5 seconds was used. From each of the audio segments, it calculates the following three features – Mel Spectrogram, Mel Frequency Cepstral Coefficient (MFCC), and Chromagram. Mel Spectrogram and MFCC provide a perception of the frequencies present in the audio tone in the Mel scale. In contrast, Chromagram or the Chroma feature provides a representation of the 12 different pitch classes. These tonal and pitch features are used to identify the vocal emotion using a Multi-layer Perceptron (MLP) having one hidden layer with 300 nodes. We trained the MLP model using a subset of the IEMOCAP dataset [24] that contains 5 scripted or improvised dialogue sessions between 10 different actors. It can be noted that the complete set of vocal emotional labels available in this dataset is not required for our purpose, as many of those labels are not feasible in the context of online meeting scenarios. We selected 5678 annotated audio utterances corresponding to 6 vocal emotional labels – *angry*, *happy*, *sad*, *neutral*, *excited* + *surprised*, and *fearful*[6]. The pre-trained MLP model using this

[5] https://en.wikipedia.org/wiki/Words_per_minute (accessed: Friday 11th August, 2023)

[6] Although the facial emotional model as discussed earlier in Section 4.4 uses a *contempt* class, this class is absent in the IEMOCAP dataset. On the other hand, the *disgust* class is highly imbalanced. So, we could not use those two classes for vocal emotion extraction.

dataset is used in *EmotiConf* to detect the vocal emotion of every audio segment into one of the above six labels.

Text processing for Speech Intent Detection

To detect the intent of the speech of meeting discussion, we first use the *Google Speech-to-Text API*[7] to convert the speech into text. For the identification of conversational intent, *EmotiConf* first segment the textual blocks and put proper punctuation based on an approach proposed in [238]. This work uses a bi-directional Recurrent Neural Network (RNN) to consider variable-length contexts in both the directions (previous and later) for a current word in the textual block. The model has been pre-trained with 40M words from the English Europarl dataset [108] containing statements from the European Parliament's Proceedings.

Once the model segments and punctuates the textual blocks, the next task is to classify each sentence into one of the three categories – *statement*, *command*, and *question*. For this purpose, we use a simple CNN with a single convolution layer with ReLu activation, a pooling layer followed by a fully connected layer that classifies the input to either of the three categories. This model has been trained with two datasets – Stanford Question Answering Dataset (SQuAD)[8], and the Speech Act Annotated Dialogues (SPAADIA) Dataset[9]. These two datasets collectively contain $80,000$ statements, $11,000$ commands, and 131000 questions.

Detection of Active Speakers and Head Movement Tracking

EmotiConf uses the vocal emotions and the intent of the speech for the active speakers. To detect the active speakers, we utilize the indicator variable $\mathcal{I}(i,p)$ (denoting whether there is a lip movement) computed during the synchronous module of the framework. Let $[F_k, F_{k+n}]$ be a window of n consecutive frames. If $\sum_{i=k}^{k+n} \mathcal{I}(i,p) > \frac{n}{2}$, then we mark the participant p as an active speaker for that window. It can be noted that n is a tunable parameter depending on the frame rate of the video; we set $n = 50$ that corresponds to approximately 1sec of video duration.

Further, *EmotiConf* correlates the instances as head movements with the multitask instances to determine whether the participant is actively engaged with the meeting's discussion. Based on the position of the nose-tip (landmark 31), *EmotiConf* tracks the participant's head movement in either of the four directions – *Up*, *Down*, *Left*, *Right*. For each consecutive frame, we calculate the head shift direction based on the nose tip position in the previous frame. Let, $\{X_{31}^n, Y_{31}^n\}$ be the x and y coordinates for the detected nose-tip over frame F_n. If $|X_{31}^{n+1} - X_{31}^n| << |Y_{31}^{n+1} - Y_{31}^n|$, i.e. the nose-tip movement in the vertical direction is significantly more than its movement in

[7] https://cloud.google.com/speech-to-text (accessed: Friday 11th August, 2023)
[8] https://rajpurkar.github.io/SQuAD-explorer/ (Access: Friday 11th August, 2023)
[9] http://martinweisser.org/spaadia_release_v01.zip (Access: Friday 11th August, 2023)

the horizontal direction, then we mark it as the *Up* movement. Similarly, we compute the head movements in other three directions.

Rule Mapping

Finally, *EmotiConf* classifies the instances of visual multitasking as follows. We use a rule-based approach by correlating the detected multitask instances of a participant with the vocal emotion of the speaker and the intent of the speech. For marking a multitask instance at time t, we use the following rules.

1. If (a) the vocal expression of the speaker during the time interval $[t - \Delta, t + \Delta]$ (Δ is a configurable parameter, we set the value as 0.5sec based on empirical observations during the experiments) matches with the facial expression (emotional state) of the participant, and (b) the speaker is commanding/questioning, then the multitasking instance is marked as positive. The intuition is that whenever the speaker instructs the participants to do something, the attentive participants (indicated by a match between the vocal emotion of the speaker and the facial emotion of the participant) may trigger another task (like taking a note or searching something over the Internet), which is positive multitasking.

2. If the participant has been an active speaker within the time interval $[t - \Delta, t + \Delta]$, the multitask instances are marked as positive multitasking. The intuition is that if a participant actively participates in a meeting discussion, it will be difficult for them to continue with a completely unrelated activity for a long duration.

3. If we observe a significant facial movement of the participant within the duration $[t-\Delta, t+\Delta]$, and the facial expression of the participant mostly matches with the vocal expression of the speaker, then the multitasking instance for that participant is marked as positive.

Although the above three rules look simple heuristic and may not hold for 100% cases, we observe that these are the general behaviors of engaged participants. An attentive participant typically performs activities that engage them with the meeting's discussion. Such activities are supported by their active involvement in terms of expression change, assertive indications like head nodding, etc. *EmotiConf* captures such behavioral instances to generate the above rules. However, we concur that this set of rules may not be exhaustive to identify all the cases of positive multitasking, and new rules can always be included in the design of *EmotiConf*.

4.6 Deployment and Experiments

We developed *EmotiConf* using Python toolboxes, which works as a wrapper on top of two existing meeting platforms – *Appear* (now *Whereby*)[10] and *Zoom*[11]. Figure 5.10 shows the screenshots of the application working in synchronous and asynchronous modules. Figure 4.8a (top) shows the "in-meeting" processing of expressions and active participation in the estimation of instantaneous attention. At the bottom, is the post-meeting result for the particular meeting attendee, displaying his/her distribution of facial expression throughout the meeting (pie chart on the top left corner), average attention on the top right corner (out of 100), average attention with respect to the average attention of the other participants (bottom left) and the average match of expressions with other participants (bottom right). The graph also shows the attention score curve and the speech curve of the participant throughout the meeting. Moreover, figure 4.8b shows the feature graph of 2 participants and the post-meeting analysis of the multitasking instances and their classes (positive peak implies positive task and negative peak indicates negative task). We evaluate the platform's performance on two different experimental setups – (i) a pilot study over 30 different meetings to asses the performance of different components of *EmotiConf*, (ii) an *in-the-wild study* with 96 separate users to assess the usability of the platform.

4.6.1 Pilot Study (30 Online Meetings, 3–12 Participants per Meeting)

We considered three types of meetings. Table 4.1 enlists the meeting types, setup, duration of each meeting type, number of participants in each type of meeting and the number of sessions recorded for each type. For the 7 classroom scenarios (meeting type T3), although there were $\sim$ 55 students, only 5 of them agreed to share the ground-truth annotations. So, we have validated *EmotiConf* on the data obtained from those 5 students during online classes. Out of all the participants, approximately 48% were men, and the remaining were women; $\sim$ 50% were wearing glasses. 60% of the participants belonged to the age group of 24-35 years, and the remaining belonged to the age group of 35-60 years. 70% of the participants were graduate students or research scholars, 20% were industry professionals, and the remaining were faculty members at academic institutions.

4.6.2 In-the-wild Study (96 Individual Participants)

For the in-the-wild usability study of *EmotiConf*, we have made the system publicly available along with a demo video that explains the working steps of the platform. As *EmotiConf* is entirely a client-side application, all the meeting participants don't have to use the platform.

[10] https://whereby.com/ (Accessed: Friday 11th August, 2023)
[11] https://zoom.us/ (Accessed: Friday 11th August, 2023)

(a) Attention estimation over synchronous module (b) Multitask analysis over asynchronous module

Figure 4.8: The application screenshots: The synchronous module (left side) inscribes the attention estimation along with emotional ground truth over participants' video feed. The asynchronous module (right side) shows the HTM prediction pattern from which the multitasking analysis is done offline.

With this advantage, individual participants can try the solution and provide their feedback on the system's usability.

To obtain the feedback, we have used the standard questionnaire from the System Usability Scale (SUS) [23], where the participant needs to give a rating on a scale of 1 (*Strongly Disagree*) to 5 (*Strongly Agree*) against 10 different usability statements. The statement and the formula to calculate the SUS score can be found in Appendix A.2. It can be noted that obtaining a strong agreement towards the odd statements and a strong disagreement towards the even statements would indicate high usability of *EmotiConf*. For each participant, the scaled average score was calculated using the formula shown in [23]. From the public survey, we have obtained 96 filtered responses from a total of about 104 responses from in-the-wild participants. $\approx 4.16\%$ of the responses were from UK, $\approx 7.29\%$ from the US, $\approx 2.08\%$ from Canada and the rest from India. To ensure data validity, those responses that showed random answering patterns (e.g. all questions are scored 5 or 1 and missed any feedback) were treated as outliers and removed.

4.6.3 Ground Truth Annotation

Ground truth annotation is a major challenge for evaluating *EmotiConf*, as *attentiveness* is a subjective measure. For ground truth annotation, we had asked the participants to also

Type	Setup	Duration	Participants	Sessions
T1	General discussions among a group of participants without a formal presentation	Total=5.85 hours (min =∼ 18 minutes, Max=∼ 40 mins)	Average=4 (min = 3, max = 5)	18
T2	Group discussion with a formal presentation by a participant	Total =5.5 hours, (min =∼ 50 minutes, max =1.2 hours)	Average 9 (min = 7, max = 12)	5
T3	Classroom teaching scenario	Total =7 hours (each having a duration of ∼ 1 hour)	5	7

Table 4.1: Details of different meeting types

record their computer screens using *OBS Studio*[12]. The participants' video feed, along with the recorded screen, is used for annotating the data. Additionally, we use a test-based validation of the annotated attention labels.

Annotating Participants based on Attentiveness

Based on the activities being performed by the participants during the meeting, we have marked the participants as *attentive* or *inattentive* for every 5min window duration during the meeting. For this annotation, the entire meeting is divided into 5 min windows (similar to [42]). We have recruited 4 independent annotators, out of which one is the participant herself. The annotators have been asked to carefully scrutinize the video feed of the participant and their recorded screens. Accordingly, they mark the participants as *attentive* or *inattentive*. We observe that for around 60% of the cases, the annotators agree on the cognitive attention state of the participants, with an average inter-annotator κ value of > 0.80 over a Cohen's Kappa test [157]. When the annotators have a high agreement, we use those annotations as the ground truth to evaluate *EmotiConf*. From the annotated data, we observe that for around 62% of the cases, participants were *attentive*; for remaining cases, participants were *inattentive*.

To further cross-validate the annotated data, we provided a set of questionnaires, based on Multiple Choice Questions (MCQs), to the participants just after the meeting. The questionnaire contained a set of questions carefully chosen on the discussions going on at various meeting instances. To our best, we discussed with the meeting coordinators (like the presenters in the type T2 meetings, the teacher for the type T3 meetings, etc.) a priori and ensured the minimum possibility of utilizing background knowledge of the participants to answer the questions. We observed that the validation matches for the instances that we have taken as the ground truth, i.e., the attentive participants could answer the questions, whereas the inattentive participants

[12] https://obsproject.com/ (Accessed: Friday 11th August, 2023)

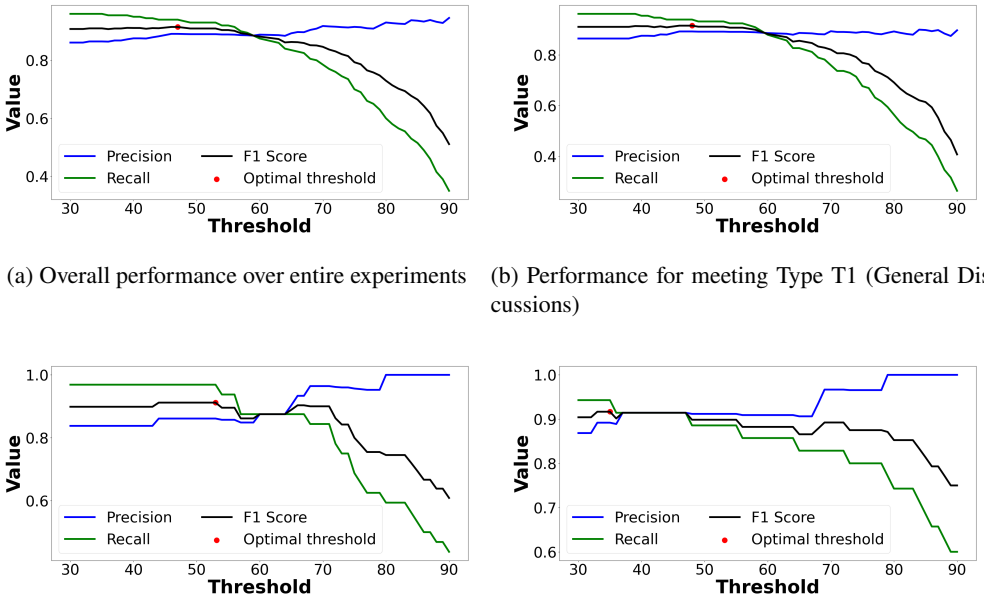
could not answer.

To estimate how quickly and efficiently, a meeting organizer might detect the inattentive participants on manual inspection, we requested 2 distinct annotators to view 5 recorded meetings and identify the inattentive participants as quickly as possible. These annotators had no prior bias towards any participant as they did not know any of the participants personally. The annotators marked each Earliest Occurrence of Inattentiveness (EOI) by closely observing the beginning and end of inattentiveness span of each participant. For example, if participant $P1$ was found to be inattentive from time 3minutes 15seconds-5minutes 30seconds and from 8minutes 30seconds-9minutes 35seconds by annotator $A1$, the EOIs of $P1$ were marked at 3minutes 15seconds and 8minutes 30seconds respectively. The exact EOI of their inattentiveness were obtained through self-annotation. The EOI were also recorded by *EmotiConf*. To understand how quickly and accurately inattentive participants could be detected both manually and by *EmotiConf*, we compared the ground truth with the manual and predicted EOIs.

For *EmotiConf*, the time required for the detection of inattentive participants was significantly less (p-value of 0.0003238 over Mann Whitney U-test) than that required by the annotators. The average delay in EOI by the annotators was found to be 1.06 minutes while that by *EmotiConf* was 7.43 seconds. To capture more nuanced reasons for this difference, we interviewed the annotators and observed the following points:
(i) Lack of definition of attentiveness: Since no clear instruction was provided to the annotators (organizers), regarding which parameter should define "attentiveness", both of them looked for visual inattentiveness. They only marked a participant to be inattentiveness if the participant gazed away from the screen. This, not only delayed the detection, but also lead to inaccurate detection. (ii) Parallel observation: The meetings consisted of 3-4 participants each and were 20 minutes long. Even though the number of participants was less, it was difficult for the organizers to observe the participants simultaneously to monitor their attentiveness. The delay was highly based on this fact. The difficulty in parallel monitoring also lead to missed detection of exact inattentive instances (45.8% of the instances were missed by the organizers). These results show the necessity of *EmotiConf*.

Annotating Multitasking

Here we work with part of the data for which we have successfully been able to annotate the cognitive attention state of the participants, as discussed above. To annotate multitasking instances, we utilized the information from the screen recordings. The recorded screen shows the different tasks the participants performed on the computer while attending the meeting. We have used four independent annotators, out of which one is the participant herself, to annotate the multitasking instances, as well as their labels as *positive* or *negative*, depending on whether the task is related to the meeting. For this annotation, we observed that for all the cases, we have



(a) Overall performance over entire experiments

(b) Performance for meeting Type T1 (General Discussions)

(c) Performance for meeting Type T2 (Group Presentation)

(d) Performance for meeting Type T3 (Class-rooms)

FIGURE 4.9: *EmotiConf*'s performance for attention estimation (Synchronous Module)

an agreement among the annotators with inter-annotator $\kappa > 0.80$ for Cohen's Kappa test [157].

It can be noted that the above annotations can correctly provide the ground truth for the on-screen multitasking instances, i.e., the participants perform it on the same primary device where the meeting app is running. If the participant takes a note on her mobile, the annotators might mark it as a *negative multitasking* although it is a *positive* one, because they do not know whether the participant is taking notes on her mobile or chatting with a friend over WhatsApp. The participants themselves can only annotate such cases; however, we did not want to rely on the participants' annotations entirely. Further, we observed that such instances where the participants do something out of their primary device are low ($< 2\%$) in our data; therefore, we did not explicitly consider those cases in our evaluation.

In addition, we collected following details from the participants – (i) screen size of the primary device, (ii) approximate screen-to-face distance, (iii) lighting condition in the room. As *EmotiConf*'s Asynchronous module relies on the reflected light from the participants' faces, we utilized this information to analyze its performance under diverse scenarios.

4.7 Results and Evaluation

We start with the analysis of the attention estimation, followed by a detailed study of multitasking detection and classification.

4.7.1 Pilot Study: Attention Estimation

EmotiConf computes an *attention score* in a scale of [1-100] based on the factors like lip movement and facial emotion matching. We then apply a threshold on this computed attention score to mark a participant as attentive or inattentive. Figure 4.9 shows the average precision, recall, and F1-score for different threshold values for four different cases – the overall scenario and the three meeting types as we mentioned earlier. Precision indicates the average percentage of inattentive (attentive) participants detected by our method, out of all the inattentive (attentive) participants. In contrast, recall indicates the average percentage of correctly detected inattentive (attentive) participants out of all the detected inattentive (attentive) participants. We compute the F1-score as the harmonic mean of precision and recall. The figure indicates that overall *EmotiConf* achieves a F1-score of 0.91 with Precision 0.89 and Recall 0.94 with a threshold of 47 (Figure 4.9a). Further, it is comforting to see that *EmotiConf* detects the inattentive (attentive) participants for the Type T1 meetings (group discussions without formal presentation) with an F1-score of 0.92 (Precision 0.89, Recall 0.94) with a threshold value of 48. For the Type T2 meetings (a group discussion with formal presentations), we observe a maximum F1-score of 0.91 (Precision: 0.86, Recall: 0.97) with a threshold value of 53. Finally for the Type T3 meetings (online classrooms), *EmotiConf* achieves a F1-score of 0.89 (Precision 0.94, Recall 0.92) with a threshold range of 35-47. In summary, we observe that a threshold value between 45-55 works well for all the cases. It was interesting to notice that *EmotiConf* shows slightly better results for meeting types T1 (many-to-many discussion setup with high rate of active communication) and T2 (one-to-many discussion setup with moderate rate of active communication) when compared to T3 (one-to-many discussion setup with low rate of active communication). It could hence be inferred that the setup had little or no effect on the system's performance, rather, the difference in the F1-score resulted from the degree of discussion. *EmotiConf* performed better for meetings with higher rate of active communication.

In order to compare the performance of *EmotiConf* in the different types of meetings, we have performed a significance analysis using *Welch's unpaired t-test* and *unpaired t-test*. For this, we have considered the scores of the participants in the MCQ tests for each of the videos and compared them with the attention score generated by *EmotiConf*. By comparing the differences (error) between these two scores for each video, it could be seen that *EmotiConf* worked the best for the meeting type T1 with an average error of 7.05. The error was significantly less than both types T2 (p-value of .000105 for the unpaired t-test and .039 for Welch's unpaired t-test) and T3 (p-value of .000000016 for unpaired t-test and .0324 for Welch's unpaired t-test). For type T2, the mean error was 14.53 which was significantly less (p-value of .041 for unpaired t-test and .046 for Welch's unpaired t-test) than type T3 which had a mean error of 25.7.

In the light of this result, the challenges and advantages of each setup with respect to *EmotiConf* need to be discussed. The design of *EmotiConf* relies on active speaker detection and

emotion matching for attention estimation, making it best suitable for meeting type T1. Under T1, all participants had an equal chance of actively participating in the discussion and were likely to express themselves more actively in an ideal scenario. This, in turn, would allow *EmotiConf* to capture the outliers in a better way. However, in a real scenario, even under a many-to-many communication setup, we found a significant imbalance in active communication among the participants in some meetings. For example, in one meeting, participant P1 introduced participant P2 to P3 and allowed them to interact. Even though P1 was fully engaged throughout the meeting, their communication was sparse. However, *EmotiConf* solved such cases by relying on the change in their facial expressions. In setup T2, the opportunity for active vocal participation was limited as the presenter was the main speaker and others could occasionally communicate with them. Moreover, during formal presentations, we observed that the changes in facial expression were not significant and mainly varied from neutral to happiness/surprise/anger primarily. However, since *EmotiConf* relies on emotion matching, instead of change rate, it could perform seamlessly. Considering the average rate of active participation and expressiveness *EmotiConf* can also be applied to type T2. The most challenging setup of *EmotiConf* is, however, type T3. The major disadvantage of such a setup is the significant (or sometimes complete) absence of communication between the teacher and students during the lessons. Even though *EmotiConf* performs well with a limited number of students under T3 setup, estimating the performance of *EmotiConf* with an increased number of students in this setup would be an interesting observation in the future.

Improvement Analysis

To analyze whether *EmotiConf* can further improve gaze-based methods by capturing the cognitive aspect of the subject, rather than mere visual focus, we used a technique similar to the one proposed in [86]. In this approach, we track the eye gaze and gaze gesture using a method proposed in [199] and then use that information to find out when the participant's mind diverts from the meeting, i.e., the participant is inattentive. Figure 4.10 shows a comparison between *EmotiConf* and gaze-based attention over the data collected during the pilot study. We observe *EmotiConf* performs better than gaze-based attention. Interestingly, while the precision for both the methods is closer, *EmotiConf* has a higher recall than the gaze-based method. While the gaze-based methods can detect most of the inattentive (or attentive) instances (therefore, the precision is high), it also results in high false positives by marking some attentive instances as inattentive. Such false positives are mainly for the cases when the participant gazes on a different application.

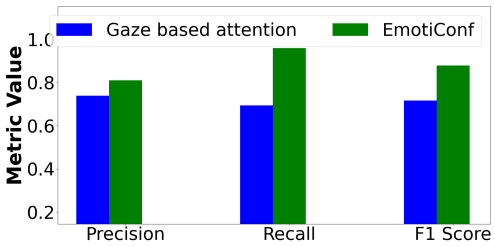


Figure 4.10: Comparing *EmotiConf* with gaze-based attention estimation

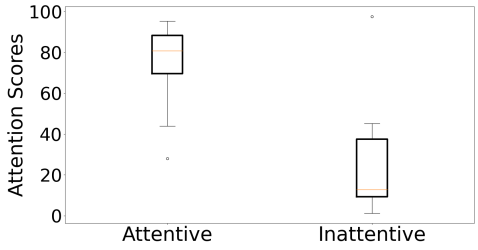


Figure 4.11: Sensitivity of *EmotiConf* in differentiating between attentive and inattentive participants

Sensitivity Analysis

The classification of the participants into *attentive* and *inattentive* depends on the value of a threshold that we use. To analyze the sensitivity of this threshold, we plot the distribution of the attention scores (Figure 5.9), as generated by the synchronous module of *EmotiConf*. The figure indicates that the scores corresponding to the attentive instances are distributed around the high values (more than 60 for the majority of the cases), whereas the scores corresponding to the inattentive instances are around the low values (less than 40 for most of the cases). Consequently, we see that *EmotiConf* produces the attention scores on a scale that indicates significantly different distributions between the two classes. Thus we can obtain a clear threshold for most cases to mark the participants as attentive or inattentive.

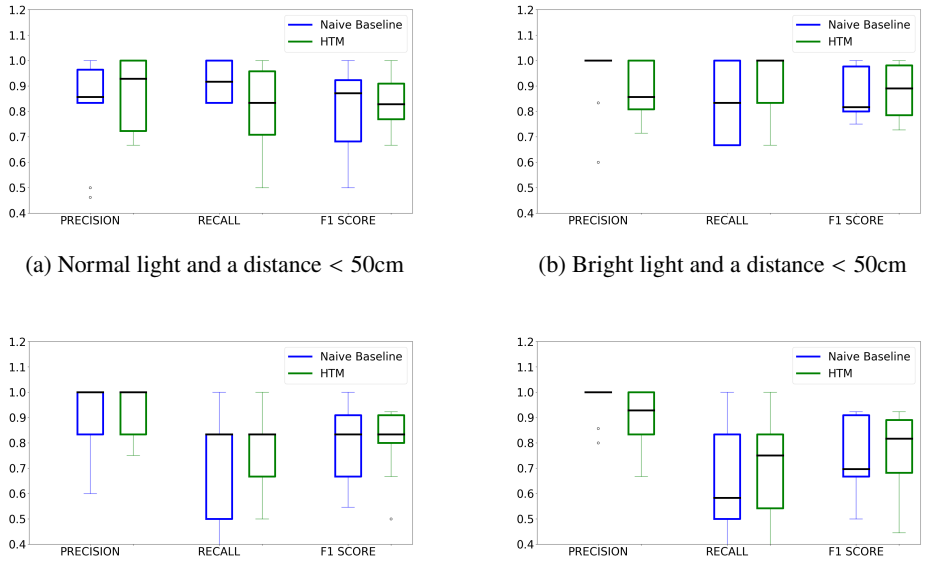
4.7.2 Pilot Study: Multitasking Detection

We use the metrics *Precision*, *Recall*, and *F1-Score* to analyze the performance of *EmotiConf* in correctly identifying the multitasking instances over the Asynchronous Module. *Precision* indicates the percentage of multitasking instances detected out of all the multitasking instances. In contrast, the *Recall* indicates the percentage of correctly detected multitasking instances out of all the detected multitasking instances. We compute the metrics for each participant over every meeting session and then find out the average values.

Design of a Naïve Baseline

To understand how well the HTM-based model used in *EmotiConf* works in comparison with other possible approaches, we design a threshold-based naïve baseline as follows. We compare the feature value (as used in HTM) for two consecutive frames. If the difference of the feature value over two successive frames is more than a threshold, we mark it as a multitasking instance. The core idea in this threshold-based approach is that the intensity difference of the ambient light over two consecutive frames should have a significant deviation whenever the

participant switches the task on her computer screen. Experimental evaluations showed that the average F1 score was ≈ .74 when a static threshold was calculated to be about twice the average difference between two consecutive features for all participants under different setups. However, choosing an optimal threshold dynamically increased the overall F1 score to ≈ .809 for the naïve baseline. As we are interested in comparing the performance of *EmotiConf* with the best possible performance of this naïve baseline, we experimentally choose the threshold value for each participant over each meeting session that results in a maximum F1-Score.



(a) Normal light and a distance < 50cm

(b) Bright light and a distance < 50cm

(c) Normal light and a distance > 50cm

(d) Bright light and a distance > 50cm

Figure 4.12: Comparison of performances between HTM and threshold-based methods

Observations

Figure 4.12 summarizes the results from four different cases with different lighting conditions and screen-to-face distance combinations. We observe that the HTM-based method used in *EmotiConf* in general performs better than the naïve threshold-based approach under a normal lighting condition. However, HTM suffers apparently a bit in terms of F1-score under bright room-lights (Figures 4.12b and 4.12d), notably because the change-points in the intensity of the reflected light from participants' faces gets subverted by the intensity of the room lights. Interestingly, we observe that under bright lighting conditions, the naïve threshold-based approach shows a high precision with a low recall. This indicates that although the naïve threshold-based method detects many multitasking instances, the false positives are very high, thus resulting in a low recall. This implies that although *EmotiConf* suffers in terms of the overall F1-score under the bright room-lights, the overall performance of *EmotiConf* is still

better than the naïve baseline as whatever multitasking instances it detects, it detects most of them correctly.

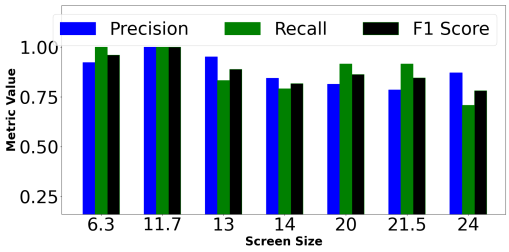


Figure 4.13: Performance of HTM-based visual multitask detection under different screen sizes

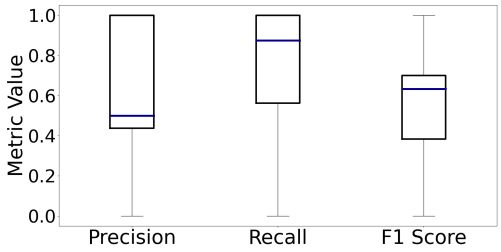


Figure 4.14: Performance of multitask classification

Impact of the Screen Size

The screen size of the primary device plays a vital role in our approach for detecting multitasking instances, as the intensity of the ambient light changes with a change in the screen size. Figure 4.13 shows the impact of the screen size on the performance of the multitasking detection module of *EmotiConf*. The figure indicates that although the performance varies marginally for different screen sizes, it is comforting to see that the F1-Score always remains more than 0.80. Further, we do not observe any specific patterns in the performance based on the screen size. From this analysis, we conclude that factors like the room's lighting condition, screen-to-face distance, the brightness of the screen, etc. influence the performance more than the screen size.

4.7.3 Pilot Study: Multitask Classification

Classification of the multitasking instances is the most challenging module in the design of *EmotiConf*, as it depends on a complex interplay of various factors. We utilized a rule-based heuristic following various general observations from the behavioral patterns of individuals during an online meeting. The rules classify the multitask instances as positive or negative or abstain when none of the rules match. Figure 4.14 shows the precision, recall, and F1-score for classifying multitask instances. The figure shows that although the precision is low, the recall is high, indicating that although the rule-based approach may not classify all the multitask instances, the positive and negative instances that it returns are mostly correct. Indeed we observe that the low precision is attributed to the low accuracy of the vocal emotion detection module, which is around 70%. Detection of vocal emotion is much more challenging than detecting facial expressions, as the audio signal gets significantly affected by the presence of the noise from the environment. This is one of the precise reasons why we avoided utilizing vocal emotions in the synchronous module.

Table 4.2: Distribution of Classification Rate for Classification of Multitasking Instances

Features	Positive instances %	Negative instances %
Vocal emotion, Statement intent, Mouth movement, Face (head) movement	**58.69%**	**66.33%**
Statement intent, Mouth movement, Face (head) movement	39.13%	41.58%
Vocal emotion, Mouth movement, Face (head) movement	19.56%	13.86%
Vocal emotion, Statement intent, Face (head) movement	2.17%	26.73%
Vocal emotion, Statement intent, Mouth movement	56.52%	14.85%

Ablation Study

We primarily used four features to develop our rule-based approach – vocal emotion of the speaker, intent of the statement (speech) from the speaker, mouth movement of the participant, and face (head) movement of the participant. We perform an ablation study here, whereby we use a combination of these features to see how it impacts the performance of *EmotiConf* in terms of multitasking classification. Table 4.2 summarizes the results. The table shows the percentage of positive and negative multitasking instances retrieved correctly by *EmotiConf* with different combinations of features. This ablation study confirms that all four features are important to maximize the percentage of correctly marking multitasking instances in the respective classes.

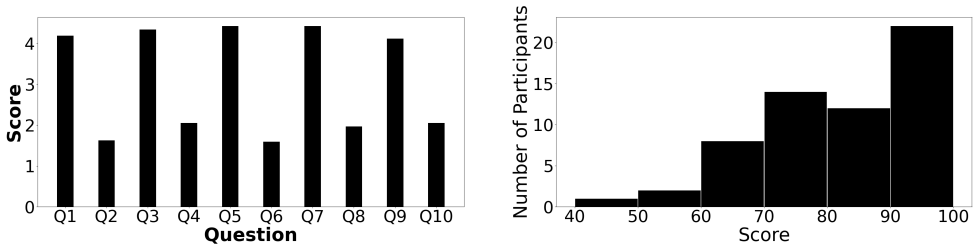
4.7.4 Runtime Performance

During the pilot study, we also compute a few runtime performance metrics of *EmotiConf*. The synchronous module takes 0.00097ms to process each frame for a video feed with a frame rate of 30fps. In general, the output is displayed with a lag of ~ 0.05 sec, which is minimal. Consequently, the attention information can be displayed in real-time. The asynchronous module of *EmotiConf* takes ~ 100 mins to process a ~ 58 mins video and display the output. During the runtime, the synchronous module takes around ~ 60MB, whereas the asynchronous module takes ~ 520 MB of memory.

4.7.5 Usability Study In-the-Wild

Figure 4.15 shows the distribution of the SUS score as obtained based on the feedback from the public deployment and survey (Cronchbach alpha=.753). It can be noticed that the even statements (mentioned in Section 4.6.2) which are negative in nature, have received lower

scores than the odd statements which are positive statements about *EmotiConf*. We get an average SUS score of 80.5, indicating that the respondents on average felt *EmotiConf* as a usable platform. Figure 4.15b shows that the majority of the participants have given an average SUS score of more than 70.



(a) Statement wise score distribution

(b) Participant wise score distribution

Figure 4.15: Statistical distribution of SUS scores: Statement-wise and Participant-wise

The survey also contained a field for providing textual feedback from the respondents. Apart from appreciating the effort, we also obtained a few exciting feedback. One of the participants mentioned *"I think it would change the experience of online meeting calls....It would make everyone as a responsible person....."*. Indeed, one of our primary objectives of developing *EmotiConf* was to design a watchdog that can oversee the participants' activities during an online meeting. Such a platform can undoubtedly boost up the responsibility of individuals during an online meeting. While this is inherent during an in-person meeting as the participants know that others are watching them; however, this has been a challenge for online meeting platforms. Another interesting feedback is – *"One other interesting application could be showing how "interesting" the content is. For example, if many people are focusing their attention on the meeting, that must mean that something interesting is going on."* While this was not in our list of objectives, identification of *interesting contents* can undoubtedly be a by-product of *EmotiConf*.

4.8 Discussion

Overall, we observed that *EmotiConf* performs satisfactorily while tested over a diverse setup. During the entire phases of design, development, and evaluation of the platform, we got some interesting insights that we summarize here.

(1) Diversity of Multitasking: *EmotiConf* performs best in detecting on-screen visual multitasking instances when the participant switches the task on her primary device where the meeting application is also running. There can be other forms of multitasking, like browsing mobiles, reading a storybook, talking over mobile, etc. Such multitasking instances can be

detected directly by applying a video-based activity detection framework such as [6]; therefore, we have not focused on them in this work. *EmotiConf*'s detection accuracy will drop if there are instances where a person's facial expression while reading irrelevant articles during the meeting coincidentally matches with the expression of the conference attendees. However, such cases are sparse and will not significantly reduce the overall accuracy of the system. Further, it isn't easy to correlate such kinds of multitasking instances with the cognitive attention of the participant. For example, a participant can chat with her friend over WhatsApp while simultaneously listening to the discussion in the meeting. Consequently, we believe that a detailed study needs to be done further to characterize such multitasking, which is currently beyond the scope of *EmotiConf*.

(2) Privacy and Permission Issues: *EmotiConf* runs entirely as a wrapper on top of the meeting client and does not use any additional data beyond the ones available with the meeting client. However, while collecting the public feedback, one respondent mentioned that *"Does not make much sense as a lot of permissions would need to be granted leading to breach of privacy."* However, as we mentioned earlier, *EmotiConf* uses the participants' video feeds (from the webcam) and the audio data which the meeting client provides. As the platform works as a wrapper, depending on the OS or the browser setting, the participants might have to give additional permission to access the webcam and the audio; however, it is not more than what is needed for an online meeting client to run.

(3) Impact on the Presenter or the Primary Speaker during the Meeting: During the public feedback, one respondent shared a very interesting observation on the applicability of the synchronous module of the platform: *"The attention of the teacher may be diverted to observing the students rather than presenting the content with interest. ... In the process of identifying the non-listeners, we may not present well to the active listeners. This will compromise the purpose."* We indeed partially agree to this point. For some cases, primarily for the teachers in an online class, such additional information might divert their minds. However, the synchronous module can always be made asynchronous by providing complete feedback about participants' attentiveness at the end of the meeting. Consequently, we may keep this module configurable, depending on whether the participant wants the information live or not.

(4) Complex Device Setup and Environmental Conditions: One inherent assumption behind the design of *EmotiConf* is that the participant uses a single computer screen during the meeting duration. However, in practice, we observed that some participants might use more than one screen and switch between the tasks over those two screens. *EmotiConf* fails to correctly detect the multitasking instances in such scenarios, as the ambient light from one screen interferes with the ambient light from the other screen. Further, a change in room lighting condition, for

example, due to the window curtain movement during the daytime, can also affect the detection and increase false positives.

(5) Utilizing Deep Learning for Multitask Classification: We used a simple rule-based approach based on behavioral modeling of the participants to classify the multitasking instances into the positive and the negative ones. One can certainly argue for using a deep learning model to perform this classification. While we agree with this view, any realistic deep learning model will need massive labeled data to train the model. To reliably generate the labeled data, we need the screen recording of the participants, which provides the ground-truth information. However, it was highly challenging for us to convince the participants to share their screen recordings. As a matter of fact, we could obtain the data from only 5 participants from the online classes, whereas the course had $\sim$ 55 students on average. However, it is comforting to see that the rule-based approach can correctly detect $\sim$ 66% of the negative and $\sim$ 59% of the positive multitasking instances. One exciting aspect of this rule-based method is that the system remains silent when none of the rules matches, resulting in low false positives. Future work will also aim at exploring meeting-oriented datasets[13] for gesture and expression detection.

(6) Multitasking over Applications Having Similar Color Contrast: One of the major criticisms of using the ambient light intensity to detect the multitasking instances can be that the approach may fail when the participant switches between two applications having similar background intensity. However, in practice, we observe that the popular online meeting applications like Zoom, Google Meet, Microsoft Teams, etc., typically use a contrasting background so that the individual video feeds from the participants or the main presentation screen becomes clearly visible. On the other hand, applications that individuals typically use simultaneously with a meeting app, like emails, browsers, notepads, etc., have a white or light-colored background. Therefore, *EmotiConf* works well for such a majority of the cases. *EmotiConf* is not affected by blurred backgrounds as the light reflection is analyzed from the facial region of the participants. However, we admit that a user can cheat the platform by changing the background color contrast of different apps they want to use simultaneously with the meeting app.

(7) Workplace Surveillance and Risks: The experiments with *EmotiConf* also included feedback from several participants. Among positive feedback like *"I think it will change the experience of online meeting calls. It would make every one as a responsible", "A good application for immense use for schools and institutions", "The system is very user friendly and the use case is appropriate for the current scenario. I found this idea remarkably worthy."*, and so on, there were some negative yet interesting feedback that depict that *EmotiConf* is not free from certain loopholes and pitfalls. The evident drawbacks of *EmotiConf* can be discussed in

[13] https://groups.inf.ed.ac.uk/ami/corpus/ (accessed: Friday 11th August, 2023)

the context of the following feedback:
"There's a simple way one can cheat this system... direct the webcam stream to a looping video of a face attentively watching a lecture."
Indeed, there are scopes for cheating the system. Even though liveness detection is not incorporated in *EmotiConf*, playing pre-recoded videos will not have an effect on the performance of the system as the expressions can only match coincidentally. However, one may cheat the system by imitating the expressions of others by observing their faces, rather than listening to the discussion itself.
"The attention of the teacher may be diverted to observing the students rather than presenting the content with interest. My opinion, no need to bother about those who are doing other tasks. Evaluate them based on some Q&A or test. In the process of identifying the non-listeners, we may not present well to the active listeners. This will compromise the purpose." If the attention scores are continuously visbile in realtime, it might both help and hinder the speaker in terms of attentiveness. As a solution, a future versions of *EmotiConf* will provide the option of viewing the scores after the meeting has ended.
"what about security", *"a lot of permissions would need to be granted leading to breach of privacy."*
The most crucial aspect of *EmotiConf* is the maintainance of participant's privacy. While, no user-centric data is stored for processing, many users do not feel comfortable with a system being a "watchdog" or being monitored continuously. In addition, one of the undeniable pitfall of *EmotiConf* is the requirement for all users to turn their camera on. Granting the permission to access their camera might not be comfortable for a lot of users (as observed during our experiments). In addition, automated systems like *EmotiConf* are rarely free from false positives. These false positives in case of multitask detection, can jeopardize the confidence, reputation, assessment and even the attention of some participants.

4.9 Summary

To the best of our knowledge, *EmotiConf* is the first of its kind that develops a platform to detect and characterize multitask instances during an online meeting to understand the cognitive attentiveness of its participants. Such a platform can certainly promote better engagement of the participants towards the meeting, even when they work from home. While the quality of an online meeting has been a serious concern among different professions, a watchdog platform like *EmotiConf* can help in making the meeting more productive. While a thorough evaluation of *EmotiConf* shows that the platform may fail in certain boundary cases; however, overall, we observe a satisfactory performance with feedback of high usability of the platform. Nevertheless, individuals need to be responsible for their duties, as technologies can always be cheated, as in the case of *EmotiConf* too.

5

Ubiquitous System for Detecting Engagement in Online Videos

Imagine an interactive smartphone application that can "sense" its user's mood when browsing social media profiles, watching a movie, or typing a long message on it. Considering the global prevalence of depressive and anxiety disorders, particularly among the young population [80], such an application can pervasively help and alert individuals early. It can even recommend remedies, such as suggesting music, a funny reel, or a comedy movie that helps change the mood. Facial expressions are one of the prominent ways to correctly infer an individual's mood [132] and thus can fulfill the above vision if the smartphone can monitor the temporal changes of its user's facial expressions. Interestingly, there have been decades of research on inferring facial expressions from video or image-based data [134, 159, 280, 16]; however, these works are not suitable to fulfill the above vision of developing a pervasive smartphone application because of the following reasons. **Firstly**, image and video processing is computationally heavy and consumes a significant amount of system resources and energy. Running a computationally heavy model on a hand-held device like a smartphone is not always feasible as it will affect the device's performance. **Secondly**, in the absence of proper lighting conditions, camera-based techniques result in missed detection. In a scenario where a person is watching a movie while sitting in a dark room, camera-based expressions detection will fail due to the lack of ambient light. **Thirdly**, the most important drawback of image and video-based systems is the privacy concern. In systems that aim to monitor all user activities continuously and operate

as a "watchdog", the users often feel uncomfortable. Moreover, continuous camera usage also depletes the battery life at an unusually faster rate.

To achieve the above vision, in this paper, we explore acoustic sensing over a commercial off-the-shelf (COTS) smartphone to identify four basic facial expressions of the user when they browse through a smartphone app (say, watching a movie on Netflix). Due to smartphones' relatively lower processing capabilities, lightweight solutions need to be developed for expression detection, and acoustic processing evokes less power consumption than image processing techniques. Since the trade-off between accuracy and system resource consumption should be optimal while designing expression detection models for smartphones, unlike that for desktops where accuracy is of prime importance, acoustic sensing becomes a suitable approach for smartphones. Summarily, developing

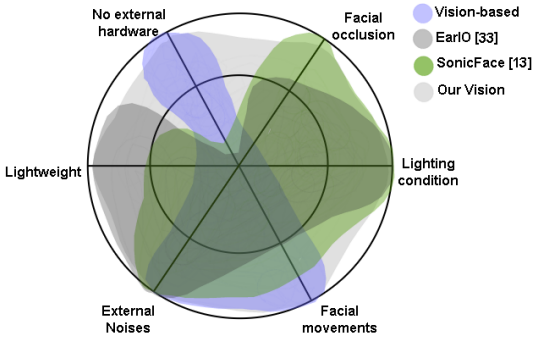


FIGURE 5.1: Our vision in contrast to the existing literature

a pervasive smartphone system for facial expression recognition has the following requirements – (1) lightweight technology with in-device computing, (2) no use of external hardware, (3) performance-invariant under different lighting conditions, (4) correct detection of different facial movements, (5) should not get impacted from occlusions, like glasses or face masks, and (6) performance-invariant from external noises, like motion or interference from other signals. Interestingly, there have been a few recent works [64, 126] that explore acoustic sensing for tracking facial movements. Vision-based approaches [159] can capture a maximum number of facial expressions while capturing the video/image with the embedded camera of the device. However, both SonicFace [64] and EarIO [126], although work based on acoustic sensing principle, they need external hardware supports (while the former needs a microphone array, the latter works on earphones). Further, SonicFace is computationally heavy and typically works on a desktop computer, while EarIO needs specific hair arrangements (ponytail hairstyle). Further, EarIO captures only some specific facial movements (like eye, mouth open/close, and their combinations) and not the expressions directly. Although the design of *ExpresSense* is inspired from [136], it novelty lies in utilizing the simple acoustic features in classifying facial expressions; rather than merely exploring the magnitude of its displacement caused by the movement

of facial areas (like blink). Moreover, *ExpresSense* eliminates the requirement of any static threshold for comparing the derived acoustic features. Figure 5.1 shows how our vision bridges these gaps and addresses the basic requirements.

5.1 How Do We Utilize Acoustic Sensing?

In general, visible facial expressions that last between 0.5–4 seconds are broadly grouped under macro expressions. Apart from these, more subtle and spontaneous micro-expressions last for less than half a second [283]. Irrespective of the type of facial expression, each expression results from a set of *Action Units* (AU) (facial muscle movements). Interestingly, facial muscle movements are the building blocks of expressions and can be categorized under *ocular* (around the eye region), *nasal* (around the nose region), and *oral* (around the mouth area) groups [228]. For example, a particular expression, say "*Happiness*", is a combination of facial muscle movements, mainly around the oral region and subtly around the ocular and nasal area. The core idea of this paper is to detect such AUs through lightweight acoustic sensing over a COTS smartphone to capture a subject's facial expression. In contrast to the complex signal processing techniques over microphone array as used in SonicFace [64] or deep learning model used in EarIO [126], we rely on simple signal processing techniques and light machine learning models that can effectively be implemented on a smartphone app, while using only the embedded smartphone hardware.

5.2 Contributions

In the essence of the above discussion, we propose *ExpresSense*, a lightweight smartphone application that utilizes near-ultrasound acoustic signals to detect four basic expressions [90, 7] of a user: *Happiness*, *Anger*, *Surprise*, and *Sadness* (or *Neutral*[1]). *ExpresSense* transmits chirps between the range of 16-19 kHz, using the inbuilt speaker of a commercial smartphone. The reflected chirps are recorded through the single microphone of the smartphone, filtered, and processed to derive the amplitude and phase of the reflected signal for different frequency bins. The frequency bins that contribute the most to predicting the expressions are selected. This process of bin pruning is followed by utilizing the phase and amplitude of the echo from the desired frequency bins and using them to predict the corresponding expression of the subject by a pre-trained ensemble of classifiers.

The significant challenges in developing *ExpresSense* arise from three primary aspects: design, development, and data. From the perspective of designing the system, the underlying characteristics of facial expressions, acoustic signals, and their intricate correlation needs to

[1]Existing literature argue that for both *Sadness* and *Neutral*, there is no visible sign of the facial gestures, and they appear to be very similar; therefore, these two expressions are used interchangeably in the literature [122, 255, 89].

be considered. Assessing whether the facial expression has an identifiable effect on the signal reflection concerning the signal features poses a challenge in designing the proposed system. Next, a major challenge is to develop a system that overcomes the limitations of commodity smartphones. We address this issue by developing a model that can thrive on any commercial smartphone's very minimal and basic capability. The main contributions of this work are as follows.

1. Development of a lightweight and camera-free smartphone application that uses acoustic signals for facial expression detection. The system works on a standalone smartphone and requires no additional hardware.

2. Experimental analysis of *ExpresSense* and its application, demonstrating its significant performance under a realistic environment for both posed and natural expressions.

We conduct a thorough lab-scaled study with 12 participants to evaluate the performance of *ExpresSense*[2] – both as a standalone platform under a controlled environment as well as an embedded application in the wild. The experimental results reveal that *ExpresSense* can work with an average accuracy of about $\approx$ 75% as a user-dependent model and performs significantly well under different conditions like angular variance, ambient sound, motion, and so on. Further, to evaluate *ExpresSense* under a realistic setup with natural expressions, we develop a smartphone app that can monitor and measure user engagement while watching a streaming video. The application matches the overall facial expressions of the subject and the video genre to provide a temporal variation in the engagement score, which we compare with the ground truth captured from questionnaires and self-assessment. From a thorough study with the 12 participants under an in-the-wild setup, we observe that the app performs with an average F1-score of 0.84. Further, we performed a large-scale study with 72 participants to test the usability of *ExpresSense* with the help of this video streaming app that monitors the users' engagement and shares a summary report with them at the end of the streaming. The study revealed a high usability score of 85, indicating the system's effectiveness.

5.3 ExpresSense Design: Opportunities and Challenges

We start with a pilot experiment to understand how the acoustic signal generated from a COTS smartphone gets impacted by the movement of facial muscles. By showing how near-ultrasound signals can detect such movement, we then discuss the challenges associated with a COTS smartphone in designing a system like *ExpresSense*.

[2]We have taken the institute's ethical committee's approval to perform all the human studies reported in this paper.

5.3.1 Pilot Study

ExpresSense considers Frequency Modulated Continuous Wave (FMCW) or chirps that have linearly increasing frequency in time. Based on the characteristic of transmitted and reflected signals, we can say that a reflected signal is only a time-delayed variation of the transmitted chirp. In [136], the authors explain how a change in the reflective surface, along with subtle movements causes a shift in both phase and amplitude of the signal. This is caused by both the nature of the reflective surface and the length of the path the transmitted signal travels before hitting the reflective surface. If the reflective surface is skin, some signals will be absorbed, causing higher attenuation. On the other hand, more reflective surfaces, like teeth, eyeballs, etc., will result in lower attenuation. This affects the amplitude of the signal. Therefore, the amplitude value of the signal when a person is laughing (teeth are visible) or surprised (eyes enlarged) will be significantly different from when they are sad (mouth closed, eyes normal). The signal path is affected by actions like an eye blink when the eyelid comes before the eyeballs and reflects the incoming signal. This causes a change in the signal phase. In terms of facial expressions, the phase of a signal will be significantly different when a person is surprised (mouth opened) or laughing (lips separated) from a scenario when the person is sad (lips cover the teeth).

Thus, in contrast to the existing approaches [64, 124, 66, 25, 129] that use complex frequency-domain signal processing techniques, we consider the time-domain amplitude and phase of a signal by intelligently and adaptively choosing a frequency bin that can capture the facial expression information. For choosing the frequency bin, we rely upon the discussed characteristic of amplitude and phase variation. If there are moving objects in the environment, like ceiling fans, moving curtains or passing vehicles, the changes in amplitude of the signals from these mediums will be less due to the static reflective surface. Based on this understanding, we select the frequency bin with the largest variance in the phase. We next analyze how the amplitude and the phase of a signal varies due to the movement of facial muscles.

Methodology

To test whether smartphone-generated acoustic signals can differentiate between relaxation and contractions of facial muscles in different regions, we performed a pilot study with two subjects. We asked the subjects to perform the following sequence of facial actions while relaxing the facial muscles in between every two actions: *Raise eyebrows/frown* (B), *blink* (E), *raise left cheek* (CL), *raise right cheek* (CR), *move mouth region (left and right)* (M), and *smile* (S). During the experiment, we placed a smartphone in front of their faces at a distance of about 30 cm and an elevation and angle of zero degrees. We continuously played FMCW signals between 16-19 kHz, through the speaker. The reflected signals are recorded through the microphone. After a series of processing (explained in Section 5.5), we derive the phase and amplitude of the

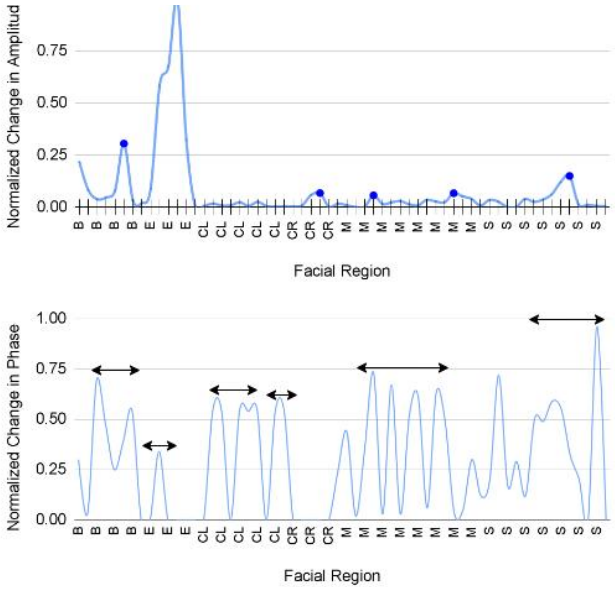


Figure 5.2: Facial Muscles and Acoustic Feature Variation

Figure 5.3: Movement of facial muscles due to Forced Expressions (FE) and Natural Expressions (NE)

reflected chirps and select the frequency bin with the most significant variation, as mentioned earlier. We then plot the absolute difference in consecutive amplitudes and consecutive phases of these signals with respect to the time.

Observations

The pilot study led to the following observation: (1) Facial AUs involving a significant muscle movement induce a change in either amplitude, phase, or both. (2) AUs that involve lower muscle movement and no change in the reflective surface do not affect the signal features. For example, the movement in the nasal region failed to change the amplitude and the phase in the selected face-bin. To further compare the degree of effect individual AU has on the phase and the amplitude, we plot Figure 5.2. We prune out the zones with no movement (face is relaxed) and nasal movement as they cause no variation in features. Then, we plot these variations for phase and amplitude individually, along with the corresponding facial AU that caused it. Figure 5.2 shows that each AU has resulted in a peak in the plot. However, it can be seen that the blink has caused the most prominent peak, whereas the AUs around the oral region

generated lower peaks. This observation infers that the amplitude and the phase of an FMCW signal reflection can be used to predict the facial expressions of a subject due to their varying characteristics and correlation with the underlying expressions.

Notably, we observe the above variations in the amplitude and phase values of the reflected signal for facial expressions that are posed. In most popular image data sets, facial expressions are posed and quite different from more subtle expressions in real life. In Figure 5.3, the difference between posed or forced expressions and natural expressions is shown in terms of displacement around ocular and oral regions. It can be seen that forced expressions are much more animated, thus causing much more movement of the AUs. However, normal expressions that can be seen in daily life are more abstruse in terms of AUs. Therefore, in *ExpresSense*, we aim to estimate how smartphones can distinguish between these challenging natural expressions, by exploring the features (amplitude and phase) of the reflected signal and then learning the subtle variations in the signal properties through light-weight machine learning approaches.

5.3.2 Challenges and Design Ideas

Although observing the time-domain signal properties for judicially selected frequency bins provides us an opportunity to develop a lightweight model for detecting facial expressions over smartphones, we still face the following challenges.

Supported Frequency Range : Majority of the COTS smartphones work over the audible frequency range (< 20 kHz). Moreover, signals above 19 kHz are very noisy and unsuitable for the purpose of sensing. Consequently, in *ExpresSense*, we use a near-ultrasonic range of 16-19 kHz to utilize the maximum possible bandwidth of 3 kHz and ensure the least overlap with the audible range. Notably, even though the audible range varies between 20 Hz to 20 kHz, most of the audible sounds lie below 16 kHz, thus ensuring minimal interference with *ExpresSense*.

Single Microphone: The popular acoustic sensing approaches [64, 124, 66, 25, 129] use a microphone array to estimate the Direction of Arrival (DoA) of the signal, thus eliminating the unwanted signal components. In contrast, most COTS smartphones use a single microphone, thus limiting the number of signal properties we can utilize. Further, approaches like SonicFace [64] uses signal fusion technique to precisely track even minor object movements, which is not possible with a single microphone due to the lack of sophisticated interference cancellation techniques. Therefore, *ExpresSense* solely relies on lightweight, intelligent signal processing and machine learning methods to judicially select the frequency bins, which can eliminate signal components reflected from the background objects.

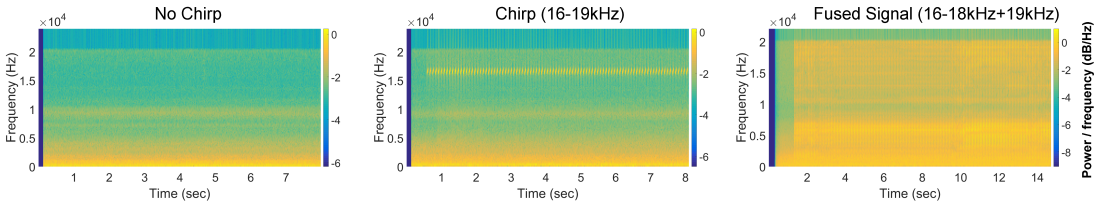


Figure 5.4: Spectrum of received signals when only no chirp is played (left), chirp is played (middle) and fused signal is played (right)

Signal Fusion: In acoustic sensing, Signal Fusion is a technique where two different signals (such as a chirp and a pure tone signal) are transmitted in sync so that both coarse-grained and fine-grained movements of a target can be tracked. However, in a standalone smartphone, Signal Fusion can cause severe interference in the recorded signals and generate impaired and noisy data. Figure 5.4 shows why *ExpresSense* will be unusable if Signal Fusion is performed. The figure is a waterfall graph of ambient sound in absence of chirps (left image), presence of chirps (middle image) and presence of fused signal (right image), as generated and recorded by a smartphone. The figure is a time series plot of ambient signal frequencies between 0-20kHz where the brighter color zones indicate higher intensity and darker colors indicate lower intensity of signals. We see that chirps only effect the frequency bins between 16-19kHz whereas, fused signals effect different frequency bins and create noise component in all bins. It also increases the intensity of the sound, thus making it unsuitable for non-intrusive applications.

Multi-Class Expression Detection : In context of the above device-related challenges, the task of expression detection becomes difficult. Although some of the previous works have considered a single or dual microphone(s) [65] and signals with frequency between 16-19kHz, for tasks like continuous tracking respiratory functions [229] or binary classification of authentic and unauthorized users [32, 109], multi-class classification of facial expressions using acoustic sensing is particularly challenging due to the following reasons: (a) The corresponding features should not only reflect deviations indicating displacement (like chest movement), but should also account for differentiating patterns induced by movement of different facial AUs, (b) without at least 2 microphones, signal enhancement processes like beamforming are not possible, making the selection of relevant frequency bins; and hence the corresponding features; more challenging.

Lightweight: The developed system should be lightweight to be deployed directly on the smartphone. Existing studies use complex frequency-domain signal processing techniques or computationally heavy deep learning models. Such methods will consume significant computational resources (like RAM) and thus can slow down other running services on the smartphone. So, we have to rely on lightweight techniques that need minimal computation and can work in

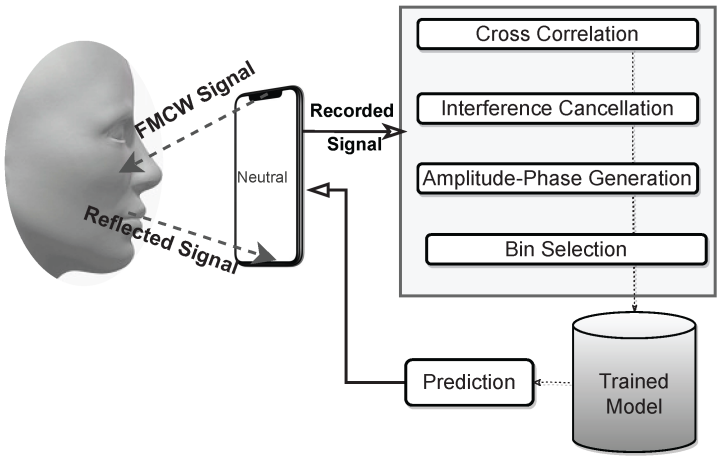


FIGURE 5.5: The overview of *ExpresSense*

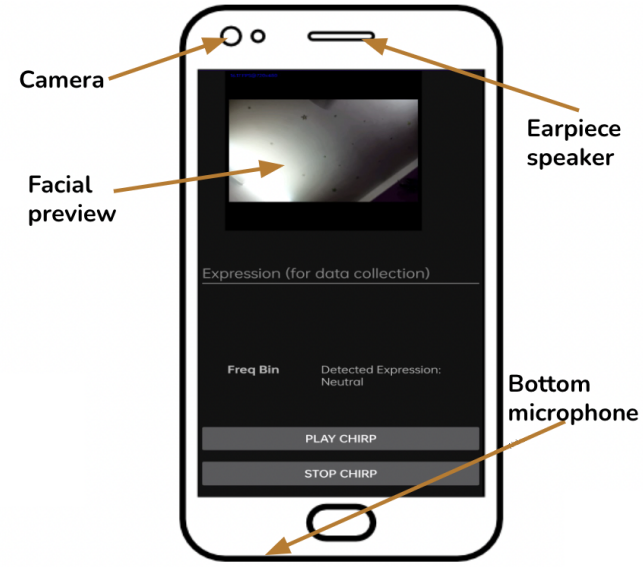


FIGURE 5.6: The interface of *ExpresSense* for data collection.

real time.

5.4 The Overview of *ExpresSense*

Figure 5.5 shows the overview of *ExpresSense* in terms of chirp transmission, reception, and in-device processing. We start with discussing an overview of the proposed architecture, followed by a discussion on the smartphone applications developed for data collection and prediction, adhering to the proposed architecture.

5.4.1 Proposed Architecture

In *ExpresSense*, a standalone commodity smartphone is used for camera-less expression detection using near-ultrasound signals. The overall idea of *ExpresSense* is to generate near-inaudible chirp signals ranging from 16-19 kHz utilizing the smartphone's speaker. The smartphone's microphone is used to capture the echo of the signals reflected from the facial regions of the subject, along with other objects in the subject's vicinity. The recorded echo is then processed to prune out the static multi-path interference. The amplitude and phase from different frequency bins are extracted from the residual signal. In the next phase, only the appropriate bin is selected, and the phase and amplitude features from this bin are passed through a learning algorithm to predict the expression of the subject's face as one of the four basic expressions: *Sadness*, *Happiness*, *Surprise*, or *Anger*. The following sections describe the individual modules of *ExpresSense* in detail.

5.4.2 The Smartphone Application

As shown in Figure 5.6, an Android phone usually has a camera on the top, besides an earpiece speaker. This speaker is responsible for transmitting in-call audio to the user. Apart from this speaker, every smartphone has a main speaker that facilitates the transmission of sounds like call alerts, music, etc. In addition, a typical smartphone generally contains one microphone at the bottom. This microphone aids in capturing voice and other ambient sounds. However, some modern smartphones are equipped with two or more microphones. In this work, we assess smartphones with minimal capabilities, thus considering only those with a single microphone.

Figure 5.6 shows the interface for data collection. We use the earpiece speaker for transmitting the chirps with low intensity. The main speaker can also be used for sending signals. The interface consists of simple play and stops buttons for starting and stopping the chirps. It also contains a text box for manually entering the ground truth label for expressions like *Happiness*, *Anger*, *Surprise*, and *Sadness*. Apart from that, we also use an automated ground-truth labeling method using image-based techniques. For this purpose, the application uses a camera preview that tracks the subject's facial region and automatically detects the expression using an image-based detection model. This camera-oriented feature has been used to validate the manually entered expression labels, as discussed in Section 5.7. Notably, we use the camera only to collect the ground-truth under a pure lab-scale controlled setup, and the system does not need it during its actual runtime.

5.5 Design Details

We now discuss the details of each sub-module used in *ExpresSense*.

5.5.1 Generation of FMCW Signals

We consider FMCW signals or chirps that sweep from f_{min} = 16 kHz to f_{max} = 19 kHz. Each chirp is played for an optimal duration of T = 40 ms; T is directly proportional to the degree of overlap in the reflected echoes. Consecutive chirps are played while *ExpresSense* remains active and are separated by a silent period of T_{sil} = 30 ms. T_{sil} ensures that a recorded chirp does not contain a part of the next chirp being played in sequence, thus reducing the interference through overlap.

For a linear chirp, like the ones generated in *ExpresSense*, the instantaneous frequency (f) is dependent on the time (t) and can be expressed as,

$$f(t) = f_{min} + ct, \tag{5.1}$$

where the chirp rate is $c = \frac{f_{max} - f_{min}}{T}$. The corresponding phase of the same signal is an integration of $f(t)$ and can be expressed as,

$$\phi(t) = \phi_{min} + 2\pi(\frac{c}{2}t^2 + f_{min}t), \tag{5.2}$$

where ϕ_{min} is the phase at $t = 0$. Now, by considering the sign of Equation (5.2), the linear chirp can be expressed as the function of time t. Hence,

$$x(t) = sin[\phi_{min} + 2\pi(\frac{c}{2}t^2 + f_{min}t)] \tag{5.3}$$

Reception of Reflected Signals

In exponential form, Equation (5.4) can also be written as,

$$x(t) = e^{-j2\pi(f_{min}t + \frac{c}{2}t^2)} \tag{5.4}$$

Let r be the reflected signal that is also a function of time t. As discussed in Section 5.3.1, a reflected signal is only a time-delayed ($t-\tau$) version of the original signal, having a time-of-flight τ. Therefore, it can be expressed as,

$$r(t) = \sum_{p=1}^{N} \alpha_p e^{-j2\pi(f_{min}(t-\tau_p) + \frac{c}{2}(t-\tau_p)^2)} \tag{5.5}$$

Notably, the reflected signal r is a superimposition of multiple signals received from various environmental reflectors. For *ExpresSense*, the primary reflector is assumed to be the subject's facial region by default. However, besides reflectors like walls, furniture, etc., the echo also contains the direct path reflections from the speaker to the microphone. In Equation (5.5), N denotes the number of such multi-path signals. α is the signal attenuation. Similar to [136],

we define the mixed signal to be recorded as a multiplication of the received signal with the complex conjugate of the transmitted signal, as follows.

$$r_m(t) = \sum_{p=1}^{N} \alpha_p e^{-j2\pi(c\tau_p t + f_{min}\tau_p - \frac{c}{2}\tau_p^2)} \tag{5.6}$$

The recorded signal is then passed through a high pass filter to allow the frequency range above 15.9kHz. This automatically removes the ambient noises that fall within the audible frequency range.

5.5.2 Processing of the Recorded Signals

The in-device signal processing begins with synchronizing the speaker and the microphone. This allows us to remove the delay between the transmitted and the received signals caused by the device's imperfection. This is achieved by measuring the signal similarity between the generated and the reflected signals at different delays through cross-correlation. Assuming there are N points in the transmitted and the reflected chirps, the normalized cross-correlation of the transmitted signal $x(t)$ and the received signal $r_m(t)$, shifted by n, can be expressed as,

$$X_{corr}(n) = \frac{\frac{1}{N}\sum_{t=0}^{N}[r_m(t) - \overline{r_m}][x(t-n) - \overline{x}]}{\{\sum_{t=0}^{N}[r_m(t) - \overline{r_m}]^2 \sum_{t=0}^{N}[x(t-n) - \overline{x}]^2\}^{\frac{1}{2}}} \tag{5.7}$$

Here, $\overline{r_m}$ and $\overline{x}$ can be estimated by taking the average of r_m over N points and x over N points, respectively. Pertaining to the nature of the direct path signal, it should be maximally correlated to the generated chirp, as it suffers from no reflection. Thus, the delay at which the correlation value is maximum is considered in *ExpresSense* to synchronize the speaker and the microphone of the smartphone. Hilbert Transform [94] is used on the reflected signal to derive the analytic signal, i.e., its representation in the exponential form. As mentioned earlier, $r_m(t)$ is a product of this analytic signal and its complex conjugate in this domain. As we are interested in only the real part, the final expression of the mixed signal is just the real part of this complex multiplication.

This step is followed by the *Fourier Transformation* of the received signal and static interference cancellation. In signal processing, static multi-path reflections are created from objects whose locations are fixed in the environment. For example, the chirp generated by the speaker will also be reflected from a wall, present behind the subject, a nearby table, and so on. Since these objects are static, the irrelevant noise induced by the reflections from these objects can be subtracted from the overall reflection capturing both static and dynamic reflections (caused by facial expressions). To achieve this, we first generate template recordings by transmitting chirps in an environment without the subject's presence. As a new session begins, where the subject is present in front of the device, the recorded template is subtracted from the reflected

chirp, thus eliminating the external interference caused by static objects in the room.

The next step is to select the frequency bins that are most likely to capture the information about different facial regions (*Facial AUs*). As we use a bandwidth of 3 kHz, any two objects or points of reflection, separated by a distance of 5.6 cm or more, will result in a reflected echo, having distinct frequencies. For example, while the signal path generated by the reflection of the transmitted chirp from the eye region will fall in a frequency bin f_1, that from a wall behind will fall in a different frequency bin f_2. To further ensure that the frequency bin capturing the information about the face is selected, we depend on the phase-amplitude variation induced by AUs. Thus, we choose the frequency bin with the maximum variance, as shown in [136].

5.5.3 Prediction of expressions

Finally, the phase and amplitude of the signal with the selected frequency are used to predict the subject's facial expression using an ensemble of three different classifiers using a majority voting technique. The three classifiers are chosen by comparing the classification accuracy of different algorithms. Empirically, we observed that the average accuracy of *Support Vector Machine* was 17.5%, *3-Nearest Neighbour* was 44.1%, *Adaboost with* 50 *estimators* was 47%, *MLP Classifiers* was 49%, *Random Forest with maximum depth of 10* was 72.72%, *Decision Tree* was 96.61%, *Logistic Regression* was 66.41% and *Naive Baye's Classifier* was 36.29%. In our implementation, the three classifiers with the highest accuracy, which were chosen for the ensemble, were – Logistic Regression with L2 Penalty [3] and Limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) solver[4], Decision Tree with Gini impurity [5] and the Random Forest with a maximum depth of 10. By using the majority voting method with these three best classifiers, we achieved an overall improvement of $\sim$ 4% in the classification accuracy, as compared to the average accuracy of the individual models (details in Section 5.7).

5.6 Implementation, Resource Profiling, and Evaluation Methodology

This section provides the implementation details and resource consumption benchmarking of *ExpresSense* and an overall discussion on how we conduct the evaluation of the proposed system in a principled way. The implementation of *ExpresSense* along with partial data (anonymized) has been made open-sourced[6].

[3] https://medium.com/@aditya97p/l1-and-l2-regularization-237438a9caa6 (Accessed: Friday 11th August, 2023)

[4] https://machinelearningmastery.com/bfgs-optimization-in-python/ (Accessed: Friday 11th August, 2023)

[5] https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33 (Accessed: Friday 11th August, 2023)

[6] Code link: https://github.com/anonymous0304/ExpresSense.git.

5.6.1 Implementation Apparatus

ExpresSense has been developed for Android Platforms, using Android Studio[7]. A sampling rate of 44100 Hz has been considered for the FMCW signals, which are encoded using Pulse Code Modulation (PCM) to represent the sampled signals. For transmitting these signals, the `AudioTrack`[8] class has been used, which allows the streaming of Pulse Coded signals. In order to observe the general capability of commodity smartphones in facilitating acoustic sensing-based facial expression detection, we have installed and used the application on different commodity smartphones (Realme and Samsung) having 4, 6 and 8 GB of RAM. The minimum Android Version considered is 9. The chipset of the tested devices include Qualcomm Snapdragon 730G, Qualcomm Snapdragon 710, Qualcomm SDM730 Snapdragon 730G (8 nm), MediaTek Helio G95 (12 nm), Qualcomm SDM675 Snapdragon 675 (11 nm), and so on. For testing the accuracy of different models in predicting expressions, we train and test the various models offline using Python sckit-learn[9] library. Finally, the trained model is uploaded to the Heroku[10] platform and connected to the smartphone application using an intermediate FLASK-API[11]. It is to be noted that the entire signal processing part is executed locally in the user's smartphone and only the numeric values of generated amplitude and phase are sent to the remote server for being predicted. The class label is then communicated back to the user's device. For FMCW signals, range resolution can be defined as the ratio between the speed of sound in the air and twice the bandwidth of the signal [136]. Hence, for a signal with a bandwidth of 3 kHz, if there are two reflected signals, caused by two objects, placed at a distance of 5.6cm or more, with respect to the sound source, then these two signals will fall in different frequency bins. Hence, we need to select a particular frequency that corresponds to the reflection from the facial region of the person. Due to some of the inherent limitations of a smartphone (discussed in Section 5.3.2), it is difficult to reduce the range resolution of the signal, without making it overlap with the audible sounds. Thus, we assume that the user's face and any other nearby object are separated at least by a distance of 5.6 cm. If there is an object very close to the face ($<$ 5.6 cm), then the selected frequency bin will also reflect the information of the second object, along with the face.

5.6.2 Profiling the Resource Consumption

ExpresSense consumes around 75-112 MB RAM during the runtime. The lightweight nature of our model is tested by fully charging a smartphone and keeping the application on till the charge

[7] https://developer.android.com/studio (Accessed: Friday 11th August, 2023)
[8] https://developer.android.com/reference/android/media/AudioTrack (Accessed: Friday 11th August, 2023)
[9] https://scikit-learn.org/stable/ (Accessed: Friday 11th August, 2023)
[10] https://www.heroku.com/ (Accessed: Friday 11th August, 2023)
[11] https://flask.palletsprojects.com/en/2.1.x/ (Accessed: Friday 11th August, 2023)

drops to 20%. The application could continuously run for more than 7 hours. Further, to ensure near real-time processing, we analyzed the time taken to process the received signals using different smartphones with 4, 6, and 8 GB RAM. For this purpose, we calculate the processing time for each chirp for different expressions. The time reflects the total time in which a reflected chirp is captured through the microphone and processed locally in the smartphone to generate the amplitude and phase values from the selected frequency bin. The average processing time was estimated to be $\approx$ 5 seconds, $\approx$ 3.5 seconds, and $\approx$ 1 second, respectively, with the three different RAM availability.

5.6.3 Evaluation Methodology

To evaluate *ExpresSense* in a principled manner, we set the following objectives to analyze the system thoroughly under different aspects.

1. How well can *ExpresSense* infer the four basic facial expressions of a subject in general?
2. How do different environmental factors, like the elevation, orientation, and tilting of the phone, motion of the subject, ambient sound, hand placement, glasses, finger movement, etc., impact the performance of *ExpresSense*?
3. How does *ExpresSense* perform under natural expressions?
4. How usable *ExpresSense* is in practice?

Evaluating *ExpresSense* under objectives (1) and (2) above is straightforward, as we can go for a controlled lab-scale setup where trained subjects can pose for different expressions under different conditions for a short and fixed duration while holding the phone in front. We can also collect the ground truth using other modalities, such as self-annotation, annotation through one or more dedicated volunteers, or well-established vision-based automated labeling techniques by capturing the subject's face through the phone's front camera. We performed controlled experiments to evaluate *ExpresSense* in a general setup, as discussed in Section 5.7.

However, evaluating *ExpresSense* under Objective (3) is challenging. First, we need third-party applications that can naturally trigger changes in the subject's facial expression. For example, a video streaming app may trigger natural changes in facial expression based on the genre of the video being streamed. Second, annotating the data is challenging as the facial expressions may change continuously, so we need precise time boundaries when the expression changes. Human annotation cannot work with this precision. Further, automated annotation using the camera may cause discomfort to the subject or may divert their attention, thus affecting their natural expressions. To solve this issue, we use an indirect way of evaluating the system by utilizing the existing research on gauging human engagement through facial expressions [51, 20, 221]. We match the temporal changes of the subject's facial expression

with the video genre and derive an engagement score for the entire duration of the video streaming session. We then collect the ground truth through questionnaires and self-assessment and match the ground truth with the computed engagement score. The underlying hypothesis is that if there is a good match between the computed and the ground-truth engagement scores, then *ExpresSense* has captured the facial expressions accurately. Section 5.8 discusses this evaluation methodology and the corresponding results in detail.

Finally, we performed a thorough usability study in the wild using the proof of concept (PoC) video streaming application that can inform the subjects about their engagement level while watching the video.

5.7 Evaluating *ExpresSense* under a Lab-Scale Controlled Environment

First, we explore the performance of *ExpresSense*, as an individual module for detecting facial expressions based on acoustic chirps from a standalone smartphone. The details follow.

5.7.1 Experimental Setup

The data collected for training and testing *ExpresSense* has been generated in a monitored setup from 10 participants (P1-P10) who volunteered in the evaluation. These 10 participants (4 females, 6 males) belonged to different age groups and professional backgrounds. Two participants belonged to the age group of 20-25 years, four participants belonged to the age group of 26-49 years, and the rest belonged to the age group of 50-65 years. To ensure professional diversity, we chose the participants in such a way that three belonged to the IT industry and were Software Engineers, two were Undergraduate students, two were home tutors, one belonged to the banking sector, one was a research scholar, and one was a retired personnel. Four of them used glasses and others had normal eyesight.

In this setup, the participants were asked to place the smartphone at a distance of $\approx$ 30 cm from their faces. The angle of elevation of the device with respect to the face was not fixed a priori; however, in all the cases, the participants preferred to place the phone at about -20^o (on the vertical axis), corresponding to the face. The azimuth angle of the smartphone with respect to the face was roughly 0^o (directly in front of the face), as this was the natural viewing angle for all the users. The subjects were asked to hold the smartphone by hand or use a smartphone holder for convenience during the session. One of the participants performed the experiment in complete darkness, while others performed under normal lighting conditions. The experiment was performed indoors, in the presence of natural ambient sounds generated by ceiling fans, outdoor noises (like cars passing by, children playing on the ground, etc.), and so on. For collecting the data, the methodology aligned with that explained in Section 5.8.2.

For each participant, the entire experiment was conducted in three different sessions, preceded by a training session. In each session, the participants reproduced the four facial expressions in different orders. In each session, every expression was captured for a duration of 1 min when the participant could render different variants of a chosen expression, followed by a pause between two expressions. The participants were free to choose the pause duration. We advised the participants to keep the expressions as natural as possible.

5.7.2 Ground Truth

The ground truth generation and validation has been conducted in a 3-layered technique. Firstly, the data was labeled manually by the participants themselves. During the pause phase between two expressions, the participants entered the expression to be produced next in a text box in the application. These expression labels acted as the ground truth. Secondly, automated labels were generated. The application constantly monitored the facial expressions of the subjects through the camera and automatically predicted the expressions using the trained MobileNet model [205], incorporated into the application. For the automatic detection of the facial expressions, we have used the MMA Facial Expression dataset[12] which contains about 92, 958 training and 17, 356 testing images from different expression categories like surprise, fear, angry, neutral, sad, disgust, happy. The MobileNet V2 [205] model has been trained for this purpose as it provides a significant accuracy (loss=.01) and performs real-time predictions on mobile devices. Thirdly, labels were verified through close monitoring. During the experiment, the participant's expressions were closely monitored by two different individuals to ensure that the correct expression is being produced with respect to the manually entered label and the expression sequence. In case of disagreement, the participants were requested to repeat the expression.

Finally, the manually entered and auto-generated labels were synchronized and compared for validated sessions. It is to be noted that, pertaining to the possibility of misclassification of expressions by MobileNet, more weight was given to the manually entered labels. For example, if for one minute of "*Happy*" (manually entered) expression, the MobileNet predicted expression was "*Happy*" for the majority of the data, then the rest of the data labels, although classified as a different expression, were also considered as "*Happy*". Finally, a simple pruning was performed to select the data points generated about 3–5 seconds after clicking the "*Start Chirp*" button and those generated before 3–5 seconds of clicking the "*Stop Chirp*" button. This is because we observed that for most of the participants, the actual expressions started a little after starting the chirps as they shifted their focus from clicking the button to creating the expression during that time. Similarly, the participants released the expressions slightly before stopping the chirps as they cognitively prepared to press the button.

[12] https://www.kaggle.com/datasets/mahmoudima/mma-facial-expression (Access: Friday 11th August, 2023)

Dataset collected. From each session, we finally collect a total of 20 data points (each of ≈ 1 min duration with labeling and pruning as discussed above) for each class in a session, i.e., a total of 80 data points per session. Each data sample is a pair of amplitude-feature values, along with the ground truth label. The entire dataset thus contains ≈ 2400 data samples from all four classes across all 10 participants combined. Apart from these regular sessions, the users were also asked to attend additional sessions where data was created in a similar method for different conditions like elevation and angular change of the device, different ambient sound levels, degree of motion, different hand positions for holding the device, different degree of finger movement, and presence of surrounding objects. In total we have collected ≈ 6880 data samples for the lab-scale controlled experiments, which have been divided into train-test splits for different test scenarios, as we explain later for individual cases.

5.7.3 Results

We next discuss the user-specific system performance of *ExpresSense* and its sensitivity analysis.

User-Specific System Performance

We first explore the performance of *ExpresSense* for individual subjects. We analyze *ExpresSense* from three perspectives – *overall performance*, *inter-session performance* and *intra-session performance*. In the overall performance testing, we mix the data from all sessions and make a 4:1 data segmentation for training and testing the model based on stratified random sampling. By shuffling the data and randomizing the segmentation, we run the learning model 10 times and present the average accuracy of the system per subject. To further analyze whether the system can capture the characteristics of the subjects from independent sessions, we perform an inter-session performance estimation. Here, we train the model using data from two sessions and predict the data derived from the third experimental session of individual subjects. This allows us to better estimate if *ExpresSense* can capture the overall characteristics of a subject from independent sessions. Finally, in the intra-session study, we trained the model with ≈ 16 out of the 20 data points for each expression class in a single session, and the rest 4 data points per class per session were used to test the model.

Figure 5.7 shows the comparison of the overall, inter-session, and intra-session accuracy of detecting expressions of individual subjects. In the *overall performance* assessment, *ExpresSense* achieves an average accuracy of ∼ 73% across all the subjects. In *inter-session performance* testing, we achieve an average accuracy of ∼ 73.5%. However, it can be seen that the individual accuracy for some subjects from the inter-session study has been improved (≈ 2.45%) than the overall accuracy (when data is shuffled). This infers that the model is able to learn the feature pattern better when individual sessions are fed to it, particularly because of the inter-session variations caused by the dislocation of positions and renewed expressions.

In *intra-session performance* testing, we achieve an average accuracy of 74.6%. The observation aligns with the expected output, as individual sessions are more likely to have distinctive features due to the lack of involuntary positional and angular shifts. However, for P7, the accuracy dropped from 96% in inter-session evaluation to 83% in the intra-session study. Notably, P7 took less pause time than other participants, which induced muscular fatigue, causing the participant to rapidly produce variations while holding individual expressions.

To further analyze the performance of *ExpresSense* in classifying individual expressions, Figure 5.8 shows the confusion matrix for the *overall performance* of the system. In this figure, the true positives, true negatives, false positives, and false negatives are derived by considering the average detection results of the 10 subjects. The figure depicts that the highest accuracy has been achieved for the *"Sad"* expressions, closely followed by *"Happiness"*. Although the accuracy for the class *"Angry"* is comparable to these classes, that of the class *"Surprise"* is less than the rest. The matrix shows that this expression has mostly been confused with *"Angry"* and vice versa. By considering the AUs that generate expressions like *"Anger"* and *"Surprise"*, we observe that both these expressions have overlapping characteristics like widening of eyes and furrowed brows (similar characteristics of ocular AU) along with distinct characteristics like tensed mouth, jaws in anger and relaxed or dropped jaws in surprise (dissimilar characteristics of oral AU). However, depending on the individual, the expression of anger can also show similarities in the characteristics of the oral AU with that of surprise. This explains the performance of the system in detecting *"Surprise"* with lower accuracy as the analysis of the underlying characteristics of AUs aligns with the observation (refer to Section 5.3.1) that ocular AUs have the highest effect on the signal features. Similarly, the overlaps between other expressions have resulted from the similarity in different AUs (e.g., partial visibility of teeth for *"Happy"* and *"Angry"*).

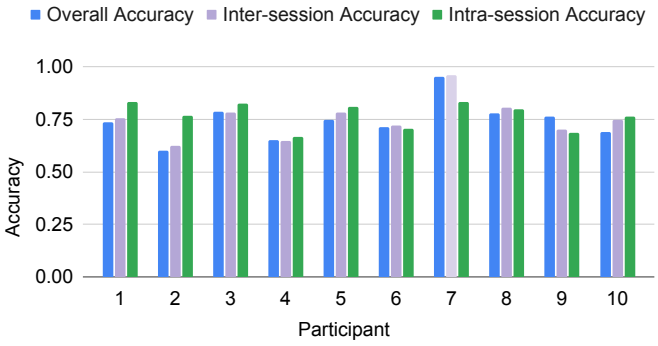


Figure 5.7: Comparison of participant-wise variation of overall, inter-session and intra-session accuracy

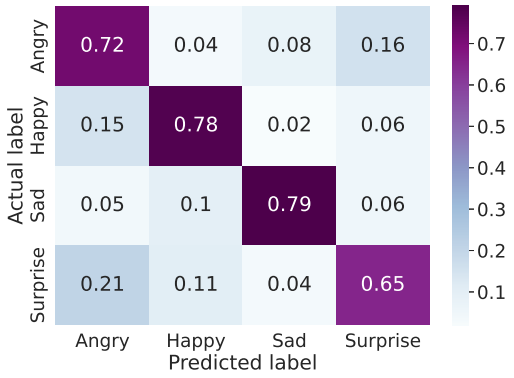


Figure 5.8: Overall classification accuracy of individual expressions

5.7.4 Sensitivity Analysis

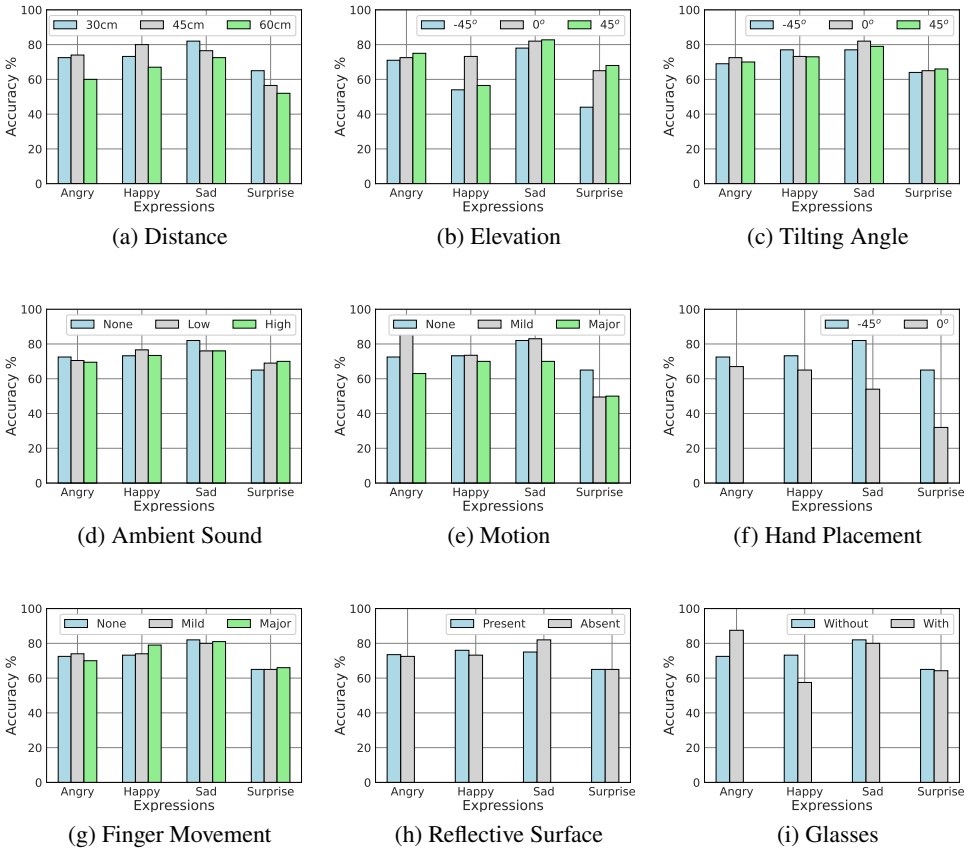
We next analyze the sensitivity of *ExpresSense* under different environmental conditions that may impact the performance of the system.

Impact of Distance

To analyze the effect of distance on the classification accuracy, we asked the subjects to place the smartphone at 30 cm, 45 cm, and 60 cm from their face, in three consecutive sessions, respectively. While recording the reflected signal from three different distances, the devices were placed at zero-degree elevation and tilting angles to the user. Figure 5.9a shows that the performance of *ExpresSense* drops (≈ 6-14%) as the distance is increased. Notably, we observe maximum drop in the performance (≈ 14%) for the *"Surprise"* class as it gets confused with *"Angry"*.

Impact of Phone's Elevation

In this study, the distance between the subject and device was kept fixed at 30cm, and the angle of elevation (vertical height) of the device was varied from −45° to +45° with respect to the subject's face. Interestingly, even though the performance of the system (Figure 5.9b) was comparable (74% and 72%, respectively) for 0° and +45° of device's elevation, at −45°, the accuracy of the system dropped to 62%. The contributing factor behind this observation was the location of the microphone. At −45°, the microphone of the smartphone, placed at the bottom of the phone, captured a noisy signal due to interference from the parts of the upper body.



(a) Distance (b) Elevation (c) Tilting Angle

(d) Ambient Sound (e) Motion (f) Hand Placement

(g) Finger Movement (h) Reflective Surface (i) Glasses

FIGURE 5.9: Sensitivity Analysis of *ExpresSense*: Impact of Various Environmental Factors

Impact of Phone's Tilting Angle

Similar to the previous study, in this analysis, the devices were kept at −45° (left), 0° (in front), and +45° (right) horizontally, with respect to the subject's face, at an elevation of 0°. However, in this case, the angle had no significant impact on the results (Figure 5.9c).

Impact of Environmental Noise

ExpresSense uses near-ultrasound signals that should not interfere with most of the audible frequency range. To test this hypothesis, we asked the subjects to produce facial expressions in three different environments, while the smartphone with *ExpresSense* was placed at 0° of vertical and horizontal angles from the user's face at a distance of 30 cm. In the first case, we aimed to eliminate all possible sources of sound. It is to be noted that complete silence (0 dBa) is not possible to attain, even in a lab-scaled study. Hence, by no sound, we indicate the absence of all audible ambient noises created by ceiling fans, keyboards, etc. In the second level, we induced noise between ∼ 15-30 dBa, which was generated by human whispers, ceiling fans,

and so on. This was marked as an environment with low ambient sound. In the third level, we incorporated a high sound level by playing background music, loud conversations, and traffic sound. Figure 5.9d shows that environmental sounds did not significantly affect the overall and class-wise accuracy of the system. This is because most of the environmental sounds fall well below the frequency of 16 kHz and are eliminated by the high pass filter of *ExpresSense*. However, the performance of the system will be adversely affected if ultrasound signals that overlap with the frequency range of the chirps in *ExpresSense* are introduced in the environment (refer to Section 5.3.2).

Impact of Motion

Next, we assess the effect of motion for three scenarios – (1) *No motion*, where the smartphone was placed on a phone holder, (2) *Mild motion*, when the phone was held by a hand, while the users remained seated and (3) *Major movement* created by allowing the users to walk in the room while holding the phone by hand. Figure 5.9e shows that only major movements decrease the system's accuracy by $\approx 9\%$ on average. Further, it had the most significant effect on the detection accuracy of the expression *"Angry"* (average decrease of 18.2%) and *"Sad"* (average decrease of 12.5%). The effect of major motion can be explained by the workflow of *ExpresSense* where bin-selection (calibration) is a one-time process. Any large change in the position and distance of the reflector (face), as indicated by the selected bin, caused by body motion or change of hands will cause the system's accuracy to be affected. However, this can be solved by re-calibrating the system as and when the smartphone's inertial sensors detect large body movements.

Effect of Hand Placement

Next, we estimate if the holding position of the phone affects the system's performance. For this study, we asked the subjects to (1) place their fingers on the side of the phone and (2) place their palm toward the bottom of the phone. Figure 5.9f shows that placing the fingers on the side allows the system to perform with an accuracy of about 74% while placing the palm towards the bottom decreases the accuracy to 55%. This is because, in the latter case, the palms cover the microphone, making it unable to capture the reflected signals completely.

Effect of Finger Motion

To analyze whether the system is affected by the movement of fingers, we asked the users to (1) restrict the movement of fingers, (2) periodically reply to text messages received in a floating window on the smartphone's screen (requires gentle finger movement), and (3) continuously chat or use the video controls (requires significant finger movement) while using the system in

the background. Figure 5.9g shows that there was no significant effect of finger movement on the overall and expression-wise accuracy of the *ExpresSense*. This observation was fascinating as it did not align with the impact of (body) motion on the system's performance. However, it should be noted that the reflector's position (face) was fixed for this experiment, and the calibration phase captured the presence of the fingers that were kept static during calibration (which takes about 4 seconds). However, if the fingers are moved during the calibration phase, the bin selection can be affected, thus decreasing the overall system performance. This overhead can be eliminated by detecting the degree of change in the signal's amplitude and phase, as a movement in facial AUs is significantly more fine-grained than finger movements.

Effect of Surrounding Objects

In this experiment, the subjects were asked to perform the same experiment while (1) a monitor (a reflective surface) was placed at about 45 cm behind the smartphone, and (2) no object was present within a distance of 60 cm from the smartphone. This experiment aimed to test whether surrounding reflectors like a monitor can affect the system's performance. Interestingly, these static objects did not affect the system's performance (Figure 5.9h). This observation can be explained through the fact that static interference cancellation is performed by *ExpresSense* in the signal processing stage.

Effect of Glasses

Finally, we present the result by considering the subjects (1) with glasses (power glasses or reading glasses) and then (2) without glasses. Figure 5.9i shows that the performance of *ExpresSense* is not affected by the presence of glasses.

5.8 Evaluating *ExpresSense* Under Natural Expressions

As discussed in Section 5.6, we developed a smartphone video streaming app that uses *ExpresSense* at its core to continuously sense the facial expressions of the subject while watching a video and measure how engaged the subject was during the session. The application has been open-sourced[13] so that participants can use it in an uncontrolled environment and record their feedback.

Figure 5.10 depicts the interface for *ExpresSense* video streaming app. The image to the left displays the **Content Display Area** where the user can view YouTube Videos. The user can also select the **length of the content** as either short (7-10 mins), medium (≈15 mins), or long

[13]Code link: https://github.com/anonymous0304/ExpresSense.git

(≈30 mins). Moreover, the user is allowed to select the **genre** of the content as either: *comedy*, *tragedy*, *horror*, *anger*, or *mixed*.

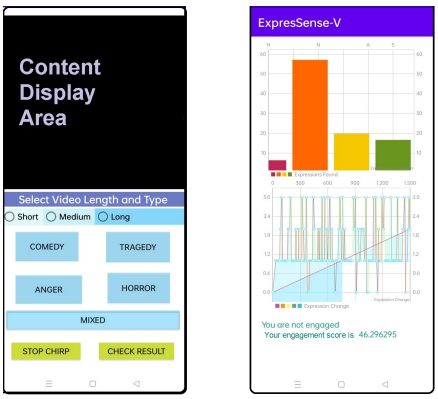


FIGURE 5.10: Interface of *ExpresSense* Video Streaming App: The Content View (left) and the Result View (right)

Once the user selects these fields, they can click on the **start chirp** button, which starts the video and simultaneously emits the chirp signals as described in *ExpresSense*. The received signals are processed using the pipeline mentioned in Section 5.5, and the final phase and amplitude values are transmitted to the trained ensemble model to predict the expression label. Thus, the application continuously monitors the instantaneous expressions and estimates the rate of expression change throughout the video. As the user completes watching the content, they can click on the **Check Result** button, which opens the interface shown to the right of Figure 5.10. In the result view, the user can see the percentage of each expression, as predicted by *ExpresSense* throughout the usage of the application. Along with this, the user can also see how their expressions varied over time. These two graphs provide the user with a visual representation of their engagement which is then summarized as detected engagement indicator and the overall engagement score. We use a rule-based approach, as discussed next, to generate the engagement indicator and engagement scores.

5.8.1 Engagement Estimation from Facial Expressions

The rules for this score generation are derived from previous works [254, 107, 95] that correlate expressions with video types. Assume that k defines the Genre ID ($k = 0$ if the genre is *comedy*, $k = 1$ if the genre is *tragedy*, $k = 2$ if the genre is *Anger*, $k = 3$ if the genre is *Horror*). Assume that an array $E[4]$ stores the total expression count for the 4 expression categories[14], as predicted during the content viewing. Let R denote the number of times the expressions have changed during the content length of l minutes. Let the function `indexOf` return the index of an element from the array $E[]$ (-1 if the element doesn't exist in the array), and $\wedge$ denote the logical AND operator. Then, the engagement indicator ($\mathcal{E}$) is estimated as follows.

1. $\mathcal{E}$ = True if k = `indexOf`$(\max(E))$: The strongest and most frequent expression matches with the content genre. For example, if the subject was mostly happy while watching a

[14]We consider *Sadness* as *Neutral*, as discussed earlier.

comedy video, it implies that the subject was engaged in the video.

2. $\mathcal{E}$ = True if $k \neq \texttt{indexOf}(\max(E)) \wedge \texttt{indexOf}(\max(E)) = 1 \wedge E[k] > \texttt{avg}(E[0], E[2], E[3])$: This formula is based on the observation that even if a person is engaged, they might stay *Neutral* for most of the time and occasionally show an expression that matches with the genre. It implies that even if the most frequent expression is neutral, the expression matching the content genre should be greater than the average of all expressions (other than neutral) found during the course. For the genre of tragedy, we check if the frequency of neutral (or Sadness) is greater than the average of all expressions predicted.

3. $\mathcal{E}$ = True if $k \neq \texttt{indexOf}(\max(E)) \wedge \texttt{indexOf}(\max(E)) = 1 \wedge R > 0.3 \times l)$: This formula is based on the observation that if a person is engaged, and neutral for most of the time, they will show at least some changes in the expression during the content viewing, to ensure that they are not blankly staring at the screen. However, this condition fails if some auxiliary task in the background causes a difference in the expression.

4. $\mathcal{E}$ = False, if none of the above conditions hold true.

5. For the content of mixed genre, where there is no predetermined expression to compare the predictions with, facial expression alone cannot be used to detect the person's engagement. Hence, for mixed-type contents, this score is NULL.

Computing the Engagement Score

Engagement score is taken as the percentage of the expression that matches with the genre of the content, concerning all expressions for the genre "Tragedy" or with respect to all non-neutral expressions for all other genres. The following formula generates the score.

$$A = \begin{cases} \frac{E[k]}{\sum_0^n E[n]} \times 100 & \text{such that } n = 0, 2, 3 \text{ if } k \neq 1 \\ \frac{E[k]}{\sum_0^n E[n]} \times 100 & \text{such that } n = 0, 1, 2, 3 \text{ if } k = 1 \end{cases} \tag{5.8}$$

For mixed genre, the engagement score is displayed as the percentage of each found expression.

5.8.2 Experimental Setup

We recruited 12 volunteers for this study. 10 of them (P1-P10) were the same participants who volunteered in the previous experiment (Subsection 5.7.1). As discussed earlier, *ExpresSense* works best as a user-dependent model. Therefore, to test the system's performance for new subjects, we further considered two additional participants (P11, P12) – one male being an IT Professional and one female teacher between 31 and 52 years of age, respectively.

Content

For this study, we considered 15 YouTube videos from the categories of *Comedy*, *Tragedy*, *Anger*, *Horror* and *Mixed* for testing *ExpresSense*. Each category contained three videos of different duration. The short videos had an average duration of 7 minutes, medium ones had an average duration of 15 minutes, and long videos had an average duration of 30 minutes. We intentionally avoided the usage of any video longer than 45 minutes to avoid the possibility of a major involuntary drop in the sustained attention level of the participants due to disinterest. To ensure that the videos were engaging, their ratings were considered for selecting them. Most of the videos had a YouTube rating of more than 25k. The comedy videos consisted of clips from Mr. Bean and other popular movies; tragedy videos were tagged under “Sad stories", *Horror* videos ranged from short stories to long animated horror stories. Mixed-type videos were selected so that they could not be categorized as any of the categories distinctly. For example, a tutorial on “Machine Learning", a documentary on the best historical places in the world, or a video explaining Drake’s Equation. All these videos could give rise to amazement, amusement, surprise, confusion, or no emotion at all. The horror stories had an element of surprise, like “Jump scares” that were not only meant to invoke fear but also surprise. While these four categories had distinctive characteristics, it was difficult to choose videos under the category of “Anger" due to the inadequacy of videos under such a tag and also due to the uncertainty of the emotion the videos invoke amongst the viewers. For example, a video displaying a major social issue might cause anger in one viewer during empathy in another. We carefully selected three videos on social injustice like animal cruelty, bullying, and elder abuse that are likely to cause anger in the viewer. Since YouTube videos have been found to cause expressions like anger [121], the selection of videos for this category was also based on the selection of representative emotional terms [33] like “angry,” “force,” etc. from the video’s title, and comments [210][206].

Methodology

The total experiment was conducted in two non-consecutive sessions – an initial training session of 15 minutes and a session for testing *ExpresSense*streaming app. In the first session, the entire experiment was explained to the participants. The participants could choose to watch a video using *ExpresSense* streaming app to understand its basic functionality. To eliminate forced attention, we assured the participants that there would be no penalty for lower scores of engagement or disengagement. After the training session, the participants were asked to use the *ExpresSense* streaming app for the second session. This session s_2 was divided into 5 sub-sessions – $s_2^1 ... \ s_2^5$, where s_2^1 was dedicated to viewing the 3 videos from the genre Comedy, s_2^2 was dedicated to tragedy and so on. Within each sub-session s_2^n, the participant had to watch 3 videos of different duration and were interviewed after each video. The participants were

free to take breaks (at least for 15 mins) within or between these sub-sessions to remove eye fatigue. Each sub-session s_2^n was planned to have a duration of 1.5 hours as the total viewing time for short, medium, and long videos were $\approx$ 60 minutes, and the total interview time was 30 minutes. However, considering the breaks within the sub-sessions, the maximum time was increased to 2-3 hours. Hence, for each participant to complete viewing all the videos in *ExpresSense* streaming app from each of the 5 categories, a total of 10-15 hours were estimated. Each participant completed these experiments over multiple consecutive days based on their interests. Also, they were free to choose which video they wanted to see at a point in time from the list of available videos. We delegated these choices to the participants to ensure that they could watch the videos freely and that watching consecutive videos does not put much cognitive load on them, which might affect their engagement level.

There was no restriction on the sitting position of the participants; however, the placement of the device was kept optimal based on the viewing preference of individual participants. The participants were requested to minimize significant body movements during the sessions, and the ambient noise was minimal to avoid disturbance. In addition, we added secondary tasks like showing funny, sad, and scary images (that did not match with the instantaneous video genre) on a screen behind the device for Participant 7 during the experimental sessions. This was to test if *ExpresSense* streaming app could also capture low engagement levels caused by secondary tasks. For each video genre, the predictions of *ExpresSense* were generated for a total of $\approx$ 930 data points while each participant watched the short ($\approx$ 126 data points), medium ($\approx$ 270 data points) and long ($\approx$ 540 data points) videos.

Interview Mechanism

After each experimental session, the interview consisted of two question-answer sessions.

(1) *Questions from the video content*: These questions were asked based on the video viewed before the interview and were selected in such a way that captured the overall engagement of the participant. For example, for a 30 minutes long video, the first question was based on the first 3 minutes of the video, the second question was from 3rd to 6th minute of the video, and so on. For short videos, medium, and long videos, there were two, five, and ten questions, respectively. The questions did not come with options to avoid the possibility of guessing. Based on whether the participant answered the question correctly, we marked it as 0 or 1. The final **ground truth engagement score** was estimated by taking the percentage of the correct answers for each video under each category. A total of 85 questions were asked to each participant for all the videos and genres combined.

(2) *Self-assessment*: Self-perception of engagement refers to whether the user considers themselves engaged or disengaged. Self-perception is essential in the case of content rating. For example, if they feel that the content is engaging, their ratings of the content will be high. Suppose the estimation generated by *ExpresSense* can relate to this personal or self-assessment of engagement. In that case, it can promote automated feedback by eliminating the requirement of manual rating or feedback, as manual feedback can be biased or influenced by different compelling factors[15]. Whether the user perceives themselves as engaged depends on two underlying factors – Whether the viewer pays attention to the content (F1) and whether the viewer found the content interesting (F2); in some cases, F2=F1. For example, if the user is bored with the content, their sustained attention will drop. Conversely, exciting content will promote engagement. In other cases, F1 and F2 can be unrelated. For example, the user may find the content boring but still choose to pay attention to it. Thus, it is essential to analyze each of the two factors independently to understand the overall engagement of the user.

Based on this understanding, after each video, the participants were asked if they paid attention to the video shown in the application or were focused on something else. They were also asked if they found the video interesting. These two questions were yes(1)/no(0) type. The **self-engagement indicator** for each video was calculated by taking the logical AND between these two answers.

5.8.3 Hypotheses

This study hypothesizes that *ExpresSense*-generated scores will highly correlate with the ground truth engagement score and the self-engagement indicator. From theoretical evidence, we also hypothesize that most facial expressions will be Neutral, as, in real-world scenarios, content-invoked expressions are sparse and aperiodic. Therefore, changes in the expressions would be natural depending on the video content, and the models should still be able to correlate with the manual (ground truth engagement score and self-engagement indicator) scores. If we observe a high accuracy in this prediction, *ExpresSense* could likely identify the natural expressions correctly during the session.

5.8.4 Results

In this subsection, we discuss the performance of *ExpresSense* through the analysis and comparison of engagement scores and engagement indicators.

[15] https://factorialhr.com/blog/bias-in-performance-reviews/#types-of-bias-in-performance-reviews (Accessed: Friday 11th August, 2023)

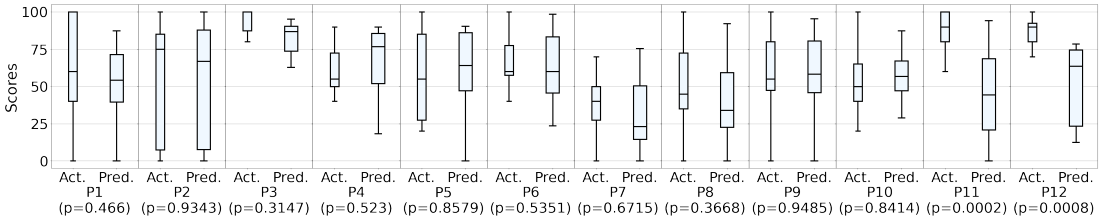


FIGURE 5.11: Distribution of participant-wise actual and predicted engagement scores in *ExpresSense*. The null hypothesis is that the two score distributions are similar. The graph shows that, for P1-P10, the null hypothesis is accepted (p>0.05). For P11-P12, the alternate hypothesis that the scores are different, is accepted.

Analysis of Engagement Score

Figure 5.11 depicts the distribution of the ground truth engagement scores generated from the interview with that of the scores generated by *ExpresSense* streaming app. The graph compares the distribution of these two scores for each participant by considering all the videos from the 4 video genres – comedy, tragedy, anger, and horror. We infer the following observations from these results.

1. The ground truth and the predicted scores are significantly correlated (with the reported p-values of a statistical T-test between the two distributions) for most of the participants, thus proving a part of our hypothesis.

2. For P7, it can be seen that both the ground truth and the predicted scores are low. This is because P7 paid attention to the secondary task, which caused a variety of facial expressions (based on the images), thus leading to a lower predicted score. Similarly, for P3, both these scores are high as the participant was fully engaged to the videos and could answer most of the questions correctly. This proves the capability of *ExpresSense* streaming app to distinguish engaged participants from the less-engaged ones, thus indicating the success of *ExpresSense* for correctly inferring the natural facial expressions.

3. For P11 and P12, the difference in the ground truth and predicted scores were caused by the misclassification of expressions by *ExpresSense*. This is because the system is user-dependent and needs to be calibrated and trained with a few data samples from the users before being able to perform with significant accuracy. However, even though the error level for unseen participants was high, we could observe a correlation between the score level. For higher ground truth scores, the overall distribution of the predicted scores tends to be higher.

This observation leads to the inference that affective facial expressions are correlated to the attention level of a person [99]. The correlation can be estimated by analyzing its similarity with the video genre.

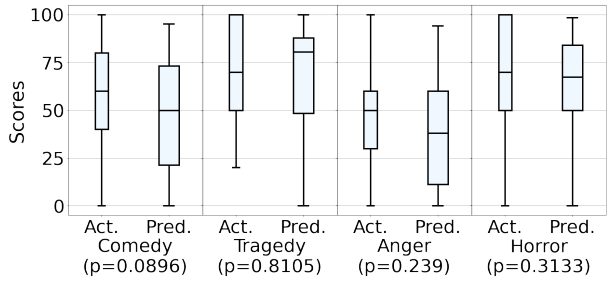


FIGURE 5.12: Distribution of genre-wise actual and predicted engagement scores; the null hypothesis is that the two score distributions are similar. The graph shows that, for all genres, the null hypothesis is accepted (p>0.05).

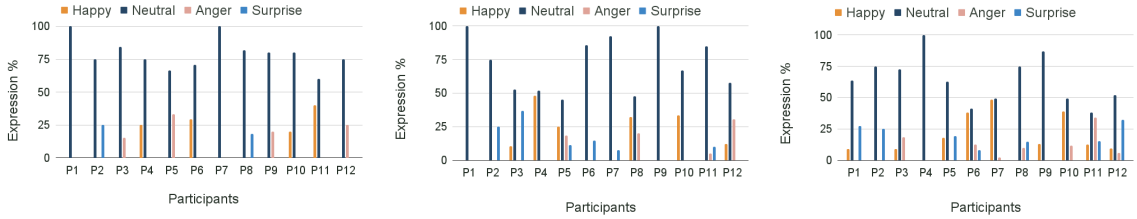
Next, Figure 5.12 shows the genre-wise distribution of the ground truth engagement score and the predicted scores. We observe a strong correlation between the two scores indicating that *ExpresSense* streaming app can perform well for various video types. However, the overall score under the anger category is lower than that of others. This is because anger videos failed to invoke the expression of anger in most of the participants. Moreover, the participants were found to be less attentive to such videos in this study. On the contrary, the expressions were mainly neutral for tragedy, but the attention levels were high. This aligns with the interchangeability of sadness and neutral expressions in *ExpresSense*. However, it was noted that some of the tragedy videos received "empathetic smiles" from the users. Such instances were mis-classified as "Happy" thus leading to some level of disagreement between the video genre and found expressions. By comparing these scores based on different thresholds, it was found that the average F1-score for optimal thresholds in each video category is .84.

Analysis of Engagement for Mixed Genre

As discussed earlier, for mixed videos, since there are no pre-determined regular expressions, engagement score based on one single expression is rather unfair. Hence, we compare the percentages of all the found expressions for these videos for each participant for the three different videos of lengths short, medium, and long. Figure 5.13 graphically shows these distributions for different video duration. The following inferences can be drawn from the figure.

1. As hypothesized, for almost all the participants and video duration, the most prevalent expression is found to be neutral.

2. A clear correlation between the video length and the number of expressions can be established. It can be seen that for medium or longer videos, the participants showed a greater number of expressions than for short videos. This indicates that for mixed-type videos of longer duration, engagement can be estimated by mapping the user's expression



(a) Short videos (b) Medium videos (c) Long videos

FIGURE 5.13: Distribution of facial expressions for different mixed type videos for individual participants

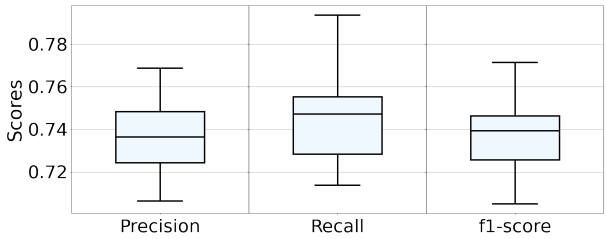


FIGURE 5.14: Comparison of Overall Precision, Recall and f1-score for self reported engagement indicator vs predicted engagement indicator for *ExpresSense*

with other video viewers.

Analysis of Engagement Indicator

We now explore *ExpresSense* streaming app's performance in terms of engagement indicator. In Figure 5.14, we compare the engagement labels as predicted by the *ExpresSense* streaming app with the self-assessment score. The graph shows a significant correlation between the predicted score and the self-assessment score. It aligns with the assumption that engagement is the underlying attention and interest of the user to the video content and that it can be quantified by considering the strength of expressions and their rate of change. Even though engagement indicator and engagement scores are estimated as separate variables, Figure 5.15 shows that if a user is marked as engaged by *ExpresSense* streaming app, they are likely to have higher engagement scores. In contrast, disengaged users will usually have significantly lower engagement scores.

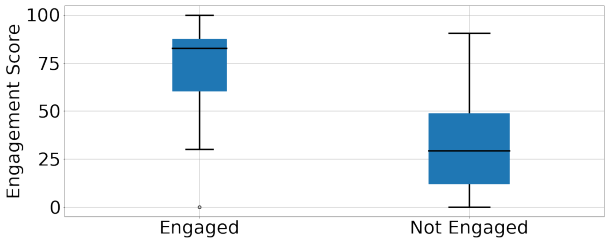


Figure 5.15: Correlation between engagement score and engagement indicator, as predicted by *ExpresSense*. The null hypothesis that the distribution of these scores are similar is rejected by the ttest as the p-value << 0.05.

5.9 Large-scale Usability Study with *ExpresSense* Streaming App

The objective of this study is to analyze how well the participants in-the-wild rate *ExpresSense* streaming app in terms of its usability in the practice. In this study, we have considered 72 subjects from different countries, professions and age groups (between 18-71). The participants were from different countries like United States (10), India (27), South Korea (4), Germany (3), United Kingdom (2), Australia (5), Bangladesh (3), Japan (3), Austria (1), Brazil (2), Canada (3), Croatia (1), Israel (1), Hong Kong (3), and China (4). 47 of the participants were male, and the rest were female. While Figure 5.16 shows the distribution of age-groups and the corresponding count of participants, Figure 5.17 shows the professional categories to which the participants belonged. The category "*Academics*" comprises of professions like Professors, Research Scientist, Research Scholars, Post Doctoral Fellows and Educators. We have grouped professions like Service, Software Engineers, Software Developers and IT under "*Industry*". Banking Services and Government employees are categorized under "*Others*". It can be noted that these participants are distinct from those who participated in the previous studies. Notably, we did not have the ground-truth information for these participants, so we only analyzed the usability ratings they submitted after experiencing the app.

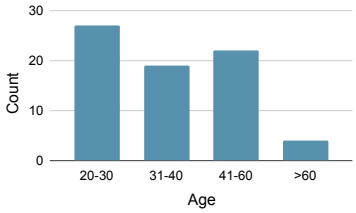


Figure 5.16: Distribution of participants based on age groups

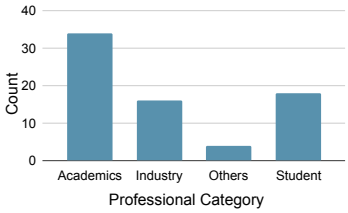


Figure 5.17: Distribution of participants based on profession

5.9.1 Methodology

In this study, we released the APK of *ExpresSense* streaming app to the participants, along with a video demonstrating the applications. The participants were requested to install the applications, use them thoroughly and provide their feedback through SUS questionnaire. Refer to Appendix A.2 for the statements and formula for SUS scores.

5.9.2 Result

From this experiment, we received an average SUS score of 85.34 which establishes the usability of a system like *ExpresSense*. To further test our hypothesis, we plot Figure 5.18 that shows the questionnaire's statement-wise average score (on the scale of 1–5), as provided by the participants. Figure 5.19 shows that the majority (61) of the responses indicated an SUS score

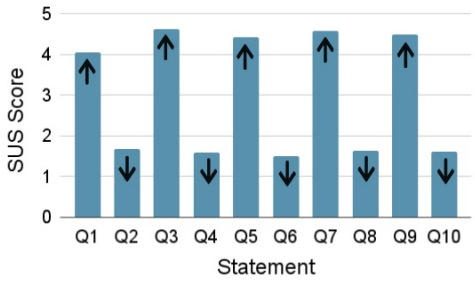


Figure 5.18: Distribution of SUS scores based on individual statements

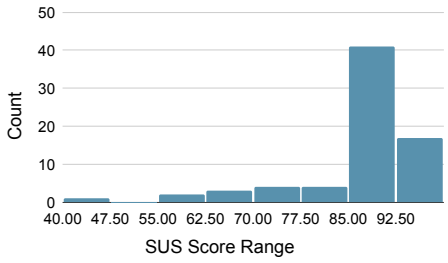


Figure 5.19: Histogram of SUS Scores

80 or more. We also found that age of the participants had no effect on the usability of the system. The age-group-wise aggregation of the scores varied between 82.1 to 90, showing the application was equally usable by people from all age groups. Similarly, for both male and female users, the average SUS score was similar (84.5 for males and 86.02 for females). Based on professions, the application was rated highest by academicians (87.5), followed by users from industry (86.09), students (81.1) and other services (80.6). The application was also found to be widely accepted by people from various cultural backgrounds. The average score based on the demographic information of the participants showed that for different countries, the SUS score ranged from 80.1 to 91.25.

5.10 Discussion

ExpresSense demonstrates how a single microphone and a single speaker present in a commercial smartphone can address the problems of camera-based facial expression detection. However, our model also suffers from a few limitations that make it suitable as a performance enhancer of camera-based techniques. However, in scenarios with a limited number of expression variations,

ExpresSense can even replace camera-based techniques. This section discusses some of the limitations and future scopes of *ExpresSense*.

5.10.1 Near-ultrasonic nature of the signals

The supported frequency range is up to 20 kHz in a commodity smartphone; the near ultrasound signals used in this model are slightly audible. This can cause mild discomfort in some users. As reported by some of the participants under feedback, the chirps, though mildly audible, created minor distractions. However, this problem can be mitigated by periodic system usage or by further narrowing the frequency range. In the future, we will aim to extend the system to the iOS platform as some iPhones can support frequencies above 20 kHz.

5.10.2 Effects of obstruction, movement and device orientation

Though the current experiments allow natural body movements, they lack large-scale movements. The performance of *ExpresSense* in the presence of significant activity is subject to further testing. However, such noises can easily be eliminated by sudden and abnormal changes in I-Q values, measured against a manual threshold. Moreover, since we use smartphones, it has been assumed that the face will be the nearest object to the device, and the reflected signals from the facial regions will be the strongest. However, the performance of *ExpresSense* might degrade if an additional obstruction (like a mask or spoon) is inserted between the device and the face. The current prototype of *ExpresSense* is developed to work in portrait (vertical orientation). Even though some users might prefer to hold the device in landscape mode, it would negatively impact the accuracy of *ExpresSense*. This is due to the fact that in this orientation, when a person hold the smartphone the microphone (as well as the speaker) will be mostly covered by the user's palm (based on the observation in section 5.7.4).

5.10.3 Validity of engagement scores

In this work, due to the lack of any standard metric, we have compared the predicted engagement score with that generated manually through interviews and questions based on the viewed content. However, such manual scores can sometimes be misleading. Since the interviews take place at the end of each video/reading session, for longer content, the users could forget the answer to a question based on the first part of the video or story. This might lead to a lower manual engagement score and a high predicted score based on continuous monitoring of expressions. However, we assume that the contents are of optimal duration, such that if a participant pays engagement, the information will be retained in their short-term memory. Moreover, misclassified expressions can be generated by *ExpresSense*, which might create lower scores. In the future, we aim to extend the data for training the model to achieve higher accuracy.

5.11 Summary

In this chapter, we present *ExpresSense*, a lightweight near-ultrasound signal-based facial expression detection system that works in real-time on a commodity smartphone. The system not only eliminates the requirement of any additional hardware or modification of the inbuilt components of any inexpensive smartphone with minimal processing capacity but also overcomes the challenges associated with camera-based techniques – such as occlusion or privacy impairment. Through rigorous lab-scaled and unconstrained testing, *ExpresSense* depicts a significant performance in classifying various facial expressions and proves the application of acoustic sensing in this domain. It also reveals the capabilities of commercial smartphones in facilitating such acoustic applications, thus proving its feasibility and scope of global acceptance.

6

Interactive System for Touch-free Control on MOOC videos by Learners

Massive Online Courses (MOOCs) have gained enormous popularity due to the evolution of traditional learning from centralized classrooms to global knowledge distribution. Moreover, the unforeseen pandemic situation due to the COVID-19 virus has led to the avoidance of mass gathering and limiting physical contact among individuals and devices. The inflation in the utility of online courses is also a direct result of this scenario.

However, in pre-recorded video lectures, the pace of delivery of the contents does not match with that of cognition for every student. Even in a real-time presentation, it is often impractical to discuss each of the personal queries. In the case of confusion, the learner often takes note of the important points so that the queries can be mediated later. The learner might often look into important terms discussed in a lecture and find relevant materials from other online resources. These may include keywords or figures shown in a video. While key terms can be noted quickly, noting down figures or descriptions while the online video is being played, requires repetitive pausing and playing involving physical contact with the device. This is further infeasible if the course is being followed by the learner in a mobile scenario. Since the process of taking notes for relevant article search is essential, the mentioned intricacies necessitate an interface that does not require physical contact like button clicks and can be controlled through facial cues for selecting the sections from the lecture and generate relevant links for the keywords.

However, such a model needs to address generic problems like the free head movement, context

switching, and restricted set of gestures to be used for controlling the interface. In this paper, a novel, automated note-generation system has been proposed which can be controlled through simple blinks. The model is an integration of simultaneous eye tracking from the user's preview and notes selection and image capture based on the blink counts, followed by a speech to text translation and keyword retrieval, generating Wikipedia links for the relevant keywords. A video instance can also be paused, captured, or closed by blinks or gazing down.

6.1 Contributions

The contributions of the paper include the depiction of a novel touch-free interface, controlled through simple blinks. This facilitates touch-free interaction of the user with the device. In doing so, we only utilise the device's camera and eliminate the requirement of any additional hardware. Further, the generation of notes and relevant web links are automated requiring no explicit search. The system is lightweight, required only a device camera and imposes no restriction on the learner's natural movement.

6.2 Proposed Model

In this section, a detailed discussion of the proposed model has been presented, along with the discussion of its individual submodules. Figure 6.1 depicts the overview of different architectural components of the system. The following sections present the details of the Blink based control module and the section processing and notes generation module, as shown in the figure.

6.2.1 Blink based control module

The execution of the system begins with this initial module which integrates a sequence of hierarchical condition checking mediated through continuous eye tracking and corresponding lecture video execution. The control actions that can be performed by the user to control the selection and viewing process are as follows:
Double blink: Each double blink marks the beginning and end of a note section in an alternate pattern. If the user forgets to end the last section selection, the end point is taken as the end point of the video. For each section, a sectional note will be generated after the program closes.
Triple blink: In some cases, the user might want to freeze a particular frame, e.g. a diagram, to follow further references to the diagram in the later parts of the video. In the proposed model, a learner can perform a triple blink when a diagram is shown in the video (or at any frame). This will capture the frame and display it in a window for reference while the video lecture plays in the window beside.

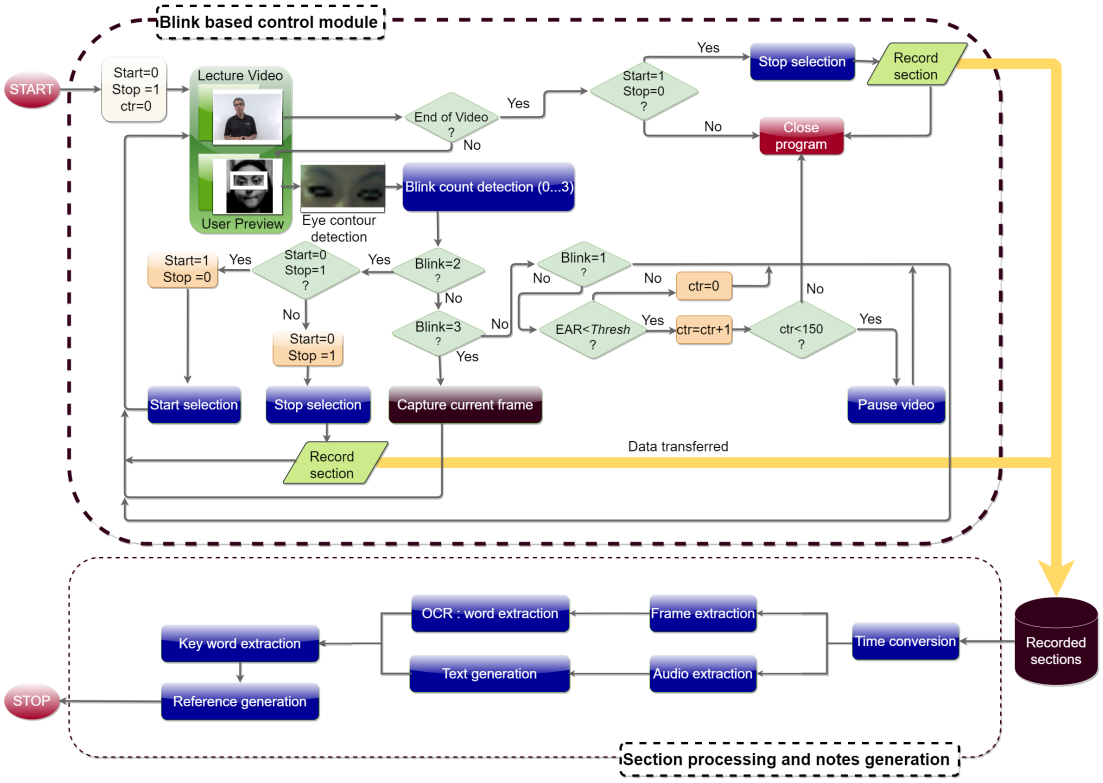


Figure 6.1: The architecture and workflow of the proposed model

Gaze down/ Moving out of the frame: If the learner temporarily looks down, closes his/her eyes for a few seconds or becomes temporarily unavailable, the video lecture will automatically pause until the learner looks back at the screen or moves back in front of the screen. This is to ensure that periodic context switching is supported by the system. However, each time the user looks down/ moves out of the frame, a program closing counter starts. At 5th second, the user is prompted with a voice command. If the user continues to look down/ stay out for 5 more seconds, the program closes. This ensures a contact-free exit mechanism at any instance of the lecture. Mathematically, if

$$EAR(Frame_n(pre), ..., Frame_{n+150}(pre)) < .15(Thresh),$$

the program is closed. However, if

$$EAR(Frame_n(pre), ..., Frame_{n+k}(pre)) < Thresh$$

such that $k < 150$, then display

$$Frame_n(lec) \forall Frame(pre) \in \{Frame_n(pre), \cdots, Frame_{n+k}(pre)\},$$

afterwards $Frame_{n+1}(lec)$ is displayed.
Here, pre and lec refers to the preview video displaying the user's face and the lecture video, respectively.

If the preview contains the face of the user, the eye region is detected using the eye landmarks and the Eye Aspect Ratio (EAR) is calculated [231]. The double and triple blink functionality of this module has been depicted in Algorithm 5. The threshold for counting a valid consecutive blink is set as 20 frames for restricting false positives or false negatives. Logically, a triple blink will always be preceded by a double blink. Since double blink invokes the start/stop function in an alternate pattern, a triple blink will always start or stop the selection process before capturing the frame. This is not particularly a problem if the triple blink invokes a section selection start. However, if the selection has already been started, a triple blink will prematurely end it. To avoid this error, a temporary stop function has been added to the system. When a double blink is encountered for an active section selection, the next 20 frames are checked for a valid blink while keeping the stop function temporary. If a blink is found within the window of the next consecutive 20 frames, the temporary closure is undone and the current frame is captured/frozen. However, finding no blinking within the consecutive window stops the session permanently. At the end of every committed stop function, the section is stored for further processing.

6.2.2 Section processing and notes generation module

This module is a sequential flow of operations performed on the total set of sections. For each section, the frame numbers are mapped to the corresponding time of the video. From each extracted frame in a section, the written words (if any) are extracted using optical character recognition. This is based on the assumption that in lectures, keywords are often written on boards by the lecturers while explaining the details verbally. For the audio sections, the audio-to-speech module generates the corresponding set of texts (notes).

The final submodule identifies the keywords from each text section and provides the related Wikipedia links to the learner. The proposed system leverages the advantage of the salient concept annotator, SWAT [191], successor to the widely popular annotator, tagme [59], to identify the conceptual entities from their mentions in the generated notes. Eg. in the sentences *"The current pandemic has increased the demand for MOOCs..."* the module correctly identifies *"pandemic", "the demand" and "MOOCs"* as the mention of the conceptual entities *"Pandemic", "Video on demand"* and *"Massive open online course"* respectively. The module then hyperlinks the mentions to the Wikipedia article corresponding to the conceptual entity. Since Wikipedia articles generally are written to be comprehensible with minimal prerequisite knowledge but have adequate references for further detailed perusal, we feel that such a convenient link in the generated notes will greatly magnify its usefulness to the user.

Algorithm 5: Blink based control

```
Bcount ← 0, CurrentFrame ← 0, FoundTwoAt ←0, start ←0, stop ←1, i ←0, sectionList[n,2] ←0;
foreach frame ∈ Lecture Video do
    CurrentFrame ← CurrentFrame+1;// Increase frame counter
    if CurrentFrame>FoundTwoAt+20 then
        Bcount=0;// Reset blink counter if no blink found for 20 frames
    end
    if (DetectBlink()) then
        if FoundTwoAt > 0 && CurrentFrame < FoundTwoAt+20 then
            // If one blink is found within 20 frames from the double blink,
               consider it a triple blink
            FoundTwoAt←FoundTwoAt-20;// To ensure that a fourth blink is not
               treated as a triple blink
            CaptureFrame(CurrentFrame);
            if start==0 && stop==1 then
                start←1;
                stop←0;
                sectionList[i,1]←0;// Undo section selection stop
                i←i-1;
            end
        end
        if Bcount==0 then
            FrameLast ← CurrentFrame;
            Bcount←Bcount+1;
            else
                if CurrentFrame<FrameLast+20 then
                    Bcount ← Bcount+1;
                    // Increase blink counter iff a blink is found within an
                       interval of 20 frames from the last blink
                end
            end
        end
        if Bcount%2==0 then
            // If blink counter is 2
            Bcount←0;
            FoundTwoAt ← CurrentFrame;
            if start==0 && stop==1 then
                sectionList[i,0] ← CurrentFrame;// Start selection is previous selection
                   has stopped
                start←1;
                stop ← 0;
            end
            else
                sectionList[i,1] ← CurrentFrame;
                // Stop selection if current selection has started
                // This will be undone if a third blink is found
                i ← i+1;
                start←0;
                stop←1;
            end
        end
    end
end
```

6.3 Experimental Results

The system has been tested on a device with Intel Core i5-4440 CPU @ 3.1GHz x 4 processor. The python libraries like dlib, SpeechRecognition and pytesseract are used for the development of the system.

6.3.1 Module estimation and real world study

The first study aimed at estimating the usability of the system in the real world. The SUS was used for this purpose on a total of 21 participants, 11 females and 10 males, belonging to the age group of 25-70 years after they watched the demonstration of the system. The survey revealed an average score of 84.17 for the system, thus proving its feasibility. Figure 6.2 shows the distribution of the scores.

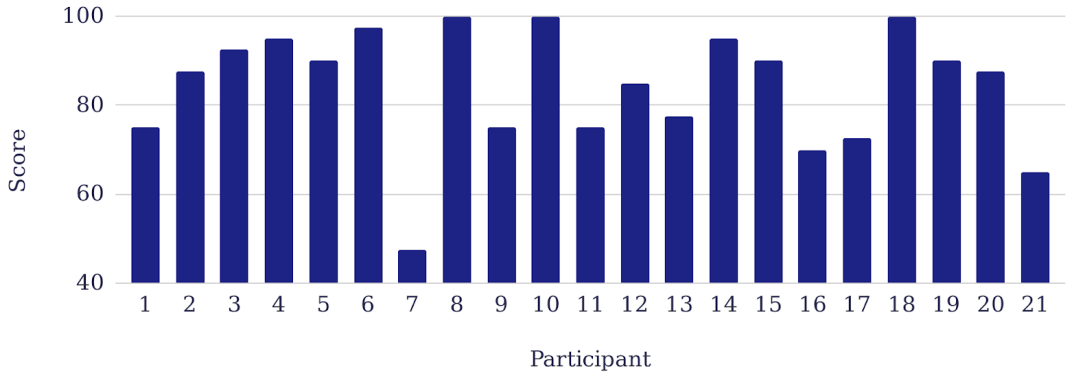


Figure 6.2: Distribution of SUS scores by participants.

The performance of individual modules were tested by performing a set of evaluations on real world users, similar to [29]. The second and third studies were conducted on 5 participants, 2 females and 3 males belonging to the age group of 25-60 years. In the second study, each participant was asked to use the system under two different light conditions: low and normal. A set of predefined blink patterns (a double blink followed by a triple blink followed by a voluntary (V) single blink) were given to the users which had to be performed by them at short intervals to time with no additional constraint. The given pattern had to be repeated for 10 times under each lighting condition. Hence, each participant had to manadatority blink for 120 times along with any additional involuntary blink (InV).

Figure 6.3a and 6.3b shows the comparison between the average misses (false negatives) and number of extra blinks detected (false positive) under the two lighting conditions. The results show that the impact of light is negligible on the performance.

The missed count can be interpreted as the V and InV that go undetected. For double and triple blinks missed indicate the number of double or triple blinks that were not/ mis-registered.

False positive accounts for the total false blinks that are generated by the system (total blinks). For double, triple blinks, it is the additional number of false triggers that results in additional section creation or image capture. It is noted that most of the false positives are single blinks, thus resulting in no additional section or capture. However, in case of some participants, sets of falsely generated blinks had falsely triggered triple blinks and hence additional image capture. Analysing the total intended blinks and correctly predicted blinks in each category, the accurate blink prediction rates are shown in figure 6.3c. Total blink accounts for the correctly detected Single (V) + Double * 2+ Triple *3 + InV.

(a) Missed Blinks

(b) False Positives

(c) Accuracy

Figure 6.3: Distribution of missed blinks (a), false positives (b) under low and normal light and the total accuracy (c) for each blink type.

Even though involuntary blinks are registered by the system, they had no practical impact on the function calls. This is due to the fact that in the implementation of the approach, some criteria were set. Firstly, if the second double/ triple blink occurs immediately after the first, the session start is reversed/ capture is not registered. Secondly, exactly two/ three blinks will invoke a function call. If a sequence of four (say) blinks are detected due to false positives or user's involuntary blink, the fourth blink will not have an impact.

In the third study, the keyword detection and overall system performance has been estimated. In this study, 10 sections are selected from 5 different lecture videos from the Computer Science (CS) domain. Each section is approximately a few minutes long. The transcript

corresponding these sections were obtained and then manually annotated by two post graduates from the CS domain, independently. The annotations typically consisted of marking the keywords from the given texts. These were then validated by two doctoral researchers from the same domain. Five users were asked to select these predefined sections from the video lecture using double blinks. The generated textual note and the keywords were validated with the previously created ground truth. The results showed that 91.56% of the transcripts were generated correctly and 70.41% of the keywords were identified and linked to Wikipedia in proper contextual form.

6.4 Summary

The chapter presents a novel approach that facilitates touch-free interaction with a video lecture to automatically generate relevant study materials and reference links based on the user-selected sections. The use of simple blink gestures facilitates easy handling of the system features by the learners. Even though experimental results demonstrate the usefulness of the proposed system, future directions will aim at improving the note generation accuracy. Moreover, board occlusion removal is a feature that can promote the quality of the lecture video. Even though freezing frames for reference is a solution for occluded boards in frames, by extracting foreground pixels and extrapolating background pixels, occluded boards can be recovered. The usability scale proves the feasibility and necessity of such a system in the practical domain and hence establishes its promising scope.

7

Interactive System for Touch-free Writing in Smartphones

The exchange of textual messages is one of the most routine functions of mobile devices. Touch-based typing systems are prevalent but are often inadequate for users experiencing medical conditions. Users suffering from conditions like dactylitis, sarcopenia, tennis elbow, or other forms of joint pains are often unable to use finger movements and touch-based text entry methods. Moreover, eyesight problems hinder the usage of the tiny keyboard layouts in a smartphone. This necessitates the development of hands-free text entry systems for these user groups. Some of the notable research directions have been driven towards gaze-based text entry systems [150, 151, 218, 117] and head orientation-based key selection [78, 269, 267]. However, most of these systems are either suitable for desktop or laptop environments or use commercial trackers to increase the accuracy. Touch-free text entry has also been aimed by the voice processing researches for the development of voice to text entry and editing systems [61, 69, 216]. However, they come with challenges like noise incorporation and privacy concern in outdoor environments.

The challenges of adopting touch-free typing in a smartphone are manifold. First, desktop computers have ample space between proximal keys in the soft-key layout. The same soft-key design, if placed on a mobile screen, will have significantly reduced inter-key space. Hence, even a small shift in the facial feature can affect the result. Second, even though powerful algorithms can identify the minor change in patterns, the computational cost restricts the application those

algorithms for mobile applications. Lightweight models are feasible in such cases that perform simple processing with high accuracy. Third, additional resources like webcams, trackers in desktops, or smartphones can be incorporated, but outdoor mobile applications need to be standalone using only the device's included capabilities. Fourth, desktop monitors are fixed. Thus a certain degree of head movement can be adjusted. Since portable device screens are prone to free movement and accelerometer noises, head and face tracking are significantly challenging. Lastly, apart from the computational and hardware restrictions, human beings tend to check the typed text for correctness while typing. In gaze-based systems, occasional glancing at the text region can shift the gaze from the key location, resulting in erroneous key selection.

In this paper, a lightweight contact-free smartphone-based text entry model, called *Nosype*, has been proposed for the clinically challenged users, which uses nose tip tracking and projection as the medium of text entry. *Nosype* uses a *"draw-n-locate"* interface for drawing alphanumeric characters and selecting whole words or punctuation marks through the nose-tip projection. While the user can move their head in a pattern to draw out a character in the air, the phone's front camera can be used to capture the pattern by tracking the nose-tip movement for some consecutive frames. Once the camera follows the character's outline, the model probabilistically predicts the character and types it. The system generates a suggestion list based on the characters typed. Words can be selected from the list by pointing the nose-tip towards the word. Similarly, punctuation marks can be entered using the system. Additionally, the system allows the users to edit the texts by changing the device's orientation, tracked through the in-built device sensors. The system uses a soft-key to switch between interfaces, which can be activated through facial overlap. The design of such an application also faces some significant challenges like prediction speed, user control, and accuracy. In *Nosype*, these challenges are addressed by optimizing the model, choosing the appropriate probability threshold for predictions, and incorporating an easy refresh feature into the application.

7.1 Contributions

The primary contributions of this paper are summarized as follows.

1. A novel touch-free lightweight typing technique is presented, which shows high accuracy and acceptable typing speed under the natural head and device movements. Such a system can be handy for the user groups who face difficulty in free finger movements.

2. Computationally expensive gaze tracking and mapping are eliminated, and the user's privacy is maintained as no data has to be stored in the back-end.

3. In-built sensor tracking is induced to allow simple device movements to be used as text editing options.

FIGURE 7.1: The interfaces in *Nosype*: *Draw* interface (left), *Locate* interface (middle, right)

The evaluation of *Nosype* has been facilitated by the development and usage of an Android application that has been profiled and tested on different smartphones. Apart from the system testing, an in-depth user study has also been performed through experimental methods over multiple sessions. In the first study, we employ 11 volunteers with clinical disabilities, preventing them from typing using a touch-based smartphone keyboard. In a semi-controlled lab-scale environment, the subjects have been asked to type a few texts using *Nosype* as well as four other gaze and finger projection-based approaches. We observe that *Nosype* results in a 70.11% improvement compared to the average of different baselines in terms of projection accuracy. Further, the study reveals that the participants can type 6.31 words per minute with the autocomplete suggestion in *Nosype* and reports an average key-stroke per character of 1.06. In the second study, we make the application public, where any user can download the app and starts using it. Once the user types at least 30 words using the app, the system probes the user to fill up a survey form concerning the system's usability. We have collected 60 such surveys in the wild (including 10 participants having clinical disabilities, which does not overlap with the 11 volunteers from the lab-scale study) and observe that the system indicates a good score of 77.708 on the SUS.

7.2 Application Overview

The primary idea of *Nosype* is to allow the users to move their head freely in the air while holding the smartphone in front of their face. Alternatively, they can also move the smartphone, keeping the head still. The device's front camera continuously tracks the nose-tip movement as the user moves the head or the device. By controlling the movement's direction, the user can draw out letters and numbers in the air. Simultaneously, the application displays the drawn character on the screen and predicts them as valid alphanumeric characters. Each predicted character is added to the text to form the complete message. After each prediction of a drawn character, the system recommends a list of dictionary words to the user based on the letters of the current word predicted so far. The user can select these words by merely pointing their nose-tip towards the chosen word. The nose-tip is projected on the screen, and the word from that screen's location is added to the current message. This eliminates the requirement of drawing each letter of a word, which is difficult if it is too long. Similarly, the user can select

punctuation marks to be added to the text using nose-tip projection.

Figure 7.1 shows the different interfaces of *Nosype*. The necessity of two different interfaces arises from multiple factors. **Firstly**, alphabets in the upper or the lower cases and the numbers, if placed within a fixed soft-key layout, will have proximal placement as the smartphone's screen is small. This makes their selection using precise nose-tip location difficult for the user as minor shakes result in wrong selection. Hence, the *Draw* interface is developed. Since the system primarily aims at facilitating users with clinical conditions, certain issues like Vertigo can lead to dizziness, neck pain, etc. after prolonged head rotation while using this interface. However, this can easily be avoided by drawing the characters by moving the smartphone with respect to a static head position. In this case, even though there is no head movement, the relative location of the nose-tip will move with respect to the smartphone camera due to the device's movement, thus preventing nausea or dizziness for users with particular conditions. **Secondly**, longer words are difficult to draw. If the application can predict an entire word based on one or a few characters, the user can select the word using the projected nose-tip. This greatly reduces the effort, drawing complexity and increases the writing speed. **Thirdly**, punctuation marks require discontinuous strokes to be drawn. To detect these pauses while drawing would require advanced computation, thus making the system unsuitable for smartphones. Since there is a limited set of punctuation marks required for text messages, they can be placed in a grid layout on the screen, and the users can select one of them using a focused projection. Due to these factors, the *Locate* interface is developed.

7.2.1 The *Draw* Interface

Nosype uses a **dynamic canvas** for this interface. This concept allows the user to draw a letter anywhere within the screen and in any font size. Since the font size is proportional to the degree of the free head movement, the application automatically detects the specific area of the canvas (screen) where the letter is drawn, based on the nasal position detected from the user's preview through the device's front camera, and extracts it (Figure 7.2). It is then scaled to a uniform size for prediction.

The interface also uses an **adaptive dwell time** to mark the end of a drawn letter. Since users with different medical conditions and age will have different drawing speeds, it is crucial to identify the drawing time. Keeping the drawing time for each letter static makes the system unusable for many people. Hence, *Nosype* allows the users to draw letters at their paces. The drawing is marked completed and sent to the prediction pipeline if there is no movement (shift less than 5 pixels) of nose-tip for the last 5 frames.

The drawn letter is predicted with a **prediction score**. If the score is more than a threshold, the drawn letter is accepted as a valid prediction. The score-based filtering reduces the chance of misprediction and allows some level of distortion in the drawing, which is inevitable while

moving the nose in the air. The **auto-complete suggestion** (auto-suggestion) list is refreshed after each prediction and displayed to the user. In our implementation, there can be up to 19 suggested words in the display and an option to close the suggestion if the user does not find the required word.

Figure 7.2: User drawing 'T' using the *Draw* interface

7.2.2 The *Locate* Interface

The *Locate* interface has two different utilities – (1) to select words from the auto-suggestion list, and (2) to choose punctuation marks (Figure 7.1(right)). For any of these selections, the user has to point the nose-tip towards the display location where the word/mark is shown. The nose-tip coordinates are projected on the screen. The user might move their face to adjust the focus and hold the position for a fraction of a second to locate the corresponding word or mark to be added to the text. **Soft button activation** is required to switch to/from the punctuation grid from/to the *Draw* interface. This punctuation grid button is shown in Figure 7.1 (left). To activate this contact-free interactive button, the user should move their face towards the button. When the detected facial region overlaps slightly with the button, it is activated.

Both the interfaces use a common **orientation-based editing**. Even though space after each word is automatically added if the word is selected from the suggestion list, the user can add space by tilting the device 90^o to the right. Similarly, to backspace out a character, the user can tilt the device 90^o to the left. To refresh a half-drawn letter, the user can tilt the device upwards or move their face away from the screen for 1 second.

7.2.3 Features of *Nosype*

Following points briefly summarize some of the other features of *Nosype*.

1. *Nosype* is not affected by the presence of glasses or facial hair as nose-tip is used as the medium of text entry.

2. Free head movement is allowed in *Nosype*. Simultaneously, the head movement can be restricted by the user based on the font size they choose or if they choose to move the

device with respect to their faces. This makes *Nosype* suitable for users with conditions like vertigo.

3. Flickering trajectory and projection point is controlled by further extrapolation for undetected frames and restricting the deviated projections for minor movements.

4. Uppercase letters are easier to draw. Hence, *Nosype* has auto-conversion feature for upper-lower case.

5. *Nosype* is a completely contact-less text-entry method, which is lightweight enough to be used in a standalone smartphone with no additional resources.

7.3 Methodology: Nose Tracking-based Text Generation

This section discusses the architecture and the functional sub-modules of *Nosype*. The core idea of *Nosype* is "prediction and projection" of characters through head movements captured by nose-tip trajectory tracking. Prediction is used in the *Draw* interface to predict an alphanumeric character from the nose-tip trajectory. Projection is used in the *Locate* interface to project the nose-tip on an auto-suggested word or punctuation mark to select the corresponding word or punctuation. *Nosype* uses a pre-trained CNN model to predict the character from the nose-tip trajectory. The main reasons for using prediction and projection for the *Draw* and the *Locate* interfaces, respectively, are as follows.

1. Deep neural networks like CNN, trained with popular datasets like EMNIST [38], of feasible size, showed significant accuracy in detecting handwritten characters or digits. Therefore a pre-trained CNN model can be used to predict the character written through nose-tip trajectory. This is advantageous over using large-scale datasets to map eye-gazes to a particular key on the soft keyboard layout.

2. The reason for using projection for punctuation, instead of drawing, is that they require discontinuous curves or strokes to be drawn. For example, ';' has to be drawn by adding a gap between '.' and ','. Unlike the uppercase alphanumeric characters, which can be drawn continuously, discontinuous tracking is complicated and needs additional processing. Hence, projection is more feasible and easier to control for the punctuation marks.

The core system architecture of *Nosype* is shown in Figure 7.3. The following sub-sections discuss the Prediction and Projection sub-modules in detail.

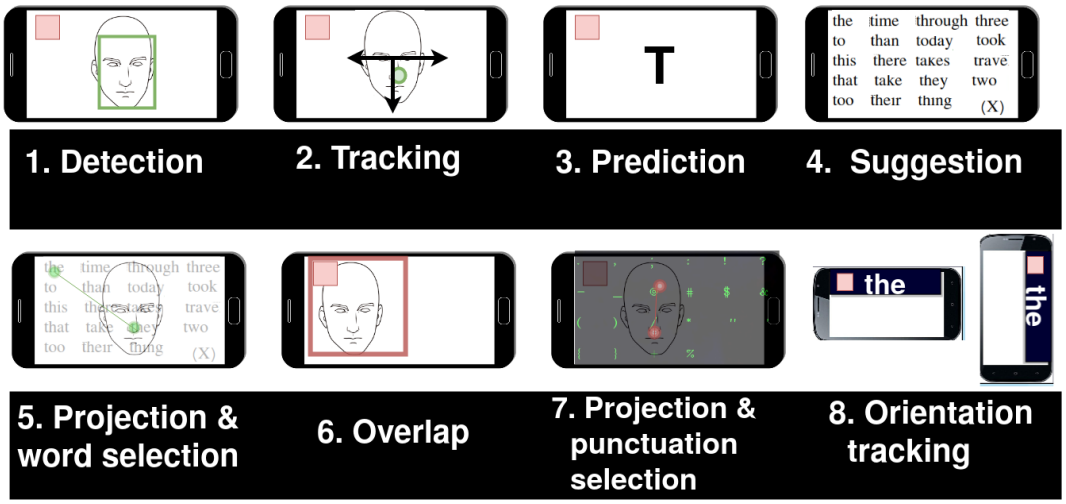


FIGURE 7.3: Overview of the Proposed Model

7.3.1 Nose-tip Trajectory Tracking and Character Prediction

In the *Draw* interface, the user draws the uppercase alphanumeric character using head movement, and the system predicts the corresponding character using a deep neural network-based approach. The detailed steps follow.

Detection of Facial Gestures

Nosype uses the device's front camera to capture the preview of the user's face. The initial stage is to detect the facial ROI and prune the ROI to the nasal area of the face. To detect the nose-tip precisely, we use the approach proposed in [198] to detect 68 landmarks (landmarks 0–67) from the user's facial preview. In this approach, initially, the feature mapping is performed by a series of local feature mapping functions by a Random Forest [22]-based regression for individual landmarks, using 300 facial images in-the-wild [203]. These locally obtained binary features are then integrated to learn the global projection and identify the landmarks. After the lower nose tip is detected from the landmarks, it is tracked in each consecutive frame, and the relative nasal shift between the two consecutive frames is recorded for each new frame.

Tracking Nasal Movements

The next task is to track the nasal movements to extract the character drawn on the screen. *Nosype* uses an adaptive **dwell time** defined as the maximum time a user is provided to draw out each character on the screen. In *Nosype*, the default dwell time is set to be 100 frames (denoted by the variable $MaxDraw$) which is approximately 3secs. However, this dwell time

is too high for most participants as most of the characters can be drawn within a second or less. Therefore, a lower dwell time value is more practical, although some users might take a long duration to draw the character. To balance the constant high dwell time and reasonable low dwell time, an **adaptive dwell time** is employed in the model. For each new frame, the nose-tip location indicates its relative movement to the previous frame. When the user completes drawing a letter, they are supposed to stop moving the head. This marks a zero or low shift of nose-tip in the consecutive frames. If the average shift of the nose-tip in 5 consecutive frames is less than 5 pixels, the adaptive dwell time is stopped, and the nose-tip trajectory is considered as the drawn character. A 5 pixel relaxation is provided instead of checking for strict zero-shift because a slight head movement can be present even if the user has completed drawing the character. To summarize the estimation of dwell time of each character drawing, $MaxDraw$ is set to either 100 or n frames where the absolute shifts between $\{\{n-5, n-4\}, \{n-4, n-3\}, \{n-3, n-2\}, \{n-2, n-1\}, \{n-1, n\}\}$ frames are less than 5 pixels, whichever is less. On the expiry of the $MaxDraw$ limit, the drawn character is sent to the prediction pipeline, and the $MaxDraw$ counter ($DrawFrame$) is refreshed and set to 1 for the next character drawing.

Character Prediction

Once we get the nose-tip trajectory indicating the drawn character over the smartphone screen, the next task is to predict the actual English character from the drawn one and then suggest the possible dictionary words to the user based on the text's context typed so far. For this purpose, we first extract the **canvas area** from the smartphone screen. While the user moves his head to draw a character, the minimum (top-left) and the maximum (bottom-right) coordinates from the set of traced coordinates are identified. These points on the screen denote the current canvas area on a screen region containing the entire drawn character. This dynamic canvas area depends entirely on the size of the illustrated character and the degree of nose movement. Making the canvas dynamic allows free facial movement for the users, thus facilitating convenient drawing. It is to be noted that the entire screen can be used as a canvas by the user, but only the part of the screen which contains the drawn character is extracted as the canvas area. This area is presented with a white background and black font color for the drawn character. To cope with the effect of varying canvas size, each canvas, after extraction, is resized to 28×28 pixel and is inverted in color for prediction.

Once we get the canvas, we use a pre-trained CNN model for alphanumeric character recognition from handwritten texts. Here, our basic intuition is that the drawn character on the canvas should look-alike a handwritten character with some noise. The CNN used for this prediction is shown in Figure 7.4. This CNN model is trained using the EMNIST [38] dataset containing 731, 668 training and 82, 587 testing data for English alphanumeric characters

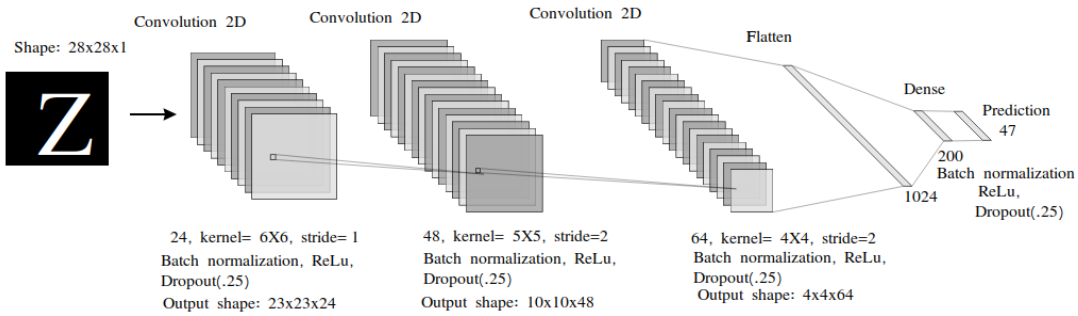


FIGURE 7.4: CNN Architecture for Character Prediction: The drawn character image is resized to 28×28 pixels with a single channel. This 3 dimensional image is fed into the first Convolution 2D (conv2d) layer. Since this layer accepts a 3D image with the kernel striding by 1 in two dimensions, the Convolution 2D layer is required. In this layer, the number of filters is 24, and kernel size is 6×6 pixels. This layer produces an output of shape $23 \times 23 \times 24$. Each of the 3 conv2d layers is followed by batch normalization, an activation function (ReLu), and a dropout layer with a rate of 0.25 to avoid overfitting. The second conv2d layer has 48 filters with a kernel size of 5×5 striding by 2 units, producing an output of shape $10 \times 10 \times 28$. The third conv2d layer uses 64 kernels of size 4×4 and produces an output of shape $4 \times 4 \times 64$. This is flattened in the next layer producing a vector of size 1024. A fully connected layer of size 200 is used, followed by batch normalization, activation, and dropout. The final layer is used to predict the input image into one of the 47 classes corresponding to the English alphanumeric characters.

belonging to 47 different classes. Each sample is a 28×28 image with a black background and white font color. For our model, the EMNIST-balanced class having $112,800$ training and $18,800$ test samples is used to train the CNN with the layered architecture, as shown in Figure 7.4. This pre-trained model is imported to the *Nosype* mobile application for the prediction process. We have tested the CNN model for character prediction with all the English alphanumeric characters drawn by the 11 volunteers (details in Section 7.4.2). We observe an average accuracy of 87.23% in predicting the characters from the drawn image over the smartphone screen.

Word Suggestion

To increase the typing speed, *Nosype* incorporates an *auto-suggestion* feature where a list of possible words is suggested to the users based on typing of the initial few characters, and the user can select a word from the list, if available (Step 4 in Figure 7.3). To build up this suggestion, as a proof of concept, we use a corpus of 1200 frequently used English words, as reported by Wikipedia and other articles[1]. It can be noted that the entire dictionary can also be incorporated; however, it will increase the search time. In this approach, we use a simple linear search of words based on the current set of drawn letters in a word. We use an existing

[1] https://www.ef.com/in/english-resources/english-vocabulary/top-1000-words (Accessed: Friday 11th August, 2023)

approach for word suggestion based on dictionary matching [78], and *Nosype* displays the 19 most relevant words in the dictionary order, along with a close option, to close the suggestion list. A set of 19 words are displayed to minimise the number of letters drawn per word.

In *Nosype*, the word selection from the above list and selecting the punctuation marks are based on the projection of the nose-tip over a choice, which is incorporated in the *Locate* interface. The details of the projection method used in *Nosype* is discussed next.

7.3.2 Projection based Selection by Nose-tip Mapping

In the projection interface, the user can either select a word from the suggestion list, overlap the facial area with the touch-free buttons to activate them, or select punctuation marks from the punctuation grid.

Projection of Nose-tip and Word Selection

The first use of projection is to select the entire word from the suggestion list. For this purpose, *Nosype* uses six landmarks over the face – nose, chin, and the four eye corners. The nose-tip is projected on the screen based on the facial yaw, pitch, and roll, computed from the above six landmarks. The projection of the nose-tip on the image plane (screen) is shown in Figure 7.5. We use the detected nose-tip and the minor facial yaw to map the projected screen-coordinates with the suggestion list index.

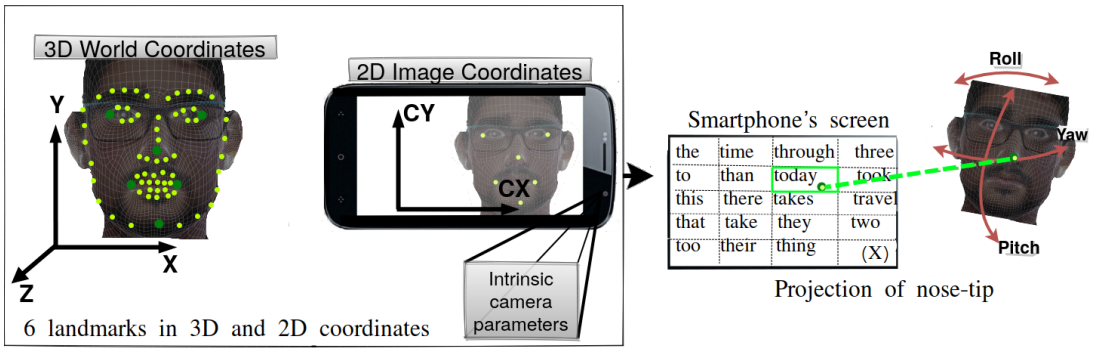


FIGURE 7.5: Facial Projection: The intrinsic camera parameters (approximate optical center from image center, approximate focal length from image width), word coordinates of 6 facial landmarks are estimated. Using these parameters, the facial translation, yaw, pitch and roll are estimated. Using these vectors, the respective projection point of all the landmarks on the screen are estimated.

Overlap of Facial Region with Interactive Button

Nosype uses one interactive, touch-free button for opening the punctuation grid. The activation of these buttons works based on an overlap of the user's displayed facial area on the screen. By

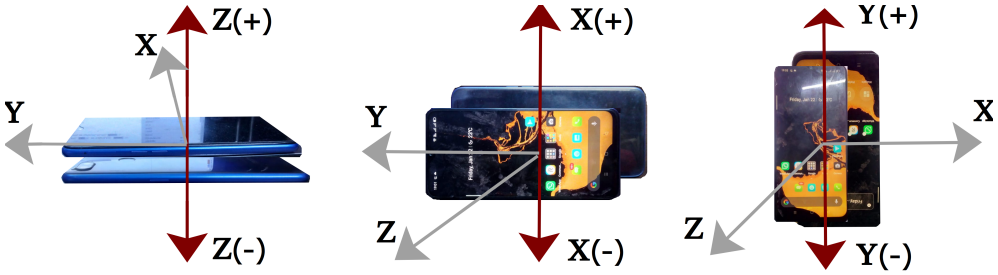


Figure 7.6: Position (orientation) of smartphone and most effective changes in the corresponding axis in accelerometer reading: placing the phone on a surface, with its front side up and down results in highest and lowest Z values, respectively. Placement of the phone facing the front and back towards the user in horizontal position results in highest and lowest X values, respectively. Keeping the top of the phone upwards and downwards in portrait mode leads to the highest and the lowest Y-axis value.

shifting the face towards the top left corner, the user can overlap the facial area with the button's area to open or close the punctuation grid.

Projection of Nose-tip and Punctuation Mark Selection

The same function, described previously in Section 7.3.2, is used when the user opens the punctuation grid. Figure 7.8(right) depicts the 4×6 grid layout of the punctuation mark placements. On selecting the virtual punctuation button, this grid is displayed on the screen. The projection function using the camera parameters, facial landmarks, and translation vectors is executed to calculate the projected nose-tip coordinate on the screen. Each grid is of $maxX$ pixels in width and $maxY$ pixels in height, depending on the screen resolution. Each grid cell's central coordinate is taken as reference points to calculate the projected point's shift from the actual point. If the shift is less than the accepted $maxX$ and $maxY$, the corresponding punctuation mark is selected. If the shift is more than that, the adjacent grid cells are considered. Here, the dwell time for selecting a punctuation mark is kept static to 5 frames, followed by a 5 frames of the gap in the nose projection so that the user can shift to the next desired location on the screen. The dwell time for selecting a word from auto-complete suggestion is kept as 20 frames to provide the users sufficient time to switch visual context to check the words displayed in the list and then choose the corresponding option/index. The dwell time can be increased based on the user's convenience. However, increasing the dwell time for selecting a key can reduce the typing speed.

Continuous Sensor Reading and Orientation Tracking.

Almost every commercial smartphone houses several built-in sensors, among which the accelerometer is the most common one. A smartphone's orientation can be measured by the

instantaneous accelerometer reading in the X, Y, and Z directions. Typically, a change of orientation along any of these axis triggers a sharp change in the reading in that direction. Figure 7.6 shows how the direction and phone's position are correlated. Since *Nosype* works in landscape mode, the change in the Y direction is used for the editing function. While using *Nosype* in landscape mode, if the user tilts the device to the right (accelerometer reading for Y-axis > 7) for 1 second, a space is added to the current text. Conversely, if the device is tilted to the left (accelerometer reading for Y-axis < -7) for 1 second, the text's last character is deleted. A span of 1 second is maintained to ensure that the change in orientation is voluntary and implied for editing. This orientation tracking is a continuous process to register the desired editing at any instant of the application's usage. Additionally, if the user moves the device away from the face and no facial region is detected for 30 consecutive frames, the current trajectory T is refreshed and set to null.

7.3.3 Handling the Corner Cases

For both the modules of *Nosype*, we considered events like the possibilities of periodic partial occlusion of face, motion blurring, or failed detection of nose-tip, etc. To compensate for these effects, extrapolated detection on missed frames is considered. In this case, where a nose area goes undetected for an intermediate frame, the last detected nose-tip position is stored, and its distance from the detected facial boundaries is considered. In a frame where the nose is not detected, the extrapolated nose-tip is placed at the same location, relative to the current facial area, at which the last recorded tip was detected concerning that immediate facial area. This extrapolation is based on the assumption that the relative nasal location for a user's face will remain unchanged unless the face significantly rotates. This ensures continuous disruption-free drawing and locating of nose-tips.

7.4 Experimental Setup for Lab-Scale Evaluation

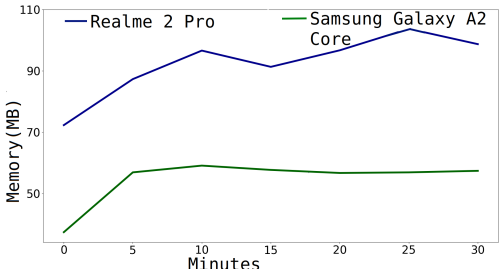
In this section describes the details of *Nosype*'s implementation, in terms of Software, Hardware and environment setup. The detail follows.

7.4.1 *Software & Devices*: Implementation and Profiling of *Nosype*

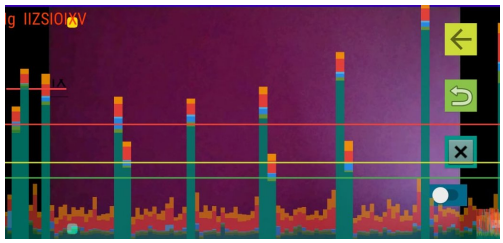
Nosype has been implemented in Android Studio using OpenCV library[2]. The CNN model's offline training is performed in Python, and the trained model is imported to the mobile application as a `.tflite` model. The application has been tested with 10 different smartphone models having a memory (RAM) from 1GB to 8GB, three different Android versions (8.1, 9, and 10), and with minimum resolution as 540×960 pixels, and maximum resolution as 1080×2340

[2] https://opencv.org/ (Access: Friday 11th August, 2023)

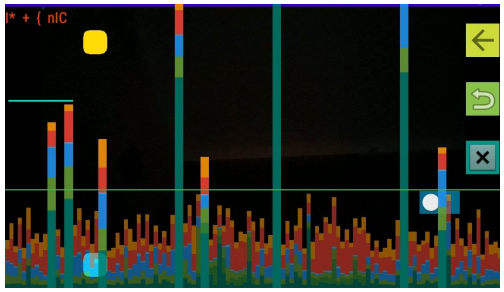
pixels. The application profiling is done on two different smartphones having different RAM – *Realme 2 Pro* and *Samsung Galaxy A2 Core*. Figure 7.7a shows a low memory usage by the application on both the devices over a time frame of 30 minutes. Figures 7.7b and 7.7c show that the application can work well in real-time (≥ 30 frames per second (fps) frame rendering). Each vertical bar in the graph depicts the time required to render a frame. The horizontal lines depict the time-frame. The green line marks 16ms, the time required for each frame to be rendered at 60fps. The yellow and red lines in *Realme 2 pro* (not available for Samsung Galaxy A2 Core) depict a time-frame of about 21ms and 31ms, respectively. Figure 7.7d shows the different components[3] of each of the horizontal bars shown in Figures 7.7b and 7.7c. Component 1 (*Swap buffer*) is proportional to the amount of GPU processing required by the application. Component 2 (*Command Issue*) is the total time for the execution of display lists. *Sync & Upload* (Component 3) indicates the usage of graphical components, and *Draw* (Component 4) indicates the time for creating or updating the view list. Component 5 is the *Measure or Layout* component that indicates the time taken by the view hierarchy. *Animation* (Component 6) depicts the time for running any animation component in the application. Components 7 (*Input handling*) and 8 (*Misc time*) indicate the time to process user input and UI threads, respectively.



(a) Memory Usage of Nosype



(b) GPU Rendering in Realme 2 Pro



(c) GPU Rendering in Galaxy A2 Core

(d) GPU Rendering : color codes

Figure 7.7: Profiling *Nosype* Smartphone Application for Resource Usage Analysis

To test the application's battery usage, the *Samsung Galaxy A2 Core* was charged to 100%, and the application was started and kept open till the battery dropped to 5%. The application kept running continuously for ~ 6 hours on the device before dropping the battery charge to

[3]Check https://developer.android.com/topic/performance/rendering/inspect-gpu-rendering (Access: Friday 11th August, 2023) for detailed meaning of the color codes.

5%.

7.4.2 Participant Details

For the lab-scale evaluations, 11 participants with one or more medical issues like dactylitis, a neurological disorder causing essential tremor, vertigo, visual problems, and joint pains, were considered. The participants were provided with a demonstration of *Nosype* and four other applications developed for baseline comparison. The participants were allowed to use all the applications initially under practice session. This allowed them to be familiar with the unconventional interfaces and control the text entry methods. These discontinuous practice sessions varied from 30–60 minutes depending on the user's convenience and the degree of comfort. After the practice session and the user's acquaintance with the applications' usability, we conduct the experiments and record the results. The users were not intimated with *Nosype*'s advantages before the experiments to avoid any possible bias.

The participants' selection aimed at including people with different medical conditions and considered their comfort in performing the experiments. The specifics of each participant are shown in Table 7.1.

7.4.3 Evaluation Methodology

Nosype combines a projection-based method to select appropriate words from the suggestions or the punctuation marks (the 'Locate' interface), along with a 'Draw' module to use nose-tip movements to draw characters. We evaluate these two components of *Nosype* separately and compare its performance with other baselines. The details of these evaluations have been discussed in the following sections.

Table 7.1: Details of participants with clinical issues

Participant	Age	Gender	Profession	Medical Condition	Additional remark
1	62	Female	Retired Teacher	Visual issues, vertigo and dactylitis	
2	63	Male	Retired corporate professional	Poor eyesight at night, thumb stiffness and tennis elbow	
3	58	Male	Lawyer	Myopia and essential tremor	
4	29	Female	Scholar	Vertigo, ganglion cyst on the right hand's wrist	Even though the cyst was treated, she has been prescribed to restrict conventional texting or gaming activities
5	30	Male	Software developer	Severe migraines after prolonged viewing of texts with tiny fonts or using mobile keypads.	
6	32	Male	IT Professional	Early arthritis, occasional stiffness of fingers and hands	
7	27	Female	IT professional	Eyesight issue	
8	35	Male	IT professional	Migraine arising from focused and prolonged gazing at a mobile screen.	
9	58	Female	Self-employed	Arthritic issues, leading to restricted finger and hand movement, partial optical cataracts	The participant also reported less familiarity with smartphones and is not comfortable with the touch-based text entry method.
10	40	Female	Private tutor	severe migraine arising from viewing small fonts, thus having difficulty in working with smartphone keyboard	
11	40	Male	Banking professional	Vertigo and essential tremor.	

7.5 Evaluation of the *'Projection'* Interface

The objective of this experiment is to check how accurately a user can select a key over the mobile soft-keyboard using the projection method as discussed earlier. We evaluate the 'Locate' (Projection) interface of *Nosype* as follows.

7.5.1 Baselines

We compare the performance of *Nosype* 'Locate' interface with four baselines, three using eye-gaze and eye gesture-based, and the fourth one using fingertip projection-based. The details follow.

Vanilla Eye Projection (VEP): We developed this customized projection-oriented text entry application by using a slightly modified version of the projection function used in *Nosype*'s 'locate' interface. This application solely uses a projection interface for selecting alphabets from a layout grid (Figure 7.8 (left)). However, instead of projecting nose-tip, eye gaze is projected for the selection of alphabets. To enter a text, the user had to gaze at a particular key on the soft grid layout.

EyeSwipe: This application is based on a calibration ratio-based approach [3] for the detection of gaze location on smartphone screen, and hence selection of individual letters from the keyboard (Figure 7.8 (left)) to construct a text.

Cascading Dwell Gaze Typing (CDGT) [168]: This approach uses a CNN-based [172] approach for localizing the gaze coordinate on the screen along with a dynamic, cascading dwell time selection. The same QWERTY layout (Figure 7.8 (left)) has been used for the text entry.

Vanilla Finger Projection (VFP): VFP is another customized application developed by us and uses the same features and functions of *Nosype*. However, instead of detecting the facial landmarks in this application, the users' hands and fingertips are detected. Using finger movement in the air, the user can draw letters or select words and punctuation marks. The punctuation grid is similar to the one use for *Nosype*, as shown in Figure 7.1 (right).

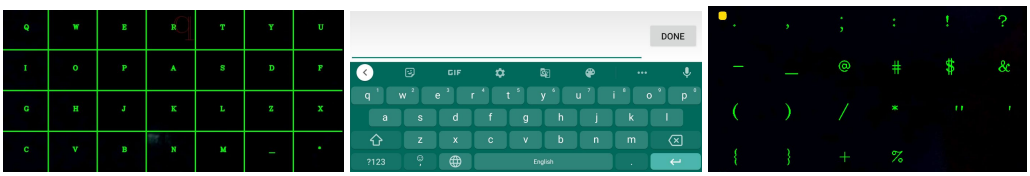


FIGURE 7.8: QWERTY Layout for VEP, EyeSwipe, and CDGT (left), TBK (middle) and punctuation grid (right) for *Nosype*

7.5.2 Evaluation Metrics

In this experiment, we compare different approaches using the following metrics.
Accuracy: Whenever a user projects the nose-tip, the fingertip, or the eye-gaze on a particular key, we record the projection for consecutive 5 frames. If the projection detects the target key for at least three out of the five frames, we consider it a correct projection. The accuracy is measured as the ratio of the number of accurate projections versus all the projection attempts during the experiment.
Shift: Let $(Key.x, Key.y)$ be the x and y coordinates of the central point of each grid, which is also the central point of each key on the screen, and $(Prediction.x, Prediction.y)$ be the x and y coordinates of the predicted or projected gaze/nose-tip/fingertip location on the smartphone screen. Then the metric Shift (S) is defined as follows.

$$S = \sqrt{(Key.x - Prediction.x)^2 + (Key.y - Prediction.y)^2} \tag{7.1}$$

The above metric estimates how well the user can control the key selection using the respective facial or fingertip attributes. A shift (S) close to zero will prove the system's accuracy and the degree of control of the user on the system, as the participants are asked to target the central location of the key on the screen for projection.

7.5.3 Study Design

The 11 participants were presented with the five different applications (*Nosype* and the other four baselines), as discussed above, in the given sequence.

In this **training session**, each of the participant was asked to select the complete set of all keys once using each of the applications. In case a participant felt discomfort or failed to complete a round, they were instructed to take a 5 minutes break after the partial round and restart the round. In total, each participant completed 5 complete rounds and 0–6 additional partial rounds of key selection using the 5 applications for in the practice session.

In the **experiment session**, they were asked to install these applications on their smartphones. Each participant completed 6 experiment sessions. The first session was used to explain the tasks that they have to perform. However, the specific details (like what each keyboard does, which one is our model, etc.) have been hidden from the participants to eliminate the bias. The next five sessions involved the experiments with the five different applications. The first session lasted for 30 minutes on average, and the remaining five sessions lasted for 1 hour each, on average. Session 2 involved experiments with *Nosype*, where the users have been asked to select one of the keys from the punctuation grid layout by looking firmly on that key. As the user looks at a specific key, the nose-tip gets projected on that key using the method. In this experiment, the user had to select each key five times from the punctuation grid. The

remaining four sessions involved experiments with eye-gaze detection and fingertip projection.

In the experiments with eye gaze, the users have been asked to choose a key highlighted on the screen by looking (gazing) on that key and hold the gaze until a beep sound is played. Once the beep sound is played[4], the user has been asked to select another key by gazing at that. The users have been instructed to choose punctuation marks in the fingertip projection session, similar to the *Nosype* interface. In all these experiments, the participants were seated comfortably in a good lighting condition to eliminate the errors due to the surrounding environments, which might affect the comparability among the four applications. Further, we have given at least a 30 minutes gap between two different experimental sessions to eliminate one experiment's influence on another. We have repeated the experiments at least 5 times on different days for each of the participants and finally computed the average performance.

7.5.4 Task Ordering for Evaluation

The order of tasks (conducting the experiments with five different applications) might have an influence on the overall performance of typing with individual applications, as the memorization while typing with one application might affect the typing with another. Further, the fatigue from one task might also affect the next task. To eliminate such bias from the task ordering, we employ a Latin Square method to order the applications and allocate a task sequences to a set of participants. In this case, we repeat the same tasks with a randomized sequence based on the Latin Square. Following this, the orders of tasks and the corresponding participants are mapped in Table 7.2.

TABLE 7.2: Latin Square-based task scheduling to counterbalance the impact of fatigue in evaluating the 'Locate' Interface

Participants ↓\ Task Order ⟶	1	2	3	4	5
1, 6, 11	Nosype	VEP	EyeSwipe	CDGT	VFP
2,7	VEP	EyeSwipe	CDGT	VFP	Nosype
3,8	EyeSwipe	CDGT	VFP	Nosype	VEP
4,9	CDGT	VFP	Nosype	VEP	EyeSwipe
5,10	VFP	Nosype	VEP	EyeSwipe	CDGT

7.5.5 Results

Figure 7.9 shows the dominance of our proposed approach in key selection in terms of accuracy of key selection using the *Projection* interface. To check whether the task ordering impacts the performance, we compare the results obtained from the Latin Latin Square (LS) task ordering

[4]The beep sound is played after recording the projection for five consecutive frames. We compute the accuracy based on this as discussed earlier.

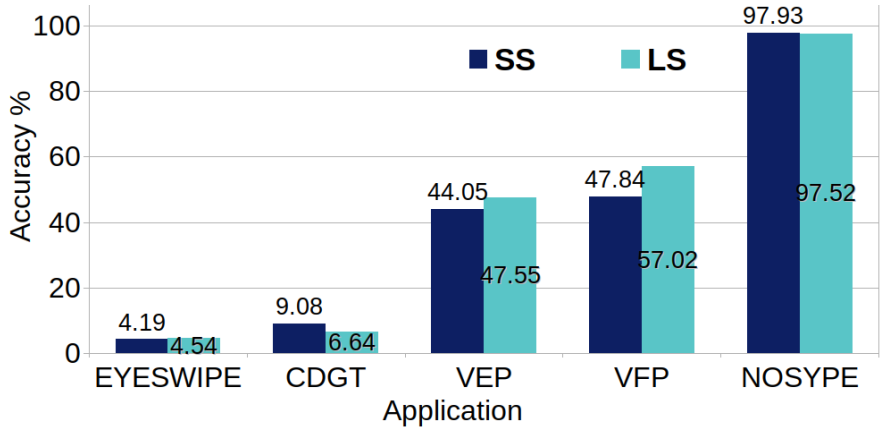
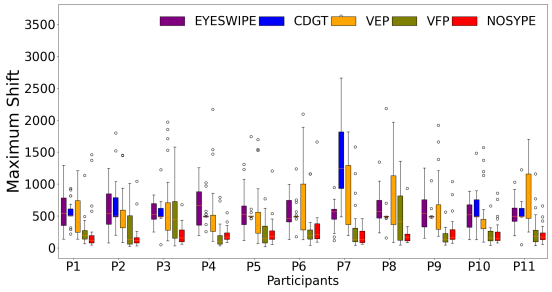


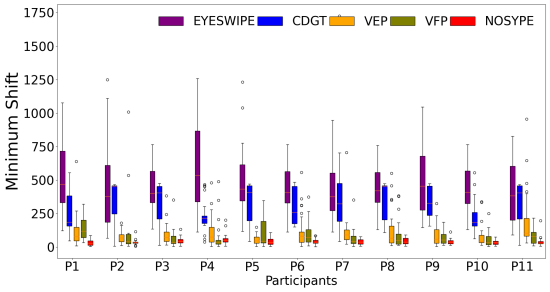
Figure 7.9: Comparison of prediction accuracy with SS and LS task orderings

with a Single Sequence (SS) task ordering method. In the *Single Sequence* task ordering, all the participants have used the same sequence while executing the individual tasks – VEP, EyeSwipe, CDGT, VFP, and Nosype. We experience that the participant could select almost all the keys using *Nosype*. Interestingly, Figure 7.9 indicates a similar trend in the result for both the Latin Square and Single Sequence task ordering-based experiments. The closest competitors are, however, the VEP and VFP models. The difference in the prediction accuracy for these projection models resulted from the fact that the eye was more challenging to control for key selection than nose tip-based selection, and finger movements were difficult for the participants as they have difficulty in flexibly using the fingers. Indeed, we observe that the majority of the participants' finger projection were shaky, resulting in poor accuracy in selecting the keys. The accuracy of EyeSwipe and CDGT are very low since it is challenging to gaze on a particular key over a small display screen. It indicates that even though eye and gaze projection is suitable for the approximate gaze region selection over a large screen system like a desktop, accurate coordinate mapping is often erroneous over a small smartphone screen, thus requiring an alternative solution like nose-tip projection.

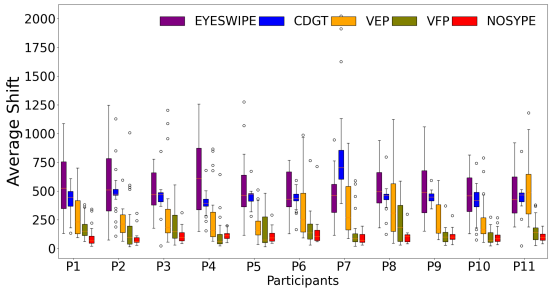
To further investigate the mapped coordinates' accuracy on the smartphone screen, we plot the shift values calculated according to the Equation (7.1) from the experiments with the Latin Square-based task ordering. For each character, the maximum Shift, the minimum Shift, and the average Shift were calculated from the 5 recorded frames. Figures 7.10a, 7.10b, and 7.10c compare the maximum, the minimum, and the average pixel shifts (error) in predicting each key by the 11 participants (P1–P11) over the five applications. We also perform a pairwise statistical testing to check the significance of *Nosype* over other baselines using *Mann-Whitney U test*. In all the cases, the prediction error (Shift) in *Nosype* is significantly less (with $P < 0.05$ over Mann-Whitney U test) than the other four approaches. Interestingly, we observe that glasses significantly increased the vanilla eye projection's error rate, thus implying another



(a) Comparison of maximum errors in prediction



(b) Comparison of minimum errors in prediction



(c) Comparison of average errors in prediction

FIGURE 7.10: Comparison of Baseline Approaches - *NoSype* works better than other baselines ($P < 0.05$ in the significance test)

disadvantage of using eye-based models.

7.6 Evaluating the Messaging Speed

In this experiment, we evaluate how quickly one can input texts using *Nosype*. The tasks primarily aim at understanding the utility of the 'Draw' interface of *Nosype* along with the auto-suggestion feature. In this experiment, the same set of 11 participants was provided with the instructions for the tasks to be performed. The details follow.

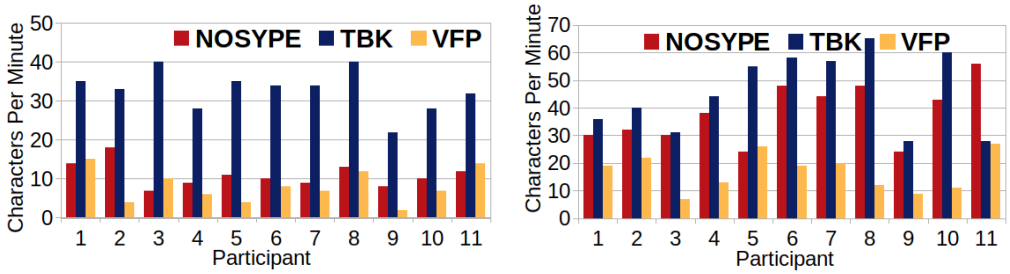
7.6.1 Baselines

We use the **VFP** and a **Touch based keyboard (TBK)** as the baselines for this experiment. The *TBK* is a simple smartphone application that uses the standard Google keyboard for the text entry, as shown in Figure 7.8 (middle). This application helps compare the performance of *Nosype* with conventional smartphone text entry systems for clinically challenged users.

7.6.2 Study Design

This study was conducted in two sessions. In the first session, each participant was given a set of sentences from [147] and was asked to type any 3 of them by using the following two approaches – (a) drawing individual characters of the sentence without using auto-suggestion to select the word, and (b) selecting words from auto-suggestion lists, as quickly as possible. They could redraw a letter in case of the wrong prediction. Selecting a word from the auto-suggestion automatically appends a space after the word in the typed sentence, thus eliminating the requirement of inserting a space explicitly by the user. Moreover, the user could choose to draw the alphabets in either upper or lower case. On selecting the word from the suggestion list, the application automatically converts an entire common English word in lower case. This task identified the typing speed in terms of words/characters per minute. The task was performed with all the 3 applications : *Nosype*, *VFP*, and *TBK*. To make a fair comparison, auto-suggestion was turned on over the other two applications, *VFP* and *TBK*, as well, when we used the auto-suggestion feature of *Nosype*.

In the second session, the participants were given 3 sentences from the same sentence corpus and asked to draw out each letter in the sentence at their own pace. However, they were instructed not to select words from the suggestion list (auto-suggestion was explicitly turned off). This estimates the accuracy of the prediction of letters from the 'Draw' interface. Each sentence consisted of 14 to 40 characters without punctuation and had additional white spaces. This task was also performed with the three applications: *Nosype*, *VFP*, and *TBK*. All the letters, spaces, and punctuation typed by the user were stored in textual log files in the back-end during these two tasks.



(a) Comparison of typing speeds without auto-suggestion (b) Comparison of typing speeds with auto-suggestion

Figure 7.11: Comparison of typing speeds

7.6.3 Task Ordering for Evaluation

To test the possibility of ordering effect on the results, we reorganize the order of the applications used by the participants using Latin Square, using the similar approach as discussed earlier. The orders of tasks and the corresponding participants are mapped in Table 7.3. This method of task ordering counterbalances the effect of possible fatigue on the participants. For each of the applications, we also change the order of keeping auto-suggestion on and off.

Table 7.3: Latin Square-based task scheduling to counterbalance the impact of fatigue in evaluating the speed of text input

Participant ↓ \ Task Order ⟶	1	2	3
1, 4, 7, 10	Nosype	VFP	TBK
2, 5, 8, 11	VFP	TBK	Nosype
3, 6, 9	TBK	Nosype	VFP

7.6.4 Results

Here we summarize the results from the above experiments.

Analysis of Writing Speed

From the first session of this experiment, the typing speed using the draw module in terms of average Words Per Minute (WPM) is computed as 6.31, across all the 11 participants using auto-suggestion in *Nosype*. This calculated value is slightly less than gaze-based models using commercial trackers [168] but more than desktop-based eye gaze oriented text entry systems [10]. Figures 7.11a and 7.11b show the text entry speed of the individual participants using the three applications with two scenarios – without auto-suggestion and with auto-suggestion. By comparing characters typed per minute with and without using the auto-suggestion feature, an average increase of about 27.76 characters, including spaces, has been

observed for *Nosype*, when the auto-suggestion is used. The average characters entered by participants in a minute without auto-suggestion was 11 and that with auto-suggestion was 37.9. We observe that the participants had to draw 1.72 characters on average to get a correct suggestion from the auto-suggestion feature of *Nosype*. For TBK and VFP, the average words per minute without suggestion were 6.9 and 1.6 respectively, and that with auto-suggestion were 8.9 and 3.3 respectively. For TBK the average characters per minute without suggestion is 32.8 and with suggestion is 45.6. For VFP the average characters per minute without suggestion is 8.0 and with suggestion is 16.8. In *Nosype*, without auto-suggestion, the average word per minutes is 2.9. However, the users reported discomfort in using the TBK and VFP applications due to their clinical conditions. Even though TBK has a marginally higher speed, the speed-comfort trade-off has been eliminated by *Nosype's* dominance in term of accuracy. Further, it is comfortable to see that the performance of *Nosype* is close to a standard TBK when the auto-suggestion feature is turned on.

Analysis of Error Percentage during Typing without Auto-Suggestion

From the second session of this study, where participants had to draw out each character without using the auto-suggestion feature, we measured the system's accuracy based on the 'Draw' module. The error percentage in each sentence is evaluated using the standard formulas for 9 different metrics as shown in [230] – Minimum string distance (MSD) error rate (M1), Keystrokes per character (KSPC) (M2), Correction Efficiency (M3), Participant's Conscientiousness (M4), Utilized Bandwidth (M5), Wasted Bandwidth (M6), Total error rate (M7), Non corrected error rate (M8), and Corrected error rate (M9). To derive these standard metrics, some of the following independent parameters are first derived by comparing the original sentence from the given corpus, and the user's typed text. Correct (C) refers to the number of characters, including spaces present in both the original text and the typed text. Incorrect but not fixed (INF) refers to the characters or spaces which are present in the typed text but not there in the original text or the number of characters or spaces that are missing in the typed text but present in the original text. Incorrect but fixed (IF) accounts for the intermediate incorrectly typed characters or spaces that were backspaced and edited and thus fixed (F), referring to the number of editing keys pressed (for example, backspace button). Since IF and F do not appear in the final typed text, we derive these counts from the log files. Table 7.4(left) depicts the formulas to derive the metrics M1-M9 along with their significance to the text entry system.

Table 7.4 shows the average values of these metrics for the three applications. Green cells depict the best, and Blue cells depict the Second Best performance out of the 3 applications, with respect to each metric. $M8$ and $M9$ are not compared as they are proportional to the total error rates.The average typing error for *Nosype* is 6.9 and that of VFP and TBK are 35.7 and 20.6, respectively. The average, standard deviation and 95% Confidence Intervals (CI) of

Table 7.4: Details of the performance metrics (top) and corresponding values (bottom) for the textual evaluation (Green and Blue cells depict the Best and the Second Best performance, respectively, out of the 3 applications, with respect to each metric. $M8$ and $M9$ are not compared as they are proportional to the total error rates.). P indicates 'Participant'.

Metric	Formula	Significance
M1	$(\frac{INF}{(C+INF)}) \times 100$	Indicates the rate of error in the typed text.
M2	$\frac{(C+INF+IF+F)}{(C+INF)}$	Proportional to the cost related to the error and fixes.
M3	$\frac{IF}{F}$	Refers to the effort required for correcting the errors.
M4	$\frac{IF}{(IF+INF)}$	Refers to the user's attentiveness and perfection.
M5	$\frac{C}{(C+INF+IF+F)}$	Rate of transfer of useful information in the system.
M6	$\frac{(INF+IF+F)}{(C+INF+IF+F)}$	Rate of transfer of wasted information.
M7	$(\frac{(INF+IF)}{(C+INF+IF)}) \times 100$	Rate of combined errors, both corrected and uncorrected.
M8	$(\frac{INF}{(C+INF+IF)}) \times 100$	Rate of errors that were corrected.
M9	$(\frac{IF}{(C+INF+IF)}) \times 100$	Rate errors that were not corrected.

P	Model	M1(%)	M2	M3	M4	M5	M6	M7(%)	M8(%)	M9(%)
P1	Nosype	8.82	1.00	-	0.00	0.91	0.09	8.82	8.82	0.00
	VFP	47.06	1	-	0	0.53	0.47	47.06	47.06	0
	TBK	0	1	-	-	1	0	0	0	0
P2	Nosype	3.92	1.07	1	0.5	0.9	0.1	7.33	3.81	3.52
	VFP	44.44	1.11	0	0	0.5	0.5	44.44	44.44	0
	TBK	15.15	1	-	0	0.85	0.15	15.15	15.15	0
P3	Nosype	3.33	1.13	0.83	0.75	0.87	0.13	8.02	3.03	4.99
	VFP	43.75	1	-	0	0.56	0.44	43.75	43.75	0
	TBK	10	1.1	1	0.33	0.82	0.18	14.29	9.52	4.76
P4	Nosype	5.71	1.04	1	0.17	0.91	0.09	7.45	5.6	1.85
	VFP	35.29	1	-	0	0.65	0.35	35.29	35.29	0
	TBK	5.88	1.24	0.33	0.5	0.76	0.24	11.11	5.56	5.56
P5	Nosype	7.79	1.09	0.75	0.28	0.85	0.15	11.02	7.47	3.55
	VFP	35.29	1	-	0	0.65	0.35	35.29	35.29	0
	TBK	3.13	1.28	0.8	0.8	0.76	0.24	13.89	2.78	11.11
P6	Nosype	2.08	1.1	0.67	0.5	0.9	0.1	5.59	2.08	3.51
	VFP	23.53	1.12	1	0.2	0.68	0.32	27.78	22.22	5.56
	TBK	42.11	1	-	0	0.58	0.42	42.11	42.11	0
P7	Nosype	13.5	1.04	1	0.11	0.83	0.17	15.14	13.29	1.85
	VFP	37.5	1	-	0	0.63	0.38	37.5	37.5	0
	TBK	29.41	1.71	1	0.55	0.41	0.59	47.83	21.74	26.09
P8	Nosype	9.71	1.04	-	0.11	0.87	0.13	11.34	9.49	1.85
	VFP	58.82	1	-	0	0.41	0.59	58.82	58.82	0
	TBK	41.18	1.06	1	0.07	0.56	0.44	42.86	40	2.86
P9	Nosype	7.98	1	-	0	0.92	0.08	7.98	7.98	0
	VFP	25	1	-	0	0.75	0.25	25	25	0
	TBK	26.32	1.11	1	0.17	0.67	0.33	30	25	5
P10	Nosype	10.83	1.1	0.5	0.25	0.81	0.19	13.75	10.63	3.13
	VFP	11.76	1.24	1	0.5	0.71	0.29	21.05	10.53	10.53
	TBK	17.65	1.47	0.6	0.5	0.56	0.44	30	15	15
P11	Nosype	2.5	1.06	1	0.5	0.92	0.08	5.28	2.5	2.78
	VFP	31.25	1	-	0	0.69	0.31	31.25	31.25	0
	TBK	36.36	1	-	0	0.64	0.36	36.36	36.36	0

each of the metrics (M1-9) for each of these applications are depicted in Table 7.5. VFP's and TBK's error rates are high, particularly because the participants were uncomfortable using their fingers for typing. The performance metrics show significant accuracy in the *Nosype's* 'Draw' module's performance of the proposed approach and are comparable to popular gaze-based text entry methods [217]. Interestingly, we observe that *Nosype* performs either the best or the second best among all the participants across all the metrics. Thus, we infer that nose-tip-based

Table 7.5: Average (Avg.), Standard Deviation (Stdv.) and 95% Confidence Intervals of performance metrics (Green and Blue cells depict the Best and the Second Best performance, respectively, out of the 3 applications, with respect to each metric. $M8$ and $M9$ are not compared as they are proportional to the total error rates.)

Model		M1(%)	M2	M3	M4	M5	M6	M7(%)	M8(%)	M9(%)
Nosype	Avg.	6.93	1.06	0.84	0.29	0.88	0.12	9.25	6.79	2.46
	Stdv.	3.72	0.04	0.19	0.24	0.04	0.04	3.19	3.68	1.54
	95% CI	4.73,9.13	1.04,1.08	0.73,0.95	0.15,0.43	0.86,0.9	0.1,0.14	7.36,11.14	4.62,8.96	1.55,3.37
VFP	Avg.	35.79	1.04	0.67	0.06	0.61	0.39	37.02	35.56	1.46
	Stdv.	12.9	0.08	0.58	0.16	0.1	0.1	10.96	13.26	3.44
	95% CI	28.17,43.41	0.99,1.09	0.33,1.01	- 0.03,0.15	0.55,0.67	0.33,0.45	30.54,43.5	27.72,43.4	- 0.57,3.49
TBK	Avg.	20.65	1.18	0.82	0.29	0.69	0.31	25.78	19.38	6.4
	Stdv.	15.3	0.23	0.26	0.29	0.17	0.17	15.68	14.98	8.16
	95% CI	11.61,29.69	1.04,1.32	0.66,0.98	0.12,0.46	0.59,0.79	0.21,0.41	16.51,35.05	10.53,28.23	1.58,11.22

text entry is more feasible for the specific user group with clinical disabilities in finger usage.

7.7 Human Study in the Wild

This subjective evaluation of *Nosype* aims at understanding the usability of the application in the real world. This system's objective, even though, is to primarily benefit the users with clinical issues, this evaluation aimed at understanding if a model like *Nosype* can be globalized and be used as a substitute the generic text entry systems. The study hypothesizes that if the common masses can use the application, it can be used under scenarios where traditional text entry methods get cumbersome, even beyond clinical issues. For example, such an application can also be useful for a straphanger carrying commodities in one hand and using the other hand only to send a text message or a user using a smartphone with a damaged or unresponsive touch screen.

7.7.1 Methodology

The users could install *Nosype* in their smartphones and use it for entering a textual message. The participants were instructed to enter at least 30 words using *Nosype* and then provide feedback for the application by filling up a questionnaire based on the popularly used SUS [23].

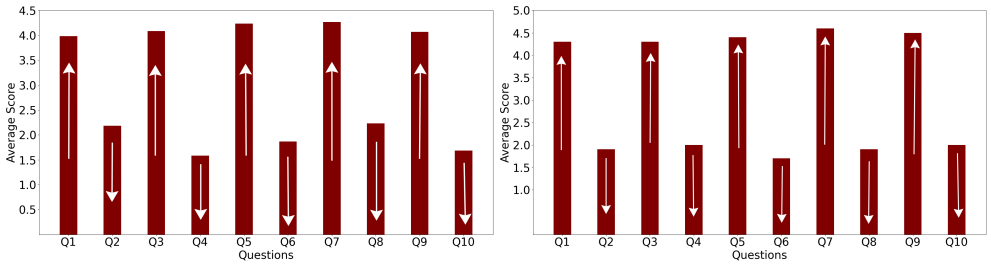
7.7.2 Participant Details

We received responses from 60 participants, including 10 participants with medical conditions like dactylitis and sarcopenia that prevent them from using the regular touch-based text entry applications using small soft-keyboards. It can be noted that in this study, we have not included the previous 11 volunteers who participated in the lab-scale study, as they had been trained explicitly to use the system. Among the 60 participants, 56.7% were male, and 43.3% were female. By categorizing them according to the age groups, 18.33% were found to be under 25, 43.33% belonged to the age group of 25-40, and 38.33% were above 40. The lower and upper age limits were 23 and 90, respectively. The participants also belonged to a wide variety of professions. To categorize them professionally on a higher level, users belonging to the IT industry, sales profiles, and banking services were grouped under 'Service.' 33.33% of the participants belonged to this category. Faculties and research scholars formed the 'Academics' category, which was 15% of the participants. 20% were undergraduate students, and 31.67% were either retired professionals, doctors, lawyers, homemakers, businessmen, or self-employed, forming the "Others" category. 5 of the 10 participants with clinical disabilities were male, and the rest were female. All these 10 participants were senior citizens and belonged to the specific age group of 65–90. Three of them have been diagnosed with dactylitis and the rest with sarcopenia.

7.7.3 Survey Outcome

The System Usability Scale (SUS) revealed an average score of 77.708 over all the 60 participants, thus showing our proposed model's feasibility. Individual average scores of the 10 questions have been displayed in Figure 7.12a. From the figure, we observe that the odd questions have higher scores than the even ones, as expected, thus proving the application's usability. However, we observed that the users are a bit concerned about the application's complexity, and they preferred a training session before using it. This is mostly because people are not very convenient in drawing characters using head movements. Participants who are less than 25 years of age gave an average score of 69.5, those between 25-40 provided 73.8, and people above 40 scored *Nosype*, 84.02. It is interesting to note that older people found the system most usable. Indeed, the older people may not feel very comfortable with finger typing over the tiny soft keyboard on a smartphone, so they found *Nosype* to be a usable system.

The average scores given by the professional groups *Service, Academics, Students* and *Others* were 81.5, 70.27, 66.04, and 84.60, respectively. We found that the service and other people saw the application as most useful. The service people need to type short messages a lot during busy hours when they can use *Nosype* in case the hands are blocked. On the other hand, retired and old-age people feel uncomfortable with finger typing on smartphones; thus *Nosype* was more comforting to them. Gender-wise scores were nearly balanced (Male 77.94 and Female 77.40). From this analysis, it could be inferred that the system was widely popular among people from different backgrounds, ages, and gender.



(a) Question-wise average scores by all the users (b) Question-wise scores by users with clinical condition

FIGURE 7.12: SUS : Average question-wise scores

Figure 7.12b shows the average scores for the 10 SUS questions for the 10 participants with clinical disabilities. We observe that these participants have found the application to be very useful while reporting usability scores between 75 to 87.5. It is comforting to see that *Nosype* can be beneficial for senior citizens and disabled people of the society who can take advantage of smartphone features by using their head movements for interaction.

7.8 Summary

In this work, a novel touch-free writing model has been proposed for mobile devices that use no additional resources and allow flexible head movement, making its applicability more feasible. The approach presents the feasibility of using a nose tip, a rather unconventional facial feature, for touch-free interactive text entry systems. The approach can be beneficial to many users, especially the physically challenged or visually impaired users. Since the model is free from calibration, restricted movement, and erroneous eye tracking due to reflective glasses, it can be used under various indoor and outdoor scenarios. The subjective evaluation and usability study using the standard measurement metrics reveal the model's advantages over popular touch-free text entry models.

Interestingly, to explore whether *Nosype* can perform in the presence of a face-mask (which is inevitable during COVID-19 Pandemic), we also asked 2 of the 11 participants to draw all 26 alphabets and 10 numbers using the Draw interface and select 22 punctuation marks using the Locate interface of *Nosype*, once wearing a mask and then without wearing the mask. Without the mask, the overall accuracy % of *Nosype* for entering all these 58 alphanumeric characters and marks is 93.61%. With mask, even though *Nosype* can perform, its accuracy reduces to 70.83%. As a future scope of this work, *Nosype* can include mask detection technique to further enhance its accuracy. Despite the propitious results, this primary version of the proposed model can be improved by incorporating features like automated numeric and alphabetical mode changing, using n-gram models for the generation of focused suggestion list, using Trie data structure for storing the entire Dictionary , mask detection sub-module and anti-shake or motion-reduction aspects. To this end, this model serves significantly well for short messages. By examining the results, this model's scope seems promising as it will find immense usage among users belonging to all age groups.

8

Conclusion and Future Scopes

The idea of an Utopian future is meshed around the core of educational advancements in the society. We believe that the true essence of societal and personal progress can be experienced through proper education. Education, however, should not be constricted by classroom boundaries, predetermined tenures, and limited accessibility. Therefore, this thesis focuses on smart education, especially it's online mode, where the learners can overcome the drawbacks of classroom-based lectures. Even though self-motivation, and hence attention is a personal responsibility for absorbing the contents of the course, it is often difficult for an online meeting participant to maintain attention. Such inattentional spans are caused by different factors like boredom, lack of comprehension, personal priorities leading to mind wandering, scope for parallel activities, lack of active communication, etc. In this thesis, we address this problem of online education through the development of non-intrusive assistive systems for MOOCs, live meetings and open access YouTube videos. Further, mental disruptions can be an indirect effect of visual context switching. In attempt of taking notes continuously, while attending an online lecture, a learner might get fatigued and miss some key points. On another hand, a learner with certain clinical challenges might not even interact with the device in a convenient way. These challenges, in addition to the COVID 19 protocols of social distancing and touch-free interactivity, led to the development of assistive systems that can promote interactivity between users and devices in a non-conventional, yet seamless manner. The thesis also discusses these novel systems in details.

In this chapter, we first discuss the summary of the entire thesis and then discuss some

future scopes with respect to the proposed models.

8.1 Summary

The thesis begins with a broad discussion on the concepts of attention and human-computer interactivity. Through this exploration, we discuss the challenges associated with automated attention estimation and novel interactive systems. However, the inherent challenges, along with the aim of addressing them motivates the contributions of this thesis. To present a better understanding of the correlated works, the thesis then presents an elaborate study of the existing literature in the field of smart education, thus discussing major research directions like sensing types, educational setups, data modalities and interactivity. These works, along with their future scopes help us understand some research questions regarding whether it is possible to develop some improved assistive models to promote the quality of education. Next, we discuss the major research contributions as follows:

Firstly, we aim at automatically estimating the very basic level of attention– visual attention, through gaze gesture tracking. Based on the observation that a significant proportion of the global population uses smartphone sot attend MOOCs, we develop a smartphone application *GestAtten*, that uses gaze and visual patterns to infer upon the learner's visual attention. In particular, the system checks if (a) the learner is looking at the screen to follow the concept, (b) visually following the content and (c) visually focused on particular objects of interest. Through thorough human evaluations, we prove that the system not only accurately distinguishes pupils with higher visual attention from those with low attention, but also establishes a direct correlation of visual attention with short term memory (high-level cognition). Further, the system shows that manual inspections are perspective in nature and can often be inaccurate. In such cases, an automated estimator is reliable and accurate.

Secondly, we solve the problem of estimating a deeper level of attention– cognition, through facial expression, active communication and vocal cues of presenters/speakers and audiences of live online meetings. Considering the fact that formal meetings are more likely to be attended using laptops/desktops, we propose an application *EmotiConf* that ubiquitously tracks the participants' cognitive focus. Further, through ambient light sensing and human-behaviour modelling, the system identifies whether the participants have opened a different tab to read-/watch another relevant/ irrelevant article during the meeting. The lab and large-scaled human evaluations show that such a system can accurately work for different types of online meetings– formal, informal, presentation-based. Although he system works on client side and is completely secured, it uses the device's camera to track the facial cues.

Thirdly, we propose a system *ExpresSense*, that eliminates the requirement of opening a camera for tracking the facial cues of the learners. This smartphone-based system utilises near-ultrasound signal features to capture the movement of facial AUs and hence the overall

expression. Through thorough evaluation, we show how, the system is robust and work well under external noise, finger movements, distance, elevations, angles, etc. On proving the system's performance, we then utilise the predictions in understanding the engagement of the users, as they watch YouTube videos of different genres.

To solve the challenge of disruption due to visual context switching required for taking manual notes from online courses, we next develop *AutoNotes*, an application that allows the users to simply blink for recording different sections of the lecture. The system then automatically generates textual notes along with blink-based screenshots and auto-marked keywords with their corresponding Wikipedia references. The experimental evaluation show that the system can perform significantly well even in the presence of natural eye-blinking tendencies of the learners.

Finally, we propose a system called *Nosype* that addresses the problems of touch-based text entry in smartphones. The system requires the users to simply move their head in air to draw alphabets. The nose-tip trajectory is hence captured by the application through the smartphone's front camera and reproduced on the device's screen as entered alphabets. From the corresponding auto-generate suggestions and punctuation list, the user can then point their face to a particular word or mark to complete the entire sentence. Through user study, we found that such a system is highly useful for users of different age groups, suffering from Sarcopenia, Dactylitis and other clinical issues. Such a system can be used for taking quick notes while watching course videos in different indoor or outdoor environments by all types of users.

8.2 Future Scopes

Finally, we discuss some of the future scopes of the works presented in this thesis.

8.2.1 Capturing Guessing Behavior in Online Examinations

Online examination is another crucial aspect of online education. While many works have aimed at promoting the quality of lectures, developing recommendation systems for learners, feedback systems for teachers, understanding the guessing behaviour of learners in online MCQ based tests have remained unexplored. Our finding regarding ubiquitous attention estimation from facial cues, along with aspects like response time measurement can be used to understand whether a particular answer has been guessed by the student. Moreover, future work can also aim at classifying guesses as either random, intellectual or analytical guessing based on the learner's answer patterns, interface activities and physiological signals and data, collected solely from the smartphone or laptop's inbuilt hardware.

8.2.2 Tracking Macro and Micro-activities during Online Lectures

In our current works, we have dealt with visual multitasking. However, a learner often practices micro activities like involuntary body movements that can account for their stress levels. Moreover, macro activities like playing games, eating, chatting can lead to disengagement. In future, we will aim at utilising ultrasound chirps for detecting and classifying such activities for fine-tuned estimation of a learner's cognition.

8.2.3 Tracking Typing Speeds and Smartphone Addiction among Students

While exploring the robustness of *ExpresSense*, we realised that not only can finger movements be detected through acoustic sensing, but their patterns can also be detected. Based on this observation, our future works will aim at solving a major social challenge of mobile phone addiction amongst young adults. The system will not only be helpful for the users to control their smartphone usage, but can also be extended to promote their performance in educational courses.

8.2.4 Expression-based Text Generation in Online Classes

Finally, we aim at extending *Nosype*, to develop different text entry systems that can work with facial expressions. Further, such systems can be used to sense sarcasm or predict the mood of the user. The system can also be useful for children with early signs of autism and can be used as a medium for learning various expressions.

References

[1] Abdelrahman, Y., Khan, A. A., Newn, J., Velloso, E., Safwat, S. A., Bailey, J., Bulling, A., Vetere, F., and Schmidt, A. Classifying attention types with thermal imaging and eye tracking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3*, 3 (sep 2019). 21

[2] Ahmad, S., Lavin, A., Purdy, S., and Agha, Z. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing 262* (2017), 134–147. 76

[3] Ahmed, F., Rayhan, M. S. S., Rahman, S., Benazir, N., Chowdhury, A. E., and Al Imran, M. Controlling multimedia player with eye gaze using webcam. In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (2019), IEEE, pp. 152–156. 153

[4] Ahuja, K., Kim, D., Xhakaj, F., Varga, V., Xie, A., Zhang, S., Townsend, J. E., Harrison, C., Ogan, A., and Agarwal, Y. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3*, 3 (2019), 1–26. 15

[5] Ahuja, K., Shah, D., Pareddy, S., Xhakaj, F., Ogan, A., Agarwal, Y., and Harrison, C. Classroom digital twins with instrumentation-free gaze tracking. CHI '21, Association for Computing Machinery. 17

[6] Akdemir, U., Turaga, P., and Chellappa, R. An ontology based approach for activity recognition from video. In *Proceedings of the 16th ACM international conference on Multimedia* (2008), pp. 709–712. 92

[7] Almeida, L. M., Silva, D. P. d., Theodório, D. P., Silva, W. W., Rodrigues, S. C. M., Scardovelli, T. A., Silva, A. P. d., and Bissaco, M. A. S. Altriras: A computer game for training children with autism spectrum disorder in the recognition of basic emotions. *International Journal of Computer Games Technology 2019* (2019), 1–16. 97

[8] Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., and Laeng, B. Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of vision 14*, 4 (2014), 1–1. 61

[9] Arce-Lopera, C., Cardona, J. J., and García, F. Acoustic monitoring system for teacher and student engagement evaluation. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (2020), IEEE, pp. 1–4. 16

[10] ASHTIANI, B., AND MACKENZIE, I. S. Blinkwrite2: an improved text entry method using eye blinks. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications* (2010), pp. 339–345. 159

[11] AVRAHAMI, D., VAN EVERDINGEN, E., AND MARLOW, J. Supporting multitasking in video conferencing using gaze tracking and on-screen activity detection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (2016), pp. 130–134. 21

[12] BABAEI, E., SRIVASTAVA, N., NEWN, J., ZHOU, Q., DINGLER, T., AND VELLOSO, E. Faces of focus: A study on the facial cues of attentional states. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–13. 19, 61

[13] BÂCE, M., STAAL, S., AND BULLING, A. Quantification of users' visual attention during everyday mobile device interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–14. 15

[14] BAO, Y., CHENG, Y., LIU, Y., AND LU, F. Adaptive feature fusion network for gaze tracking in mobile tablets. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), IEEE, pp. 9936–9943. 18

[15] BARRETT, L. F., ADOLPHS, R., MARSELLA, S., MARTINEZ, A. M., AND POLLAK, S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest 20*, 1 (2019), 1–68. 70

[16] BEN, X., REN, Y., ZHANG, J., WANG, S.-J., KPALMA, K., MENG, W., AND LIU, Y.-J. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 9 (2022), 5826–5846. 95

[17] BIAN, Z.-P., HOU, J., CHAU, L.-P., AND MAGNENAT-THALMANN, N. Facial position and expression-based human–computer interface for persons with tetraplegia. *IEEE journal of biomedical and health informatics 20*, 3 (2015), 915–924. 22

[18] BLAIR, M. R., WATSON, M. R., WALSHE, R. C., AND MAJ, F. Extremely selective attention: eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition 35*, 5 (2009), 1196. 61

[19] BOHME, M., HAKER, M., MARTINETZ, T., AND BARTH, E. Head tracking with combined face and nose detection. In *2009 International Symposium on Signals, Circuits and Systems* (2009), IEEE, pp. 1–4. 24

[20] BOSCH, N., D'MELLO, S. K., OCUMPAUGH, J., BAKER, R. S., AND SHUTE, V. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS) 6*, 2 (2016), 1–26. 109

[21] BOZHILOVA, N. S., MICHELINI, G., KUNTSI, J., AND ASHERSON, P. Mind wandering perspective on attention-deficit/hyperactivity disorder. *Neuroscience & Biobehavioral Reviews 92* (2018), 464–476. 5

[22] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32. 144

[23] BROOKE, J., ET AL. Sus-a quick and dirty usability scale. *Usability evaluation in industry 189*, 194 (1996), 4–7. 81, 163

[24] BUSSO, C., BULUT, M., LEE, C.-C., KAZEMZADEH, A., MOWER, E., KIM, S., CHANG, J. N., LEE, S., AND NARAYANAN, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation 42*, 4 (2008), 335–359. 77

[25] CAI, C., PU, H., WANG, P., CHEN, Z., AND LUO, J. We hear your pace: Passive acoustic localization of multiple walking persons. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5*, 2 (2021), 1–24. 20, 99, 101

[26] CALVO, R. A., AND D'MELLO, S. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing 1*, 1 (2010), 18–37. 15

[27] CAO, H., LEE, C.-J., IQBAL, S., CZERWINSKI, M., WONG, P. N., RINTEL, S., HECHT, B., TEEVAN, J., AND YANG, L. Large scale analysis of multitasking behavior during remote meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–13. 21, 61

[28] CHANG, J.-J., LIN, W.-S., AND CHEN, H.-R. How attention level and cognitive style affect learning in a MOOC environment? based on the perspective of brainwave analysis. *Computers in Human Behavior* (2018). 26

[29] CHAU, M., AND BETKE, M. Real time eye tracking and blink detection with usb cameras. Tech. rep., Boston University Computer Science Department, 2005. 135

[30] CHEN, D., TANG, X., OU, Z., AND XI, N. A hierarchical floatboost and mlp classifier for mobile phone embedded eye location system. In *Proceedings of the Third International Conference on Advnaces in Neural Networks - Volume Part II* (2006), pp. 20–25. 18

[31] CHEN, H.-R. Assessment of learners' attention to e-learning by monitoring facial expressions for computer network courses. *Journal of Educational Computing Research 47*, 4 (2012), 371–385. 19

[32] CHEN, Y., NI, T., XU, W., AND GU, T. Swipepass: Acoustic-based second-factor user authentication for smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6*, 3 (sep 2022). 102

[33] CHEN, Y.-L., CHANG, C.-L., AND YEH, C.-S. Emotion classification of youtube videos. *Decision Support Systems 101* (2017), 40–50. 120

[34] CHONG, E., RUIZ, N., WANG, Y., ZHANG, Y., ROZGA, A., AND REHG, J. M. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 383–398. 61

[35] CHOWDHURY, S. R., KAR, P., CHATTOPADHYAY, M., BHATTACHARYA, M., AND CHATTOPADHYAY, S. Mobile enabled content adaptation system for pdf documents. In *2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)* (2017), IEEE, pp. 1–8. 22

[36] CHUNG, J. S., AND ZISSERMAN, A. Out of time: automated lip sync in the wild. In *Asian conference on computer vision* (2016), Springer, pp. 251–263. 19

[37] COETZEE, D., FOX, A., HEARST, M. A., AND HARTMANN, B. Should your MOOC forum use a reputation system? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), ACM, pp. 1176–1187. 26

[38] COHEN, G., AFSHAR, S., TAPSON, J., AND VAN SCHAIK, A. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), IEEE, pp. 2921–2926. 143, 145

[39] COMMODARI, E. Novice readers: The role of focused, selective, distributed and alternating attention at the first year of the academic curriculum. *i-Perception 8*, 4 (2017), 2041669517718557. 5

[40] DALAL, N., AND TRIGGS, B. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (2005), vol. 1, IEEE, pp. 886–893. 73, i

[41] DANISMAN, T., BILASCO, I. M., DJERABA, C., AND IHADDADENE, N. Drowsy driver detection system using eye blink patterns. In *2010 International Conference on Machine and Web Intelligence* (2010), IEEE, pp. 230–233. 22

[42] DAS, S., CHAKRABORTY, S., AND MITRA, B. Quantifying students' involvement during virtual classrooms: A meeting wrapper for the teachers. 82

[43] DAZA, R., MORALES, A., FIERREZ, J., AND TOLOSANA, R. Mebal: A multimodal database for eye blink detection and attention level estimation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (2020), pp. 32–36. 61

[44] DE OLIVEIRA DIAS, M., LOPES, R. D. O. A., AND TELES, A. C. Will virtual replace classroom teaching? lessons from virtual classes via zoom in the times of COVID-19. *Journal of Advances in Education and Philosophy* (2020). 60

[45] DE PAOLIS, L. T., AND DE LUCA, V. The effects of touchless interaction on usability and sense of presence in a virtual environment. *Virtual Reality 26*, 4 (2022), 1551–1571. 22

[46] DEDE, C., BROWN-L'BAHY, T., KETELHUT, D., AND WHITEHOUSE, P. Distance learning (virtual learning). *The internet encyclopedia* (2004). 2

[47] DI LASCIO, E., GASHI, S., AND SANTINI, S. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2*, 3 (sep 2018). 16

[48] DIAZ-TULA, A., AND MORIMOTO, C. H. Augkey: Increasing foveal throughput in eye typing with augmented keys. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (2016), pp. 3533–3544. 24

[49] DISALVO, B., BANDARU, D., WANG, Q., LI, H., AND PLÖTZ, T. Reading the room: Automated, momentary assessment of student engagement in the classroom: Are we there yet? *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6*, 3 (sep 2022). 16

[50] DREWES, H., DE LUCA, A., AND SCHMIDT, A. Eye-gaze interaction for mobile phones. In *Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology* (2007), pp. 364–371. 18, 39

[51] DUBBAKA, A., AND GOPALAN, A. Detecting learner engagement in moocs using automatic facial expression recognition. In *2020 IEEE Global Engineering Education Conference (EDUCON)* (USA, 2020), IEEE, pp. 447–456. 109

[52] DUKIĆ, D., AND SOVIC KRZIC, A. Real-time facial expression recognition using deep learning with application in the active classroom environment. *Electronics 11*, 8 (2022), 1240. 15

[53] EKMAN, P. Facial expression and emotion. *American psychologist 48*, 4 (1993), 384. 70

[54] EKMAN, P., AND FRIESEN, W. V. A new pan-cultural facial expression of emotion. *Motivation and emotion 10*, 2 (1986), 159–168. 15

[55] EMMOREY, K., THOMPSON, R., AND COLVIN, R. Eye gaze during comprehension of american sign language by native and beginning signers. *Journal of Deaf Studies and Deaf Education 14*, 2 (2008), 237–243. 43

[56] EPP, C. D., MUNTEANU, C., AXTELL, B., RAVINTHIRAN, K., ALY, Y., AND MANSIMOV, E. Finger tracking: facilitating non-commercial content production for mobile e-reading applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2017), ACM, p. 34. 26

[57] ESTERMAN, M., AND ROTHLEIN, D. Models of sustained attention. *Current opinion in psychology 29* (2019), 174–180. 5

[58] ESTEVES, A., VELLOSO, E., BULLING, A., AND GELLERSEN, H. Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (2015), ACM, pp. 457–466. 18

[59] FERRAGINA, P., AND SCAIELLA, U. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management* (2010), pp. 1625–1628. 23, 133

[60] FIALA, M., GREEN, D., AND ROTH, G. A panoramic video and acoustic beamforming sensor for videoconferencing. In *Proceedings. Second International Conference on Creating, Connecting and Collaborating through Computing* (2004), IEEE, pp. 47–52. 20

[61] FURUI, S., KIKUCHI, T., SHINNAKA, Y., AND HORI, C. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing 12*, 4 (2004), 401–408. 138

[62] GAO, N., RAHAMAN, M. S., SHAO, W., JI, K., AND SALIM, F. D. Individual and group-wise classroom seating experience: Effects on student engagement in different courses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6*, 3 (2022), 1–23. 16

[63] GAO, N., SHAO, W., RAHAMAN, M. S., AND SALIM, F. D. n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4*, 3 (2020), 1–26. 16

[64] GAO, Y., JIN, Y., CHOI, S., LI, J., PAN, J., SHU, L., ZHOU, C., AND JIN, Z. SonicFace: Tracking facial expressions using a commodity microphone array. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5*, 4 (2021), 1–33. 19, 96, 97, 99, 101

[65] GAO, Y., JIN, Y., LI, J., CHOI, S., AND JIN, Z. Echowhisper: Exploring an acoustic-based silent speech interface for smartphone users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4*, 3 (sep 2020). 102

[66] GARG, N., BAI, Y., AND ROY, N. Owlet: Enabling spatial information in ubiquitous acoustic devices. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services* (New York, NY, USA, 2021), MobiSys '21, Association for Computing Machinery, p. 255–268. 20, 99, 101

[67] GASHI, S., DI LASCIO, E., AND SANTINI, S. Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3*, 1 (mar 2019). 16

[68] GEORGE, A., AND ROUTRAY, A. Fast and accurate algorithm for eye localisation for gaze tracking in low-resolution images. *IET Computer Vision 10*, 7 (2016), 660–669. 24

[69] GHOSH, D., LIU, C., ZHAO, S., AND HARA, K. Commanding and re-dictation: Developing eyes-free voice-based interaction for editing dictated text. *ACM Transactions on Computer-Human Interaction (TOCHI) 27*, 4 (2020), 1–31. 138

[70] GIZATDINOVA, Y., ŠPAKOV, O., AND SURAKKA, V. Face typing: vision-based perceptual interface for hands-free text entry with a scrollable virtual keyboard. In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)* (2012), IEEE, pp. 81–87. 23

[71] GIZATDINOVA, Y., ŠPAKOV, O., TUISKU, O., TURK, M., AND SURAKKA, V. Gaze and head pointing for hands-free text entry: applicability to ultra-small virtual keyboards. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (2018), pp. 1–9. 24

[72] GOLDBERG, P., SÜMER, Ö., STÜRMER, K., WAGNER, W., GÖLLNER, R., GERJETS, P., KASNECI, E., AND TRAUTWEIN, U. Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review 33* (2021), 27–49. 15

[73] GORODNICHY, D. O. On importance of nose for face tracking. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition* (2002), IEEE, pp. 188–193. 24

[74] GORODNICHY, D. O., AND ROTH, G. Nouse â€~use your nose as a mouseâ€™perceptual vision technology for hands-free games and interfaces. *Image and Vision Computing 22*, 12 (2004), 931–942. 24

[75] GRAHAM, C. R. Blended learning systems. *The handbook of blended learning: Global perspectives, local designs 1* (2006), 3–21. 2

[76] GUPTA, A., JI, C., YEO, H.-S., QUIGLEY, A., AND VOGEL, D. Rotoswype: Word-gesture typing using a ring. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12. 24

[77] HAN, S., YANG, S., KIM, J., AND GERLA, M. EyeGuardian: a framework of eye tracking and blink detection for mobile device users. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications* (2012), ACM, p. 6. 37

[78] HANSEN, J. P., TØRNING, K., JOHANSEN, A. S., ITOH, K., AND AOKI, H. Gaze typing compared with input by head and hand. In *Proceedings of the 2004 symposium on Eye tracking research & applications* (2004), pp. 131–138. 138, 147

[79] HARSTON, J., CHAINANI, R., AND FAISAL, A. Gaze grammars-is there an invariant hierarchical sequential structure of human visual attention in natural tasks? *Journal of Vision 22*, 14 (2022), 3894–3894. 15

[80] HAWES, M. T., SZENCZY, A. K., KLEIN, D. N., HAJCAK, G., AND NELSON, B. D. Increases in depression and anxiety symptoms in adolescents and young adults during the covid-19 pandemic. *Psychological medicine 52*, 14 (2022), 3222–3230. 95

[81] HOWARD, A. G., ZHU, M., CHEN, B., KALENICHENKO, D., WANG, W., WEYAND, T., ANDREETTO, M., AND ADAM, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR abs/1704.04861* (2017). 33

[82] HUANG, J., RATHOD, V., SUN, C., ZHU, M., KORATTIKARA, A., FATHI, A., FISCHER, I., WOJNA, Z., SONG, Y., GUADARRAMA, S., AND MURPHY, K. Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR abs/1611.10012* (2016). 35

[83] HUNT, A. R., AND KINGSTONE, A. Covert and overt voluntary attention: linked or independent? *Cognitive Brain Research 18*, 1 (2003), 102–105. 5

[84] HURST-HILLER, O., AND FARAGO, J. Searching for content using voice search queries, Mar. 2 2010. US Patent 7,672,931. 22

[85] HUTT, S., KRASICH, K., R. BROCKMOLE, J., AND K. D'MELLO, S. Breaking out of the lab: Mitigating mind wandering with gaze-based attention-aware technology in classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–14. 15

[86] HUTT, S., MILLS, C., BOSCH, N., KRASICH, K., BROCKMOLE, J., AND D'MELLO, S. "out of the fr-eye-ing pan" towards gaze-based models of attention during learning with technology in the classroom. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (2017), pp. 94–103. 15, 86

[87] ISLAM, B., RAHMAN, M. M., AHMED, T., AHMED, M. Y., HASAN, M. M., NATHAN, V., VATANPARVAR, K., NEMATI, E., KUANG, J., AND GAO, J. A. Breathtrack: Detecting regular breathing phases from unannotated acoustic data captured by a smartphone. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5*, 3 (sep 2021). 20

[88] ITO, K. Video conference system, Apr. 28 1998. US Patent 5,745,161. 60

[89] ITO, K., ONG, C. W., AND KITADA, R. Emotional tears communicate sadness but not excessive emotions without other contextual knowledge. *Frontiers in Psychology 10* (2019), 878. 97

[90] JACK, R. E., SUN, W., DELIS, I., GARROD, O. G., AND SCHYNS, P. G. Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General 145*, 6 (2016), 708. 97

[91] JAMES, W., BURKHARDT, F., BOWERS, F., AND SKRUPSKELIS, I. K. *The principles of psychology*, vol. 1. Macmillan London, 1890. 4

[92] JIANG, H., DYKSTRA, K., AND WHITEHILL, J. Predicting when teachers look at their students in 1-on-1 tutoring sessions. In *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition* (2018), IEEE, pp. 593–598. 15

[93] JOE LOUIS PAUL, I., SASIREKHA, S., UMA MAHESWARI, S., AJITH, K., ARJUN, S., AND ATHESH KUMAR, S. Eye gaze tracking-based adaptive e-learning for enhancing teaching and learning in virtual classrooms. In *Information and Communication Technology for Competitive Strategies*. Springer, 2019, pp. 165–176. 5

[94] JOHANSSON, M. The hilbert transform. *Mathematics Master's Thesis. Växjö University, Suecia. Disponible en internet: http://w3. msi. vxu. se/exarb/mj_ex. pdf, consultado el 19* (1999). 106

[95] JOHO, H., STAIANO, J., SEBE, N., AND JOSE, J. M. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications 51*, 2 (2011), 505–523. 118

[96] KALYUGA, S. Cognitive load theory: How many types of load does it really need? *Educational Psychology Review 23*, 1 (2011), 1–19. 61

[97] KAMILALI, D., AND SOFIANOPOULOU, C. Microlearning as innovative pedagogy for mobile learning in MOOCs. *International Association for Development of the Information Society* (2015). 26

[98] KANGAS, J., AKKIL, D., RANTALA, J., ISOKOSKI, P., MAJARANTA, P., AND RAISAMO, R. Gaze gestures and haptic feedback in mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 435–438. 18

[99] KAR, P., CHATTOPADHYAY, S., AND CHAKRABORTY, S. Bifurcating cognitive attention from visual concentration: Utilizing cooperative audiovisual sensing for demarcating inattentive online meeting participants. *Proc. ACM Hum.-Comput. Interact. 6*, CSCW2 (nov 2022). 123

[100] KASSNER, M., PATERA, W., AND BULLING, A. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication* (2014), ACM, pp. 1151–1160. 18, 51

[101] KAZEMI, V., AND SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 1867–1874. 71, 73, i

[102] KELLER, A. S., DAVIDESCO, I., AND TANNER, K. D. Attention matters: How orchestrating attention may relate to classroom learning. *CBE—Life Sciences Education 19*, 3 (2020), fe5. 4

[103] KENNEDY, G., COFFRIN, C., DE BARBA, P., AND CORRIN, L. Predicting success: how learners' prior knowledge, skills and activities predict MOOC performance. In *Proceedings of the fifth international conference on learning analytics and knowledge* (2015), ACM, pp. 136–140. 52

[104] KHAN, S. S., SUNNY, M. S. H., HOSSAIN, M. S., HOSSAIN, E., AND AHMAD, M. Nose tracking cursor control for the people with disabilities: An improved hci. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)* (2017), IEEE, pp. 1–5. 24

[105] KIM, M., WANG, O., AND NG, N. Convolutional neural network architectures for gaze estimation on mobile devices. 24

[106] KNUDSEN, E. I. Fundamental components of attention. *Annual review of neuroscience 30*, 1 (2007), 57–78. 4

[107] KO, M., LI, J., AND LEE, C. Learning minimal intra-genre multimodal embedding from trailer content and reactor expressions for box office prediction. In *2019 IEEE International Conference on Multimedia and Expo (ICME)* (Los Alamitos, CA, USA, jul 2019), IEEE Computer Society, pp. 1804–1809. 118

[108] KOEHN, P., ET AL. Europarl: A parallel corpus for statistical machine translation. In *MT summit* (2005), vol. 5, Citeseer, pp. 79–86. 78

[109] KONG, C., ZHENG, K., WANG, S., ROCHA, A., AND LI, H. Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones. *IEEE Transactions on Information Forensics and Security 17* (2022), 3238–3253. 102

[110] KOSMYNA, N., SARAWGI, U., AND MAES, P. Attentivu: Evaluating the feasibility of biofeedback glasses to monitor and improve attention. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (2018), pp. 999–1005. 16, 60

[111] KRAFKA, K., KHOSLA, A., KELLNHOFER, P., KANNAN, H., BHANDARKAR, S., MATUSIK, W., AND TORRALBA, A. Eye tracking for everyone. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). 18

[112] KRAFKA, K., KHOSLA, A., KELLNHOFER, P., KANNAN, H., BHANDARKAR, S., MATUSIK, W., AND TORRALBA, A. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2176–2184. 18

[113] KRITHIKA, L. B., AND GG, L. P. Student emotion recognition system (sers) for e-learning improvement based on learner concentration metric. *Procedia Computer Science 85* (2016), 767–776. 17

[114] KRÓLAK, A., AND STRUMIŁŁO, P. Eye-blink detection system for human–computer interaction. *Universal Access in the Information Society 11* (2012), 409–419. 22

[115] KUMAR, A., PAEK, T., AND LEE, B. Voice typing: a new speech interaction model for dictation on touchscreen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 2277–2286. 23

[116] KUMAR, A., SRIVASTAVA, K., YADAV, K., AND DESHMUKH, O. Multi-faceted index driven navigation for educational videos in mobile phones. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (2017), ACM, pp. 357–361. 26

[117] KUMAR, C., HEDESHY, R., MACKENZIE, I. S., AND STAAB, S. Tagswipe: Touch assisted gaze swipe for text entry. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–12. 138

[118] KUZMINYKH, A., AND RINTEL, S. Classification of functional attention in video meetings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13. 21, 60

[119] LAMME, V. A. Why visual attention and awareness are different. *Trends in cognitive sciences 7*, 1 (2003), 12–18. 5, 61

[120] LANDER, C., KOSMALLA, F., WIEHR, F., AND GEHRING, S. Using corneal imaging for measuring a human's visual attention. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (2017), pp. 947–952. 14

[121] LEE, C. S. Exploring emotional expressions on youtube through the lens of media system dependency theory. *New media & society 14*, 3 (2012), 457–475. 120

[122] LEE, E., KANG, J. I., PARK, I. H., KIM, J.-J., AND AN, S. K. Is a neutral face really evaluated as being emotionally neutral? *Psychiatry research 157*, 1-3 (2008), 77–85. 97

[123] LEMLEY, J., KAR, A., DRIMBAREAN, A., AND CORCORAN, P. Efficient cnn implementation for eye-gaze estimation on low-power/low-quality consumer imaging systems. *arXiv preprint arXiv:1806.10890* (2018). 24

[124] LI, D., LIU, J., LEE, S. I., AND XIONG, J. *FM-Track: Pushing the Limits of Contactless Multi-Target Tracking Using Acoustic Signals*. Association for Computing Machinery, New York, NY, USA, 2020, pp. 150–163. 20, 99, 101

[125] LI, K., ZHANG, R., LIANG, B., GUIMBRETIÈRE, F., AND ZHANG, C. Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 6*, 2 (jul 2022). 19

[126] LI, K., ZHANG, R., LIANG, B., GUIMBRETIÈRE, F., AND ZHANG, C. EarIO: A low-power acoustic sensing earable for continuously tracking detailed facial movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6*, 2 (2022), 1–24. 96, 97

[127] LI, S., LAJOIE, S. P., ZHENG, J., WU, H., AND CHENG, H. Automated detection of cognitive engagement to inform the art of staying engaged in problem-solving. *Computers & Education 163* (2021), 104114. 15

[128] LI, Y., CHANG, M.-C., AND LYU, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE international workshop on information forensics and security (WIFS)* (2018), IEEE, pp. 1–7. 22

[129] LIAN, J., LOU, J., CHEN, L., AND YUAN, X. Echospot: Spotting your locations via acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5*, 3 (2021), 1–21. 20, 99, 101

[130] LIAN, J., YUAN, X., LI, M., AND TZENG, N.-F. Fall detection via inaudible acoustic sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5*, 3 (2021), 1–21. 20

[131] LIAN, Y. Smart education: Education reform in the age of intelligence. In *2021 5th International Conference on Education and E-Learning* (New York, NY, USA, 2022), ICEEL 2021, Association for Computing Machinery, p. 41–45. 2

[132] LIETZ, R., HARRAGHY, M., CALDERON, D., BRADY, J., BECKER, E., AND MAKEDON, F. Survey of mood detection through various input modes. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (New York, NY, USA, 2019), PETRA '19, Association for Computing Machinery, p. 28–31. 95

[133] LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P., AND ZITNICK, C. L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision* (2014), Springer, pp. 740–755. 33

[134] LISETTI, C. L., AND SCHIANO, D. J. Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragmatics & cognition 8*, 1 (2000), 185–235. 95

[135] LIU, F., AND FANG, J. Multi-scale audio spectrogram transformer for classroom teaching interaction recognition. *Future Internet 15*, 2 (2023), 65. 16

[136] LIU, J., LI, D., WANG, L., AND XIONG, J. Blinklistener: "listen" to your eye blink using your smartphone. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5*, 2 (jun 2021). 18, 96, 99, 105, 107, 108

[137] LIU, W., ANGUELOV, D., ERHAN, D., SZEGEDY, C., REED, S. E., FU, C., AND BERG, A. C. SSD: single shot multibox detector. *CoRR abs/1512.02325* (2015). 33

[138] LIU, Z., HUANG, W., ZHENG, Y., AND SUN, M. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (2010), pp. 366–376. 23

[139] LORENZ, O., AND THOMAS, U. Real time eye gaze tracking system using cnn-based facial features for human attention measurement. In *VISIGRAPP (5: VISAPP)* (2019), pp. 598–606. 5

[140] LU, Y., YU, C., FAN, S., BI, X., AND SHI, Y. Typing on split keyboards with peripheral vision. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12. 24

[141] LU, Y., YU, C., YI, X., SHI, Y., AND ZHAO, S. Blindtype: Eyes-free text entry on handheld touchpad by leveraging thumb's muscle memory. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1*, 2 (2017), 1–24. 24

[142] LUCEY, P., COHN, J. F., KANADE, T., SARAGIH, J., AMBADAR, Z., AND MATTHEWS, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops* (2010), IEEE, pp. 94–101. 22, 74

[143] LYONS, K., KIM, H., AND NEVO, S. Paying attention in meetings: Multitasking in virtual worlds. In *First Symposium on the Personal Web, Co-located with CASCON* (2010), vol. 2005, p. 7. 21, 64

[144] MA, S., ZHOU, T., NIE, F., AND MA, X. Glancee: An adaptable system for instructors to grasp student learning status in synchronous online classes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022), CHI '22, Association for Computing Machinery. 20

[145] MACK, A. Inattentional blindness: Looking without seeing. *Current directions in psychological science 12*, 5 (2003), 180–184. 5, 61

[146] MACKENZIE, I. S., AND ASHTIANI, B. Blinkwrite: efficient text entry using eye blinks. *Universal Access in the Information Society 10*, 1 (2011), 69–80. 23

[147] MACKENZIE, I. S., AND SOUKOREFF, R. W. Phrase sets for evaluating text entry techniques. In *CHI'03 extended abstracts on Human factors in computing systems* (2003), pp. 754–755. 158

[148] MACKENZIE, I. S., AND ZHANG, S. X. The design and evaluation of a high-performance soft keyboard. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (1999), pp. 25–31. 23

[149] MAFFEI, A., AND ANGRILLI, A. Spontaneous eye blink rate: An index of dopaminergic component of sustained attention and fatigue. *International Journal of Psychophysiology 123* (2018), 58–63. 61

[150] MAJARANTA, P., AHOLA, U.-K., AND ŠPAKOV, O. Fast gaze typing with an adjustable dwell time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009), pp. 357–360. 138

[151] MAJARANTA, P., AND RÄIHÄ, K.-J. Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 symposium on Eye tracking research & applications* (2002), pp. 15–22. 138

[152] MARIAKAKIS, A., GOEL, M., AUMI, M. T. I., PATEL, S. N., AND WOBBROCK, J. O. Switchback: Using focus and saccade tracking to guide users' attention for mobile task resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), pp. 2953–2962. 17

[153] MARK, G., CZERWINSKI, M., AND IQBAL, S. T. Effects of individual differences in blocking workplace distractions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–12. 61

[154] MARK, G., IQBAL, S. T., CZERWINSKI, M., JOHNS, P., AND SANO, A. Neurotics can't focus: An in situ study of online multitasking in the workplace. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (2016), pp. 1739–1744. 61

[155] MARLOW, J., VAN EVERDINGEN, E., AND AVRAHAMI, D. Taking notes or playing games? understanding multitasking in video communication. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (2016), pp. 1726–1737. 21

[156] MASSÉ, B., BA, S., AND HORAUD, R. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE transactions on pattern analysis and machine intelligence 40*, 11 (2017), 2711–2724. 61

[157] MCHUGH, M. L. Interrater reliability: the kappa statistic. *Biochemia medica 22*, 3 (2012), 276–282. 69, 82, 84

[158] MEHTA, L., MUSTAFA, A., AND AKAD. Prediction and localization of student engagement in the wild. In *Digital Image Computing: Techniques and Applications (DICTA), 2018 International Conference on, IEEE* (2018). 19

[159] MIAO, Y., DONG, H., JAAM, J. M. A., AND SADDIK, A. E. A deep learning system for recognizing facial expression in real-time. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 15*, 2 (2019), 1–20. 95, 96

[160] MIRJAFARI, S., MASABA, K., GROVER, T., WANG, W., AUDIA, P., CAMPBELL, A. T., CHAWLA, N. V., SWAIN, V. D., CHOUDHURY, M. D., DEY, A. K., ET AL. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3*, 2 (2019), 1–24. 60, 61

[161] MIRSKY, A. F., ANTHONY, B. J., DUNCAN, C. C., AHEARN, M. B., AND KELLAM, S. G. Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology review 2*, 2 (1991), 109–145. 21

[162] MISSIMER, E., AND BETKE, M. Blink and wink detection for mouse pointer control. In *Proceedings of the 3rd international conference on pervasive technologies related to assistive environments* (2010), pp. 1–8. 22

[163] MOHAMED, Z., EL HALABY, M., SAID, T., SHAWKY, D., AND BADAWI, A. Characterizing focused attention and working memory using eeg. *Sensors 18*, 11 (2018), 3743. 5

[164] MOHAMMED, A. A. A., ET AL. Efficient eye blink detection method for disabled-helping domain. *International Journal of Advanced Computer Science and Applications 5*, 5 (2014). 22

[165] MONKARESI, H., BOSCH, N., CALVO, R. A., AND D'MELLO, S. K. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing 8*, 1 (2016), 15–28. 16

[166] MOORE, J. L., DICKSON-DEANE, C., AND GALYEN, K. e-learning, online learning, and distance learning environments: Are they the same? *The Internet and higher education 14*, 2 (2011), 129–135. 2

[167] MORRIS, T., BLENKHORN, P., AND ZAIDI, F. Blink detection for real-time eye tracking. *Journal of Network and Computer Applications 25*, 2 (2002), 129–143. 22

[168] MOTT, M. E., WILLIAMS, S., WOBBROCK, J. O., AND MORRIS, M. R. Improving dwell-based gaze typing with dynamic, cascading dwell times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 2558–2570. 153, 159

[169] MUKHOPADHYAY, M., PAL, S., NAYYAR, A., PRAMANIK, P. K. D., DASGUPTA, N., AND CHOUDHURY, P. Facial emotion detection to assess learner's state of mind in an online learning system. In *Proceedings of the 2020 5th international conference on intelligent information technology* (2020), pp. 107–115. 22

[170] MUTLU, B., FORLIZZI, J., AND HODGINS, J. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots* (2006), IEEE, pp. 518–523. 15

[171] NAGAMATSU, T., YAMAMOTO, M., AND SATO, H. MobiGaze: Development of a gaze interface for handheld mobile devices. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (2010), pp. 3349–3354. 18, 51

[172] NAMBI, A. U., MEHTA, I., GHOSH, A., LINGAM, V., AND PADMANABHAN, V. N. Alt: towards automating driver license testing using smartphones. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems* (2019), pp. 29–42. 153

[173] NGOC ANH, B., TUNG SON, N., TRUONG LAM, P., PHUONG CHI, L., HUU TUAN, N., CONG DAT, N., HUU TRUNG, N., UMAR AFTAB, M., AND VAN DINH, T. A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences 9*, 22 (2019), 4729. 15

[174] NGUYEN, P., BUI, N., NGUYEN, A., TRUONG, H., SURESH, A., WHITLOCK, M., PHAM, D., DINH, T., AND VU, T. Tyth-typing on your teeth: Tongue-teeth localization for human-computer interface. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services* (2018), pp. 269–282. 24

[175] NOMURA, K., IWATA, M., AUGEREAU, O., AND KISE, K. Estimation of student's engagement using a smart chair. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (New York, NY, USA, 2018), UbiComp '18, Association for Computing Machinery, p. 186–189. 16

[176] NORIEGA, L. Multilayer perceptron tutorial. *School of Computing. Staffordshire University* (2005). 72

[177] OBERAUER, K. Working memory and attention—a conceptual analysis and review. *Journal of cognition* (2019). 4

[178] OEPPEN, R. S., SHAW, G., AND BRENNAN, P. A. Human factors recognition at virtual meetings and video conferencing: how to get the best performance from yourself and others. *British Journal of Oral and Maxillofacial Surgery 58*, 6 (2020), 643–646. 64

[179] OHNO, T., MUKAWA, N., AND KAWATO, S. Just blink your eyes: A head-free gaze tracking system. In *CHI'03 extended abstracts on Human factors in computing systems* (2003), pp. 950–957. 22

[180] OTHMAN, M., AMARAL, T., MCNANEY, R., SMEDDINCK, J. D., VINES, J., AND OLIVIER, P. Crowdeyes: crowdsourcing for robust real-world mobile eye tracking. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2017), ACM, p. 18. 28

[181] PANNING, A., AL-HAMADI, A., AND MICHAELIS, B. A color based approach for eye blink detection in image sequences. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (2011), IEEE, pp. 40–45. 22

[182] PAPOUTSAKI, A., SANGKLOY, P., LASKEY, J., DASKALOVA, N., HUANG, J., AND HAYS, J. Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence-IJCAI 2016* (2016). 18

[183] PARANJPE, D. Learning document aboutness from implicit user feedback and document structure. In *Proceedings of the 18th ACM conference on Information and knowledge management* (2009), pp. 365–374. 23

[184] PEI, J., AND SHAN, P. A micro-expression recognition algorithm for students in classroom learning based on convolutional neural network. *Traitement du Signal 36*, 6 (2019). 15

[185] PELL, M. D., AND KOTZ, S. A. On the time course of vocal emotion recognition. *PLoS One 6*, 11 (2011), e27256. 77

[186] PETRUSHIN, V. A. Emotion recognition in speech signal: experimental study, development, and application. In *Sixth international conference on spoken language processing* (2000). 77

[187] PETTERSSON, K., MÜLLER, K., SOKKA, L., AND PAKARINEN, S. Capturing attentional problems with smart eyewear. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (2019), pp. 643–646. 15, 60

[188] PHAM, P., AND WANG, J. Adaptive review for mobile MOOC learning via implicit physiological signal sensing. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (2016), ACM, pp. 37–44. 19, 26

[189] PICCINNO, F., AND FERRAGINA, P. From tagme to wat: a new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation* (2014), pp. 55–62. 23

[190] PILARCZYK, J., KUNIECKI, M., WOŁOSZYN, K., AND STERNA, R. Blue blood, red blood. how does the color of an emotional scene affect visual attention and pupil size? *Vision Research 171* (2020), 36–45. 61

[191] PONZA, M., FERRAGINA, P., AND PICCINNO, F. Swat: A system for detecting salient wikipedia entities in texts. *Computational Intelligence 35*, 4 (2019), 858–890. 23, 133

[192] POSNER, M. I. Cognition: An introduction. 61

[193] POSNER, M. I. Orienting of attention. *Quarterly journal of experimental psychology 32*, 1 (1980), 3–25. 5, 61

[194] POSNER, M. I., AND BOIES, S. J. Components of attention. *Psychological review 78*, 5 (1971), 391. 4

[195] POWERS, D. M. W. The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (2012), pp. 345–355. 69

[196] PRAETORIUS, A.-K., KLIEME, E., HERBERT, B., AND PINGER, P. Generic dimensions of teaching quality: The german framework of three basic dimensions. *ZDM 50* (2018), 407–426. 15

[197] RÄIHÄ, K.-J., AND OVASKA, S. An exploratory study of eye typing fundamentals: dwell time, text entry rate, errors, and workload. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2012), pp. 3001–3010. 23

[198] REN, S., CAO, X., WEI, Y., AND SUN, J. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1685–1692. 144

[199] ROBAL, T., ZHAO, Y., LOFI, C., AND HAUFF, C. Webcam-based attention tracking in online learning: A feasibility study. In *23rd International Conference on Intelligent User Interfaces* (2018), pp. 189–197. 86

[200] RUAN, S., WOBBROCK, J. O., LIOU, K., NG, A., AND LANDAY, J. Speech is 3x faster than typing for english and mandarin text entry on mobile devices. *arXiv preprint arXiv:1608.07323* (2016). 23

[201] RUSSELL, J. A. Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies. *Psychological bulletin 115*, 1 (1994), 102. 70

[202] RUTTER, L. A., DODELL-FEDER, D., VAHIA, I. V., FORESTER, B. P., RESSLER, K. J., WILMER, J. B., AND GERMINE, L. Emotion sensitivity across the lifespan: Mapping clinical risk periods to sensitivity to facial emotion intensity. *Journal of experimental psychology: general 148*, 11 (2019), 1993. 22

[203] SAGONAS, C., TZIMIROPOULOS, G., ZAFEIRIOU, S., AND PANTIC, M. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2013), pp. 896–903. 144

[204] SAINI, M. K., AND GOEL, N. How smart are smart classrooms? a review of smart classroom technologies. *ACM Computing Surveys (CSUR) 52*, 6 (2019), 1–28. 2

[205] SANDLER, M., HOWARD, A., ZHU, M., ZHMOGINOV, A., AND CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (USA, 2018), IEEE, pp. 4510–4520. 111

[206] SARAKIT, P., THEERAMUNKONG, T., HARUECHAIYASAK, C., AND OKUMURA, M. Classifying emotion in thai youtube comments. In *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)* (USA, 2015), IEEE, pp. 1–5. 120

[207] SARCAR, S., PANWAR, P., AND CHAKRABORTY, T. Eyek: an efficient dwell-free eye gaze-based text entry system. In *Proceedings of the 11th asia pacific conference on computer human interaction* (2013), pp. 215–220. 23

[208] SARKAR, A., RINTEL, S., BOROWIEC, D., BERGMANN, R., GILLETT, S., BRAGG, D., BAYM, N., AND SELLEN, A. The promise and peril of parallel chat in video meetings for work. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–8. 21

[209] SAVCHENKO, A. V., SAVCHENKO, L. V., AND MAKAROV, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing 13*, 4 (2022), 2132–2143. 19

[210] SAVIGNY, J., AND PURWARIANTI, A. Emotion classification on youtube comments using word embedding. In *2017 international conference on advanced informatics, concepts, theory, and applications (ICAICTA)* (USA, 2017), IEEE, pp. 1–5. 120

[211] SAWAHATA, Y., KHOSLA, R., KOMINE, K., HIRUMA, N., ITOU, T., WATANABE, S., SUZUKI, Y., HARA, Y., AND ISSIKI, N. Determining comprehension and quality of tv programs using eye-gaze tracking. *Pattern Recognition 41*, 5 (2008), 1610–1626. 43

[212] SCHALKWYK, J., BEEFERMAN, D., BEAUFAYS, F., BYRNE, B., CHELBA, C., COHEN, M., KAMVAR, M., AND STROPE, B. "your word is my command": Google search by voice: A case study. *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics* (2010), 61–90. 22

[213] SCHENK, S., DREISER, M., RIGOLL, G., AND DORR, M. Gazeeverywhere: enabling gaze-only user interaction on an unmodified desktop pc in everyday scenarios. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 3034–3044. 23

[214] SEAN, V., CIBRIAN, F., JOHNSON, J., PASS, H., AND BOYD, L. Toward digital image processing and eye tracking to promote visual attention for people with autism. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (2019), pp. 194–197. 15, 60

[215] SEN, R., SIRIAH, P., AND RAMAN, B. Roadsoundsense: Acoustic sensing based road congestion monitoring in developing regions. In *2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks* (USA, 2011), IEEE, pp. 125–133. 20

[216] SENGUPTA, K., BHATTARAI, S., SARCAR, S., MACKENZIE, I. S., AND STAAB, S. Leveraging error correction in voice-based text entry by talk-and-gaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–11. 138

[217] SENGUPTA, K., MENGES, R., KUMAR, C., AND STAAB, S. Gazethekey: Interactive keys to integrate word predictions for gaze-based text entry. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion* (2017), pp. 121–124. 161

[218] SENGUPTA, K., MENGES, R., KUMAR, C., AND STAAB, S. Impact of variable positioning of text prediction in gaze-based text entry. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (2019), pp. 1–9. 138

[219] SHARMA, K. Gaze analysis methods for learning analytics. Tech. rep., EPFL, 2015. 27

[220] SHARMA, K., ALAVI, H. S., JERMANN, P., AND DILLENBOURG, P. A gaze-based learning analytics model: In-video visual feedback to improve learner's attention in moocs. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (New York, NY, USA, 2016), LAK '16, ACM, pp. 417–421. 27

[221] SHARMA, K., NIFORATOS, E., GIANNAKOS, M., AND KOSTAKOS, V. Assessing cognitive performance using physiological and facial features: Generalizing across contexts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4*, 3 (2020), 1–41. 109

[222] SHARMA, V. K., MURTHY, L., AND BISWAS, P. Enabling learning through play: Inclusive gaze-controlled human-robot interface for joystick-based toys. In *Social Robotics: 14th International Conference, ICSR 2022, Florence, Italy, December 13–16, 2022, Proceedings, Part II* (2023), Springer, pp. 452–461. 15

[223] SHI, Y., XIE, W., XU, G., SHI, R., CHEN, E., MAO, Y., AND LIU, F. The smart classroom: merging technologies for seamless tele-education. *IEEE Pervasive Computing 2*, 02 (2003), 47–55. 2

[224] SHIMADA, A., OKUBO, F., YIN, C., AND OGATA, H. Automatic summarization of lecture slides for enhanced student previewtechnical report and user study. *IEEE Transactions on Learning Technologies 11*, 2 (2018), 165–178. 22

[225] SILVA, L. P., ZAVAN, F. H. D. B., BELLON, O. R., AND SILVA, L. Follow that nose: tracking faces based on the nose region and image quality feedback. In *Conf. on Graphics, Patterns and Images-W. Face Processing* (2016). 24

[226] SINGH, H., AND MIAH, S. J. Smart education literature: A theoretical analysis. *Education and Information Technologies 25*, 4 (2020), 3299–3328. 2

[227] SINGH, S., GUPTA, A., AND PAVITHR, R. Automatic classroom monitoring system using facial expression recognition. In *International Conference on Artificial Intelligence and Sustainable Engineering: Select Proceedings of AISE 2020, Volume 1* (2022), Springer, pp. 151–165. 15

[228] SONG, X., HUANG, K., AND GAO, W. Facelistener: Recognizing human facial expressions via acoustic sensing on commodity headphones. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)* (USA, 2022), IEEE, pp. 145–157. 97

[229] SONG, X., YANG, B., YANG, G., CHEN, R., FORNO, E., CHEN, W., AND GAO, W. Spirosonic: Monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking* (New York, NY, USA, 2020), MobiCom '20, Association for Computing Machinery. 20, 102

[230] SOUKOREFF, R. W., AND MACKENZIE, I. S. Metrics for text entry research: an evaluation of msd and kspc, and a new unified error metric. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2003), pp. 113–120. 160

[231] SOUKUPOVA, T., AND CECH, J. Eye blink detection using facial landmarks. In *21st computer vision winter workshop, Rimske Toplice, Slovenia* (2016), p. 2. 133

[232] SPELKE, E., HIRST, W., AND NEISSER, U. Skills of divided attention. *Cognition 4*, 3 (1976), 215–230. 5

[233] STEVENS, R. H., GALLOWAY, T., AND BERKA, C. Eeg-related changes in cognitive workload, engagement and distraction as students acquire problem solving skills. In *User Modeling 2007: 11th International Conference, UM 2007, Corfu, Greece, July 25-29, 2007. Proceedings 11* (2007), Springer, pp. 187–196. 15

[234] SÜMER, Ö., GOLDBERG, P., D'MELLO, S., GERJETS, P., TRAUTWEIN, U., AND KASNECI, E. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing* (2021). 15

[235] SUN, K., YU, C., SHI, W., LIU, L., AND SHI, Y. Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (2018), pp. 581–593. 24

[236] SUNAR, A. S., WHITE, S., ABDULLAH, N. A., AND DAVIS, H. C. How learners' interactions sustain engagement: a MOOC case study. *IEEE Transactions on Learning Technologies 10*, 4 (2017), 475–487. 26

[237] TAO, Y., MITSVEN, S. G., PERRY, L. K., MESSINGER, D. S., AND SHYU, M.-L. Audio-based group detection for classroom dynamics analysis. In *2019 International Conference on Data Mining Workshops (ICDMW)* (2019), IEEE, pp. 855–862. 16

[238] TILK, O., AND ALUMÄE, T. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *INTERSPEECH* (2016). 72, 78

[239] TIMM, F., AND BARTH, E. Accurate eye centre localisation by means of gradients. *Visapp 11* (2011), 125–130. , 18, 28, 47, 49, 50, 51

[240] TONSEN, M., STEIL, J., SUGANO, Y., AND BULLING, A. Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1*, 3 (2017), 106. 18

[241] TS, A., AND GUDDETI, R. M. R. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and information technologies 25*, 2 (2020), 1387–1415. 15

[242] TSAI, S., AND MACHADO, P. E-learning, online learning, web-based learning, or distance learning: Unveiling the ambiguity in current terminology. *E-learn Magazine* (2002). 2

[243] TUISKU, O., MAJARANTA, P., ISOKOSKI, P., AND RÄIHÄ, K.-J. Now dasher! dash away! longitudinal study of fast text entry by eye gaze. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (2008), pp. 19–26. 23

[244] TURNEY, P. D. Learning algorithms for keyphrase extraction. *Information retrieval 2* (2000), 303–336. 23

[245] UNSWORTH, N., AND ROBISON, M. K. Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience 16*, 4 (2016), 601–615. 61

[246] URBINA, M. H., AND HUCKAUF, A. Alternatives to single character entry and dwell time selection on eye typing. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (2010), pp. 315–322. 23

[247] VELIYATH, N., DE, P., ALLEN, A. A., HODGES, C. B., AND MITRA, A. Modeling students' attention in the classroom using eyetrackers. In *Proceedings of the 2019 ACM Southeast Conference* (New York, NY, USA, 2019), ACM SE '19, Association for Computing Machinery, p. 2–9. 15

[248] VELIYATH, N., DE, P., ALLEN, A. A., HODGES, C. B., AND MITRA, A. Modeling students' attention in the classroom using eyetrackers. In *Proceedings of the 2019 ACM Southeast Conference* (2019), pp. 2–9. 15

[249] VIOLA, P., AND JONES, M. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2001), vol. 1, IEEE Comput. Soc, pp. I–511. 37

[250] VISURI, A., AND VAN BERKEL, N. Attention computing: overview of mobile sensing applied to measuring attention. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers* (2019), pp. 1079–1082. 14, 60

[251] WALTHER, D., RUTISHAUSER, U., KOCH, C., AND PERONA, P. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding 100*, 1-2 (2005), 41–63. 43

[252] WAN, H., SHI, S., CAO, W., WANG, W., AND CHEN, G. Resptracker: Multi-user room-scale respiration tracking with commercial acoustic devices. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications* (USA, 2021), IEEE Press, p. 1–10. 20

[253] WANG, J.-G., SUNG, E., AND VENKATESWARLU, R. Eye gaze estimation from a single image of one eye. In *Proceedings of the IEEE International Conference on Computer Vision* (2003). 18

[254] WANG, M., AND SHAO, Y. *The Google Challenge: Video Genre Classification*. Citeseer, 2010. 118

[255] WANG, Y., ZHAO, S., ZHANG, Z., AND FENG, W. Sad facial expressions increase choice blindness. *Frontiers in Psychology 8* (2018), 2300. 97

[256] WARD, D. J., AND MACKAY, D. J. Fast hands-free writing by gaze direction. *Nature 418*, 6900 (2002), 838–838. 23

[257] WEIWEI ZHANG, MURPHEY, Y. L., TIANYU WANG, AND QIJIE XU. Driver yawning detection based on deep convolutional neural learning and robust nose tracking. In *2015 International Joint Conference on Neural Networks (IJCNN)* (2015), pp. 1–8. 24

[258] WHITEHILL, J., MOHAN, K., SEATON, D., ROSEN, Y., AND TINGLEY, D. MOOC dropout prediction: How to measure accuracy? In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale* (2017), ACM, pp. 161–164. 26

[259] WILSON, P. I., AND FERNANDEZ, J. Facial feature detection using haar classifiers. *Journal of Computing Sciences in Colleges 21*, 4 (2006), 127–133. 37

[260] WOOD, E., AND BULLING, A. EyeTab: Model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (2014), pp. 207–210. 18, 51

[261] XIAO, X., AND WANG, J. Towards attentive, bi-directional MOOC learning on mobile devices. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (2015), ACM, pp. 163–170. 26

[262] XIAO, X., AND WANG, J. Context and cognitive state triggered interventions for mobile MOOC learning. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (2016), ACM, pp. 378–385. 19, 26, 52

[263] XIAO, X., AND WANG, J. Understanding and detecting divided attention in mobile MOOC learning. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (2017), pp. 2411–2415. 19, 26, 52

[264] XIAO, X., AND WANG, J. Understanding and detecting divided attention in mobile mooc learning. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (2017), pp. 2411–2415. 21

[265] XU, X., DANCU, A., MAES, P., AND NANAYAKKARA, S. Hand range interface: Information always at hand with a body-centric mid-air input surface. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2018), pp. 1–12. 23

[266] YANAGIDA, Y., KAWATO, S., NOMA, H., TETSUTANI, N., AND TOMONO, A. A nose-tracked, personal olfactory display. In *ACM SIGGRAPH 2003 Sketches & Applications*. 2003, pp. 1–1. 24

[267] YANG, Z., YU, C., YI, X., AND SHI, Y. Investigating gesture typing for indirect touch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3*, 3 (2019), 1–22. 23, 138

[268] YIN, L., AND BASU, A. Nose shape estimation and tracking for model-based coding. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)* (2001), vol. 3, IEEE, pp. 1477–1480. 24

[269] YU, C., GU, Y., YANG, Z., YI, X., LUO, H., AND SHI, Y. Tap, dwell or gesture? exploring head-based text entry techniques for hmds. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 4479–4488. 138

[270] ZALETELJ, J., AND KOŠIR, A. Predicting students' attention in the classroom from kinect facial and body features. *EURASIP journal on image and video processing 2017*, 1 (2017), 1–12. 16

[271] ZHANG, L., ZHOU, F., LI, W., AND YANG, X. Human-computer interaction system based on nose tracking. In *International Conference on Human-Computer Interaction* (2007), Springer, pp. 769–778. 24

[272] ZHANG, Q., WANG, D., ZHAO, R., YU, Y., AND SHEN, J. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5*, 3 (2021), 1–30. 20

[273] ZHANG, W., HUANG, X., WANG, S., SHU, J., LIU, H., AND CHEN, H. Student performance prediction via online learning behavior analytics. In *2017 International Symposium on Educational Technology (ISET)* (2017), IEEE, pp. 153–157. 52

[274] ZHANG, X., PARK, S., BEELER, T., BRADLEY, D., TANG, S., AND HILLIGES, O. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16* (2020), Springer, pp. 365–381. 18

[275] ZHANG, X., SUGANO, Y., AND BULLING, A. Everyday eye contact detection using unsupervised gaze target discovery. *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (2017), 193–203. 28

[276] ZHANG, X., SUGANO, Y., FRITZ, M., AND BULLING, A. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 4511–4520. 18

[277] ZHANG, Y., LI, X., ZHU, L., DONG, X., AND HAO, Q. *What Is a Smart Classroom? a Literature Review*. Springer Singapore, Singapore, 2019, pp. 25–40. 2

[278] ZHAO, S., BURY, G., MILNE, A., AND CHAIT, M. Pupillometry as an objective measure of sustained attention in young and older listeners. *Trends in hearing 23* (2019), 2331216519887815. 5

[279] ZHAO, Y., ROBAL, T., LOFI, C., AND HAUFF, C. Stationary vs. non-stationary mobile learning in MOOCs. In *Proceedings of the Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (2018), ACM, pp. 299–303. 26

[280] ZHAO, Z., LIU, Q., AND ZHOU, F. Robust lightweight facial expression recognition network with label distribution training. *Proceedings of the AAAI Conference on Artificial Intelligence 35*, 4 (May 2021), 3510–3519. 95

[281] ZHU, S., ZHENG, J., ZHAI, S., AND BI, X. i'sFree: Eyes-free gesture typing via a touch-enabled remote control. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12. 24

[282] ZHU, Z., AND JI, Q. Eye and gaze tracking for interactive graphic display. *Machine Vision and Applications 15*, 3 (Jul 2004), 139–148. 18

[283] ZONG, Y., ZHENG, W., HONG, X., TANG, C., CUI, Z., AND ZHAO, G. Cross-database micro-expression recognition: A benchmark. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (New York, NY, USA, 2019), ICMR '19, Association for Computing Machinery, p. 354–363. 97

A Appendix

A.1 Facial Landmark Detection

The detection of facial region is facilitated by using the method proposed in [40], which is based on Histogram of Oriented Gradients (HOG) and Linear Support Vector Machine (SVM). In this approach, the input image is divided into a grid of cells containing several pixels. Histograms are derived from the gradient magnitude and the orientation of these cells, indicating the degree and direction of change for each pixel location in the cell. For making the system more tolerant towards variations in lighting conditions, these cells are grouped into blocks of overlapping cells for the final estimation of the HOG features. Using these features, the SVM is trained for the detection of facial regions. From the detected facial regions, 68 facial landmarks (shown in Figure A.1) are extracted using the method proposed in [101], that uses an iterative training of a cascade of regressors, each using a gradient boosting tree algorithm. The *iBug 300-W dataset*[1] has been used for training this model. This dataset is particularly suitable for *EmotiConf* due to two major reasons. Firstly, this dataset contains landmark annotations for the different facial images, taken from categories like "conference" etc. which is necessary in *EmotiConf*. Moreover, facial occlusion is common in online conferences as participants might voluntarily or involuntarily move out of frame partially. The dataset works significantly well for occluded faces (distribution of occluded facial images is 29.83%), which is essential for natural video conferences.

[1] https://ibug.doc.ic.ac.uk/resources/facial-point-annotations/ (Access: Friday 11th August, 2023)

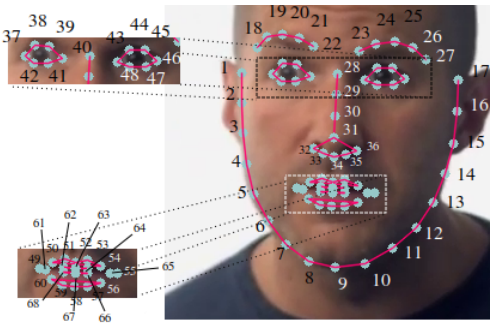


Figure A.1: 68 facial landmarks on a video frame from Youtube-8M dataset

A.2 System Usability Scale Questions

Table A.1 enlists the 10 System Usability Scale Questionnaire and their indication towards the usability of the system.

Number	Question	Type	Ideal Score
Q1	I think that I would like to use this system frequently	Positive	5
Q2	I found the system unnecessarily complex	Negative	1
Q3	I thought the system was easy to use	Positive	5
Q4	I think that I would need the support of a technical person to be able to use this system	Negative	1
Q5	I found the various functions in this system were well integrated	Positive	5
Q6	I thought there was too much inconsistency in this system	Negative	1
Q7	I would imagine that most people would learn to use this system very quickly	Positive	5
Q8	I found the system very cumbersome to use	Negative	1
Q9	I felt very confident using the system	Positive	5
Q10	I needed to learn a lot of things before I could get going with this system	Negative	1

Table A.1: System Usability Scale–questions and types

The final SUS score is calculated as : `((QA1-1)+(5-QA2)+(QA3-1)+(5-QA4)+(QA5-1)+(5-QA6)+(QA7-1)+(5-QA8)+(QA9-1)+(5-QA10))*2.5`, where `QAn` is the score to statement `Qn`, provided by a participant.

B

List of Acronyms

Pragona Kaur